nature
REVIEWS

GENETICS

TO PASTURES NEW

Complex trait analysis in livestock

Getting it together

Remodeling and chromatin genes

nature publishing group

▶ **COVER:** 'Grazing a trail' by Patrick Morgan, inspired by the Review on p381.

LOUISA FLINTOFT

TANITA CASCI

MARY MUERS

MEERA SWAMI

Model organisms continue to be fundamental to almost every aspect of genetic research. Three articles in this month's issue highlight the breadth of areas in which animal models provide insights, but also emphasize the need to use these organisms in the most effective way.

The Review from Lessing and Bonini on page 359 illustrates how one classic genetic model can be used to shed light on pathways that are relevant to disease. The wide range of genetic tools that can be used in *Drosophila melanogaster* has allowed the identification of many genes that are required to maintain neuronal integrity — over half of which have mouse or human counterparts that lead to neurodegeneration when disrupted.

Despite the advantages of flies and other non-mammalian organisms, the mouse remains the leading genetic model for disease. In a Review on page 371, Beckers and colleagues discuss three key areas in which the use of mouse models needs to be improved for maximum utility in both disease and basic research. They stress the need for genotypes more similar to those in human populations, more comprehensive and coordinated phenotyping and — the biggest challenge — methods to analyse environmental effects.

The choice of model organisms for evo–devo research is the topic of the Opinion article by Sommer on page 416. The author argues that focusing on a few key models with extensively developed genetic toolkits would be beneficial for tackling several aspects of this field, rather than studying a wider range of organisms in less detail.

The issue also features the first of two interviews with the recipients of this year's March of Dimes Prize in Developmental Biology (page 351). The second interview will appear in our July issue.

EDITORIAL OFFICES

LONDON NatureReviews@nature.com
The Macmillan Building, 4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0)20 7843 3620;
Fax: +44 (0)20 7843 3629
CHIEF EDITOR: Louisa Flintoft
SENIOR EDITOR: Tanita Casci
ASSOCIATE EDITOR: Mary Muers
ASSISTANT EDITOR: Meera Swami
COPY EDITOR: Elizabeth Neame
SENIOR COPY EDITORS: Isobel Barry,
Craig Nicholson, Man Tsuey Tse, Gillian Young
SENIOR ART EDITOR (NRG): Patrick Morgan
ART CONTROLLER: Susanne Harris
SENIOR ART EDITOR: Vicky Summersby
MANAGING PRODUCTION EDITOR:
Judith Shadwell
SENIOR PRODUCTION EDITOR:
Simon Fenwick
PRODUCTION CONTROLLER:
Natalie Smith
SENIOR EDITORIAL ASSISTANT: Laura Firman
EDITORIAL ASSISTANT: Jacques Smit
WEB PRODUCTION MANAGER:
Deborah Anthony
MARKETING MANAGERS: Tim Redding,
Leah Rodriguez

MANAGEMENT OFFICES

LONDON nature@nature.com
The Macmillan Building, 4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0)20 7833 4000;
Fax: +44 (0)20 7843 4596/7
OFFICE MANAGER: Kiersty Darnell
PUBLISHER: Stephanie Diment
MANAGING DIRECTOR: Steven Inchcoombe
EDITOR-IN-CHIEF, NATURE PUBLICATIONS:
Philip Campbell
ASSOCIATE DIRECTORS:
Jenny Henderson, Tony Rudland
EDITORIAL PRODUCTION DIRECTOR:
James McQuat
PRODUCTION DIRECTOR: Yvonne Strong
DIRECTOR, WEB PUBLISHING: Timo Hannay
HEAD OF WEB PRODUCTION:
Alexander Thurrell
NATUREJOBS PUBLISHER: Della Sar

NEW YORK nature@natureny.com
Nature Publishing Group,
75 Varick Street, 9th floor, New York,
NY 10013-1917, USA
Tel: +1 212 726 9200;
Fax: +1 212 696 9006
CHIEF TECHNOLOGY OFFICER:
Howard Ratner
HEAD OF INTERNAL SYSTEMS DEVELOPMENT:
Anthony Barrera
HEAD OF SOFTWARE SERVICES:
Luigi Squillante
HEAD OF GLOBAL ADVERTISING, SALES AND SPONSORSHIP: Dean Sanderson
HEAD OF NATURE RESEARCH & REVIEWS MARKETING: Sara Girard
BUSINESS DEVELOPMENT EXECUTIVE:
David Bagshaw

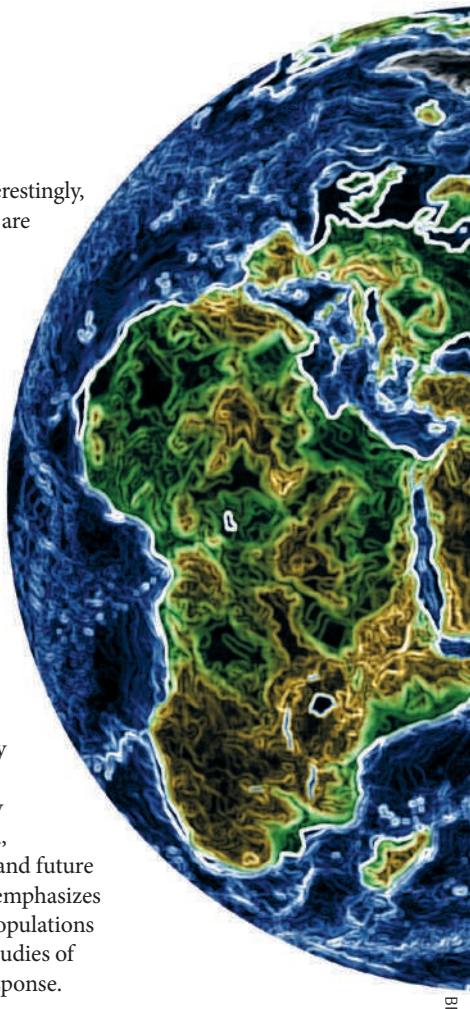
TOKYO nature@natureasia.com
Chiyoda Building 5F, 2-37-1 Ichigayatamachi,
Shinjuku-ku, Tokyo 162-0843, Japan
Tel: +81 3 3267 8751; Fax: +81 3 3267 8746
ASIA–PACIFIC PUBLISHER: Antoine E Bocquet
MANAGER: Koichi Nakamura
ASIA–PACIFIC SALES DIRECTOR:
Kate Yoneyama
SENIOR MARKETING MANAGER:
Peter Yoshihara
MARKETING/PRODUCTION MANAGER:
Takesh Murakami

INDIA 5A/12 Ansari Road, Daryganj,
New Delhi 110 002, India
Tel/Fax: +91 11 2324 4186
SALES AND MARKETING MANAGER, INDIA:
Harpal Singh Gill

Printed in Wales by Cambrian Printers
on acid-free paper

POPULATION GENETICS

Genetic landscapes out of Africa



BRANDX

As the birthplace of modern humans, Africa holds unique significance for human population genetics. The continent is now home to approximately 900 million people in more than 2,000 different ethnolinguistic groups, and yet patterns of genetic variation in African populations have remained largely uncharacterized. A major study of genetic diversity in African and African American populations now helps to unravel the complex evolutionary history of these groups and human populations worldwide, and might lead to improvements in understanding population-specific disease risk.

Tishkoff and colleagues analysed variation in a panel of over 1,300 microsatellites and insertions–deletions in 121 African, 4 African American and 60 non-African populations. This work has increased the scale of genetic analysis of African populations to include a much wider range of groups in a genome-wide study than has previously been done. This gave the authors sufficient data to explore patterns of genetic diversity both within and between populations, to construct phylogenetic trees to look at the genetic distances between populations (which could be compared with the geographic distances), and to explore the relationship between genetic and cultural distinctions in greater detail than before.

African and African American populations stand out in the worldwide comparisons as having the highest levels of within- and between-population diversity. This is consistent with divergent ancestral populations in Africa, compared with much smaller founder groups that were formed as humans migrated away from the continent. Indeed, diversity within populations generally decreases with increased distance from Africa. The authors found genetic evidence of 14 ancestral population clusters. Importantly, these analyses revealed more substructure within African populations than had previously been observed, which has implications for the appropriate design and interpretation of association studies for diseases and other traits.

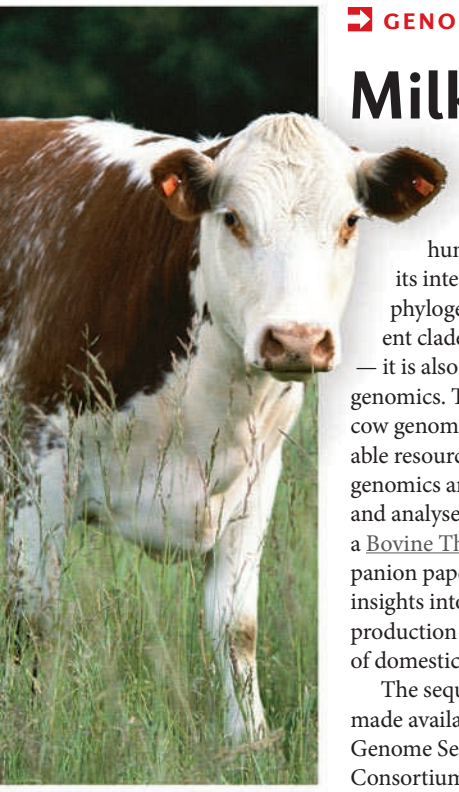
The authors also used their genetic data to explore the population history of many ethnolinguistic groups, giving further insight into where migrations have occurred and which populations have remained relatively isolated. For example, considerable Niger-Kordofanian ancestry was found in nearly all populations, which probably reflects the spread and mixing with local populations of the farming Bantu-speaking group in the last ~5,000 years. Interestingly, although genetic clusters are generally consistent with language groups there are some exceptions — such as the Maasai and Pygmies — for which cultural distinction has been robust despite genetic mixing.

A central theme that emerges from these analyses is the genetic complexity of African populations. The significance of this is twofold: it confirms that further genotyping and resequencing of African genomes is likely to be highly informative for dissecting the history of human evolution; and, importantly for current and future inhabitants of Africa, it emphasizes that ethnically diverse populations need to be included in studies of disease risk and drug response.

Mary Muers

ORIGINAL PAPER Tishkoff, S. A. *et al.* The genetic struture and history of Africans and African Americans. *Science* 30 Apr 2009 (doi:10.1126/science.1172257)
FURTHER READING Pagel, M. Human language as a culturally transmitted replicator. *Nature Rev. Genet.* 12 May 2009 (doi:10.1038/nrg2560)

Nature Reviews Genetics | AOP, published online 7 May 2009; doi:10.1038/nrg2598



GENOMICS

Milking the cow genome

The domestic cow, *Bos taurus*, is not only an established part of human agriculture but with its interesting position in the phylogenetic tree — in a different clade to humans and rodents — it is also favoured in comparative genomics. Two assemblies of the cow genome now provide a valuable resource for both evolutionary genomics and livestock breeding, and analyses of the data (including a Bovine Thematic Series of companion papers) are already providing insights into the evolution of milk production and the genomic impact of domestication and breeding.

The sequence data have been made available by the Bovine Genome Sequencing and Analysis Consortium, who have published an analysis of their genome assembly. In a simultaneous publication, Zimin *et al.* have published an alternative assembly using the primary sequence data. Both assemblies represent ~90% of the full genome sequence but differ, for example, in the number of segmental duplications.

The authors of the consortium paper estimate that the cattle genome contains at least 22,000 protein-coding genes, with over 16,000 having orthologues in other placental mammals. One interesting finding from this assembly is that genes associated with reproduction are overrepresented in segmental duplications, which might have contributed to ruminant-specific aspects of maternal adaptation and fetal growth. They also found extensive duplication and divergence of innate immune system genes, which might reflect either the high exposure to microorganisms that occurs in the cow rumen or selection owing to the rapid transmission of disease that can occur within herds.

Given the importance of dairy farming, it is unsurprising that the availability of the bovine genome has already triggered new research into the genetics of milk production and lactation. For example, Lemay and colleagues compared the cow genome with other mammalian genomes and identified 197 unique milk protein genes in cattle, but found that, on average, genes involved in lactation are highly conserved among mammals and evolve slowly. Copy number variation seems to have made a significant contribution to the diversity of milk composition between species. This comprehensive catalogue of bovine milk genes might also help in the search for candidate genes within milk-trait QTLs — a significant step for enhancing yield.

In a companion paper to the genome assembly, the Bovine HapMap Consortium analyse SNP variation in different cattle populations, including from the humpless (taurine) breeds and the humped (indicine) breeds. Their studies of genetic diversity reveal that, overall, domestic cattle had a large ancestral population, so the population bottlenecks that are commonly associated with domestication and breed formation were not as severe in cattle as they are in species such as the dog. However, indicine breeds had a much larger ancestral population than taurine cattle (which have a similar level of SNP diversity to humans). Significantly, for cattle breeders, there has been a very recent rapid decline in genetic diversity, which is probably due to selection.

These studies suggest that the cow will provide new perspectives on mammalian genome evolution, as well as revealing the genetic impact of past and future prospects for cattle breeding.

Mary Muers

ORIGINAL RESEARCH PAPERS The Bovine Genome Sequencing and Analysis Consortium *et al.* The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**, 522–528 (2009) | Zimin, A. V. *et al.* A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol.* 24 Apr 2009 (doi:10.1186/gb-2009-10-4-r42) | Lemay, D. G. *et al.* The bovine lactation genome: insights into the evolution of mammalian milk. *Genome Biol.* 24 Apr 2009 (doi:10.1186/gb-2009-10-4-r43) | The Bovine HapMap Consortium. Genome-wide survey of SNP variation uncovers genetic structure of cattle breeds. *Science* **324**, 528–532 (2009)

WEBSITE

BioMed Central Bovine Thematic Series: http://www.biomedcentral.com/series/bovine

Nature Reviews Genetics | AOP, published online 7 May 2009; doi:10.1038/nrg2599

COMPLEX DISEASE

Autism clues from genome-wide studies

Our understanding of the genetic contribution to autism spectrum disorders (ASDs) has expanded rapidly, but remains far from complete. Despite good evidence for roles of rare and *de novo* variants in some cases, the genetic basis of most cases remains unexplained and the involvement of common genetic variants is poorly understood. Two genome-wide studies now alter this picture. One provides the first firm evidence for a role of common SNPs in ASDs, whereas the other expands our knowledge of the involvement of copy number variation.

Wang and colleagues carried out genome-wide studies using two large cohorts, one comprising ASD families and the other population based. To define the cohorts, stringent diagnostic criteria were used — an important step given the clinical heterogeneity of these conditions. The authors tested for association with more than 550,000 SNPs and achieved reliable results for 780 families as well as 1,204 cases and 6,491 controls. Six SNPs showed association at genome-wide significance across the two cohorts. The SNPs — which were replicated in a third cohort — lie between two genes that encode the cell adhesion molecules cadherin 9 (*CDH9*) and cadherin 10 (*CDH10*). This is an interesting finding given that altered neuronal cell adhesion has been implicated in ASDs; furthermore, *CDH10* is expressed in the frontal cortex, an area that is known to be affected in ASD cases.

'Pathway-based' approaches — which combine data from SNPs to look for differences in statistical significance between certain groups of genes and the rest of the genome — highlighted a group of 25 cadherin genes and 8 neurexin genes, providing further evidence for a role of altered neuronal cell adhesion in ASDs.

In a second study, Glessner and colleagues looked for copy number variants (CNVs) that are involved in ASD susceptibility. Again, strict diagnostic assessments were made and large numbers of cases and controls were used — 1,246 and 1,409, respectively. Using the same genotyping platform as Wang *et al.*, the authors made a total of 78,490 CNV calls. The results provided additional support for the involvement of some CNVs that have already been implicated in ASDs and implicate nine new variants. The genes that are associated with these CNVs again suggest the importance of neuronal cell adhesion in ASD, and also highlight a role for the ubiquitin pathway — another function that has been implicated in previous genetic studies of ASDs.

As well as adding to our understanding of the genetic architecture of ASD susceptibility by implicating common variants, these studies provide clues to the biological functions affected in these conditions. Studying the expression patterns and functions of these genes will be a crucial next step.

Louisa Flintoft

ORIGINAL RESEARCH PAPERS Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* 28 Apr 2009 (doi:10.1038/nature07999) | Glessner, J. T. *et al.* Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 28 Apr 2009 (doi: 10.1038/nature07953)

FURTHER READING Abrahams, B. S. & Geschwind, D. H. Advances in autism genetics: on the threshold of a new neurobiology. *Nature Rev. Genet.* **9**, 341–355 (2008)

Nature Reviews Genetics | AOP, published online 12 May 2009; doi:10.1038/nrg2601

IN BRIEF

GENOME INSTABILITY

Chromosome instability is common in human cleavage-stage embryos

Vanneste, E. *et al. Nature Med.* 26 Apr 2009 (doi:10.1038/nm.1924)

This study describes the presence of a large amount of genomic instability early in human embryogenesis. The authors used new array-based methods to analyze genome-wide changes in copy number and loss of heterozygosity in multiple single cells of *in vitro* fertilization embryos taken from fertile women. Only 2 of the 23 cleavage stage embryos were chromosomally normal; the others were mosaic for deletions, duplications, amplifications and aneuploidies. These frequent and complex rearrangements might account for the low fertility rate in humans, or for the high rate of miscarriage.

EPIGENETICS

A *C. elegans* LSD1 demethylase contributes to germline immortality by reprogramming epigenetic memory

Katz, D. J. *et al. Cell* **137**, 308–320 (2009)

Some histone modifications, such as dimethylation of histone H3 at lysine 4 (H3K4me2), are thought to help cells to 'remember' patterns of transcription, and they are erased in the germ line to allow normal development. These authors found that mutation of *spr-5*, which encodes the *Caenorhabditis elegans* orthologue of LSD1 and removes H3K4me2, leads to increasing sterility across subsequent generations. Sterility correlates with increased H3K4me2 and with aberrant expression of spermatogenesis genes. This work provides insight into potential mechanisms of epigenetic memory and reprogramming.

REPLICATION

Transcription initiation activity sets replication origin efficiency in mammalian cells

Sequeria-Mendes, J. *et al. PLoS Genet.* **5**, e1000446 (2009)

This study shows that DNA replication initiates preferentially at sites of active transcription in mouse embryonic stem cells, and that nearly half of all replication origins (ORIs) are at promoters. ORIs at promoters in CpG islands are the most efficient at initiating replication. The association between ORIs and transcriptional units is maintained in other cell types and is in agreement with earlier work in human cells. These findings provide further evidence of an intimate relationship between transcription and replication, and could reflect co-evolution of their regulatory regions.

GENE EXPRESSION

Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans

Calvo, S. *et al. Proc. Natl Acad. Sci. USA* 13 Apr 2009 (doi: 10.1073/pnas.0810916106)

Approximately half of all human and mouse transcripts contain upstream ORFs (uORFs) — mRNAs that originate in the 5′UTR of a gene but are out of frame with the ORF. An expression analysis of ~11,600 matched mouse mRNAs and proteins shows that uORFs significantly reduce downstream protein expression. Mutations that alter the uORFs of some disease-associated genes reduce the expression of the downstream protein, indicating that uORFs could affect disease phenotypes.

Nature Reviews Genetics | AOP, published online 28 April 2009; doi:10.1038/nrg2595

CANCER GENOMICS

A modular approach to signalling

Instead of considering signalling in terms of a linear sequence, the concept of modules as units of signalling activity is a useful way to represent complex biological networks, such as those involved in cancer. A recent study describes an approach to dissect oncogenic signalling pathways into functional modules on the basis of gene expression signatures, which can then be used to analyse disease outcome and responses to therapeutics.

Joseph Nevins and colleagues reasoned that whole genome expression data could be used to define oncogenic pathway modules. Using the Ras and E2F signalling pathways as examples, they defined a core set of genes for each pathway; for the oncoprotein Ras, these are genes that encode proteins that directly bind to Ras, and those with one degree of separation from Ras in a protein–protein interaction network. Using the previously generated NCI-60 data set (which is composed of expression profiles of human cancer cell lines from a range of different tissues) as a source of expression data, the authors then used statistical analyses to identify genes related to the core pathway that showed similar variation in their expression as the core genes. This approach allowed them to generate signatures that correspond to sets of genes that share expression patterns.

The authors identified 20 gene signatures in the Ras pathway and 8 signatures in the E2F pathway. By comparing these signatures with the signatures of mutants that selectively activate downstream effectors or to signatures from cells that are sensitive to drugs that target specific pathway members, they could assign the signatures to specific signalling effectors, such as Raf or phosphatidylinositol 3-kinase for Ras signalling and S phase or mitotic events for E2F. This allowed them to define signalling pathway modules on the basis of expression signatures.

Can module signatures be used to predict clinical outcome? Chang *et al.* analysed the response of colon cancer patients to the epidermal growth factor receptor (EGFR)-specific therapy cetuximab. They derived a set of 20 gene expression signatures for EGFR from the NCI-60 expression data, and then compared the EGFR, Ras and E2F signatures to see if they could differentiate between the gene signatures of patients who responded or did not respond to cetuximab. Only the EGFR signatures could distinguish between the two sets of patients, indicating the specificity of each set of signatures for a particular oncogenic signalling pathway. Therefore, the oncogenic module approach can be used to identify clinically relevant tumour phenotypes. In a broader context, a modular model of pathway structure could also be valuable for studying the way that information is transmitted through cellular networks and the relationships between signalling modules and phenotypes.

Meera Swami

DIGITAL VISION

ORIGINAL RESEARCH PAPER Chang J. T. *et al.* A genomic strategy to elucidate modules of oncogenic pathway signaling networks. *Mol. Cell* **34**, 104–114 (2009).
FURTHER READING Nevins, J. R. & Potti, A. Mining gene expression profiles: expression signatures as cancer phenotypes. *Nature Rev. Genet.* **8**, 601–609 (2007).

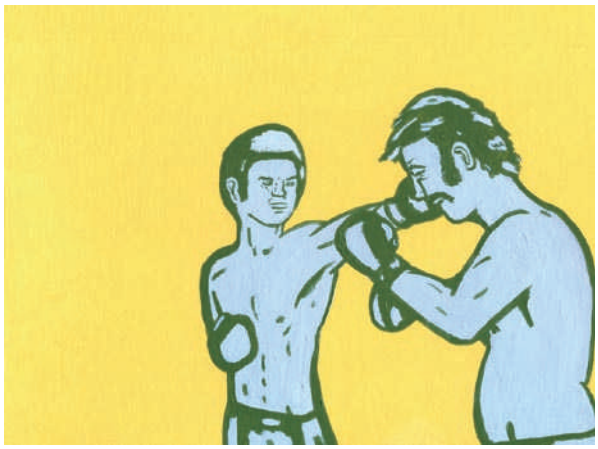
Nature Reviews Genetics | AOP, published online 28 April 2009; doi:10.1038/nrg2596

GENE EXPRESSION

Structure versus codon bias

It is well known that codon bias and gene expression are correlated. The established explanation is that mRNAs with a high codon adaptation index (CAI) — that is, with a high number of 'preferred' codons — are translated more efficiently because there are more tRNAs that match the codons. Now, however, Plotkin and colleagues show that it is mRNA structure not CAI that affects expression levels.

The authors generated 154 gene constructs that encoded the same green fluorescent protein (GFP) under the control of a T7 promoter, but in each construct they introduced random synonymous mutations in the third base positions of up to 180 codons. When constructs were put into *Escherichia coli* cells, their fluorescence levels varied 250-fold. However, surprisingly, there was no correlation between expression of the construct and its CAI.



Plotkin and colleagues looked at whether the folding energy of each GFP mRNA correlated with fluorescence. Although the structure of the entire mRNA had no bearing on expression, the folding energy of nucleotide positions –4 to +37 explained over half of the variation: the tighter the folding, the lower the level of expression. These findings support the hypothesis that strong secondary structure at the 5′ end of an mRNA blocks ribosome binding and delays translation initiation.

To test this idea, the authors added an identical stretch of 28 codons with weak mRNA secondary structure to the 5′ end of 72 of the GFP constructs. As expected, the tagged constructs produced consistently high levels of expression. The reduction in translation efficiency for mRNAs with strong folding at the ribosome binding site is consistent with previous studies that suggested that initiation, not elongation, is the rate-limiting step in mRNA translation.

How can these results be reconciled with the well-known link between CAI and expression level? The authors suggest that selection for efficient translation at the global level, rather than at the gene level, has led to an indirect link between expression and CAI among endogenous genes. In their model, high CAI speeds up elongation of a gene but does not affect its expression level. However, faster elongation means that fewer ribosomes are sequestered on the mRNA. This increases the total rate of protein synthesis in the cell, thereby providing a selective advantage in terms of an increased rate of cell growth.

Elizabeth Neame

ORIGINAL RESEARCH PAPER Kudla, G. *et al.* Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**, 255–258 (2009)

Nature Reviews Genetics | AOP, published online 12 May 2009; doi:10.1038/nrg2602

SYNTHETIC BIOLOGY

Towards off-the-shelf networks

Synthetic biologists dream of constructing gene expression networks with predictable functions. However, they come up against a frustrating problem: because the assembly parts (ORFs and control regions) are poorly characterized and limited in number, each new circuit has to be painstakingly tweaked until it behaves as intended. To overcome this hurdle Ellis and colleagues have generated and characterized a library of components, and then used computer modelling to inform how these components — in this case, promoters — should be assembled for particular uses.

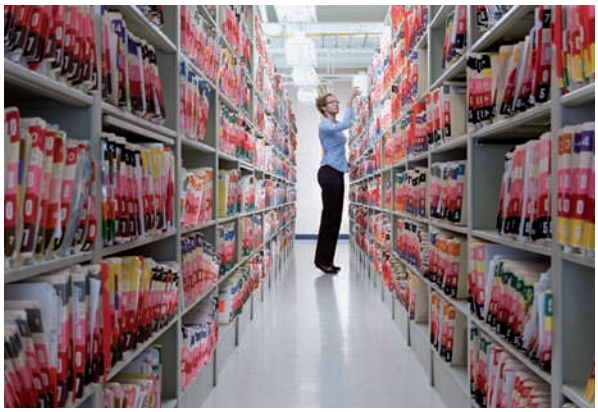


IMAGE SOURCE

To produce the promoter library a set of *Saccharomyces cerevisiae* promoters that can be regulated by TetR was generated using a synthesis protocol that specifically alters non-essential sequences. The library components were then classified by their expression output — as inferred from a reporter gene, *EGFP* — depending on the concentration of TetR.

An *in silico* prediction method was used to select which of the 20 promoters would be the most appropriate to use in any particular type of circuit. For example, in a feed-forward loop two repressors (LacI and TetR) feed onto an output gene, but with TetR also inhibiting *LacI*. The challenge was to select, from the library, the *LacI* promoter that would yield a predicted expression landscape in response to varying concentrations of TetR and LacI inducers. The input–output landscape of the three experimentally constructed networks correlated well with computational predictions.

The approach was also effective when applied to a more complicated circuit, a genetic timer: here, expression levels switch (toggle) from one state to another depending on the relative concentration of opposing repressor molecules. Two libraries of promoters were used in this application, and the model was derived by combining first principles with a single reconstructed circuit; however, as before, the experimental circuits behaved as expected.

The predictions associated with the genetic timer were applied to control the timing of activation of a yeast gene that triggers the clumping and sedimentation of yeast cells. This process allows cells to be removed easily from fermented liquid, and is therefore of interest in beer and wine production.

This off-the-shelf strategy promises to accelerate the rate of progress of synthetic biology by removing the fiddly adjustments that currently hamper the post-construction stage. And promoters are just the beginning — the same principles can be extended to other biomolecular components, such as RNAs or proteins.

Tanita Casci

ORIGINAL RESEARCH PAPER Ellis, T. *et al.* Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nature Biotech.* 19 Apr 2009 (doi:10.1038/nbt.1536)

IN BRIEF

HUMAN GENETIC VARIATION

The diversity present in 5,140 human mitochondrial genomes

Pereira, L. *et al. Am. J. Hum. Genet.* 7 May 2009 (doi:10.1016/j.ajhg.2009.04.013)

High-throughput sequencing is rapidly increasing the amount of data on human mitochondrial genetic variation. These authors developed a new computational tool, mtDNA-GeneSyn, which analyses diversity among mitochondrial genomes. Using this tool to analyse all mitochondrial DNA (mtDNA) data currently in GenBank, the authors provide an overall picture of human mtDNA diversity. The software is free to download, allowing other researchers to perform similar analyses as more data is deposited, and to assess mtDNA diversity in specific populations.

TRANSLATION

Bases in the anticodon loop of tRNA prevent misreading

Murakami, H. *et al. Nature Struct. Mol. Biol.* **16**, 353–358 (2009)

A sequence element that tunes *Escherichia coli* $tRNA^{Ala}_{GGC}$ to ensure accurate decoding

Ledoux, S. *et al. Nature Struct. Mol. Biol.* **16**, 359–364 (2009)

Using *in vitro* assays for peptide synthesis these papers show that the interaction between the anticodon on a tRNA molecule and its cognate mRNA codon is labile, as some variants lead to relaxed constraints on codon–anticodon interactions and to the insertion of a wrong amino acid. Such translational infidelity is caused by particular combinations of interacting base pairs (positions 32 and 38) on either side of the anticodon loop. The structural basis for the infidelity is unknown, but the suggestion that mutations at positions 32 or 38 are causal to several progressive human diseases is intriguing.

GENE REGULATION

Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development

Ji, Z. *et al. Proc. Natl Acad. Sci. USA* **106**, 7028–7033 (2009)

This paper reveals that an important means of regulating gene expression post-transcriptionally might be provided by the progressive lengthening of the 3′UTR of mRNAs. Such lengthening — which was observed *in vivo* and *in vitro* — is attributed to alternative adenylation, permitted by weak polyadenylation signals; the resulting AU-rich 3′UTRs would make them better targets for microRNAs, among other regulators.

COMPLEX DISEASE

Multilocus Bayesian meta-analysis of gene-disease associations

Newcombe, P. J. *et al. Am. J. Hum. Genet.* 30 Apr 2009 (doi:10.1016/j.ajhg.2009.04.001)

In meta-analyses of gene–disease association studies the widely used approach of pooling data for each SNP is inefficient because, often, only a subset of studies provide data about a particular marker. This study reports a generally applicable Bayesian, multimarker approach to meta-analysis that uses all data for a region or gene, irrespective of the specific markers that have been typed, to make efficient use of data from all constituent studies.

Nature Reviews Genetics | AOP, published online 7 May 2009; doi:10.1038/nrg2597

SMALL RNAS

A tiny stabilizer of development

A new paper provides the first experimental evidence that microRNAs (miRNAs) confer robustness, by showing that a *Drosophila melanogaster* miRNA buffers a developmental process against environmental fluctuation.

Li and colleagues investigated the role of miR-7 in sensory organ development. Previous studies had shown that, under uniform conditions, loss of miR-7 function has little impact on either the expression levels of its regulatory targets or on developmental outcome. Does miR-7 have a role in stabilizing sensory organ development in less stable environments?

Using transgenes, reporter assays and mutant analyses, the authors carefully dissected the interactions between miR-7, its regulators and its targets during the differentiation of photoreceptors and proprioreceptors. In both cases, miR-7 functions as part of one or more feedforward or feedback loops in a way that is predicted to stabilize the expression of key determinants of cell fate. Indeed, *mir-7* mutant larvae that were exposed to temperature fluctuations showed altered expression of these genes and defects in the specification and patterning of sensory organs.

Interestingly, although miR-7 is highly conserved from flies to humans, its functions are not; for example, it is not involved in vertebrate sensory organ development. The authors suggest that conserved miRNAs might be recruited into new regulatory interactions during evolution specifically to provide robustness to regulatory networks.

Louisa Flintoft

ORIGINAL RESEARCH PAPER Li, X. *et al.* A microRNA imparts robustness against environmental fluctuation during development. *Cell* **137**, 273–282 (2009).
FURTHER READING Flynt, A. S. & Lai, E. C. Biological principles of microRNA-mediated regulation: shared themes amid diversity. *Nature Rev. Genet.* **9**, 831–842 (2008)

TECHNOLOGY

The holy grail for plant biologists

Plant biologists and biotechnologists have long suffered from the lack of an efficient and sequence-specific method for gene targeting. Two recent reports of successful gene targeting in maize and tobacco come as a welcome improvement on the laborious conventional mutagenesis or transgenesis approaches. For the first time, endogenous plant loci can be targeted at high efficiency.

Both studies rely on zinc-finger nucleases (ZFNs) — engineered enzymes that create double-stranded breaks at specific loci and that have previously been used to modify, *in vitro*, plant transgenes and endogenous genes in human cells. ZFNs are a fusion between an endonuclease domain and a zinc-finger-based DNA recognition domain; this latter domain can be designed to recognize almost any DNA sequence, and therefore gives ZFNs their specificity.

Working in maize, Shukla and colleagues used a panel of pre-validated ZFNs that were designed against two independent endogenous loci: *IPK1*, the product of which catalyzes the final step in phytate biosynthesis in seeds, and *ZP15*, which encodes a seed protein. Following pre-screening to determine the relative efficiency of their ZFNs, the authors detected modification in plants by selecting for the insertion of a herbicide tolerance gene or by amplifying and sequencing the targeted locus. As transgene insertion is very specific the authors were able to generate several independent lines of fertile plants that transmitted the modification to the next generation.

In an independent effort, Townsend and Wright *et al.* report the modification of multiple acetolactate synthase loci in tobacco plants; when inactivated, these genes render a plant resistant to two types of herbicide. For their ZFN design they relied on a publicly available resource from the Zinc Finger Consortium, and selected the molecules with highest specificity by pre-screening them in bacteria and subsequently in yeast. Interestingly, in a small proportion of cases they also saw targeting over 1.3 kb from the site of cleavage. This suggests that it should be possible to modify plant genes even if their surrounding genomic sequence is not optimal for ZFN targeting.

The precision and efficiency of the ZFNs offer clear advantages for dissecting gene function in plants as well as for plant engineering for food or fuel. Given the versatility of ZFNs, this approach can be used in any plant species as long as it is amenable to DNA delivery. In the future, targeted modifications could be identified in DNA-sequence-based screens; considering the advances in sequencing technologies, high-throughput plant genomic engineering could be just around the corner.

Magdalena Skipper, Senior Editor, Nature

ORIGINAL RESEARCH PAPERS Shukla, V. K. *et al.* Precise genome modification in the crop species of *Zea mays* using zinc-finger nucleases. *Nature* 28 Apr 2009 (doi:10.1038/nature07992) | Townsend, J. A. & Wright, D. A. *et al.* High-frequency modification of plant genes using engineered zinc-finger nucleases. *Nature* 28 Apr 2009 (doi:10.1038/nature07845)

AN INTERVIEW WITH...

Kevin Campbell

The 2009 March of Dimes Prize in Developmental Biology has been awarded jointly to Kevin Campbell of the University of Iowa and to Louis Kunkel of Harvard Medical School and The Children's Hospital, Boston, for their pioneering work in identifying the genes and proteins that are disrupted in muscular dystrophies. The prize recognizes researchers whose work has contributed to our understanding of the science that underlies birth defects. We talked to the winners about their scientific careers and their views on biomedical research. This month's interview is with Kevin Campbell, who spoke to Louisa Flintoft. The interview with Louis Kunkel will appear next month.

Your bachelor's degree is in physics. Have you found that useful as a biologist?
The problem-solving aspect is what I find really helpful today. Especially early on in your career I think it's important not to be too specialized, and having a diverse scientific background is really helpful. You never know where a research topic is going to lead.

How did your work on ion channels lead to your discovery of the dystrophin–glycoprotein complex?
When I moved to Iowa my goal was to clarify the channels involved in excitation–contraction coupling. I thought I was set and nothing would change. I sent a paper in with a grant renewal to the Muscular Dystrophy Association (MDA), and in that paper we showed that the purified ryanodine receptor was a very large protein. Don Wood, who ran the research programme at MDA, had been at a meeting where Lou Kunkel had reported the mRNA size for the gene, and it was very large. Don Wood put Lou and I together and we quickly exchanged antibodies and within 2 weeks we knew that the ryanodine receptor and dystrophin were two separate proteins. That was the beginning of my work on muscular dystrophy (MD). We probably would have given up on it, but the next summer we were trying to isolate a combined receptor complex, and it was working with the calcium channel but this other protein was coming along that wasn't working with the antibodies to the ryanodine receptor. Finally one day I stained with antibodies to dystrophin and dystrophin was there. We purified the proteins and ended up with the dystrophin–glycoprotein complex.

Was it exciting to make the link with a disease?
Disease really gave us the opportunity and the reagents to further the basic science. We went to Duchenne's muscular dystrophy biopsies because we wanted to see what happened to proteins in the complex. I was teaching the red cell cytoskeleton for a graduate course and I had an idea that maybe if one protein was missing the whole complex would be lost. In the red cell cytoskeleton you can get fragile cells if you lose ankyrin or spectrin, and I was thinking that loss of these other proteins might cause MDs. So some of the work we do is using these clinical reagents, but it was initially because there weren't that many mouse models at the time. We were getting new biopsies in and looking at new dystrophies and we got good at that, and now we have a whole centre dedicated to it. If 20 years ago I had made mouse or fly mutants then you would consider it pure basic science, but instead we're using the natural resource of the MD genes that are out there in the human population.

How important is it for you to discuss your work with practising physicians?
Very important. We work together with physicians to try to identify new genes. At one meeting I presented a mouse model that's defective for neuronal nitric oxide synthase (nNOS) [Campbell and his group recently published a paper in *Nature* using mouse models, which suggests that the muscle fatigue seen in patients with neuromuscular conditions is due to defective nNOS signalling] and a physician said "They look just like my Becker patients." That was one of several hints with that study. We always knew patients with Becker's syndrome were missing nNOS, but a lot of people thought the fatigue was mainly due to the muscle weakness.

You direct an institute that focuses on therapeutic approaches. Does the way that biomedical research is funded encourage the clinical translation of basic research?
We're lucky in that we're very well funded. But it can be more difficult to get funding for work that's considered pure applied research. There probably needs to be new ways to evaluate that work. It can be difficult. That kind of work doesn't always lead to outstanding papers, *Nature* papers. Often, that work is done by companies. To get it done in academic institutions can be harder.

Will traditional gene therapy live up to its potential?
I agree with the statement of Harold Varmus that gene therapy will be used to cure disease in the next 10 to 100 years. It's obviously the way to go to put the gene back in, but it's very complicated and there are a lot of things that need to be worked out. In a number of cases, at least in the MD field, the challenge is secondary problems: delivery problems, potentially immune problems. If you put the DNA into muscle directly it incorporates it and you get a few fibres that stain quite nicely. But a few cells are not really going to help and you need to hit all the muscles. There are muscles like the diaphragm that are hard to get the gene into. So it's going to take time. In terms of other avenues, exon skipping looks very promising, and so do therapies where you allow read-through of stop codons.

Does the media report disease-related research in a useful way?
The media is really important in getting information to the general public. Sometimes it gets inflated and that's scary. Even scientifically I think we're having a problem. If you search for "rescue for mdx" there are so many papers, but in most cases those are not going to be directly translated into therapies. I think that leads to a lot of people thinking that these diseases are about to be cured. I try to make sure that we don't do that.

New insights into the aetiology of colorectal cancer from genome-wide association studies

Albert Tenesa and Malcolm G. Dunlop

Abstract | Genome-wide association studies have recently identified ten common genetic variants associated with colorectal cancer susceptibility, several suggesting the involvement of components of the transforming growth factor beta (TGFβ) superfamily signalling pathway. To date, no causal sequence variants have been identified, and risk seems to be mediated through effects on gene regulation. Several markers are located close to poorly characterized genes or in gene deserts, raising challenges for elucidating mechanisms of susceptibility. Disease-associated common genetic variation offers the potential to refine risk stratification within populations and enable more targeted disease prevention strategies.

Over 1 million new cases of colorectal cancer are diagnosed worldwide each year, and incidence seems set to rise with the progressive 'westernization' of lifestyles in Asian and African populations. It is the third most common malignancy and the fourth biggest cause of cancer mortality[1]. Incidence rates closely parallel economic development, reflecting a westernized lifestyle and attendant risk factor exposures. However, manifestation of colorectal cancer burden can also relate to longer life expectancy in developed populations, along with better diagnostic and recording tools. International cancer registry data indicate that the overall mortality rate is ~50%, with the single most important arbiter of survival outcome being extent of tumour progression at diagnosis. In Scotland, the 5-year survival rate among young patients exceeds 80% for those with localized tumours, but is only ~40% for those with metastatic disease at presentation[2]. Compelling evidence indicates that early detection and prevention by removal of premalignant polyps can reduce colorectal cancer mortality. Randomized trials of population screening have shown reduced mortality in subjects with an average risk[3], and improved survival is also observed in genetically defined high-risk groups who undergo more intensive surveillance[4].

Colorectal cancer is a complex trait influenced by genetic and environmental factors and their interactions. The concept of familial colorectal cancer reflects one end of the plausible risk spectrum of contributory genetic variants. Population genetics theory predicts that the distribution of allelic effects influencing complex traits is L shaped[5,6], with a small number of variants having a large effect on phenotype and a large number of variants having an individually small effect. Rare variants of large effect contribute predominately in the subgroup of patients with disease segregating in families. However, despite these large effects, the low allele frequency means their overall contribution to disease burden is small[7]. Growing evidence suggests that an appreciable component of the genetic contribution to 'sporadic' colorectal cancer is due to common variants with individually small effects, thereby invoking the common disease–common variant paradigm in colorectal cancer. Analysis of phenotype concordance in twins estimates the heritability of colorectal cancer on the liability scale to be around 0.35 (REF. 8).

Until mid-2007, no common variants contributing to colorectal cancer risk had been identified and consistently replicated. However, ten common low-penetrance variants contributing to colorectal cancer risk have since been identified using genome-wide association (GWA) studies, and replicated through genotyping tens of thousands of individuals. This has opened the door to unprecedented advances in our understanding of the role of common genetic variation in colorectal cancer. Initial results indicate that these variants exert only subtle effects on cancer risk, probably through influences on gene regulatory regions. The findings also offer the possibility of developing multi-locus prediction models of genetic risk that could be combined with conventional risk factors. Such risk stratification models could be used to tailor the intensity or frequency of screening to the level of genetic risk, with the ultimate aim of reducing colorectal cancer mortality. Furthermore, the development of drugs that target specific pathways, for example, the transforming growth factor beta (TGFβ) superfamily signalling pathway[9], might also enable rational drug discovery for both established cancer and for chemoprevention. Here, we review what has been learned during the last 18 months in the contribution of common genetic variation to the aetiology of colorectal cancer, and we discuss the challenges and opportunities that lie ahead.

GWA studies for colorectal cancer

Despite the known genetic contribution to colorectal cancer, mapping the contributing loci has been challenging. However, recent progress through the application of GWA approaches has identified a number of common variants involved in the aetiology of colorectal cancer. A number of key enabling factors have made this feasible. First, recent years have seen the assembly of large sample sets from well-characterized colorectal cancer case–control series, with sufficient power to detect small effect sizes and account for the large number of statistical tests performed. Second, the HapMap project[10] has enabled efficient selection of tagging SNPs (tSNPs) from across the genome. Finally, technological advances in high-order genotyping platforms have facilitated rapid, cost-effective and reproducible genotyping of large numbers of markers in large sample sets.

Table 1 | **Loci associated with colorectal cancer risk from GWA studies**

Gene* (or locus)	Chr	SNP	Study population for GWA study	Number of phases	Sample size (cases/controls)		Effect size: OR (95% CI)	Allele frequency	PAR‡ (%)	Refs
					GWA study	Total				
POU5F1P1, DQ515897, MYC	8	rs6983267	England	2	940/965	8,264/6,206	1.21 (1.15–1.27)	0.51	9.7	15
POU5F1P1, DQ515897, MYC		rs10505477	Canada	5	1,226/1,239	7,480/7,779	1.17 (1.12–1.23)	0.50	7.8	16
POU5F1P1, DQ515897, MYC		rs7014346	Scotland	3	1,012/1,012	14,500/13,294	1.19 (1.14–1.24)	0.37	6.6	13
SCG5, GREM1, FMN1	15	rs4779584	England	2	718/960	7,922/6,741	1.26 (1.19–1.34)	0.18	4.5	18
SMAD7 (intron 3)	18	rs4939827	England	2	940/965	8,413/6,949	1.18 (1.12–1.23)	0.52	8.6	17
SMAD7 (intron 3)		rs4939827	Scotland	3	1,012/1,012	14,500/13,294	1.20 (1.16–1.24)	0.51	9.2	13
LOC120376, FLJ45803, c11orf53, POU2AF1	11	rs3802842	Scotland	3	1,012/1,012	14,500/13,294	1.12 (1.07–1.17)	0.29	3.4	13
c8orf53, EIF3H	8	rs16892766	England	4	940/965	18,831/18,540	1.25 (1.19–1.32)	0.07	1.7	19
FLJ3802842	10	rs10795668	England	4	940/965	18,831/18,540	1.12 (1.10–1.16)	0.67	7.4	19
BMP4	14	rs4444235	United Kingdom	3	1,952/1,977	20,288/20,971	1.11 (1.08–1.15)	0.46	4.8	20
CDH1	16	rs9929218	United Kingdom	3	1,952/1,977	20,288/20,971	1.10 (1.06–1.12)	0.71	6.6	20
RHPN2	19	rs10411210	United Kingdom	3	1,952/1,977	20,288/20,971	1.15 (1.10–1.20)	0.90	11.9	20
BMP2	20	rs961253	United Kingdom	3	1,952/1,977	20,288/20,971	1.12 (1.08–1.16)	0.35	4.0	20

*In each case the genes or ORFs given are within the linkage disequilibrium (LD) block tagged by the associated SNPs or there is circumstantial evidence of proximity and even location within the genomic structure of the gene (as is the case with *SMAD7*). However, considerable experimentation will be required to definitively establish which genes are responsible. It is even feasible that the associated SNP is tagging a regulatory region, but the functional effects are in a distant gene that is not part of that LD block. ‡The population attributable risk [PAR = AF(OR – 1)/(AF*(OR – 1) + 1)] (AF, allele frequency) is defined as the reduction in the incidence of the disease if the population were not exposed to the risk allele. Although widely used in epidemiology it has dubious interpretation and questionable practical benefit in the genetics context. Chr, Chromosome; CI, confidence interval; OR, odds ratio.

General issues of GWA design, including choice of genotyping platform, SNP selection and stringent quality control, are discussed elsewhere[11,12]. GWA studies in colorectal cancer have employed multi-phased designs to increase statistical power in conjunction with selection strategies for cases in the first phase of the study. Rigorous selection of phenotype can help to boost power. For example, early onset patients are likely to be enriched for genetic contribution as environmental exposure accumulates with age[13]. Larger environmental variance constrains statistical power by reducing the contribution of genetic variance in older people. Using information about family history can also increase power, by enriching for cases in which genetic factors explain proportionately more of the trait variance. Thus, selection of cases with an affected first degree relative may reduce the required sample size by 50%, whereas increasing this to two affected first-degree relatives could lead to a fourfold reduction[14]. Similarly, power may be enhanced through the use of 'super-controls', such as individuals who are actively screened for bowel tumours, very elderly unaffected subjects or those without a family history of colorectal cancer. However, it remains to be confirmed whether two-phased designs in combination with selection based on enriched phenotypes provides any benefit in practice. Indeed, enriching for a given phenotype in only one phase of a two-phased design might be counterproductive if the genetic contribution to the two subgroups is markedly different.

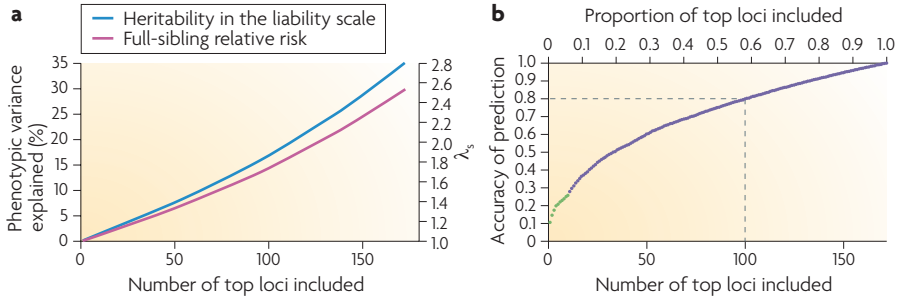
Insights into genetic architecture. To date, ten new loci have been identified through three GWA studies carried out in England[15], Scotland[13,16] and Canada[16]. The first phase of these studies involved modest sample sizes (~1,000 cases and ~1,000 controls genotyped for ~0.5 million tSNPs). Large international collaborative efforts were required to replicate the original findings to a sufficient level of statistical stringency (TABLE 1), amounting to >15,000 samples per risk locus identified. The first round of GWA data analysis resulted in identification of six associations, the reporting of which was fast-tracked in a series of publications[13,15–19]. However, the effect size of these genetic variants was modest (odds ratio (OR) ≈ 1.2). A meta-analysis of all UK GWA data[20] involved

a comprehensive analysis of 1,632 cases and 1,977 controls genotyped for 550,000 tSNPs (phase 1), followed by a second phase in which additional case–control sets were genotyped for the best-supported SNPs. A total of 13,315 individuals were genotyped for 38,710 SNPs that were common to phases 1 and 2, harvesting all associations from phase 1 with a *p* value of less than 0.039. All SNPs showing an association with colorectal cancer in the meta-analysis ($p < 10^{-5}$) were then systematically assessed in additional, independent case–control series from various populations, yielding four novel risk loci. As might be anticipated, these had even smaller effect sizes (OR ≈ 1.1)[20] than those that had been fast-tracked previously. Taking all ten of these loci together, these explain approximately 6% of the full-sibling relative risk (λ_s; one way of measuring excess familial risk), corresponding to ~1.26% of the phenotypic variance in the liability scale and 0.04% in the observed scale. Further collaborative efforts that involve larger sample sets and combine available GWA data will hopefully lead to the identification of new variants with even smaller effects. Variants with larger effects that were not captured in phase 1 of the GWA studies to date could also be identified in such studies.

For illustrative purposes, and making the assumption that common genetic variants account for all the genetic contribution, we constructed a model from Scottish GWA data and estimated that up to ~170 common independent variants explain the observed genetic contribution to colorectal cancer (BOX 1). ~170 markers is only an indicative number of common variants (minor allele frequency (MAF) > 0.05), because the contribution of rarer, or private, variants and their effect on risk are unknown. In a multi-locus model, an estimated 100 SNPs are required to achieve 80% accuracy of prediction of genetic risk. These SNPs would explain ~17% of the phenotypic variance in the liability scale, thereby providing useful predictive value of genetic risk. Hence, striving to identify all of the genetic variance is not necessarily required to provide potential public health benefits.

Pathways to colorectal cancer susceptibility. The ten genetic associations identified to date have the potential to lead to novel insights into the molecular aetiology of colorectal cancer. The causal variants have yet to be identified at any of these loci but, intriguingly, none of the tSNPs is in a coding region. Furthermore, common coding

Box 1 | **Number of loci and prediction of genetic risk**



The number of loci that account for all the common genetic variance of colorectal cancer can be estimated. This requires the plausible assumption that top-ranking SNPs from our recent genome-wide association (GWA) study[13] in a Scottish population-based case–control series do indeed contribute to the genetic variance of the disease. For the purposes of providing this indicative number of responsible variants, we also assume that common genetic variants (minor allele frequency > 0.05) account for the vast majority of the genetic aetiology.

Heritability (h_O^2) of the observed scale can be expressed as[31]:

$$h_O^2 = \left(\frac{K}{1-K}\right) \times \left(\frac{\prod_{i=1}^{n}[1 + p_i(\lambda_i^2 - 1)]^2}{\prod_{i=1}^{n}[1 + p_i(\lambda_i - 1)]^4} - 1\right)$$

where *K* is colorectal cancer prevalence in the population (0.004 in Scotland, from the Information Service Division Scotland colorectal cancer data 2005); p_i and λ_i are the frequency and relative risk of the risk allele, respectively — estimated from our Scottish GWA study data[13]. Derivation of this formula was based on a multiplicative model on the risk scale, both within and between loci (additive model on the log(risk) scale).

Hence, we estimate that ~172 common SNPs account for all of the genetic variance for colorectal cancer. Rare or private mutations would serve to change this estimate, depending on the overall number and effect size of each of such variants in the population.

Number of loci required for genetic risk prediction in populations

Next, we estimated the accuracy of genetic risk prediction using increasing number of loci. We estimated the true probability of disease using effect sizes and allele frequencies from our GWA study[13]:

$$P(D_i \mid G_i) = f_0 \prod_{j=1}^{n} \lambda_j^{x_{ij}}$$

We also calculated the same probability when including only a subset of those SNPs in the model:

$$\hat{P}(D_i \mid G_i) = f_0 \prod_{j=1}^{m} \lambda_j^{x_{ij}}$$

where f_0 is the probability of disease for a person with wild-type alleles at all loci (assumed to be 1), *n* is the total number of true risk loci (which in this case is 172, see above) and *m* is the number of true risk loci that are included in the genetic risk prediction model sorted by the strength of stastical support (*m* is always less than or equal to *n*), and x_{ij} is the number of risk alleles for person *i* at locus *j*.

The ten loci published so far have an estimated accuracy of prediction of only 26% and explain 1.26% of the phenotypic variance in the liability scale (full-sibling relative risk, $\lambda_s \approx 1.06$) (see the figure, part **a**), corresponding to 0.04% in the observed scale. Furthermore, an estimated 100 SNPs are required in a model to achieve 80% accuracy of prediction of genetic risk, and these SNPs would explain ~17% of the phenotypic variance in the liability scale. Part **b** of the figure shows the accuracy of prediction of genetic risk with increasing number of loci. The contribution to the accuracy of prediction for the ten SNPs so far published from GWA data is represented by the green segment of the curve. It is manifest from the plateau of the plot in part **b** that there is only limited incremental value in identifying >100 SNPs.

The accuracy of prediction was estimated by calculating the correlation between the logarithms of true and predicted disease probabilities[31]. The relationship between the narrow sense heritability on the unobserved and presumably normally distributed liability scale and the heritability in the observed scale is:

$$h_O^2 = \frac{h^2 z^2}{K(1-K)}$$

where *z* is the height of the standard normal curve at the threshold that truncates the proportion *K*[43]. For simplicity, we assume that all the genetic variance for colorectal cancer is additive.

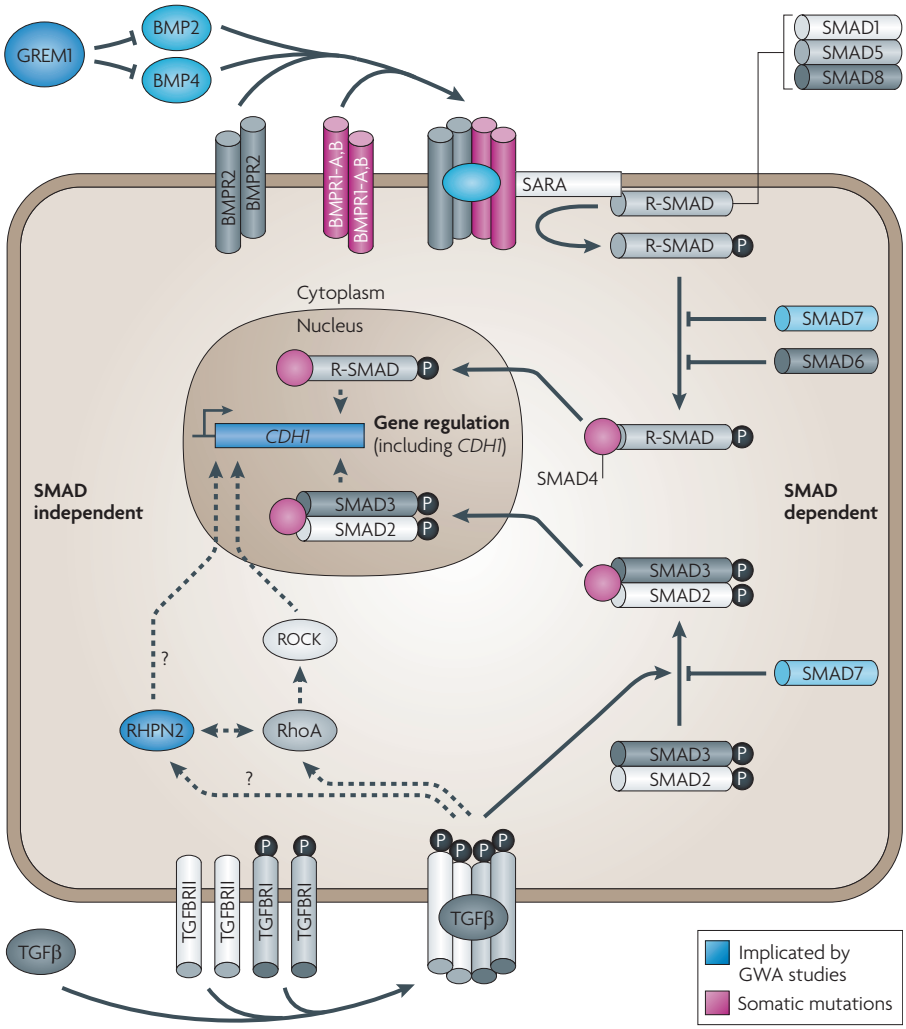


Figure 1 | **The TGFβ signalling pathway and its role in colorectal cancer.** When activated, transforming growth factor beta (TGFβ) isoforms bind cell surface receptors (TGFBRI, TGFBRII, TGFBRIII) in a highly cooperative process — only the first two are illustrated here, for simplicity — which act in consort to induce intracellular phosphorylation (P) of SMAD2 and SMAD3. These in turn bind the coSmad SMAD4, and translocate to the nucleus to drive Smad-responsive gene expression. TGFβ can also initiate Smad-independent pathways[34]. These include the mitogen-activated protein kinase (MAPK) and PP2A–p70S6 pathways (not shown here) and the activation of Rho-like GTPases, such as Rho-A. ROCK is a Rho-associated kinase, and RHPN2 is a Rho-A effector[35] that seems to regulate gene expression responses to TGFβ signalling. Silencing of Rho-A expression in cancer cell lines has been shown to increase levels of E-cadherin[36], the protein product of *CDH1*, another gene shown to be associated with colorectal cancer in genome-wide association (GWA) studies. Another, mitogenic, component of TGFβ–Smad signalling is the bone morphogenetic proteins (BMPs). BMP2 and BMP4 can initiate cell signalling by binding to their type I receptors BMPR1A (also known as ALK3) and BMPR1B (also known as ALK6) or to a higher-affinity heteromeric complex formed by the type I receptors and BMPR2. Ligand binding is not sufficient to activate signalling, but requires the phosphorylation of the glycine–serine domain in the BMPRI receptors by the BMPRII receptors. The BMP receptor complex binds and induces phosphorylation of SMAD1, 5 and 8 (various combinations of these three Smads form the receptor Smad, R-SMAD, which in turn can bind SMAD4 and translocate to the nucleus where it drives transcription). Inhibitory Smads have also been characterized, including SMAD6, which predominantly inhibits BMP stimuli, and SMAD7, which inhibits TGFβ signals. SMAD7 binds to the type I receptors and prevents recruitment and phosphorylation of R-SMADs. In addition, GREM1 is a BMP antagonist that influences bioavailability of BMP2 and BMP4. The genes in the TGFβ pathway demonstrated by recent GWA efforts to influence colorectal cancer risk are shaded blue. It is important to note that there is a requirement for a variety of transcription factors and co-factors that are not discussed further here. Genes previously implicated in the colorectal cancer susceptibility syndrome juvenile polyposis are shaded pink, namely SMAD4 (REF. 37) and BMPR1A[38]. In addition, somatic mutations in colorectal cancer have been described for SMAD4 (REF. 39), TGFBR2γ and TGFBR1 (REF. 42).

sequence variants have so far not been identified by SNP fine mapping or by resequencing the relevant loci[20]. This suggests that the associated SNPs are tagging variants that influence gene expression. The effect size on cancer risk is modest, and so it is possible that the molecular consequences of the causative variation are too subtle to allow detection. However, it is also plausible that low-risk genetic variation could be associated with functional effects that are relatively large, but that have a small impact at the level of cancer risk to humans (due, for example, to pathway redundancy, counterbalancing effects of other variants and gene–environment interactions). Thus, experimentation is underway to determine the effects of genetic variation on individual gene expression. It is important to stress that the GWA studies have identified loci associated with colorectal cancer susceptibility and these need not necessarily contribute to tumour progression. However, five of the ten SNPs identified so far tag linkage disequilibrium (LD) blocks that include, or are near to, genes of the TGFβ superfamily signalling pathway, which has been previously implicated in tumour biology. These TGFβ signalling components include *SMAD7* (REFS 13,17), *GREM1* (REF. 18), the bone morphogenetic protein genes *BMP2* and *BMP4*, and *RHPN2* (REF. 20) (FIG. 1).

This overrepresentation of TGFβ components suggests a key role for perturbations of this pathway in colorectal cancer susceptibility, implicating for the first time the TGFβ pathway in common inherited predisposition to colorectal cancer. The hypothesis that the TGFβ superfamily (TGFβ proteins, BMPs and activins) and affiliated proteins have a role in colorectal cancer is attractive, because TGFβ superfamily proteins are known to play an important part in developmental biology, cell proliferation, differentiation and migration. Rare, high-penetrance variants in *SMAD4* and the BMP receptor *BMPR1A* (also known as *ALK3*) are responsible for juvenile polyposis, an autosomal dominant colorectal polyposis syndrome with a high risk of colorectal cancer[21]. Nonrandom somatic mutations of *SMAD4* and the TGFβ receptor *TGFBR2* have been identified in colorectal cancer tissue; somatic *TGFBR1* mutations are also reported, and allele-specific expression of *TGFBR1* may contribute to germ line susceptibility[22]. Several TGFβ superfamily components have been ascribed tumour suppressor roles in view of their induction of cell cycle arrest and inhibition of cell proliferation. Interestingly, the cancer initiation properties seem to be distinct from

those of progression, as tumour cells become resistant to TGF-β signalling, eventually overexpressing components of that pathway leading to enhanced tumour growth and metastatic potential[23].

There are considerable challenges in elucidating the effect of common genetic variation on expression of individual components of the TGFβ signalling pathway, as well as unravelling the complex functional consequences of perhaps quite subtle perturbations of the pathway at multiple points. Nonetheless, such insights should inform future rational drug development, probably by exploiting 'pathway medicine' approaches that target multiple components of the TGFβ signalling pathway.

Loci tagged by other SNPs also highlight candidate cancer susceptibility genes, including the cadherin *CDH1* (REF. 24), eukaryotic translation initiation factor 3 (*EIF3H*)[25,26] and the predicted gene *FLJ3802842* (REF. 27). The association at chromosome 8q24 is worthy of specific mention, owing to implication of the locus in several cancer types[28,29]. The LD block associated with colorectal cancer includes *POU5F1P1*, a pseudogene of the candidate stem cell gene *POU5F1*(also known as *OCT4*). However, no gene product of *POU5F1P1* or causative sequence variant has yet been identified. Another possible causal mechanism could be through effects on *MYC* expression (~340 kb distal), but there is no apparent relationship between 8q24 genotype and *MYC* expression in tumours or in HapMap lymphoblastoid cell lines[16]. There is little doubt that elucidating the functional consequences of genetic variants will be challenging. However, novel insights into disease causation are already being revealed and will lead to greater understanding of the complexities of colorectal cancer risk at the biological level.

Clinical and public health implications

There are two ways in which understanding the molecular genetic aetiology of colorectal cancer could affect chemoprevention strategies. In the longer term, understanding the functional effects on the pathways involved could lead to identification of novel small molecule targets. However, in the more immediate future, identifying population groups at higher risk could result in a widening of the therapeutic index of established chemopreventative agents, such as aspirin. The 'number needed to prevent' could be reduced, thereby balancing the risk of gastrointestinal haemorrhagic complications against the need to reduce a predicted high risk of colorectal cancer.

Another tangible outcome that is relevant to public health is the potential for genetic risk stratification within populations. Thus, the invasiveness or frequency of surveillance could be tailored to the predicted level of risk imparted by genotypes at multiple loci within population subgroups. Using plausible assumptions, we estimate that the ten common variants currently identified have an accuracy of prediction of genetic risk of only 26% (BOX 1). By comparison, the accuracy of prediction of genetic risk using the family history of parents is $\sqrt{0.5h^2}$ (REFS 30,31), which for colorectal cancer is ~42% (h^2 represents heritability). However, using only parental disease history for risk prediction in the offspring assigns equal genetic risk to all offspring. As siblings inherit different alleles and combinations thereof from the four parental chromosomes, genomic profiling using a sufficient number of genetic markers could, in principle, provide better predictions of individual genetic risk than those based on family history alone. We estimate that prediction models of colorectal cancer risk that are based on common genetic variants will require ~100 SNPs for an accuracy of genetic risk prediction of 80% (BOX 1). Environmental risk factors are major contributors to disease variance, and so inclusion of variables that reflect these environmental exposures (for example, age and gender) in the prediction model will further improve the overall predictive value of disease risk.

Challenges and opportunities

Despite advances in our understanding of the role of common genetic variation in colorectal cancer aetiology, considerable challenges lie ahead. These include: identifying the causative variants responsible for association signals; definitively establishing the role of particular genes and elucidating the functional consequences of genetic variation; and determining the contribution of these loci to cancer risk in different ethnic groups. Furthermore, elucidating most of the genetic component, which is currently undiscovered, will be challenging and will probably be resistant to currently available methodologies.

The identification of causal variants will require a multidisciplinary approach involving a range of expertise, including molecular and cell biology, animal models, statistical genetics, population genetics and bioinformatics. More than one gene is frequently tagged in the LD block indicated by the SNP association. As SNPs are highly correlated, it is unlikely that the causative variant can be distinguished from neighbouring SNPs by genotyping large sample sets. Combining functional analysis with deep resequencing of implicated regions using multi-ethnic cohorts with different population histories (and therefore differing LD structures) might be a useful approach, especially in black African populations. Within 100 kb of the SNP that is at the peak of the 11q23 association signal for colorectal cancer susceptibility there are four ORFs and a predicted microRNA binding site. Functional analysis of these genes and deep resequencing of Asian and European populations might help to pinpoint the causative variant in view of the population-specific differences in colorectal cancer risk[13]. The latest generation of high-throughput resequencing technologies offer the potential to identify multiple low-frequency causative variants. Furthermore, massive parallel sequencing of RNA will allow systematic study of the transcriptome in colorectal epithelium in relation to genotypes at variants that have been shown to contribute to cancer risk[32].

In terms of the genetic architecture of colorectal cancer susceptibility, the variants responsible for the extremes of the L-shaped distribution of effects have begun to be elucidated. That is, highly penetrant low-frequency variants and common variants with low penetrance have been identified.

Glossary

Excess familial risk
The increased risk of developing the disease in a relative of an affected individual. It is usually, and more appropriately, referred to for a specific type of relative. For example, the full-sibling relative risk (λ_s) is the increased risk of developing a disease for a full sibling of an affected person compared with the risk of a person from the general population.

Heritability
The proportion of phenotypic variance that is explained by inherited genetic factors.

Liability scale
The assumed and unobserved normally distributed risk scale. In the case of a disease, those individuals with a liability score above a specific threshold will have the disease.

Observed scale
For a disease trait, the observed scale of the phenotype is either disease or non-disease. The heritability in the observed scale depends on the disease prevalence.

Odds ratio
A measure of the effect size. It is the odds of exposure (that is, a specific allele) among the cases divided by the odds of exposure among the controls. In case–control studies, the odds ratio is used as an approximation to the relative risk.

However, the middle part of the distribution is so far completely unexplored. One reason for this is because current SNP arrays are poor at tagging low-frequency variants and so have low power to detect moderately rare variants, even those with moderate to large effects. Indeed, rare low-penetrance variants may prove almost undetectable. Technical constraints have also limited our exploration of this middle zone, because sequencing strategies thus far have required isolation of specific regions of interest, an approach that does not lend itself to analysis of large numbers of samples. However, these problems are progressively being overcome, and identifying 'rare-ish' variants with moderate effects (that is, MAF ≈ 0.01; OR ≈ 5) will also benefit from large-scale resequencing and genotyping efforts in large consortia, and from the 1000 Genomes Project. Owing to cost constraints, these efforts might initially be concentrated on regions that are identified by GWA studies[33] because they are good candidates. As the costs of whole-genome sequencing come down, systematic searches for private mutations in a clinical genetics setting could become a possibility in the foreseeable future.

Albert Tenesa and Malcolm G. Dunlop are at the Colon Cancer Genetics Group, University of Edinburgh and MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, 4th Floor MRC Human Genetics Unit, Western General Hospital, Crewe Road South, Edinburgh EH4 2XU, UK.

e-mails: Albert.Tenesa@ed.ac.uk; Malcolm.Dunlop@hgu.mrc.ac.uk

doi:10.1038/nrg2574
Published online 12 May 2009

1. World Health Organization. *World Cancer Report* (eds Stewart B. W. & Kleihues P.) 13 (IARC, Lyon, 2003).
2. Barnetson, R. A. *et al.* Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. *N. Engl. J. Med.* **354**, 2751–2763 (2006).
3. Towler, B. P., Irwig, L., Glasziou, P., Weller, D. & Kewenter, J. Screening for colorectal cancer using the faecal occult blood test, hemoccult. *Cochrane Database Syst. Rev.* **2007**, CD001216 (2000).
4. Jarvinen, H. J. *et al.* Controlled 15-year trial on screening for colorectal cancer in families with hereditary nonpolyposis colorectal cancer. *Gastroenterology* **118**, 829–834 (2000).
5. Bost, B., de Vienne, D., Hospital, F., Moreau, L. & Dillmann, C. Genetic and nongenetic bases for the L-shaped distribution of quantitative trait loci effects. *Genetics* **157**, 1773–1787 (2001).
6. Mackay, T. F. C. The genetic architecture of quantitative traits. *Annu. Rev. Genet.* **35**, 303–339 (2001).
7. Foulkes, W. D. Inherited susceptibility to common cancers. *N. Engl. J. Med.* **359**, 2143–2153 (2008).
8. Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer — analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78–85 (2000).
9. Yingling, J. M., Blanchard, K. L. & Sawyer, J. S. Development of TGF-β signalling inhibitors for cancer therapy. *Nature Rev. Drug Discov.* **3**, 1011–1022 (2004).
10. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
11. Kruglyak, L. The road to genome-wide association studies. *Nature Rev. Genet.* **9**, 314–318 (2008).
12. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9**, 356–369 (2008).
13. Tenesa, A. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nature Genet.* **40**, 631–637 (2008).
14. Antoniou, A. C. & Easton, D. F. Polygenic inheritance of breast cancer: implications for design of association studies. *Genet. Epidemiol.* **25**, 190–202 (2003).
15. Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature Genet.* **39**, 984–988 (2007).
16. Zanke, B. W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature Genet.* **39**, 989–994 (2007).
17. Broderick, P. *et al.* A genome-wide association study shows that common alleles of *SMAD7* influence colorectal cancer risk. *Nature Genet.* **39**, 1315–1317 (2007).
18. Jaeger, E. *et al.* Common genetic variants at the *CRAC1* (*HMPS*) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nature Genet.* **40**, 26–28 (2008).
19. Tomlinson, I. P. *et al.* A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nature Genet.* **40**, 623–630 (2008).
20. Houlston, R. S. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature Genet.* **40**, 1426–1435 (2008).
21. Howe, J. R. *et al.* The prevalence of *MADH4* and *BMPR1A* mutations in juvenile polyposis and absence of *BMPR2*, *BMPR1B*, and *ACVR1* mutations. *J. Med. Genet.* **41**, 484–491 (2004).
22. Valle, L. *et al.* Germline allele-specific expression of *TGFBR1* confers an increased risk of colorectal cancer. *Science* **321**, 1361–1365 (2008).
23. Blobe, G. C., Schiemann, W. P. & Lodish, H. F. Role of transforming growth factor beta in human disease. *N. Engl. J. Med.* **342**, 1350–1358 (2000).
24. Guilford, P. *et al.* E-cadherin germline mutations in familial gastric cancer. *Nature* **392**, 402–405 (1998).
25. Okamoto, H., Yasui, K., Zhao, C., Arii, S. & Inazawa, J. *PTK2* and *EIF3S3* genes may be amplification targets at 8q23-q24 and are associated with large hepatocellular carcinomas. *Hepatology* **38**, 1242–1249 (2003).
26. Savinainen, K. J. *et al.* Expression and copy number analysis of TRPS1, EIF3S3 and MYC genes in breast and prostate cancer. *Br. J. Cancer* **90**, 1041–1046 (2004).
27. Shima, H. *et al.* Loss of heterozygosity on chromosome 10p14-p15 in colorectal carcinoma. *Pathobiology* **72**, 220–224 (2005).
28. Haiman, C. A. *et al.* A common genetic risk factor for colorectal and prostate cancer. *Nature Genet.* **39**, 954–956 (2007).
29. Ghoussaini, M. *et al.* Multiple loci with different cancer specificities within the 8q24 gene desert. *J. Natl Cancer Inst.* **100**, 962–966 (2008).
30. Falconer, D. S. & Mackay, T. F. C. *Introduction to Quantitative Genetics* (Longman, 1996).
31. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–1528 (2007).
32. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
33. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genet.* **40**, 695–701 (2008).
34. Derynck, R. & Zhang, Y. E. Smad-dependent and Smad-independent pathways in TGF-β family signalling. *Nature* **425**, 577–584 (2003).
35. Peck, J. W., Oberst, M., Bouker, K. B., Bowden, E. & Burbelo, P. D. The RhoA-binding protein, rhophilin-2, regulates actin cytoskeleton organization. *J. Biol. Chem.* **277**, 43924–43932 (2002).
36. Chang, Y. W., Marlin, J. W., Chance, T. W. & Jakobi, R. RhoA mediates cyclooxygenase-2 signaling to disrupt the formation of adherens junctions and increase cell motility. *Cancer Res.* **66**, 11700–11708 (2006).
37. Howe, J. R. *et al.* Mutations in the *SMAD4*/*DPC4* gene in juvenile polyposis. *Science* **280**, 1086–1088 (1998).
38. Huang, S. C. *et al.* Genetic heterogeneity in familial juvenile polyposis. *Cancer Res.* **60**, 6882–6885 (2000).
39. Woodford-Richens, K. L. *et al.* SMAD4 mutations in colorectal cancer probably occur before chromosomal instability, but after divergence of the microsatellite instability pathway. *Proc. Natl Acad. Sci. USA* **98**, 9719–9723 (2001).
40. Kodach, L. L. *et al.* The bone morphogenetic protein pathway is inactivated in the majority of sporadic colorectal cancers. *Gastroenterology* **134**, 1332–1341 (2008).
41. Parsons, R. *et al.* Microsatellite instability and mutations of the transforming growth factor β type II receptor gene in colorectal cancer. *Cancer Res.* **55**, 5548–5550 (1995).
42. Pasche, B. *et al.* Somatic acquisition and signaling of *TGFBR1**6A in cancer. *JAMA* **294**, 1634–1646 (2005).
43. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nature Rev. Genet.* **9**, 255–266 (2008).

Acknowledgements
Work that forms the basis of this discussion is supported by Cancer Research UK (C348/A8896, C48/A6361), and the Scottish Executive Chief Scientist's Office (CZB/4/449) — a centre grant from CORE as part of the Digestive Cancer Campaign. We acknowledge those in the Colon Cancer Genetics Group who have contributed to the work reviewed, particularly S. Farrington and H. Campbell, as well as our collaborators R. Houlston and I. Tomlinson and their groups. We thank R. Wilson and N. Cartwright, all of who worked on the COGS and SOCCS administrative teams, R. Cetnarskyj and the research nurse teams who recruited in Scotland, and all clinicians throughout Scotland at collaborating centres. We acknowledge N. Wray and P. Visscher for comments on BOX 1 and for sharing unpublished data on prediction models of genetic risk.

FURTHER INFORMATION
Malcolm G. Dunlop's homepage:
http://www.hgu.mrc.ac.uk/people/m.dunlop.html

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Maintaining the brain: insight into human neurodegeneration from *Drosophila melanogaster* mutants

Derek Lessing and Nancy M. Bonini*‡*

Abstract | The fruitfly *Drosophila melanogaster* has enabled significant advances in neurodegenerative disease research, notably in the identification of genes that are required to maintain the structural integrity of the brain, defined by recessive mutations that cause adult onset neurodegeneration. Here, we survey these genes in the fly and classify them according to five key cell biological processes. Over half of these genes have counterparts in mice or humans that are also associated with neurodegeneration. Fly genetics continues to be instrumental in the analysis of degenerative disease, with notable recent advances in our understanding of several inherited disorders, Parkinson's disease, and the central role of mitochondria in neuronal maintenance.

RNAi
RNAi in the adult fly is achieved with transgenic constructs expressing an inverted repeat sequence targeted to the mRNA of interest. The expressed dsRNA is processed *in vivo* into short interfering RNAs, which lead to degradation of the target gene transcripts for a loss-of-function mutant effect.

Mosaic
An animal comprised of tissue of different genotypes. In flies, mosaics are generated by site-specific recombination, to yield homozygous mutant tissue or cells in an otherwise heterozygous animal.

**Howard Hughes Medical Institute and*
‡Department of Biology, University of Pennsylvania, 306 Leidy Laboratoriess, 433 South University Avenue, Philadelphia, Pennsylvania 19104, USA.
Correspondence to N.M.B
e-mail:
nbonini@sas.upenn.edu
doi:10.1038/nrg2563
Published online 12 May 2009

Neurodegenerative diseases typically afflict adults in mid-life, and are characterized by motor or cognitive symptoms that get progressively worse with age and that usually reduce life expectancy. Human neurodegenerative disease can result from a variety of environmental and genetic causes. Genetic factors in particular have been instrumental in developing our understanding of the aetiology and progression of such diseases, and can range from mutations that increase the risk for a particular disorder, to mutations that are the sole, direct cause of a disease. As with cancer, which is another collection of related diseases, neurodegeneration can result from dominant or recessive mutations. In this Review, we focus on the role of *Drosophila melanogaster* in characterizing 'neurodegeneration suppressor genes', which we see as analogous to cancer tumour suppressor genes. Defined by recessive loss-of-function mutations that cause neurodegeneration, such genes are required for maintaining the integrity of the adult central nervous system (CNS).

D. melanogaster has been a key tool in much of the work to identify genes involved in neuronal integrity and to discover their functions. Gene discovery is quick and straightforward in the fly, as is the analysis of how separate genes function together, two points that we expand on in the following section. Other useful *D. melanogaster* techniques are also addressed below: the expression of transgenes that can be directed precisely in space and time, including RNAi constructs; and genetically mosaic flies, which are useful for identifying the location of gene function.

Although superficially different, humans and flies are remarkably similar in key respects. Crucial signalling pathways in development, cancer and innate immunity are conserved between the fly and humans[1,2]. The CNS of invertebrates and vertebrates share a common evolutionary origin[3], and the fly has been used successfully for the genetic analysis of complex behaviours ranging from sleep[4] to learning and memory[5] to aggression[6]. Of the human protein sequences associated with disease in the Online Mendelian Inheritance in Man (OMIM) database, 74% have highly related sequences in the fly[7] (see also the Homophila web site). Moreover, a number of dominantly inherited human neurodegenerative diseases, such as those caused by polyglutamine repeat expansions, have been successfully modelled in *D. melanogaster* by transgenic expression of the human disease genes (reviewed in REF. 8). Subsequent screens for fly genes that modify the effects of toxic human proteins have been enormously successful, leading to new insight into these diseases and demonstrating the parallels between human and fly neurodegeneration.

Here we focus on a complementary approach, distinct from transgenic modelling of human neurodegenerative diseases: the identification of loss-of-function mutations in endogenous fly genes that cause brain degeneration. More than half of such genes either have human orthologues linked to disease or provide information about conserved processes that are required for maintaining the structural integrity of the CNS — a tissue with complex cell types that, for the most part,

undergo no cell division or renewal during the lifetime of the animal. The last few years in particular have seen rapid advances in this field with screens to identify new fly neurodegeneration genes, and to elucidate the roles of genes previously known to be crucial in human neurodegeneration. Below, we highlight examples showing how fly genetics can be used to address human disease. From a survey of currently known fly neurodegeneration genes, five key cell biological processes stand out as vital for maintaining CNS integrity. Finally, we address insights from *D. melanogaster* mutants into the role of glia and cell–cell interactions in neuronal integrity. Parallels between the fly, mouse and human underscore the conservation of gene function in maintaining the nervous system, and suggest that further investigations in the fly will reveal additional insight relevant to the entire field.

Of screens and genes

Genes required for the maintenance of the adult fly brain have been discovered by three approaches: screens, candidate genes and fortuitous mutations. In screens to identify genes associated with neurodegeneration, the classic approach has been to select viable adult mutant fly lines with a behavioural defect, and then screen by histology for CNS degeneration. Interestingly, the first assay devised for the genetic analysis of fly behaviour was instrumental in the first discovery of a mutation that causes neurodegeneration in the adult fly, in the gene *drop-dead* (*drd*)[8,9]. Later screens in Seymour Benzer's laboratory focused on mutants with shortened lifespans[10,11]. In the late 1970s, Heisenberg and Bohl screened for fly mutants defective in phototaxis and then performed a type of high-throughput mass histology on fly heads[12]. Recent screens have looked for degeneration in mutant flies that become paralyzed with a change in temperature[13] or with mechanical stress[14], or that have altered circadian rhythms[15]. The candidate gene approach is an alternative to screens — fly orthologues of mouse or human genes that are known to cause neurodegeneration have been identified and characterized using this technique. Finally, many fortuitous mutations isolated in unrelated studies have shown unexpected loss of integrity of the brain.

Currently, 44 genes required for CNS integrity in *D. melanogaster* have been characterized (Supplementary information S1 (table); an expanded, updated version of the table is available at the Bonini Laboratory website). A subset of these genes, those discussed most extensively in this Review, is shown in TABLE 1. A gene is included in Supplementary information S1 (table) if the recessive, loss-of-function mutation causes progressive, adult onset histological abnormalities in the fly brain. Retinal degeneration, a related topic, is addressed in BOX 1. With two exceptions, the genes in Supplementary information S1 (table) have readily identifiable orthologues in mice and humans, and over half are related to mouse or human genes that are also associated with neurodegeneration. Below, three genes from the table are used to illustrate the advantages of *D. melanogaster* genetics that are crucial in addressing neurodegenerative disease.

swiss cheese. One of the mutations discovered in the Heisenberg and Bohl screen[12] was named *swiss cheese* (*sws*), after the holes discovered in sections of mutant brains[16] (see BOX 2 for a discussion of techniques). Characterization of *sws* in 1997 revealed that it encodes what was then a novel protein[16], but 1 year later the human orthologue was found to be neuropathy target esterase[17] (NTE; also known as PNPLA6). In mammalian cells, the phospholipase activity of NTE breaks down the membrane lipid phosphatidylcholine to glycerophosphocholine[18]; *sws* mutant flies have excess phosphatidylcholine[19], which may affect membrane properties in a deleterious manner. SWS and NTE also share a conserved domain that acts as a non-canonical regulatory subunit of cyclic AMP-dependent protein kinase (PKA)[20]. Inactive PKA is a tetramer of two regulatory subunits and two catalytic subunits; when the regulatory subunits bind cAMP, they release the catalytic subunits, which then become active. An N-terminal transmembrane domain anchors SWS (NTE in mammals) to the cytoplasmic face of endoplasmic reticulum membranes[18] and thus may sequester PKA catalytic subunits there. The PKA regulatory activity of SWS (discussed further in a later section) has at least a partial role in neurodegeneration, as exogenously expressed SWS with a mutation in the binding domain cannot fully rescue the *sws* mutant defect[20]. Presumably, the partial rescue observed with this transgene is mediated through the intact phospolipase activity.

Biochemical activities of SWS and NTE seem to be conserved, but is this true of the role of NTE in maintaining CNS integrity with age? Knockout of *Nte* in the mouse is lethal, but if loss of the gene is restricted to the CNS, then the mouse survives to adulthood but suffers from neurodegeneration[21]. Underscoring functional conservation, mutations in human *NTE* cause spastic paraplegia 39, a hereditary motor neuron degenerative disease[22]. Furthermore, NTE is a target of organophosphates, a class of compounds that includes many pesticides and the neurotoxin sarin. Following long-term exposure, these compounds bind to the catalytic serine residue of NTE and cause axonal degeneration[18]. Thus, characterization of the *sws* mutation in flies led to a series of studies that defined biochemical activities of the NTE protein and revealed its role in disease and in response to environmental toxins.

The example of *sws* demonstrates one reason why studies of neurodegeneration in the fly are valuable: the unbiased, forward genetics screen is a classic method in *D. melanogaster* that has led directly to the identification of new neurodegeneration genes in mice and humans.

Pink1 *and* park. At least three inherited forms of parkinsonism are caused by recessive mutations. In two of these forms, loss of *PINK1* (PTEN-induced putative kinase 1) or *PARK2* (Parkinson disease 2, parkin) function causes familial juvenile onset parkinsonism[23]. Several groups have taken a candidate gene approach to studying the roles of these Parkinson's disease genes in *D. melanogaster*. One defect in *Pink1* or *parkin* (*park*) mutant flies is abnormal wing posture. Ostensibly unrelated to

Glia
Support cells for neurons.

Phototaxis
Movement towards a light source. A behaviour often used in flies to test locomotor activity and eye function.

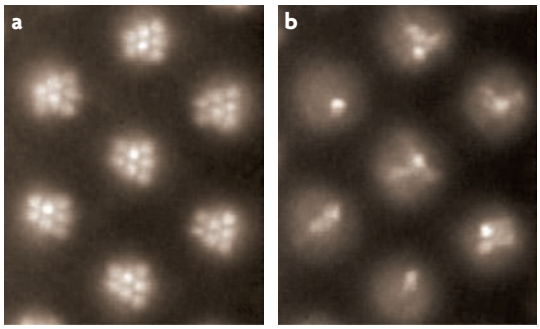
Parkinsonism
Showing symptoms characteristic of Parkinson's disease (tremor, rigidity, slowing of movement, postural instability and shuffling gait) that respond to treatment with dopamine.

Table 1 | **Neurodegeneration due to loss-of-function mutations**

Fly gene	Protein	Mouse gene: knockout	Human gene: disease	Notes
Ace	Acetyl cholinesterase	*Ache*: delayed postnatal development, early death.	*ACHE*	• Fly: null allele is lethal. Large mutant brain clones induced in embryogenesis show degeneration[62]. Temperature-sensitive allele causes paralysis and quick death at 32°C and abnormal neuropil at 29°C[63]
Atpα	Na+/K+ ATPase α subunit	*ATP1A3**:dystonia 12. Heterozygous mice: learning defects, hyperactivity.	*Atp1a3*: neonatal lethal.	• Fly: degeneration with both recessive and dominant alleles[65] • Mouse: β-subunit gene *Atp1b2** knockout causes neurodegeneration and death at 17–18 days[94] • Human diseases caused by dominant allele
ATP6	F_1F_0-ATP synthase subunit	*mt-Atp6*	*MTATP6**: Leigh syndrome	• Fly: subtle thoracic ganglia defects; no gross histological defects in aged brains. Enhances *sesB* phenotype. Thoracic muscle degeneration. Mechanical stress sensitivity. Abnormal mitochondrial ultrastructure[68]
bubblegum (bgm)	Fatty acid CoA synthetase	*Acsbg1*	*ACSBG2*	• Fly: histological defects in optic lobe only[11] • Human: adrenoleukodystrophy is caused by a recessive X-linked *ABCD1** allele, affecting a peroxisomal transporter involved in importation or anchoring of the synthetase
Cystein string protein (Csp)	Hsp40-like; component of synaptic vesicles	*Dnajc5b*	*DNAJC5B*	• Fly: subtle synaptic degeneration seen at TEM level[58] • Mouse: knockout of paralogue *Dnajc5** dies within 3 months, with progressive neuromuscular junction degeneration and behavioural abnormalities[59]
dare	Ferredoxin reductase	*Fdxr*	*FDXR*	• Fly: hypomorphic allele; null alleles are larval lethal. Severe uncoordination[43]
drop-dead (drd)	Membrane	No significantly similar gene	No significantly similar gene	• Fly: abnormal glial morphology in young adults[76]
Eaat1	Glutamate transporter	*Slc1a3*: ataxia, abnormal Purkinje cell innervation by climbing fibres[95]	*SLC1A3*	• Fly: RNAi. Behavioural defects. Increased sensitivity to paraquat[64] • Mouse: knockout of neuronally-expressed paralogue *Slc1a1** causes neurodegeneration[96]
easily shocked (eas)	Ethanolamine kinase	*Etnk1*	*ETNK1*	• Fly: mechanical shock causes brief hyperactivity and then paralysis; possible epilepsy model. Electrophysiological defects in giant fibre pathway[33,97]
fumble (fbl)	Pantothenate kinase	*Pank1*	*PANK1*	• Fly: hypomorphic alleles. Flight and climbing defects. Sensitive to paraquat[36] • Mouse paralogue *Pank2** knockout causes retinal degeneration • Human: mutation of paralogue *PANK2** causes pantothenate kinase-associated neurodegeneration
futsch	Microtubule-associated protein 1B	*Mtap1b*: central nervous system developmental defects	*MAP1B*	• Fly: hypomorphic alleles; null alleles are lethal. Learning defect observed before brain histological defects. Partial rescue by Tau[98] • Human: dominant splicing mutation in the paralogue *MAPT** (Tau) causes frontotemporal dementia with parkinsonism
levy	Cytochrome *c* oxidase (COX) subunit VIa	*Cox6a1*	*COX6A1*	• Fly: histological defects limited to retina and optic lobes[99] • Human: Leigh syndrome can be caused by mutation of *LRPPRC** or *COX10**
parkin (park)	E3 ubiquitin ligase	*Park2**: dopaminergic neuron loss (variable), behavioural defects	*PARK2**: Parkinson disease 2	• Fly: principal pathology is in mitochondria of sperm and flight muscles. Dopaminergic neurons are smaller in size[100,101]
Pink1	Mitochondrial protein kinase	*Pink1*: normal number of dopaminergic neurons; mitochondrial defects	*PINK1**: Parkinson disease 6	• Fly: conflicting data on dopaminergic neuron loss. Male sterility and wing, muscle and mitochondrial defects[24–26]
reverse polarity (repo)	Homeodomain transcription factor	*Alx4*: developmental skeletal defects	*ALX4*: parietal foramina 2	• Fly: hypomorphic allele. Apoptotic loss of both neurons and glia limited to optic lobe[75]
SNF4Aγ	AMP-activated protein kinase γ subunit	*Prkag2*	*PRKAG2*: Wolff–Parkinson–White syndrome	• Fly: mutation affects one of three isoforms[39,102] • Human disease caused by a dominant allele
swiss cheese (sws)	Membrane lipid esterase; PKA regulatory subunit	*Nte**: hippocampus, thalamus and cerebellar degeneration[21]	*NTE**: spastic paraplegia 39	• Fly: increased apoptosis. Both neuronal and glial death[12,16,19] • Mouse: neurodegeneration owing to conditional brain-specific knockout

A subset of the *Drosophila* genes associated with progressive, adult-onset histological defects in the fly brain, shown in full in Supplementary information S1 (table). *Mouse and human genes associated with neurodegeneration and that are orthologues of the respective fly genes, called by Flybase or InParanoid or by NCBI or Homologene. Mouse knockout phenotypes are from the Mouse Genome Informatics website; human diseases from Online Mendelian Inheritance in Man. TEM, transmission electron microscopy.

Box 1 | **Retinal degeneration in *Drosophila melanogaster***



Phototransduction — the conversion of light into neural signals — occurs through a G protein-coupled pathway. Mutations in many genes in this pathway in the fly typically induce light-dependent retinal degeneration, but not neuronal loss elsewhere in the fly[82]. However, as vision in the fly is not essential for viability or reproduction, the eye is an attractive tissue in which to model general neurodegeneration. A number of tools are available for such purposes. Genetic tricks include transgenic RNAi constructs targeted specifically to the developing eye or to mature photoreceptor neurons, and the generation of homozygous mutant eyes in an otherwise heterozygous animal[83]. Either method allows perturbation of the eye without affecting the viability of the animal as a whole. Methods for assaying photoreceptor degeneration include a simple behavioural assay, such as phototaxis (healthy flies run robustly towards light), or the electroretinogram, which measures the amplitude of phototransduction and the efficacy of the synaptic output of photoreceptor neurons. Photoreceptor loss can also be seen directly by histology or, more easily, by optical neutralization, commonly called the pseudopupil preparation. This visualization of the light-gathering organelles of the photoreceptors can be performed on unfixed, intact heads, allowing easy quantification[84]. The figure shows pseudopupil preparations of a normal retina (**a**) and a degenerate retina (**b**). The table highlights six genes associated with retinal degeneration that are unlikely to be directly involved in phototransduction. Roles for these genes in maintaining central nervous system integrity in the fly have not yet been addressed but are likely to exist, owing to their widespread expression or association of orthologues with neurodegenerative disease. Images reproduced from REF 85.

Fly gene	Protein	Human gene: disease
Apc[86]	Binds β-catenin and microtubules	*APC:* familial adenomatous polyposis coli
nmnat[87]	Nicotinamide mononucleotide adenylyltransferase	*NMNAT1*: (in the *Wld*^{s} mouse, slowed axonal Wallerian degradation is caused by a gene fusion that includes *Nmnat1*)
rhomboid-7[88]	Mitochondrial fusion	*PARL*: no known disease
Scs-fp (also known as *SdhA*)[89]	Succinate dehydrogenase	*SDHA:* Leigh syndrome owing to mitochondrial complex II deficiency
Sod[90]	Cu/Zn superoxide dismutase	*SOD1:* familial amyotrophic lateral sclerosis (dominant mutations)
Tefu (also known as *ATM*)[91]	Protein kinase	*ATM:* ataxia-telangiectasia
torp4a[92]	Chaperone-like glycoprotein	*TOR1A*: dystonia 1, torsion, autosomal dominant

loss of brain integrity, the wing effect has been revelatory as it is caused by problems in underlying muscles with a high demand for energy, thus focusing attention on the role of mitochondria in Parkinson's disease.

Loss of either *Pink1* or *park* in flies results in remarkably similar effects in addition to the wing defect, including structural defects in mitochondria, muscle degeneration, shortened lifespan and male infertility[24–26]. A breakthrough in the understanding of these genes came from a simple genetic experiment in the fly: the forced expression of the normal *park* gene in a *Pink1* mutant background. This showed that *park* gene function can rescue all *Pink1* mutant defects[24–26]. This is a classic example of an epistasis experiment — the study of combinations of gene activities, through either recessive loss-of-function or dominant gain-of-function mutations — and shows that PARK functions downstream of PINK1. The result has been confirmed in human cells: expression of *PARK2* in cells derived from two patients with different *PINK1* mutations can rescue the mitochondrial defect[27].

PINK1 is a protein kinase principally localized to mitochondria, whereas PARK is principally cytoplasmic, suggesting the intriguing possibility that aberrant signalling between the mitochondria and cytoplasm may have a role in disease. Mitochondria are not static structures but are in a constant flux of fusion and fission in response to cellular conditions. Further evidence that PINK1 and PARK work together to regulate mitochondrial structure comes from genetic manipulation of mitochondrial dynamics in flies: *Pink1* and *park* mitochondrial defects can be rescued by either upregulating a mitochondrial fission protein (DRP1) or knocking down mitochondrial fusion proteins (OPA1 or MARF; also known as MFN2)[28,29]. In humans, *MFN2* (mitofusin 2) mutations cause the neurodegenerative disease Charcot–Marie–Tooth type 2A and, in the mouse, loss of *Mfn2* in the cerebellum causes aberrant mitochondrial structure and function in Purkinje cells, resulting in their degeneration[30].

The work with *Pink1* and *park* demonstrates a second reason why neurodegeneration studies in the fly are valuable: epistasis experiments are routine in *D. melanogaster* and are a powerful approach to defining functional gene order in pathways conserved in the mouse and human.

Purkinje cells
Vertebrate neurons with huge, dense dendrites that integrate complex inputs in the cerebellum and project axons to the deep motor nuclei of the brain.

Box 2 | **Techniques used to study neurodegeneration in *Drosophila melanogaster***

Histology

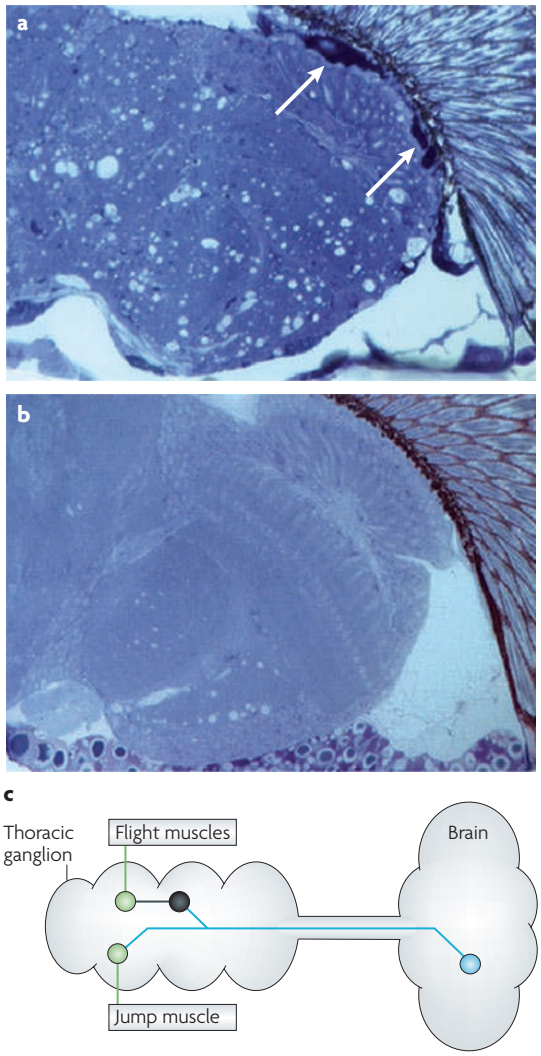
The most common approach used to assess neurodegeneration in the fly is examination of sections of the brain, usually embedded in paraffin or plastic, typically treated with a nonspecific stain or contrast such as toluidine blue or autofluorescence (see the Flybrain website). The most striking lesions in these histological sections are the vacuoles or holes in the central neuropil or outer rind where most neuron cell bodies reside, as shown in the figure in a 20-day-old *swiss cheese* mutant brain (**a**) compared with the age-matched normal brain (**b**). Arrows indicate dying glial cells. An advantage of this method is that it is a direct assay of CNS defect; a disadvantage is its low resolution as commonly practiced, as one cannot determine which neurons are dying.

Electron microscopy

A specialized form of histology, electron miscroscopy affords unparalleled resolution of sub-cellular structures and is the definitive technique for identifying autophagic defects. A disadvantage is that it is labour intensive, and antibody labelling can be difficult.

Electrophysiology

Many preparations are commonly used for electrophysiological recordings of neural activity. The giant fibre system (see the figure, part **c**) mediates the fly's escape response. With a mix of electrical, cholinergic and glutaminergic synapses, this preparation is particularly useful for testing central nervous system function in the adult[93]. The giant fibre neuron (blue) is located in the brain and, in the thoracic ganglion, it synapses with a motor neuron (green) that directs the jump muscle and with an interneuron (black) that in turn synapses with five motor neurons (only one shown for clarity) that direct a set of flight muscles. Neurons and muscles are bilaterally symmetrical and only one side is shown for simplicity. Recordings can be made from the muscles, or intracellularly from the giant fibre axon or from the motor neurons. An advantage of this technique is that it provides a direct assay of neuronal dysfunction; a disadvantage is that is labour intensive.



Behaviour

Well-characterized behaviours in the fly range from mating to aggression. Climbing (sometimes called negative geotaxis) is the most commonly assayed behaviour with respect to fly neurodegeneration for three reasons: as a test of mobility, it can reflect the ataxia that is common in human neurodegenerative diseases; large numbers of flies can be tested quickly; and the behaviour is robust (for most strains, ~90% of flies will immediately start climbing the walls of a container after they have been tapped to the bottom). An advantage of this technique is that it correlates with human disease. Disadvantages are that a loss of climbing can be due to factors other than neurodegeneration, and the behaviour can vary significantly with genetic background.

Lifespan

Not all flies that die early do so from brain degeneration, but all neurodegenerative mutants have shortened lifespans. Therefore, a straightforward first look at a mutant can be obtained with a survival curve, which can be coupled with feeding toxins such as paraquat. Advantages of this technique are that it is easy to perform and correlates with human disease; disadvantages are that it is time consuming, there are possible causes other than neurodegeneration, it can vary significantly with genetic background, and it provides little insight into pathology.

See BOX 1 for techniques associated specifically with the eye. Parts **a** and **b** of the figure are reproduced, with permission, from REF. 16 © (1997) Society for Neuroscience.

Ubiquitin–proteasome system
Members of a large family of E3 ubiquitin ligases recognize specific substrate proteins, tagging them by polyubiquitination for degradation in the proteasome, a large cylindrical protein complex.

Autophagy
More precisely, macroautophagy — the engulfment of protein aggregates or organelles by vesicles with double-bilayer membranes, which then fuse with lysosomes for degradation of their contents.

Processes crucial to CNS integrity

The genes in Supplementary information S1 (table) can be classified according to five cell biological processes, although many of these genes have multiple roles (FIG. 1). Although this analysis stems from *D. melanogaster* neurodegeneration mutants, an important theme is the commonality of these processes between flies and mammals. The protein homeostasis process includes genes that function in both the ubiquitin–proteasome system and in autophagy or lysosomal degradation, a topic reviewed elsewhere[31,32]. Genes classified in the cytoskeleton process have roles both in the function of

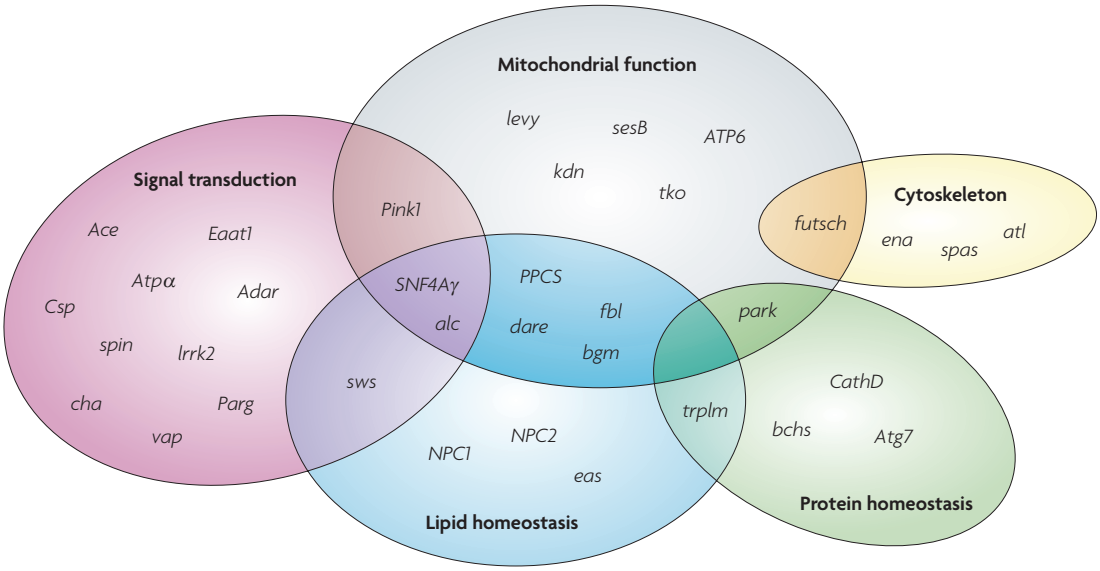


Figure 1 | **Cellular processes implicated by neurodegeneration genes.** A Venn diagram showing the relationships between five cellular processes and a suggested classification of the neurodegeneration genes from Supplementary information S1 (table), many of which have multiple roles.

microtubules (which are the principal structural element of long neurites) and in the actin-based cytoskeleton, which is vital for synapse formation and plasticity. Below, we highlight recent work on aspects of the other three processes: lipid homeostasis, signal transduction and mitochondrial function.

Lipid homeostasis. Lipids serve as the key constituent of membranes, as a source of energy and as signalling molecules. A number of *D. melanogaster* neurodegeneration mutants have been linked to pathways of lipid homeostasis. Ethanolamine kinase, encoded by *eas*[33], catalyses the first step in one pathway for synthesizing phosphatidylethanolamine, a phospholipid that in mammals is enriched in neuronal and mitochondrial membranes[34]. Exactly how phospholipid composition could affect degeneration is unclear, but one possibility involves membrane properties such as fluidity, which could influence channel or neurotransmitter functions. Furthermore, different organelles can have characteristic phospholipids in their membranes, and so altering phospholipid composition could affect, for example, organelle trafficking[35].

Three other genes linked to neurodegeneration are involved in fatty acid catabolism (FIG. 2). *fbl* (*fumble*) and *Ppcs* (phosphopantothenoylcysteine synthetase) have recently been characterized and they encode enzymes in the pathway for synthesis of coenzyme A, which is required for the first step in the degradation of fatty acids[36]. *bgm* (*bubblegum*) encodes a fatty acid coenzyme A ligase that seems to be specific for very long chain fatty acids. Degeneration in *bgm* mutants can be partially rescued by feeding the flies a fatty acid component of 'Lorenzo's Oil', a putative preventive treatment for adrenoleukodystrophy, perhaps by inhibiting the synthesis of very long chain fatty acids[11]. Together, *fbl*, *Ppcs* and *bgm* could affect neuronal membrane properties or energy availability. If the latter function is vital, then an interaction would be predicted between these genes and those encoding the mitochondrial fatty acyl translocation apparatus (FIG. 2), which consists of a translocase and enzymes which attach and remove fatty acids from the carrier molecule carnitine[37].

Cholesterol is an essential constituent of animal membranes and has been linked to Alzheimer's disease. Deposition of amyloid-β peptide into amyloid plaques, the hallmark of Alzheimer's disease, is enhanced by high cholesterol levels through unknown mechanisms, perhaps through changes in membrane properties[38] — APP, the amyloid-β precursor, is a transmembrane protein. In the fly, *SNF4Aγ* mutants have a 40% reduction in cholesteryl ester levels, which may increase free cholesterol, and loss of *Appl*, the fly version of APP, enhances degeneration in *SNF4Aγ* mutants[39]. *SNF4Aγ* encodes a subunit of AMP-activated protein kinase (AMPK), discussed further below. Although flies do not synthesize cholesterol *de novo*, they do have HMG CoA reductase, the enzyme that catalyses the rate-limiting step in cholesterol synthesis in vertebrates and that is directly phosphorylated and inhibited by AMPK[40]. Loss of one copy of the gene that encodes HMG CoA reductase (*Hmgcr*) partially rescues the degeneration in *SNF4Aγ* mutant flies, and overexpression of *Hmgcr* enhances degeneration[39]. Statins — a class of compounds that target HMGCR and are used to treat high blood cholesterol levels — seem to reduce the risk of Alzheimer's disease[38], and feeding a statin to *SNF4Aγ* mutant flies decreases degeneration[39]. Cholesterol is a versatile lipid: it is the precursor of hormones (see below), an adduct of signalling proteins[41] and a regulator of membrane fluidity with a key role in synapse function[42]. The relative importance of each of these roles in neurodegeneration remains to be determined.

Neurite
General term for axons and dendrites.

Amyloid
Protein aggregates that accumulate as fibres of 7–10 nm in diameter with common structural features including β-pleated sheet conformation and resistance to detergents and proteases.

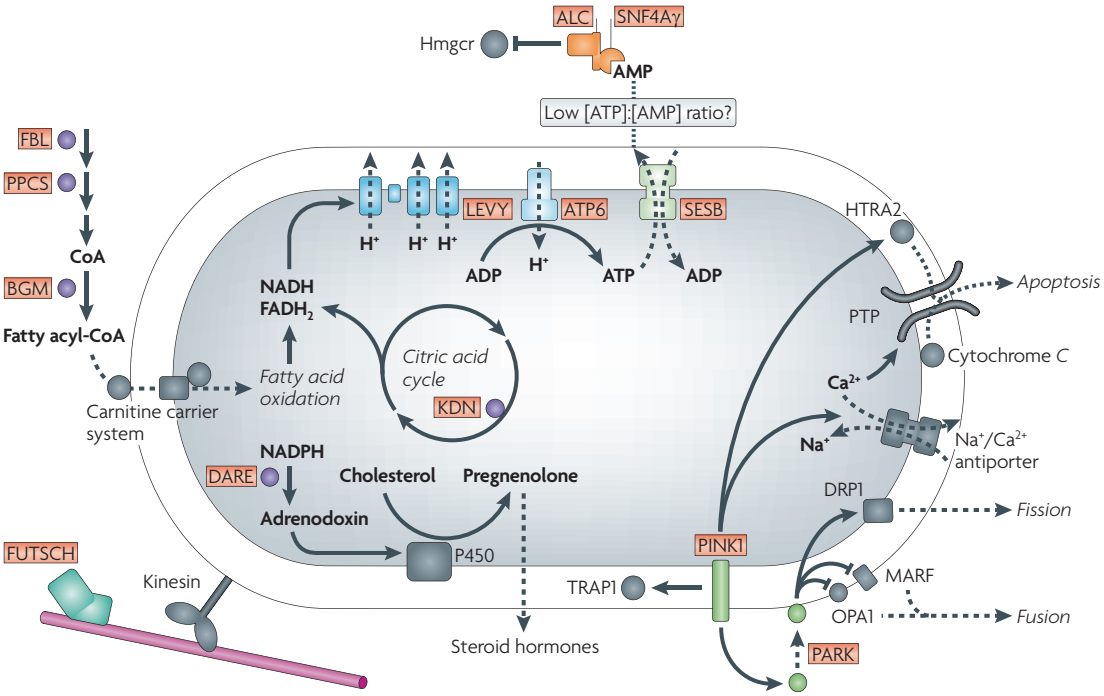


Figure 2 | **Neurodegeneration proteins associated with the mitochondrion.** Gene products from Supplementary information S1 (table) are highlighted in orange. In black are proteins not in the table that genetically interact (or are predicted to interact) with the neurodegeneration proteins. Small molecules are shown in bold type. At the bottom left, transport of mitochondria into neurites depends on the (+) end-directed motor kinesin and on microtubules (pink), which are stabilized by FUTSCH. The *dare* product transfers electrons from NADPH to adrenodoxin, which in turn transfers them to a cytochrome P450, which catalyses the first step in steroid hormone synthesis. On the left, fatty acids are activated by a number of steps before importation (black dashed arrow; see text) into the matrix for oxidation. Together with the citric acid cycle, these pathways generate the electron carriers NADH and $FADH_2$, which in turn power the complexes of the electron transport chain (blue). Proton flux across the inner membrane is indicated by dashed arrows. Incoming protons drive synthesis of ATP, which is transported out of the matrix (dashed arrow) by the ATP–ADP translocator (light green). Low levels of ATP generation result in activation of AMP-activated protein kinase (AMPK; made up of ALC and SNFAγ). At the bottom right, PINK1 regulates TRAP1 and the localization of Parkin (PARK; see text); *Pink1* and *park* interact with genes that regulate mitochondrial fission and fusion, consistent with these gene products acting downstream of PARK, as shown here, although there is no direct evidence for this yet. PINK1 also regulates another parkinsonism protein, HTRA2, and a Na^+/Ca^{2+} antiporter activity (see text). High Ca^{2+} levels in the matrix trigger formation of the permeability transition pore (PTP), through which cytochrome *c* can be released, activating caspases and apoptosis. Not shown are apoptosis-inducing factor and members of the BCL2 family, proteins that also regulate apoptosis and translocate between the cytoplasm and the outer surface of mitochondria[72].

In flies, *dare* activity is required in mitochondria for the synthesis of ecdysteroids from cholesterol (FIG. 2). Feeding larvae ecdysone rescues an early developmental defect caused by null *dare* mutations[43] and, as neuronal expression of *dare* can rescue adult neurodegeneration[44], there may be a role for ecdysteroid synthesis in CNS maintenance. The mammalian gene *NPC1* presents an interesting parallel to the *dare* story. Niemann–Pick disease can be caused by loss of *NPC1* and is characterized by misregulated cholesterol and sterol trafficking. Mice lacking *Npc1* have many of the same neurodegenerative symptoms as patients with Niemann–Pick disease, as well as deficits in the neuron-specific synthesis of steroid hormones[45]. Strikingly, these effects can be rescued by feeding the *Npc1*-null mice a brain-specific steroid[45]. The *dare* and *Npc1* results raise the intriguing possibility that steroid hormone signalling in the adult brain plays a part in maintaining morphological integrity.

Ecdysteroids
Steroids that are similar in structure to ecdysones, found in arthropods and some plants.

Ecdysone
Steroid hormone found in arthropods. In insects, 20-hydroxyecdysone stimulates moulting and metamorphosis.

Signal transduction. A number of the genes classified in FIG. 1 implicate known signal transduction pathways in neurodegeneration, but at present few molecular details are known about how dysfunction of these pathways causes degeneration. *alc* and *SNF4Aγ*, encoding the β- and γ-subunits of AMPK, respectively, are the only genes that can be classified into three of the processes depicted in FIG. 1, thus seeming to be at a nexus of cellular pathways involved in neurodegeneration[39,46]. AMPK is a crucial control point for metabolism and is composed of the catalytic α-subunit, the β-subunit (which seems to be a scaffold), and the γ-regulatory subunit, which binds AMP[40]. Rising AMP levels indicate an energy deficit, which activates the kinase[40] (FIG. 2). The *sws* mutation suggests a role for PKA in maintaining CNS integrity (discussed above); PKA functions in a well-characterized pathway that is required during development and for learning and memory[47],

and that plays a part in axon regeneration[48]. Both AMPK and PKA signalling are essential for the development and function of many cell types, and so uncovering the details of how these specific pathways affect brain integrity will require the use of tools for investigating their roles specifically in mature neurons.

Recent mammalian cell culture experiments that have grown out of the *Pink1–park* epistasis experiments have shed further light on this mitochondrial pathway. Although PARK is typically cytoplasmic, coexpression of PINK1 causes translocation of PARK to mitochondria; translocation depends on PINK1 kinase activity[49]. Further experiments have suggested that PINK1 may directly phosphorylate PARK[49]. In mammalian cells, PARK can mediate the autophagic engulfment of dysfunctional mitochondria[50]. How this is done is unclear, although PARK has alternative ubiquitin ligase activities[51] that are associated not with proteasomal degradation but rather with other processes, such as protein trafficking[52]. A likely direct substrate of PINK1 was recently shown to be the serine protease HTRA2 (REF. 53), which is released from mitochondria during apoptosis and has pro-apoptotic activity[54]. However, loss-of-function mutations of *HTRA2* in humans can cause Parkinson's disease, and knockout of *Htra2* in mice causes parkinsonism-like defects[54]. Pro-apoptotic activity for *D. melanogaster HtrA2* has been shown recently[55,56], but a role for the fly gene in neurodegeneration is unknown at this point, and the mechanism of HTRA2's role in CNS integrity remains unclear.

Many of the neurodegeneration-associated signal transduction genes play a direct part in neuronal activity, either by maintaining membrane excitability or by mediating signals across the synapse. For example, CSP is a synaptic vesicle protein that is required for proper neurotransmitter release[57]. Since its initial characterization in flies[58], loss of CSP has been shown to cause neurodegeneration in the mouse[59] and, surprisingly, this can be rescued by upregulation of α-synuclein[60] (dominant mutations of which, including increased copy number of the normal gene, are a cause of Parkinson's disease[61]). These studies reveal a potential overlap in biological function between CSP and α-synuclein at the synapse, and the susceptibility of neurons to toxicity from abnormal synaptic function.

Ace encodes acetylcholine esterase, which degrades the neurotransmitter acetylcholine; work on *Ace* in the fly 30 years ago thus implicated excitotoxicity as a mechanism for neurodegeneration[62,63]. Similar to acetylcholine esterase, the transporter encoded by *Eaat1* buffers an excitatory neurotransmitter, in this case glutamate[64]. Furthermore, excessive neuronal firing is observed in neurodegenerative, dominant mutations of *Atpα*[65], which encodes the α-subunit of the Na^+/K^+ ATPase. Collectively, these studies highlight the role of excitotoxicity in neural integrity. By contrast, other electrophysiological studies show decreased transmitter release and reduced phototransduction and synaptic transmission in *spin* mutants[66,67]. Perhaps CNS integrity in the adult requires that neurons be maintained in a 'Goldilocks state' of neither abnormally high nor abnormally low levels of activity, a phenomenon parallel to vertebrate neural development, in which excess neurons are trimmed in a manner that is dependent on their level of activity.

Excitotoxicity
The over-stimulation of excitatory neurotransmitter receptors, which causes an influx of calcium in the postsynaptic neuron.

Mitochondrial function. Mitochondria are central to the processes affecting neurodegeneration (FIG. 1): the set of genes affecting mitochondria overlaps each of the other four sets described. FIGURE 2 illustrates the functions of gene products from Supplementary information S1 (table) that are associated with mitochondria. The synthesis of the cellular energy currency ATP is the principle purpose of mitochondria, and is directly implicated in neuronal integrity by mutants in *levy* and *ATP6*, which encode components of the electron transport chain, and by the *sesB* product, which transports newly synthesized ATP to the cytoplasm. *ATP6* is the only gene implicated so far in *D. melanogaster* neurodegeneration that resides in the mitochondrial genome. The *ATP6* mutation arose spontaneously in a *sesB* mutant background, and it has a mild effect on neuronal integrity when it is the only mutation[68]. Interestingly, patients with the human disease associated with the *sesB* orthologue (progressive external opthalmoplegia 2) also have multiple, varying mitochondrial DNA deletions, raising the possibility that nuclear gene mutations that cause mitochondrial dysfunction also leave mitochondrial DNA in a state more vulnerable to lesions[69]. Reactive oxygen species such as superoxide (O_2^-) are toxic products of mitochondrial respiration, are associated with DNA damage and have long been suspected as a cause of ageing and neurodegeneration[70]. For example, the *Sod* product superoxide dismutase scavenges O_2^- (BOX 1). TRAP1, a mitochondrial chaperone that is phosphorylated in response to oxidative stress, is a recently described *in vivo* substrate of PINK1 in human cells[71]. PINK1 can protect these cells against oxidative stress, an activity that depends on TRAP1 (REF. 71).

Mitochondria are not simply passive passengers in eukaryotic cells. They regulate their morphology in response to specific cellular needs and their own level of functionality, as discussed above with respect to PINK1 and PARK in *D. melanogaster*. Furthermore, mitochondria integrate apoptotic signals and facilitate apoptotic cell death. A key event to initiate apoptosis is the opening of the permeability transition pore[72]; this allows release into the cytoplasm of cytochrome *c*, apoptosis-inducing factor and perhaps HTRA2 (see above). The pore is opened following increased Ca^{2+} concentration in the mitochondrial matrix. Recently, a Na^+–Ca^{2+} antiporter activity, which has not yet been molecularly characterized in mammals or in flies, was shown to be blocked in cultured mammalian *Pink1* mutant neurons, resulting in increased mitochondrial Ca^{2+} levels[73]. Thus, misregulated apoptosis may be a contributing cause of neuronal death in *Pink1*-dependent parkinsonism. Indeed, the proximal cause of neuronal death for many of the mutants in Supplementary information S1 (table) is apoptosis, although in many cases apoptosis may be triggered simply in response to an underlying pathology.

Glia and cellular interactions

The discussion above is largely focused on cell-autonomous ways that neurons can die, that is, from intrinsic problems such as dysfunctional mitochondria. But neurodegeneration can occur through aberrant cell–cell signalling as well, implicated by genes that play a part in synaptic transmission and that suggest a role for neuronal steroid hormones. Glial interactions are a clear example in which cells are necessary for the proper maintenance of neighbouring neurons. Below we consider fly genes that have implicated glia in neurodegeneration, and how genetically mosaic flies can uncover roles for cell–cell signalling in neuronal maintenance.

Glia carry out essential functions in vertebrates, such as modulation of synapses, formation of the blood–brain barrier, basic immune system duties and protection of long axonal tracts. All of these roles have clear parallels in the fly[74]. For example, although flies do not have a vascular circulatory system, glia form a blood–brain barrier analogue by sealing off neurons from the surrounding haemolymph. Flies do not have an adaptive immune system, but fly glia perform some of the immune tasks of vertebrate microglia, such as engulfing dead neurons. Finally, although flies do not have myelin, they do have a glial cell type that ensheaths axons — comparable to Schwann cells, which myelinate long axonal tracts in mammals. Thus, glia are essential for proper neuronal function in flies, and indeed four genes in TABLE 1 implicate glial function in the maintenance of the brain.

The expression of two of these genes — *repo* and *Eaat1* — is restricted to the glia. The REPO protein is a transcriptional regulator that is required for glial development; null alleles are lethal, reflecting the essential role for glia in the animal. A partial loss-of-function allele of *repo* reveals a role for the gene in maintaining the adult CNS[75], although the specific defects in these mutant glia that cause neuronal loss are unclear. *Eaat1*, the transporter mentioned above, probably scavenges excess glutamate, which is a neurotransmitter. The transporter is enriched in glial membranes at synaptic clefts, and thus this gene implicates *D. melanogaster* glia in modulating synaptic activity[64].

Flies mutant for *drd* seem to be normal for ~1 week, then suddenly become uncoordinated, dying within hours with holes throughout the brain. Glia in *drd* mutants have structural defects, and there are subtle glial abnormalities in young adult mutants, which are apparent before the onset of behavioural defects[76]. The location of *drd* function was addressed in the early 1970s in genetically mosaic flies that are partly mutant for *drd* and partly normal[8]. Strikingly, most flies with heads that were half mutant and half normal behaved like normal flies and had normal brain morphology. That is, in these animals brain histology was normal on the side of the head that was genotypically mutant. Therefore, the function of *drd* mutant tissue can be rescued by adjacent normal tissue, implying the existence of *drd*-dependent diffusible signals.

Young adult *sws* mutants also have an early glial defect: multiple glial membranes wrap adjacent neurites that, normally, are wrapped by only a single layer[16]. Eventually, glia die along with neurons[19]. In contrast to the non-autonomy of *drd* mosaics, however, degeneration in mosaic *sws* flies corresponds to genotypically mutant parts of the brain[16], ruling out a long-range, *sws*-dependent signal. The method that is used to construct the mosaic flies was the same that was used to analyse *drd*, producing broad swatches of mutant brain that can shed little light on a possible requirement for *sws* activity in glia, which are intermingled with neurons. Other methods of making mosaic flies, such as looking at *sws* mutant glia situated adjacent to normal neurons, could address questions such as *sws*-dependent glial–neuronal interactions. The reciprocal experiment — determining the degree of rescue in a mutant animal following targeted expression of the wild-type gene — is also informative. For example, in flies, neuronal expression of normal *NPC1* can rescue *NPC1* mutant adults and, interestingly, glial expression of the normal gene also rescues the mutant phenotype, although not as efficiently[77]. In summary, as is the case in higher organisms, glial function is crucial for neuronal integrity in the fly, and the repertoire of tools for fly genetics should allow for both glial function and interneuronal signalling to be addressed systematically.

Conclusions

The genes discussed in this Review share a number of themes. First, most are widely expressed, either ubiquitously or enriched throughout the CNS (interesting exceptions being *Eaat1* and *repo*, which are restricted to glia). Second, behavioural defects are common in flies that are mutant for these genes. Third, all these *D. melanogaster* neurodegeneration mutants have shortened lifespans, which correspond with the severity and time course of degeneration (see below). These aspects of fly neurodegeneration mutants mirror observations of neurodegenerative diseases in humans: widespread expression of causative genes, neurological symptoms such as ataxia or tremor, and early death.

A variation among the genes is the spatial extent of degeneration that occurs in the brains of mutant flies. In many cases, lesions are seen throughout the brain and even in the thoracic ganglion (the large ganglion that is a thickening of the ventral nerve cord). In other mutants, lesions are seen only in specific subregions of the brain, typically the optic lobes. The time course of degeneration can also vary widely among the mutants: from hours in the case of *dare*[43] to approximately 1 month in the case of *Adar*[78]. Moreover, whereas adult *dare* mutant flies die within a few days, *Adar* mutants can live nearly as long as wild-type flies. Such variation is probably due only partly to the absolute requirement for a particular gene in maintaining CNS integrity. Other factors include possible earlier roles in development and the strength of the mutant allele.

Undiscovered neurodegeneration genes. Identifying new genes with a role in maintaining adult CNS integrity in the fly is perhaps the fastest way to identify them in humans, given the ease of genetic screens for brain integrity mutants and the range of genetic tools available for

Haemolymph
The interstitial fluid in insects, which have an open circulatory system. Unlike blood, haemolymph has only a small role in carrying O_2 and CO_2, which is principally done by the tracheal system.

Optic lobes
Large, bilaterally symmetric structures of the fly brain that process visual input.

assessing gene function. There are almost certainly many as-yet undiscovered gene activities that are required for maintaining the CNS. *D. melanogaster* screens that are specifically designed to identify neurodegeneration genes have not been any more successful than other methods in flies: 14 genes in Supplementary information S1 (table) were identified in such screens; another 16 were candidate genes; and another 14 were fortuitous mutations. Probably owing to their labour-intensive nature, none of the previous screens achieved saturation, as most of the relevant genes are represented by single alleles. Genes with a narrow role in maintaining a specific neuronal subtype may be underrepresented owing to technical limitations (see below); most of the genes surveyed here cause broad neurodegeneration when mutated. Genes that are essential for development are also underrepresented: all those discussed here that are characterized by complete loss-of-function mutations are viable. This limitation could be overcome by the study of gene activity in differentiated adult neurons using transgenic RNAi constructs and, if necessary, methods to control expression of the RNAi construct temporally[79]. Developmentally essential genes can also be analysed in genetically mosaic flies.

Fly genes and human neurodegeneration. In a number of cases the mouse or human orthologue of a fly gene in Supplementary information S1 (table) is directly associated with neurodegeneration. There are three other possible relationships between a fly gene and mammalian brain integrity. The mouse or human orthologue may be closely related to another disease-associated gene through gene duplication (for example, *futsch* in TABLE 1). Second, a fly gene may encode a protein that is part of a complex, a separate subunit of which has been shown to cause neurodegeneration in mice or humans (for example, *Atpα*). Finally, a fly gene may function in a pathway in which another gene is linked to neurodegeneration. For example, the fly gene *levy* encodes a component of the COX complex, whereas the products of the human genes *COX10* and *LRPPRC* are required for proper assembly and expression of this complex, but the mutations are functionally equivalent. Taken together, 55% of the *D. melanogaster* genes in Supplementary information S1 (table) are currently linked to mouse or human genes in pathways associated with neurodegeneration. There remain fly genes in Supplementary information S1 (table) for which, at present, no mammalian genes related by sequence or function are associated with neurodegeneration. Such genes may have functions in maintaining adult brain integrity in mammals that are still awaiting discovery.

Open questions. As discussed above, there are probably many more genes associated with neurodegeneration that are not yet identified, and it is likely that other processes will be added to the five highlighted in this Review. Furthermore, there are a number of issues that are yet to be systematically addressed for fly neurodegeneration genes. For example, how are the neurons dying? Apoptosis is involved in many but not all of the degeneration mutants. How widespread is necrotic cell loss? Do neurons die because of defective intrinsic processes or from interactions with aberrant neighbouring cells? For widely expressed genes, tools that enable one to look at requirements for gene activity in glia or in specific neurons are available and might answer the last question: RNAi can be restricted to glia, or to small subsets of neurons[80]. In addition, genetically mosaic flies can be analysed with marked mutant clones as small as a single cell[81]. What types of neurons are dying? For the most part this question has been addressed only for candidate genes for Parkinson's disease. Cell type-specific or brain structure-specific markers combined with confocal microscopy and three-dimensional reconstruction are likely to become more prevalent in the study of brain degeneration; commonly used histological techniques (BOX 2) currently cannot differentiate between affected neuronal types. Finally, in humans, age is the single largest risk factor for diseases such as Parkinson's disease and Alzheimer's disease. A challenge for the future will be to use the advantages of the fly to tease apart the role of ageing in neurodegeneration.

Although an excellent system for investigating neurodegenerative disease, flies are not humans. They lack an adaptive immune system, for example, although flies will allow detailed study of the role of innate immunity. However, striking similarities between fly and human nervous systems mean that studies in flies and mammals complement each other well. Moreover, identification of genes and pathways that are crucial for brain maintenance is straightforward in the fly and serves as a springboard for further investigation in mice and humans. Manipulation of such pathways in the fly enhances approaches to human diseases that are associated with loss of brain integrity and cognitive function.

1. Pires-daSilva, A. & Sommer, R. J. The evolution of signalling pathways in animal development. *Nature Rev. Genet.* **4**, 39–49 (2003).
2. Ben-Shlomo, I., Hsu, S. Y., Rauch, R., Kowalski, H. W. & Hsueh, A. J. W. Signaling receptome: a genomic and evolutionary perspective of plasma membrane receptors involved in signal transduction. *Sci. STKE* **2003**, re9 (2003).
3. Hirth, F. & Reichert, H. Conserved genetic programs in insect and mammalian brain development. *Bioessays* **21**, 677–684 (1999).
4. Sehgal, A. *et al.* Molecular analysis of sleep: wake cycles in *Drosophila*. *Cold Spring Harb. Symp. Quant. Biol.* **72**, 557–564 (2007).
5. Roman, G. & Davis, R. L. Molecular biology and anatomy of *Drosophila* olfactory associative learning. *Bioessays* **23**, 571–581 (2001).
6. Dierick, H. A. Fly fighting: octopamine modulates aggression. *Curr. Biol.* **18**, R161–R163 (2008).
7. Chien, S., Reiter, L. T., Bier, E. & Gribskov, M. Homophila: human disease gene cognates in *Drosophila*. *Nucleic Acids Res.* **30**, 149–151 (2002).
8. Hotta, Y. & Benzer, S. Mapping of behaviour in *Drosophila* mosaics. *Nature* **240**, 527–535 (1972).
9. Bonini, N. M. A Tribute to Seymour Benzer, 1921–2007. *Genetics* **180**, 1265–1273 (2008).
10. Min, K. T. & Benzer, S. Spongecake and eggroll: two hereditary diseases in Drosophila resemble patterns of human brain degeneration. *Curr. Biol.* **7**, 885–888 (1997).
11. Min, K. T. & Benzer, S. Preventing neurodegeneration in the *Drosophila* mutant *bubblegum*. *Science* **284**, 1985–1988 (1999).
12. Heisenberg, M. & Bohl, K. Isolation of anatomical brain mutants of *Drosophila* by histological means. *Z. Naturforsch. B* **34**, 143–147 (1979).
13. Palladino, M. J., Hadley, T. J. & Ganetzky, B. Temperature-sensitive paralytic mutants are enriched for those causing neurodegeneration in Drosophila. *Genetics* **161**, 1197–1208 (2002).
Neurodegeneration screen based on an induced paralysis phenotype.
14. Fergestad, T., Bostwick, B. & Ganetzky, B. Metabolic disruption in Drosophila bang-sensitive seizure mutants. *Genetics* **173**, 1357–1364 (2006).
15. Rezaval, C. *et al.* A functional misexpression screen uncovers a role for enabled in progressive neurodegeneration. *PLoS ONE* **3**, e3332 (2008).

16. Kretzschmar, D., Hasan, G., Sharma, S., Heisenberg, M. & Benzer, S. The *swiss cheese* mutant causes glial hyperwrapping and brain degeneration in *Drosophila*. *J. Neurosci.* **17**, 7425–7432 (1997).
17. Lush, M. J., Li, Y., Read, D. J., Willis, A. C. & Glynn, P. Neuropathy target esterase and a homologous *Drosophila* neurodegeneration-associated mutant protein contain a novel domain conserved from bacteria to man. *Biochem. J.* **332**, 1–4 (1998).
18. Glynn, P. Neuropathy target esterase and phospholipid deacylation. *Biochim. Biophys. Acta* **1736**, 87–93 (2005).
19. Muhlig-Versen, M. *et al.* Loss of Swiss cheese/neuropathy target esterase activity causes disruption of phosphatidylcholine homeostasis and neuronal and glial death in adult *Drosophila*. *J. Neurosci.* **25**, 2865–2873 (2005).
20. Bettencourt da Cruz, A., Wentzell, J. & Kretzschmar, D. Swiss cheese, a protein involved in progressive neurodegeneration, acts as a noncanonical regulatory subunit for PKA-C3. *J. Neurosci.* **28**, 10885–10892 (2008).
SWS binds to and inhibits a catalytic subunit of cAMP-dependent protein kinase, an activity that plays a part in protecting CNS integrity.
21. Akassoglou, K. *et al.* Brain-specific deletion of neuropathy target esterase/*swisscheese* results in neurodegeneration. *Proc. Natl Acad. Sci. USA* **101**, 5075–5080 (2004).
22. Rainier, S. *et al.* Neuropathy target esterase gene mutations cause motor neuron disease. *Am. J. Hum. Genet.* **82**, 780–785 (2008).
23. Thomas, B. & Beal, M. F. Parkinson's disease. *Hum. Mol. Genet.* **16**, R183–R194 (2007).
24. Clark, I. E. *et al.* Drosophila *pink1* is required for mitochondrial function and interacts genetically with *parkin*. *Nature* **441**, 1162–1166 (2006).
25. Park, J. *et al.* Mitochondrial dysfunction in *Drosophila PINK1* mutants is complemented by *parkin*. *Nature* **441**, 1157–1161 (2006).
26. Yang, Y. *et al.* Mitochondrial pathology and muscle and dopaminergic neuron degeneration caused by inactivation of *Drosophila* Pink1 is rescued by Parkin. *Proc. Natl Acad. Sci. USA* **103**, 10793–10798 (2006).
27. Exner, N. *et al.* Loss-of-function of human PINK1 results in mitochondrial pathology and can be rescued by Parkin. *J. Neurosci.* **27**, 12413–12418 (2007).
28. Poole, A. C. *et al.* The PINK1/Parkin pathway regulates mitochondrial morphology. *Proc. Natl Acad. Sci. USA* **105**, 1638–1643 (2008).
References 28 and 29 detail epistasis experiments that show rescue of *Pink1* and *park* mutant defects by upregulating *Drp1* or by downregulating *Marf* or *opa1*.
29. Deng, H., Dodson, M. W., Huang, H. & Guo, M. The Parkinson's disease genes *pink1* and *parkin* promote mitochondrial fission and/or inhibit fusion in *Drosophila*. *Proc. Natl Acad. Sci. USA* **105**, 14503–14508 (2008).
30. Chen, H., McCaffery, J. M. & Chan, D. C. Mitochondrial fusion protects against neurodegeneration in the cerebellum. *Cell* **130**, 548–562 (2007).
31. Opal, P. & Zoghbi, H. Y. The role of chaperones in polyglutamine disease. *Trends Mol. Med.* **8**, 232–236 (2002).
32. Levine, B. & Kroemer, G. Autophagy in the pathogenesis of disease. *Cell* **132**, 27–42 (2008).
33. Fergestad, T. *et al.* Neuropathology in Drosophila mutants with increased seizure susceptibility. *Genetics* **178**, 947–956 (2008).
34. Vance, J. E. Phosphatidylserine and phosphatidylethanolamine in mammalian cells: two metabolically related aminophospholipids. *J. Lipid Res.* **49**, 1377–1387 (2008).
35. Di Paolo, G. & De Camilli, P. Phosphoinositides in cell regulation and membrane dynamics. *Nature* **443**, 651–657 (2006).
36. Bosveld, F. *et al. De novo* CoA biosynthesis is required to maintain DNA integrity during development of the *Drosophila* nervous system. *Hum. Mol. Genet.* **17**, 2058–2069 (2008).
37. Rubio-Gozalbo, M. E., Bakker, J. A., Waterham, H. R. & Wanders, R. J. Carnitine-acylcarnitine translocase deficiency, clinical, biochemical and genetic aspects. *Mol. Aspects Med.* **25**, 521–532 (2004).
38. Shobab, L. A., Hsiung, G.-Y. R. & Feldman, H. H. Cholesterol in Alzheimer's disease. *Lancet Neurol.* **4**, 841–852 (2005).
39. Tschape, J. A. *et al.* The neurodegeneration mutant lochrig interferes with cholesterol homeostasis and Appl processing. *EMBO J.* **21**, 6367–6376 (2002).
Characterization of the γ-subunit of AMP-activated protein kinase. This reference is an early report of the role of cholesterol in fly neurodegeneration, even though flies do not synthesize this molecule *de novo*.
40. Towler, M. C. & Hardie, D. G. AMP-activated protein kinase in metabolic control and insulin signaling. *Circ. Res.* **100**, 328–341 (2007).
41. Breitling, R. Greased hedgehogs: new links between hedgehog signaling and cholesterol metabolism. *Bioessays* **29**, 1085–1094 (2007).
42. Pfrieger, F. W. Cholesterol homeostasis and function in neurons of the central nervous system. *Cell. Mol. Life Sci.* **60**, 1158–1171 (2003).
43. Freeman, M. R., Dobritsa, A., Gaines, P., Segraves, W. A. & Carlson, J. R. The *dare* gene: steroid hormone production, olfactory behavior, and neural degeneration in *Drosophila*. *Development* **126**, 4591–4602 (1999).
A possible role for steroid synthesis and signalling in the brain in maintaining CNS integrity.
44. Dobritsa, A. A. *Molecular genetics of odor reception and development in Drosophila*. Thesis, Yale Univ. (2003).
45. Griffin, L. D., Gong, W., Verot, L. & Mellon, S. H. Niemann–Pick type C disease involves disrupted neurosteroidogenesis and responds to allopregnanolone. *Nature Med.* **10**, 704–711 (2004).
46. Spasic, M. R., Callaerts, P. & Norga, K. K. *Drosophila alicorn* is a neuronal maintenance factor protecting against activity-induced retinal degeneration. *J. Neurosci.* **28**, 6419–6429 (2008).
47. Tasken, K. & Aandahl, E. M. Localized effects of cAMP mediated by distinct routes of protein kinase A. *Physiol. Rev.* **84**, 137–167 (2004).
48. Teng, F. Y. & Tang, B. L. Axonal regeneration in adult CNS neurons — signaling molecules and pathways. *J. Neurochem.* **96**, 1501–1508 (2006).
49. Kim, Y. *et al.* PINK1 controls mitochondrial localization of Parkin through direct phosphorylation. *Biochem. Biophys. Res. Commun.* **377**, 975–980 (2008).
50. Narendra, D., Tanaka, A., Suen, D.-F. & Youle, R. J. Parkin is recruited selectively to impaired mitochondria and promotes their autophagy. *J. Cell Biol.* **183**, 795–803 (2008).
51. Moore, D. J. Parkin: a multifaceted ubiquitin ligase. *Biochem. Soc. Trans.* **34**, 749–753 (2006).
52. Mukhopadhyay, D. & Riezman, H. Proteasome-independent functions of ubiquitin in endocytosis and signaling. *Science* **315**, 201–205 (2007).
53. Plun-Favreau, H. *et al.* The mitochondrial protease HtrA2 is regulated by Parkinson's disease-associated kinase PINK1. *Nature Cell Biol.* **9**, 1243–1252 (2007).
54. Vande Walle, L., Lamkanfi, M. & Vandenabeele, P. The mitochondrial serine protease HtrA2/Omi: an overview. *Cell Death Differ.* **15**, 453–460 (2008).
55. Challa, M. *et al. Drosophila* Omi, a mitochondrial-localized IAP antagonist and proapoptotic serine protease. *EMBO J.* **26**, 3144–3156 (2007).
56. Igaki, T. *et al.* Evolution of mitochondrial cell death pathway: proapoptotic role of HtrA2/Omi in *Drosophila*. *Biochem. Biophys. Res. Commun.* **356**, 993–997 (2007).
57. Umbach, J. A. *et al.* Presynaptic dysfunction in *Drosophila csp* mutants. *Neuron* **13**, 899–907 (1994).
58. Zinsmaier, K. E., Eberle, K. K., Buchner, E., Walter, N. & Benzer, S. Paralysis and early death in cysteine string protein mutants of Drosophila. *Science* **263**, 977–980 (1994).
59. Fernandez-Chacon, R. *et al.* The synaptic vesicle protein CSP alpha prevents presynaptic degeneration. *Neuron* **42**, 237–251 (2004).
60. Chandra, S., Gallardo, G., Fernandez-Chacon, R., Schluter, O. M. & Sudhof, T. C. α-Synuclein cooperates with CSPα in preventing neurodegeneration. *Cell* **123**, 383–396 (2005).
Expression of normal α-synuclein (but not with mutations associated with Parkinson's disease) rescues neurodegeneration caused by loss of CSP in mice. Knockout of α-synuclein enhances CSP-caused neurodegeneration.
61. Lee, V. M. & Trojanowski, J. Q. Mechanisms of Parkinson's disease linked to pathological α-synuclein: new targets for drug discovery. *Neuron* **52**, 33–38 (2006).
62. Greenspan, R. J., Finn, J. A. Jr & Hall, J. C. Acetylcholinesterase mutants in Drosophila and their effects on the structure and function of the central nervous system. *J. Comp. Neurol.* **189**, 741–774 (1980).
63. Hall, J. C., Alahiotis, S. N., Strumpf, D. A. & White, K. Behavioral and biochemical defects in temperature-sensitive acetylcholinesterase mutants of *Drosophila melanogaster*. *Genetics* **96**, 939–965 (1980).
64. Rival, T. *et al.* Decreasing glutamate buffering capacity triggers oxidative stress and neuropil degeneration in the *Drosophila* brain. *Curr. Biol.* **14**, 599–605 (2004).
A glutamate transporter expressed only in glial processes in the neuropil demonstrates the crucial role of glia in CNS maintenance.
65. Palladino, M. J., Bower, J. E., Kreber, R. & Ganetzky, B. Neural dysfunction and neurodegeneration in *Drosophila* Na^+/K^+ ATPase alpha subunit mutants. *J. Neurosci.* **23**, 1276–1286 (2003).
66. Sweeney, S. T. & Davis, G. W. Unrestricted synaptic growth in *spinster* — a late endosomal protein implicated in TGF-β-mediated synaptic growth regulation. *Neuron* **36**, 403–416 (2002).
67. Dermaut, B. *et al.* Aberrant lysosomal carbohydrate storage accompanies endocytic defects and neurodegeneration in *Drosophila benchwarmer*. *J. Cell Biol.* **170**, 127–139 (2005).
68. Celotto, A. M. *et al.* Mitochondrial encephalomyopathy in *Drosophila*. *J. Neurosci.* **26**, 810–820 (2006).
Characterization of two genes with vital mitochondrial functions, one of which is encoded by the mitochondrial genome.
69. Kaukonen, J. *et al.* Role of adenine nucleotide translocator 1 in mtDNA maintenance. *Science* **289**, 782–785 (2000).
70. Wallace, D. C. A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annu. Rev. Genet.* **39**, 359–407 (2005).
71. Pridgeon, J. W., Olzmann, J. A., Chin, L.-S. & Li, L. PINK1 protects against oxidative stress by phosphorylating mitochondrial chaperone TRAP1. *PLoS Biol.* **5**, e172 (2007).
72. Mattson, M. P. Apoptosis in neurodegenerative disorders. *Nature Rev. Mol. Cell Biol.* **1**, 120–129 (2000).
73. Gandhi, S. *et al.* PINK1-associated parkinson's disease is caused by neuronal vulnerability to calcium-induced cell death. *Mol. Cell* **33**, 627–638 (2009).
74. Freeman, M. R. & Doherty, J. Glial cell biology in *Drosophila* and vertebrates. *Trends Neurosci.* **29**, 82–90 (2006).
75. Xiong, W. C. & Montell, C. Defective glia induce neuronal apoptosis in the *repo* visual system of Drosophila. *Neuron* **14**, 581–590 (1995).
76. Buchanan, R. L. & Benzer, S. Defective glia in the Drosophila brain degeneration mutant drop-dead. *Neuron* **10**, 839–850 (1993).
77. Phillips, S. E., Woodruff, E. A. 3rd, Liang, P., Patten, M. & Broadie, K. Neuronal loss of *Drosophila* NPC1a causes cholesterol aggregation and age-progressive neurodegeneration. *J. Neurosci.* **28**, 6569–6582 (2008).
This fly model for Niemann–Pick disease mimics many of the human and mouse symptoms. Neuron-specific expression of NPC1a rescues the mutant defects; glia-specific expression, surprisingly, can also partially rescue these defects.
78. Palladino, M. J., Keegan, L. P., O'Connell, M. A. & Reenan, R. A. A-to-I pre-mRNA editing in *Drosophila* is primarily involved in adult nervous system function and integrity. *Cell* **102**, 437–449 (2000).
79. McGuire, S. E., Mao, Z. & Davis, R. L. Spatiotemporal gene expression targeting with the TARGET and gene-switch systems in *Drosophila*. *Sci. STKE* 2004, p16 (2004).
80. Pfeiffer, B. D. *et al.* Tools for neuroanatomy and neurogenetics in *Drosophila*. *Proc. Natl Acad. Sci. USA* **105**, 9715–9720 (2008).
Summarizes recent advances in techniques for fly genetics.
81. Lee, T. & Luo, L. Mosaic analysis with a repressible cell marker for studies of gene function in neuronal morphogenesis. *Neuron* **22**, 451–461 (1999).
82. Wang, T. & Montell, C. Phototransduction and retinal degeneration in *Drosophila*. *Pflugers Arch.* **454**, 821–847 (2007).

83. Stowers, R. S. & Schwarz, T. L. A genetic method for generating Drosophila eyes composed exclusively of mitotic clones of a single genotype. *Genetics* **152**, 1631–1639 (1999).
84. Franceschini, N. in *Information Processing in the Visual Systems of Arthropods* (ed. Wehner, R.) 75–82 (Springer, Berlin, 1972).
85. Lessing, D. & Bonini, N. M. Polyglutamine genes interact to modulate the severity and progression of neurodegeneration in *Drosophila*. *PLoS Biol.* **6**, e29 (2008).
86. Ahmed, Y., Hayashi, S., Levine, A. & Wieschaus, E. Regulation of Armadillo by a *Drosophila* APC inhibits neuronal apoptosis during retinal development. *Cell* **93**, 1171–1182 (1998).
87. Zhai, R. G. *et al. Drosophila* NMNAT maintains neural integrity independent of its NAD synthesis activity. *PLoS Biol.* **4**, e416 (2006).
88. McQuibban, G. A., Lee, J. R., Zheng, L., Juusola, M. & Freeman, M. Normal mitochondrial dynamics requires Rhomboid-7 and affects *Drosophila* lifespan and neuronal function. *Curr. Biol.* **16**, 982–989 (2006).
89. Mast, J. D., Tomalty, K. M., Vogel, H. & Clandinin, T. R. Reactive oxygen species act remotely to cause synapse loss in a *Drosophila* model of developmental mitochondrial encephalopathy. *Development* **135**, 2669–2679 (2008).
90. Phillips, J. P. *et al.* Subunit-destabilizing mutations in *Drosophila* copper/zinc superoxide dismutase: neuropathology and a model of dimer dysequilibrium. *Proc. Natl Acad. Sci. USA* **92**, 8574–8578 (1995).
91. Rimkus, S. A. *et al.* Mutations in *String/CDC25* inhibit cell cycle re-entry and neurodegeneration in a *Drosophila* model of ataxia telangiectasia. *Genes Dev.* **22**, 1205–1220 (2008).
92. Muraro, N. I. & Moffat, K. G. Down-regulation of *torp4a*, encoding the *Drosophila* homologue of torsinA, results in increased neuronal degeneration. *J. Neurobiol.* **66**, 1338–1353 (2006).
93. Matthies, H. J. G. & Broadie, K. Techniques to dissect cellular and subcellular function in the *Drosophila* nervous system. *Methods Cell Biol.* **71**, 195–265 (2003).
94. Magyar, J. P. *et al.* Degeneration of neural cells in the central nervous system of mice deficient in the gene for the adhesion molecule on Glia, the beta 2 subunit of murine Na, K-ATPase. *J. Cell Biol.* **127**, 835–845 (1994).
95. Watase, K. *et al.* Motor discoordination and increased susceptibility to cerebellar injury in GLAST mutant mice. *Eur. J. Neurosci.* **10**, 976–988 (1998).
96. Aoyama, K. *et al.* Neuronal glutathione deficiency and age-dependent neurodegeneration in the EAAC1 deficient mouse. *Nature Neurosci.* **9**, 119–126 (2006).
97. Pavlidis, P., Ramaswami, M. & Tanouye, M. A. The Drosophila *easily shocked* gene: a mutation in a phospholipid synthetic pathway causes seizure, neuronal failure, and paralysis. *Cell* **79**, 23–33 (1994).
98. Bettencourt da Cruz, A. *et al.* Disruption of the MAP1B-related protein FUTSCH leads to changes in the neuronal cytoskeleton, axonal transport defects, and progressive neurodegeneration in *Drosophila*. *Mol. Biol. Cell* **16**, 2433–2442 (2005).
***futsch*, which regulates the microtubule cytoskeleton, interacts with *Fmr1*, the fly orthologue of the causative gene for fragile X mental retardation.**
99. Liu, W. *et al.* Mutations in cytochrome c oxidase subunit VIa cause neurodegeneration and motor dysfunction in Drosophila. *Genetics* **176**, 937–946 (2007).
100. Greene, J. C., Whitworth, A. J., Andrews, L. A., Parker, T. J. & Pallanck, L. J. Genetic and genomic studies of *Drosophila parkin* mutants implicate oxidative stress and innate immune responses in pathogenesis. *Hum. Mol. Genet.* **14**, 799–811 (2005).
101. Greene, J. C. *et al.* Mitochondrial pathology and apoptotic muscle degeneration in *Drosophila parkin* mutants. *Proc. Natl Acad. Sci. USA* **100**, 4078–4083 (2003).
102. Tschape, J. A., Bettencourt da Cruz, A. & Kretzschmar, D. in *Advances in Neurodegeneration* (eds Horowski, R. *et al.*) 51–62 (Springer, Wien, 2003).

Acknowledgements
We thank M. Bland, N. Liu, Z. Yu, L.-Y. Hao and C.J. Thut for comments on the manuscript. N.M.B receives funding from the National Institute of Aging and the National Institute of Neurological Disorders and Stroke, and she is an Investigator of the Howard Hughes Medical Institute.

DATABASES
Flybase: http://flybase.org
dare | *drd* | *Eaat1* | *park* | *Pink1* | *sws*
OMIM: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
adrenoleukodystrophy| Alzheimer's disease | Charcot–Marie–Tooth type 2A | juvenile onset parkinsonism | Niemann–Pick disease | Parkinson's disease | progressive external opthalmoplegia 2

FURTHER INFORMATION
Bonini Laboratory: http://bonini.bio.upenn.edu/index.html
Flybrain: http://flybrain.neurobio.arizona.edu
Homophila: http://superfly.ucsd.edu/homophila
Mouse Genome Informatics: http://www.informatics.jax.org

SUPPLEMENTARY INFORMATION
See online article: S1 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Towards better mouse models: enhanced genotypes, systemic phenotyping and envirotype modelling

Johannes Beckers[*‡], *Wolfgang Wurst*[‡§] *and Martin Hrabé de Angelis*[*‡]

Abstract | The mouse is the leading mammalian model organism for basic genetic research and for studying human diseases. Coordinated international projects are currently in progress to generate a comprehensive map of mouse gene functions — the first for any mammalian genome. There are still many challenges ahead to maximize the value of the mouse as a model, particularly for human disease. These involve generating mice that are better models of human diseases at the genotypic level, systemic (assessing all organ systems) and systematic (analysing all mouse lines) phenotyping of existing and new mouse mutant resources, and assessing the effects of the environment on phenotypes.

Genotype
A description of the endogenous genetic information carried by an organism, as distinguished from its physical appearance (its phenotype) and external environmental factors (its envirotype).

*Institute of Experimental Genetics, Helmholtz Zentrum München, GmbH, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany.
‡Technical University Munich, Center of Life and Food Sciences, Weihenstephan, Alte Akademie 8, 85354 Freising, Germany.
§Institute of Developmental Genetics, Helmholtz Zentrum München, GmbH, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany.
Correspondence to M.H.A. and J.B.
e-mails: hrabe@helmholtz-muenchen.de; beckers@helmholtz-muenchen.de
doi:10.1038/nrg2578
Published online 12 May 2009

Following the sequencing of the human and mouse genomes, in 2001 the International Mammalian Genome Society (IMGS) declared the systematic mutagenesis of every mouse gene as a key challenge for the subsequent decade of functional genetic research[1]. Since then, the generation of the first complete map of gene functions for a mammalian genome has been progressing rapidly[2,3]. Currently, we estimate that approximately half of all mouse genes have been mutagenized, and corresponding lines established. If the resources stored in freezers are also included, most mouse genes have already been hit by one or more mutations, although many of these mutations need to be identified or transferred to a living mouse line.

However, to maximize the potential of the mouse as a model organism for human diseases and basic research, we now face challenges that go beyond simply establishing a mutant line for every gene. The focus of genetic research for the next generation of mouse models can be considered at three main levels: the genotype, the phenotype and what we term the envirotype (FIG. 1).

In terms of genotype, single engineered alleles are analysed in the context of inbred strains in most cases. This strategy has been valuable but does not take into account the specific alleles that underlie most human diseases or the effects of genetic background. Taking full advantage of existing resources that more closely resemble common human genetic variation will be particularly important in this respect. From the phenotypic point of view, mutant mouse lines have historically been generated to analyse specific pathways or biological processes. Although this approach has been and will be fruitful, the pleiotropic nature of gene functions has largely been disregarded. Therefore, two aspects of phenotyping are important in the context of maximizing the potential of the mouse as a model: systematic phenotyping to analyse all existing mutant mouse lines, and systemic phenotyping to examine all organs and study pleiotropic gene functions. The recent establishment of phenotyping centres that carry out such systematic and systemic phenotyping is a first step towards achieving these goals, although the efficiency and integration of these efforts needs to be improved. Finally, mutant mice are generally housed in standardized conditions. However, in addition to endogenous factors, external environmental factors are also major triggers for many human diseases. We have adopted the term envirotype from ecosystems research[4], and argue that the incorporation of envirotypes into experimental designs will be essential for accurately modelling human diseases in the mouse.

In this Review, we will discuss the important issues at the genotype, phenotype and envirotype levels that need to be addressed to develop improved mouse models. We describe both the existing and ongoing strategies for meeting these challenges, as well as highlighting the key areas of this field that need to be addressed in the future. We also discuss related issues, such as approaches for humanizing mouse models and archiving mutant mouse lines, which should be considered for the next generation of mouse models.

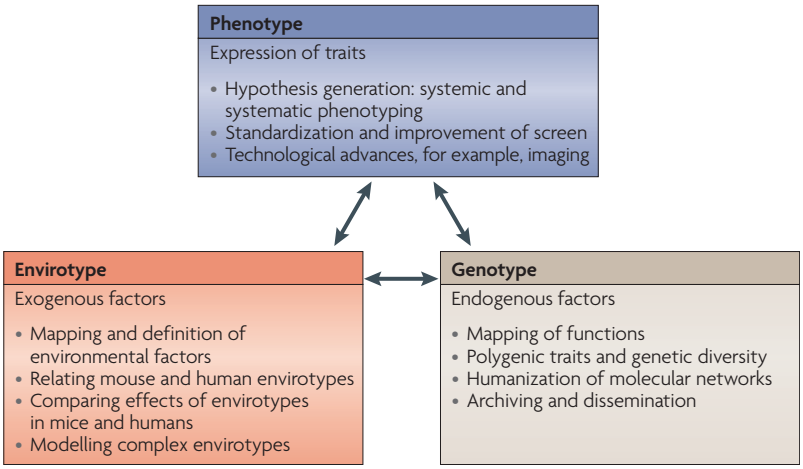


Figure 1 | **Maximizing the potential of the mouse as a model organism.** Overview of some of the major challenges in maximizing the use of the mouse as a model system in the areas of phenotyping, modelling envirotypes and generating genotypes. It is important to consider the various aspects of modelling in each of these three areas, and to appreciate the interdependent relationships between them.

Phenotype
A description of any observable (macroscopic, microscopic or molecular) trait of an individual with respect to some inherited characteristic.

Envirotype
A description of factors that are exogenous to the organism. The environmental code and the genetic code together affect the phenotype.

Pleiotropic
A situation in which a single gene has an effect on two or more distinct phenotypic characters.

Transposon
A type of mobile genetic element that consists of DNA that can move to new genomic locations conservatively (without replicating itself) or replicatively (by moving a copy of itself).

QTL
Genetic locus or chromosomal region that contributes to the variability in complex quantitative traits (such as body weight), as identified by statistical analysis. Quantitative traits are typically affected by several genes and by the environment.

Generation of mouse genotypes

The current status of mouse mutagenesis. To determine where we are in the effort to functionally annotate the mouse genome, we first summarize the databases of mutant resources and international initiatives that work towards the goals of the IMGS. The Mouse Genome Database (MGD) currently contains more than 21,000 phenotypic alleles in over 14,000 genes and markers[5]. This public resource continues to grow as a result of the combined efforts of large international consortia as well as specialized laboratories. The numbers include: spontaneous mutations; chemically, radiation- and transposon-induced mutations; transgenic mouse lines; and mutant mouse lines that were produced by gene trap insertion or homologous recombination in embryonic stem cells[6,7]. These numbers do not include ~8,000 phenotypic alleles of almost 4,000 QTLs, for most of which the underlying genetic variant still needs to be identified[5]. As recently generated and unpublished mutant mice are generally not yet included in the databases the true number of existing mutant mouse lines is much higher. Currently, large-scale efforts are underway to fully saturate the mouse genome with mutations[8]. The US-based Knockout Mouse Project (KOMP) and Texas A&M Institute of Genomic Medicine (TIGM), the European Conditional Mouse Mutagenesis Program (EUCOMM) and the North American Conditional Mouse Mutagenesis Project (NorCOMM) in Canada, together aim to generate more than 40,000 targeted and gene-trapped embryonic stem cell lines[9–12]. By 2010, more than 800 new mouse mutant lines will be derived from these embryonic stem cell resources[12]. It is thus feasible that in only a few years each mouse gene will be hit by at least one mutation — and this resource will be accessible to the scientific community.

Transposons can also be used for efficient random mutagenesis in the mouse. They can be used for genotype-driven mutagenesis and phenotype-driven mutagenesis. With phenotype-driven mutagenesis, the use of transposons is advantageous because the site of mutagenesis can easily be determined. Two major systems are currently in use: *Sleeping Beauty* and *piggyBac*[13–15]. Current developments that introduce additional genetic tools into the transposable *Sleeping Beauty* element allow conditional mobilization of the transgene[16].

Alleles relevant to human disease and basic science. The various gene targeting and trapping strategies currently used are an excellent asset to the functional genomics toolbox and are important for medical and basic research[17]. However, mutations in humans are not caused by insertions of *loxP* sites or *FRT* sites, or by selection markers or reporter genes. The next generation of mouse models will thus also require mutant alleles that more closely resemble the mutations and genetic variants that are relevant to humans. SNPs and copy number variants (CNVs) are important genetic factors in humans that influence quantitative traits and contribute to susceptibility or resistance to diseases and therapies[18–20]. Therefore, point mutations[21], deletions, duplications and translocations of genomic regions are of particular interest when generating new mouse models. Here, we discuss strategies to maximize current *N*-ethyl-*N*-nitrosourea (ENU) resources that model point mutations, and new methods of creating mouse lines that are more representative of human disease.

ENU is a highly efficient mutagen that is used to generate random point mutations in mouse spermatogonia[22]. Following mutagenesis, mutant mouse lines can be identified based on altered phenotypes (phenotype-driven ENU mutagenesis), a strategy that has been performed successfully in several research centres[21,23–26]. A major advantage of phenotype-driven screens is that they start with the desired mutant phenotype, which can directly be used as a mouse model[27,28]. However, there has been a bottleneck for identifying causative point mutations in phenotype-driven ENU screens, which involve the time-consuming and laborious generation of recombination events to narrow down the crucial interval containing the point mutation. In addition to the establishment of large mouse SNP panels for linkage analysis and their efficient detection by multiplex mass spectrometry[29], the recent advent of high-throughput sequencing will help to overcome this problem. The crucial genomic interval containing the ENU mutation can be narrowed down to the range of megabase pairs using recombination events, and candidates for the mutated gene may then be identified by sequencing this interval. Mass spectrometry for SNP detection and high-throughput sequencing are now standard in laboratories that are equipped with the corresponding infrastructures.

More recently, ENU mutagenesis in mice has also been used in approaches that aim to identify mutations in a particular gene, followed by the subsequent generation and phenotypic analysis of the corresponding mutant mouse line. Such genotype-driven screens require the establishment of large parallel archives of sperm and genomic DNA from G1 animals (the first generation offspring from ENU-injected male mice). It has previously been calculated that an archive size of 10,000 ENU mutagenized G1 animals is sufficient to give an 80%

probability of identifying five or more mutations in an average sized gene, and a 99% probability of identifying two or more mutations[30–32]. As there are currently several parallel sperm and genome archives (for example, at the RIKEN BioResource Center, the Helmholtz Centre Munich, and the Mary Lyon Centre) from over 45,000 G1 ENU-treated mice, and because each ENU-mutagenized mouse sperm is expected to carry 1,000 to 3,000 point mutations[30,33,34], it is feasible that our current resources already contain point mutations and allelic series for most or all genes. However, these resources have not been systematically exploited.

In current approaches, the archives are queried for mutations in selected genes that are requested by researchers. When the costs for next-generation sequencing decrease significantly, which is expected, it will become feasible to maximize the potential of these rich resources by systematically and fully sequencing entire genomes in these archives. This would not only allow the isolation of mutations in coding sequences, but could result in a map of millions of point mutations — including those in non-coding conserved sequences or regulatory regions[35,36] — from which researchers could select the allele or allelic series of their choice. But there is currently no concerted international effort to utilize the capacities of these combined ENU mutagenized genomes and sperm archives. Although the genome-wide approach would be the most desirable, one alternative and cheaper strategy would be to systematically sequence the transcriptomes of the ENU archives to identify alternative transcripts, non-silent mutations in coding regions and mutations in non-coding RNAs. However, although this would reduce the sequencing efforts to a few percent of the entire genome, it would only give a snapshot of the transcriptome of one tissue at one time point. If similar projects are publicly funded, the data from mutated transcripts and mutant mouse lines should be made available as a community resource.

Mutations that affect the copy numbers of chromosomal segments are also important in human disease. Engineered deletions, duplications, inversions or translocations that resemble mutations found in human disorders can be constructed using targeted meiotic recombination at *loxP* or *FRT* sites[37]. Such chromosome engineering has been used successfully in targeted approaches; for example, to generate mice that are trisomic or monosomic for a chromosomal segment that is orthologous to a region of human chromosome 21 that is associated with Down's syndrome[38]. This approach has also been instrumental for the study of gene regulatory mechanisms — for example, during embryonic development[37,39]. By contrast, systematic approaches for genome-wide screens using chromosomal engineering are probably not appropriate as many such mutations might be lethal. However, the technology may be particularly useful in modelling CNVs using targeted recombination approaches.

More recent technological developments such as the experimental use of RNA-induced silencing (through RNAi)[40–42] and designed zinc-finger nucleases (ZFNs) for mutagenesis[43–45] contribute to the variety of alleles that may be more relevant for human genotypes[46].

The experimental application of RNAi requires expression of particular forms of dsRNAs, such as small interfering RNAs (siRNAs) or short hairpin RNAs (shRNAs), which function by cleaving, destabilizing or blocking translation of their target mRNAs. In mice, this technology offers a rapid and easy method to knock down the expression of target genes. In contrast to the complete loss-of-function mutations in knockout mice, RNAi reduces the expression of the target gene — in many cases this situation may be more closely related to human disease. In addition, the generation of mutants with multiple knocked down genes can be simplified by simultaneously expressing shRNAs for multiple target mRNAs[47]. The targeting of shRNAs to the mouse genome can be designed such that experimental RNAi is performed in large-scale screening approaches. However, genome-wide RNAi approaches have, so far, been limited to *in vitro* screens in mammalian cells and *in vivo* screens in mosaic mouse models to identify new tumour suppressor genes[48]. A recent advance of the method was provided by the efficient targeting of shRNA vectors to the *Rosa26* locus[49]. This strategy allows the efficient selection of targeting events and the stable expression of shRNAs targeted against other mouse genes from the *Rosa26* locus, and conditional induction of RNAi in a tissue- and time-dependent manner using Cre/*loxP*-mediated activation. An important potential problem of RNAi is the occurrence of off-target effects[50]. In addition to experimental replication with different shRNAs, methods to reduce off-target effects include simultaneously expressing multiple shRNAs for a single target at low levels, leading to an additive effect on the target mRNA but minimal effects on off-target mRNAs.

In addition to homologous recombination in mouse embryonic stem cells, targeted and direct manipulation of genomic sequences in mammalian cells has been achieved by expressing 'designer' ZFNs[46]. ZFNs can be engineered to target specific 18-bp sequences in the genome, where they induce double-stranded breaks. Cells repair these breaks by non-homologous end joining, which is error prone and frequently results in base deletion or addition. In contrast to several existing methods for homologous recombination in mammalian cells, designer ZFNs can be used to permanently disrupt reading frames without introducing *loxP* or *FRT* sites[46]. However, off-target cleavage has been observed for ZFNs, which can generate undesired mutations[51]. In addition to addressing these issues, it needs to be demonstrated that ZFN-mediated mutagenesis has significant advantages over homologous recombination in the speed or efficiency of generating targeted mutant mouse lines. However, this method also makes it possible to use alternative methods to generate animal models in non-mouse species in which homologous recombination in embryonic stem cells is not possible[52].

Modelling polygenic traits and genetic diversity. Most human disorders are not monogenic and many recent human studies have focused on studying complex traits that are the result of a combination of many different alleles. In mouse models we have tended to eliminate the effect of genetic diversity by working with mouse inbred strains. The next generation of mouse models

Genotype-driven mutagenesis
A reverse genetic approach that starts with the targeted mutagenesis of a known gene or marker sequence. The gene targeting is followed by the analysis of the mutant phenotype. This approach is generally based on a hypothesis about a potential function of the mutated gene.

Phenotype-driven mutagenesis
A forward genetic approach that starts with the identification of a mutant phenotype caused by a random mutation in the genome. The identification of the mutated gene or marker is subsequent to the identification of the mutant mouse line. This approach makes no assumption of which genes may underlie a disease.

SNP
A type of polymorphism in which genomic segments differ by a single base pair.

Copy number variant
A type of polymorphism in which a segment of genomic DNA is present at a different copy number with respect to a reference genome.

N-ethyl-*N*-nitrosourea
A chemical mutagen that introduces point mutations in spermatogonia of male mice with high efficiency. It can be used as a mutagen in gene-driven and phenotype-driven mutagenesis.

Zinc-finger nuclease
Synthetic protein composed of a nonspecific DNA-cleaving domain and a highly specific DNA-binding domain, which comprises a string of zinc-finger motifs. Zinc-finger nucleases and subsequent DNA repair by homologous recombination can be used to mutagenize genes.

Off-target effect
These effects may compromise the specificity of RNAi and can occur if there is sequence identity between the small interfering RNA and random mRNA transcripts, causing knockdown of the expression of non-targeted genes.

will need to improve the modelling of polygenic traits and genetic diversity to provide more accurate models of human disease.

The simplest way to dissect genetic interactions is by combining independently targeted alleles, which can be used to analyse redundancy and epistasis. More recently, gene targeting and ENU mutagenesis have been combined to identify modifier loci of mutant phenotypes in sensitized mutagenesis screens[53,54]. For example, new alleles interacting with the Delta–Notch signalling pathway were recently identified in a sensitized ENU mutagenesis screen on a delta-like 1 mutant background[55]. Because sensitized screens use the same strategy as phenotype-driven ENU mutagenesis screens, they could also be performed in large-scale settings.

Many studies have shown that one mutation can have distinct phenotypes when analysed on different genetic backgrounds. This is due to the presence of different alleles at modifying loci in various inbred strains[56,57]. To support identification of such loci, genetic diversity can be increased by generating inbred strains that harbour chromosomal segments or an entire chromosome transferred from a second inbred strain. Single chromosome substitution (consomic) strains and chromosomal segment substitution (congenic) strains make the study of complex traits more efficient[58–60] and enable identification of the alleles that are the basis of the observed variability between the two original inbred strains. The first full set of 21 chromosome substitution strains was generated less than a decade ago[61] and, since then, these strains have been used to identify several QTLs related to human multigenic disorders[62,63], such as the insulin-dependent diabetes 4 (*Idd4*) and *Idd5* loci[64–66].

However, there is evidence that the epistasis that occurs between QTLs is not simple or additive[67]. Instead, the traits are highly polygenic with several modifier loci per chromosome, and individual modifiers have profound effects on quantitative traits such that the sum of effects is larger than the difference between the parental strains. The finding that epistasis is strong and pervasive suggests that it will be important to take different genetic interactions into account when analysing complex traits. It would be particularly interesting to analyse how the different QTLs analysed in the cited study affect the network of co-expressed genes, and how these in turn cause variations in the analysed quantitative traits. Similarly, a previous study showed that genome-wide expression analysis allowed the identification of the functional gene regulation underlying QTLs[68].

To further explore the effect of genetic diversity on phenotypic variation, in 2004 the International Complex Trait Consortium embarked on the generation of approximately 1,000 recombinant inbred lines derived from 8 founder strains in the so-called Collaborative Cross[69]. In the second phase the mice are currently being inbred for 23 generations to achieve 99% inbreeding[70,71]. It is estimated that each Collaborative Cross strain will capture approximately 135 unique recombination events. In this resource of inbred lines, the genetic diversity between each Collaborative Cross line will be closer to the genetic variety among humans, and it is expected that this resource will be an important tool for quantitative trait analysis and systems biology.

Archiving and dissemination of mouse models

Coordinated initiatives for the preservation and distribution of mouse strains are required and already exist. A major aim is to ensure that the mutant mouse lines generated today are still available to the scientific community in the future. Thus, the dedicated mutant mouse archives cryopreserve mouse lines and distribute them as live stocks or frozen germ cells to the scientific community. As sending live stocks is generally more expensive than sending cryopreserved germ cells or embryos, teaching the methods for archiving and re-deriving mouse lines from frozen stocks is an important aim of organizations that are responsible for archiving mutant mice. These initiatives are assembled worldwide under the umbrella of the Federation of International Mouse Resources (FIMRe)[72].

Humanization of selected pathways and organs

Although the mouse has been a useful mammalian model system, in particular for basic research, it has inherent limitations. The lack of basic knowledge on molecular mechanisms that underlie human diseases is a major gap that needs to be filled for the next generation of medicine. Several research approaches aim for humanization of the mouse model to overcome at least some of these inherent limitations. Two of the most essential differences between mice and humans are that mice are small and have a short life cycle (from the view of experimenters, these characteristics are regarded as advantages rather than limitations). As a consequence, thermoregulation in mice and humans is under different constraints owing to different surface to volume ratios. The requirements for processes such as mutation repair or stress response differ in many aspects in an animal with a life span of 2 years compared with the requirements in humans, who have a life expectancy of ~70 years. The examples of mutation repair and stress response show how fundamental some of the differences between both species are. A systematic comparison of the similarities and differences between mice and humans at all developmental stages, and for all organ functions and molecular networks, has not yet been performed, but is urgently needed for the next generation of mouse models[73]. Some major differences have already been recognized — for example, differences concerning the immune system[74] — and are currently being tackled with advanced technologies to humanize the mouse for selected biological processes.

The detailed comparison of biochemical pathways or molecular networks in specific cases has already revealed differences between mice and humans. Such a comparison may be based on knowledge about mechanisms of disease or on more general biological processes, which can then be used for targeted genetic engineering to specifically humanize mouse models. For example, the human autoimmune disease bullous pemphigoid was reproduced in a genetically engineered mouse model in which the coding sequence of the autoantigen collagen 17 (*COL17*) was humanized[75]. For this, *Col17* knockout

Redundancy
When two genes can fulfil an equivalent function. Because gene functions are frequently pleiotropic, redundancy is often partial, with two genes having overlapping rather than equivalent functions.

Epistasis
The interaction between different genes that affect the same trait. Epistasis takes place when the phenotype of one genetic allele (mutant or natural variant) is modified by one or several other genes (also called modifier genes), such that the joint phenotype differs from the one that would be produced if the two genes were acting independently.

Sensitized mutagenesis screen
A phenotype-driven mutagenesis screen in which mice carrying a targeted mutation are bred with *N*-ethyl-*N*-nitrosourea-treated males in order to provide a sensitized system for detecting dominant modifier mutations.

Consomic
Describes a mouse strain that is produced by a breeding strategy in which recombinants between two inbred strains are backcrossed to produce a strain that carries a single chromosome from one strain on the genetic background of the other.

Congenic
Describes a mouse strain that is produced by a breeding strategy in which recombinants between two inbred strains are backcrossed to produce a strain that carries a single genomic segment from one strain on the genetic background of the other.

Box 1 | **Characterization of primary, secondary and tertiary screens**

Primary screen
- Basic parameters to reveal traits of interest
- Systemic analysis covering all organs
- Non-invasive
- Efficient analysis of a large number of animals
- Power calculation based on the numbers of animals required per screen
- Examples include dysmorphological analysis (such as external observation and the click box test), X-ray analysis and bone densitometry

Secondary screen
- Used for validation
- More detailed analysis
- May be performed on smaller numbers of animals
- More time and more expensive
- Examples include peripheral quantitative computed tomography, micro-computed tomography, markers of bone metabolism and hormonal regulation, mechanical bending of long bones and complete skeleton preparation

Tertiary screen
- In-depth analysis with selected animals
- Invasive methods — for example, telemetry
- Specifically designed for the experimental question
- One example is advanced bioimaging

mice were rescued by crossing them with a transgenic line expressing human *COL17* under the control of the human keratin 14 promoter.

Another strategy for generating humanized mouse models involves engrafting human cells into immune-compromised mice, which is a widely accepted method for generating metabolic models, for toxicity testing and for humanization of the immune system. For example, genetically modified mice were used to transplant human $CD34^+$ cord blood cells into mice, where they then develop into human B, T and dendritic cells[76]. This model has been used to study the pathobiology of Epstein–Barr virus[77] or HIV-1 (REF. 78) infection. Orthotopic xenografts have been applied in mice to study human breast neoplastic development[79]. As for genetic humanization, humanization by cell grafting may be applied to specific diseases on the basis of prior knowledge, and is probably not yet applicable to large-scale studies.

Phenotyping mouse models

To make mouse models more valuable for the scientific community, a major goal is to annotate existing and new mouse models with comprehensive phenotyping data to reveal affected organ systems[80]. This systemic phenotyping is essential to distinguish between the primary and secondary effects of genetic changes. In the past, many scientists focused on specific phenotypes and may have missed phenotypes outside of their interest and expertise. Systematic phenotyping will also be necessary to analyse the phenotypes of all available mouse models.

Systemic phenotyping and pleiotropic gene functions. For several reasons systemic phenotyping is an important aspect of large-scale mutagenesis programmes and specialized laboratories. A comprehensive phenotypic description makes new mutant mouse lines more valuable for basic and medical research. It also generates scientific interest into specific mutant mouse lines (through the generation of hypotheses) for subsequent more focused research to find mechanistic explanations. The primary screens should be designed to provide an overview of affected organs and, as such, should cover general phenotypic parameters of all organs (see BOX 1 for definitions of primary, secondary and tertiary screens).

Previously, mutant mouse line analyses were performed mostly in specialized laboratories, often using protocols that were not standardized between different institutions. The mouse phenotyping centres brought the expertise of phenotyping specialists together and established standardized protocols that are validated in different and geographically separated laboratories. One of the products of the European consortium, called EUMORPHIA (European Union Mouse Research for Public Health and Industrial Applications), is the first standard set of phenotyping protocols that were validated across several laboratories. A limited number of these phenotyping standard operating protocols form the EMPReSS (European Mouse Phenotyping Resource for Standardized Screens) slim primary phenotyping screen, which was developed to form a coherent sequence of tests to fully characterize a mutant mouse line[81–83] and is the minimal standard set of tests that are performed in all European mouse clinics. These tests include: screens for changes in morphology, metabolism, neurology and behaviour; screens for changes in the cardiovascular system, bones and sensory organs; measuring haematological parameters and clinical chemical parameters; and assessing indicators of allergy and the immune system in blood and serum. All EMPReSS slim protocols and additional validated phenotyping procedures are freely available from the EMPReSS website (see the further information box). The primary screening is designed as an entry point for further phenotypic analyses that aim to explain the underlying mechanisms of the phenotypes.

Unbiased screens, such as transcriptomics or proteomics profiling approaches, would be an ideal asset to systemic phenotyping protocols[84]. Expression profiling screens may be used to identify affected organs and to help classify mutant phenotypes that might otherwise have been regarded as identical. However, systematic expression profiling screens have generally not been included in primary phenotype screens, possibly owing to cost. The associated bottlenecks here include the systematic collection of organs from at least five individual mice for statistical significance, and the high prices of microarrays and molecular reagents to analyse each collected organ.

Nevertheless, gene expression screens can be performed at least as efficiently as other screens in the context of systemic phenotyping infrastructures. One possible strategy is to obtain a large panel of organs from each mouse line and select organs for transcript profiling on the basis of either previous knowledge (of gene function, spatiotemporal gene expression and so on) or phenotype data from other primary phenotype screens. In one study, this strategy revealed molecular gene expression

Phenotype screens	Methods	Age of mice (weeks)									
		8	9	10	11	12	13	14	15	16	17	18
Pipeline 1												
Dysmorphology	Anatomical observation		•									
	DEXA, X-ray							•				
Cardiovascular	Blood pressure				•							
	Heart weight								•			
Energy metabolism	Calorimetry					•						
Clinical chemistry	Simplified IPGTT						•					
Eye	Eye size (LIB)								•			
Lung function	Plethysmography									•		
Molecular phenotyping	Expression profiling									•		
Pipeline 2												
Behaviour	Open field		•									
	Acoustic startle and PPI				•							
Neurology	Modified SHIRPA, grip strength, rotarod		•	•								
Nociception	Hot plate					•						
Eye	Ophthalmoscopy and slit lamp						•					
Clinical chemistry	Clinical chemical analysis, haematology							•			•	
Immunology	FACS analysis of PBCs, immunoglobulin concentration							•			•	
Steroid metabolism	DHEA, testosterone							•			•	
Cardiovascular	ANP, ECG or echocardiogram							•	•	•	•	
Pathology	Macro and microscope analysis											•

Figure 2 | **Scheme of the primary phenotyping protocol of the German Mouse Clinic (GMC).** This scheme includes the EMPReSS slim primary phenotyping protocol, which is a common standard of European mouse clinics. Screens such as molecular phenotyping, lung function, steroid metabolism and pathological screens are performed in addition to the EMPReSS slim protocol. The GMC primary phenotyping screen starts 2 weeks after the mutant mouse lines are imported at the age of 9 weeks. For phenotypic analysis, the mice are distributed into one of two pipelines, in which they are subjected to a defined series of tests. The primary screen ends at the age of 18 weeks. Based on the results of the screens, decisions for secondary and tertiary screens are made. ANP, atrial natriuretic peptide; DEXA, dual-energy X-ray absorption; DHEA, dehydroepiandrosterone; ECG, electrocardiogram; FACS, fluorescence-activated cell sorting; IPGTT, intraperitoneal glucose tolerance test; LIB, laser interference biometry; PBC, peripheral blood cell; PPI, pre-pulse inhibition; SHIRPA, a protocol for comprehensive behaviour assessment. Figure is modified, with permission, from REF. 111 © Humana Press (2009).

phenotypes in approximately 50% of the analysed mutant lines[85]. In terms of the frequency of phenotype identification per mutant mouse line, gene expression screens are among the most efficient screens and they can be indispensable for unravelling subtle molecular mechanisms that underlie mutant phenotypes in mice. For example, despite extensive previous phenotypic analyses, it was only possible to identify the physiological function of a membrane transporter in renal epithelial cells by a genome-wide transcriptomics approach combined with proteomics and metabolomics[86].

Recent technological advances allow us to analyse more mutant phenotypes and extend the range of phenotypes that can be examined. For example, there has been tremendous progress in the use of microscopic and optical methods for the non-invasive analysis of anatomical, functional and molecular parameters in small rodents[87]. Technical improvements in resolution and specificity have allowed clinical imaging technologies, such as X-ray tomography, magnetic resonance imaging, nuclear imaging approaches and ultrasound imaging, to be adapted for use in small animals. Transgenically expressed genes that are tagged using fluorescence or bioluminescence can be monitored non-invasively *in vivo* throughout the entire mouse body[88]. The major advantages of these imaging methods include *in vivo* monitoring of therapeutic interventions and the longitudinal observation of a single animal using repeated observations. A recent and promising technological development combines optical imaging with other modalities, such as ultrasound, X-ray and magnetic resonance imaging. We foresee that at least some of these imaging technologies will become more common for phenotyping mouse models and in a few years may be included in standard protocols of mouse clinics. For example, X-ray computed tomography and high-frequency ultrasound biomicroscopy are imaging methods that are already included in the phenotyping protocols of the German Mouse Clinic (GMC).

An example of a successful set-up for systemic phenotyping comes from the GMC, which was the first phenotyping centre dedicated to the systemic and systematic analysis of mutant mouse lines and which performs one of the most comprehensive screens[80] (FIG. 2). The GMC has so far systemically analysed over 100 mutant mouse lines (over 6,000 mice) and has identified new and unexpected phenotypes in 96% of these mutants[89–96] (FIG. 3). Approximately one-third of these mouse lines had no previously described mutant phenotypes. These findings suggest that mutant phenotypes have been overlooked even in mouse models that have been studied for many years. These data also imply that in some cases the interpretation of mutant phenotypes may have to be reconsidered, because what has previously been regarded as primary gene function may be a secondary effect. For example, the vimentin (*Vim*) knockout allele was established more than a decade ago and extensively analysed. Known mutant phenotypes included cerebellar defects, impaired motor coordination, abnormal kidneys, delayed wound healing and impaired vascular tone[97–102]. However, the GMC systemic primary screen made the novel finding that the vimentin knockout mice are characterized by decreased cytotoxic and helper T cell subsets. This phenotype may be caused by a possible defect in T cell migration and is the basis for several other phenotypic alterations in these mutant mice.

Finally, the systemic phenotyping approach is particularly important to evaluate mutant mouse lines for their suitability as models for human diseases. Many of the most prevalent human diseases — such as type II diabetes, rheumatoid arthritis, neurodegeneration and other ageing-related disorders — affect multiple organs. The systemic phenotyping of mouse models will allow us to find concordances and differences between these human disorders and the corresponding mouse models.

Gearing up for systematic phenotyping. A large number of genes and markers in the mammalian genome are still not functionally annotated through experimental data. Thus, mutations in all genes need to be established in mutant mouse lines and subsequently phenotyped. This is the systematic aspect of systemic phenotyping.

Considering the 20,000 to 25,000 mammalian genes, each with one or more mutant alleles, and the number of primary phenotypic parameters that can be measured (currently between 300 and 400 parameters per mouse line) in cohorts of male and female mice, the question arises of how an endeavour of such a scope may be accomplished. A coordinated community effort is undoubtedly required[103] and several genomics centres in Asia, Australia, Canada, Europe and the United States have already established individual research infrastructures for the standardized and thorough phenotyping of the mouse mutant resources.

An important systematic phenotyping approach was initiated for the commonly used mouse inbred strains[104]. The data of this Mouse Phenome Project[105,106] are freely accessible from the Mouse Phenome Database[107] and are voluntarily contributed by researchers all over the globe, or in some cases retrieved from open public sources.

In the European Mouse Disease Clinic (EUMODIC) consortium, the academic phenotyping centres have coordinated their efforts to undertake a primary phenotype assessment of the first 500 mutant mouse lines. These mutant mice are generated by EUCOMM, which aims to produce 14,000 conditional mutant mouse alleles. The primary phenotyping is based on standardized EMPReSS protocols[81]. To our knowledge this is the only international initiative to systematically phenotype mouse models using a common standard phenotyping protocol. The included screens have been selected to give a comprehensive multi-system phenotype of mouse lines and are performed on age-matched cohorts of male and female mice. The primary mutant phenotype data generated through the EMPReSS protocols is made accessible in the EuroPhenome database and is published in peer-reviewed journals[83].

An additional challenge associated with systematic phenotyping concerns the compatibility and accessibility of phenotype data. Mouse phenotype data is still highly fragmented into several largely independent databases around the world under different standards and in different formats. The academic initiative InterPhenome has started to develop standards using ontologies and file formats for the description of phenotyping protocols and phenotype data sets[108]. CASIMIR (Coordination and Sustainability of International Mouse Informatics Resources) is a coordination effort supported by the European Commission, which aims to coordinate and integrate multiple site databases to address the problem of data fragmentation in dispersed databases.

It is evident that the capacities of the current phenotyping centres are not sufficient to analyse all mutant mouse lines that will be produced by the large-scale mutagenesis projects in the next decade[109]. To make significant progress their throughput will need to be increased from analysing hundreds of mutant lines per year to phenotyping thousands. Although automation may in part contribute to

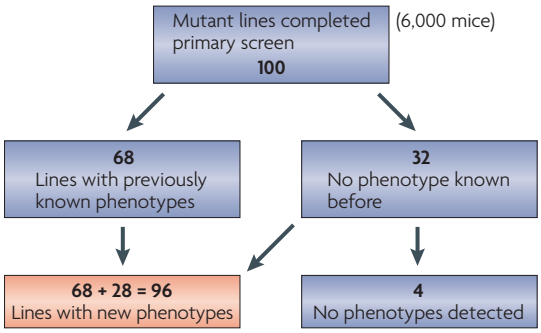


Figure 3 | **Frequency of new mutant phenotypes detected in mutant mouse lines analysed in the German Mouse Clinic (GMC).** New phenotypes have been identified in 96% of all mutant mouse lines that have been analysed in the GMC primary screen. Approximately two-thirds of the mutant mouse lines submitted to the GMC had known mutant phenotypes before the GMC primary screen. One-third of mutant mouse lines were submitted without any known mutant phenotype.

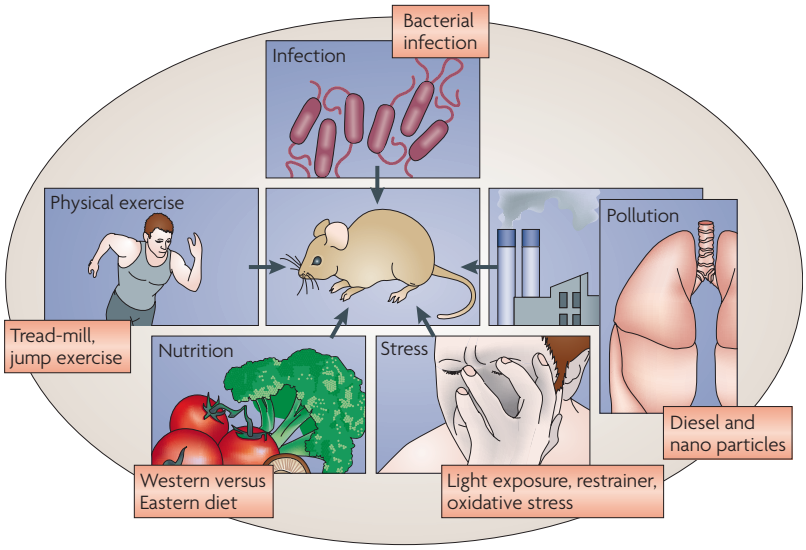


Figure 4 | **Schematic representation of the five environmental platforms currently being established at the German Mouse Clinic.** The five platforms in the blue boxes represent the major interfaces of the organism with the environment (gut, lung and skin, brain and sense organs, muscle and bone, and immune system), orange boxes give examples of environmental factors in each platform. Different test paradigms are currently being evaluated for their applicability and relevance for the mouse model system.

increased capacities, additional and larger phenotyping centres will be required in the future. The phenotyping infrastructures will also need long-term funding to have a significant effect on basic research and future medicine. In Europe, the Infrafrontier consortium is currently in the planning phase to reach a European agreement for a major upgrade, and for joint construction and implementation of the required infrastructures for mouse phenotyping and archiving of mutant mouse lines.

As the required phenotyping capacities are not yet available, one strategy is to prioritize the analysis of mutant mouse lines. A consensus of the centres that contribute to large-scale targeting and trapping of mouse genes suggested that one null allele of every gene should be phenotyped first. Conditional alleles may be given priority when knockout alleles are dominant or recessive lethal. Orthologues of human genes that have already been associated with diseases may also be prioritized. Genes that are specifically requested from the scientific community for basic research should also be given priority for systemic phenotyping.

Modelling envirotypes

Genotype and phenotype is a classical pair in genetic research. By contrast, the contribution of external factors to shaping the phenotype has been largely disregarded[4]. Similar to the map of genes and markers in the mammalian genome and their functional annotations, we will also need a map of exogenous factors for the next generation of mouse models. Sets of exogenous factors can then be used to describe and define complex envirotypes that may correspond, for example, to different human life styles (such as exercise versus resting) or social and geographical cultures (such as Western versus Eastern diet or rural versus urban culture). The extent to which the effects of particular envirotypes are the same in mice and humans remains to be determined. Indeed, most human envirotypes are not fully defined. Therefore, the characterization of envirotype–phenotype correlations will also be important to further improve the mouse as a model for human diseases. Experimental settings will be required that allow the design of investigational envirotypes that can be integrated into or associated with dedicated systemic phenotyping protocols. As an example, the GMC has started the implementation of challenging platforms that serve exactly these requirements.

The current primary systemic screening protocol at the GMC almost exclusively focuses on the analysis of mice under resting conditions in a 'protected' specified pathogen-free environment. Therefore, the primary screen will preferentially identify alterations involved in the homeostatic regulation of basic organ and cell functions and does not take into account the influence of external factors. However, some human diseases require environmental triggering factors in order to become apparent — allergic asthma, for example. Without the specific trigger, the individual might be phenotypically normal even though it carries genetic variations that potentially play a vital part in the pathophysiology of the disease. The German Mouse Clinic II (GMC II) is the first to set up standardized challenge platforms for mouse phenotyping to explore the complex relationship between the envirotype, genotype and phenotype. Challenge platforms are currently being set up that focus on major environmental risk factors for human health. Five areas — diet, air, stress, exercise and immunity — were chosen that represent the major interfaces of the organism with the environment (that is, gut, lung and skin, brain and sensory organs, muscle and bone, and immune system) (FIG. 4). For the different platforms, defined challenge conditions will be implemented for phenotypic analyses, which incorporate the latest bioimaging methods. By mimicking specific environmental exposures or life styles that have a strong impact on human health, their effects on molecular networks and on disease aetiology and progression will be determined, thereby uncovering the physiological and molecular mechanisms of interactions between the genome and environment.

The combination of exogenous factors might also shed light into the crosstalk between environmental challenges, which is poorly understood. The experimental introduction of combinations of challenges and environmental heterogeneity is hoped to provide conditions that are more comparable to those that humans experience[110]. Envirotypes that are relevant to humans can be designed as far as they are characterized and understood. However, the differences between mice and humans also have to be considered, such as those in olfactory functions, immune response and higher brain function. These differences might lead to alternative effects of environmental stimuli, which lead to different phenotypic manifestations. However, comparative analyses at all 'omics' levels might help to overcome this challenge by providing more precise classifications at the gene expression level. Finally, we note that introducing experimental envirotypes into

mouse phenotyping protocols will provide the chance to discover interventions or alterations in life style that might have positive effects on human health.

The exposure of mouse models to environmental challenges is not new. Challenge tests are used widely in the research community and have provided important results. However, envirotypes are more complex, and they need to be defined and introduced into the experimental set-up. The power of GMC II lies in the combination of challenges involving complex environmental conditions (for example, challenging diets and exercise under stress conditions). However, combining sets of exogenous factors meaningfully and efficiently remains an issue.

Conclusions

The functional study of mammalian biology in mouse models faces many important challenges that are of a much larger scope than previous genomics projects. In terms of genotyping, it is likely that the current mouse resources contain mutations for most genes, but it will be essential to generate mutant lines that more closely resemble human diseases. In addition, a systematic comparison of all organ functions and molecular networks between humans and mice will be essential to better understand and fully exploit the power of existing and future mouse models. Another major challenge is that more coordination between international phenotyping centres will be required in order to systemically and systematically phenotype all existing mouse models and guarantee accessibility and compatibility of phenotype data. Finally, the analysis of envirotypes poses a major challenge for the next generation of mouse models and is just beginning to be addressed.

1. Nadeau, J. H. *et al.* Sequence interpretation. Functional annotation of mouse genome sequences. *Science* **291**, 1251–1255 (2001).
2. Austin, C. P. *et al.* The knockout mouse project. *Nature Genet.* **36**, 921–924 (2004).
3. Auwerx, J. *et al.* The European dimension for the mouse genome mutagenesis program. *Nature Genet.* **36**, 925–927 (2004).
4. Patten, B. C. in *Eco Targets, Goal Functions, and Orientors* (eds Muller, F. & Leupelt, M.) 137–160 (Springer, Berlin, 1998).
We believe that this publication introduced the term 'envirotype'. In particular, it is mentioned that "the genotype–phenotype pair of classical genetics is an incomplete specification of determinate reproduction; an external envirotype is needed to complete the mechanism."
5. Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E. & Blake, J. A. The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res.* **36**, D724–D728 (2008).
6. Paigen, K. One hundred years of mouse genetics: an intellectual history. II. The molecular revolution (1981–2002). *Genetics* **163**, 1227–1235 (2003).
References 6 and 7 give an excellent historical overview of 100 years of mouse genetics, from its beginning to the genomic era.
7. Paigen, K. One hundred years of mouse genetics: an intellectual history. I. The classical period (1902–1980). *Genetics* **163**, 1–7 (2003).
8. Qiu, J. Animal research: mighty mouse. *Nature* **444**, 814–816 (2006).
9. Collins, F. S., Finnell, R. H., Rossant, J. & Wurst, W. A new partner for the international knockout mouse consortium. *Cell* **129**, 235 (2007).
10. Collins, F. S., Rossant, J. & Wurst, W. A mouse for all reasons. *Cell* **128**, 9–13 (2007).
11. Hansen, G. M. *et al.* Large-scale gene trapping in C57BL/6N mouse embryonic stem cells. *Genome Res.* **18**, 1670–1679 (2008).
12. Gondo, Y. Trends in large-scale mouse mutagenesis: from genetics to functional genomics. *Nature Rev. Genet.* **9**, 803–810 (2008).
13. Carlson, C. M. *et al.* Transposon mutagenesis of the mouse germline. *Genetics* **165**, 243–256 (2003).
14. Ding, S. *et al.* Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* **122**, 473–483 (2005).
15. Dupuy, A. J., Akagi, K., Largaespada, D. A., Copeland, N. G. & Jenkins, N. A. Mammalian mutagenesis using a highly mobile somatic *Sleeping Beauty* transposon system. *Nature* **436**, 221–226 (2005).
16. Keng, V. W. *et al.* A conditional transposon-based insertional mutagenesis screen for genes associated with mouse hepatocellular carcinoma. *Nature Biotechnol.* **27**, 264–274 (2009).
17. Rosenthal, N. & Brown, S. The mouse ascending: perspectives for human-disease models. *Nature Cell Biol.* **9**, 993–999 (2007).
A comprehensive and insightful recent review of the genetic tools available for the mouse and the challenges that mouse models of human diseases are facing.
18. Gieger, C. *et al.* Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.* **4**, e1000282 (2008).
19. McCarroll, S. A. & Altshuler, D. M. Copy-number variation and association studies of human disease. *Nature Genet.* **39**, S37–S42 (2007).
20. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genet.* **40**, 1166–1174 (2008).
21. Justice, M. J., Noveroske, J. K., Weber, J. S., Zheng, B. & Bradley, A. Mouse ENU mutagenesis. *Hum. Mol. Genet.* **8**, 1955–1963 (1999).
22. Soewarto, D., Klaften, M. & Rubio-Aliaga, I. Features and strategies of ENU mouse mutagenesis. *Curr. Pharm. Biotechnol.* **10**, 198–213 (2009).
23. Nolan, P. M. *et al.* Implementation of a large-scale ENU mutagenesis program: towards increasing the mouse mutant resource. *Mamm. Genome* **11**, 500–506 (2000).
24. Nolan, P. M. *et al.* A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nature Genet.* **25**, 440–443 (2000).
25. Hrabe de Angelis, M. H. *et al.* Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nature Genet.* **25**, 444–447 (2000).
26. Balling, R. ENU mutagenesis: analyzing gene function in mice. *Annu. Rev. Genomics Hum. Genet.* **2**, 463–492 (2001).
27. Vreugde, S. *et al.* Beethoven, a mouse model for dominant, progressive hearing loss DFNA36. *Nature Genet.* **30**, 257–258 (2002).
This paper of an ENU-induced mouse model was published along with an article that describes patients with the same progressive deafness phenotype caused by a mutation in the homologous human gene.
28. Lisse, T. S. *et al.* ER stress-mediated apoptosis in a new mouse model of osteogenesis imperfecta. *PLoS Genet.* **4**, e7 (2008).
29. Klaften, M. & Hrabe de Angelis, M. ARTS: a web-based tool for the set-up of high-throughput genome-wide mapping panels for the SNP genotyping of mouse mutants. *Nucleic Acids Res.* **33**, W496–W500 (2005).
30. Augustin, M. *et al.* Efficient and fast targeted production of murine models based on ENU mutagenesis. *Mamm. Genome* **16**, 405–413 (2005).
31. Coghill, E. L. *et al.* A gene-driven approach to the identification of ENU mutants in the mouse. *Nature Genet.* **30**, 255–256 (2002).
32. Michaud, E. J. *et al.* Efficient gene-driven germ-line point mutagenesis of C57BL/6J mice. *BMC Genomics* **6**, 164 (2005).
33. Quwailid, M. M. *et al.* A gene-driven ENU-based approach to generating an allelic series in any gene. *Mamm. Genome* **15**, 585–591 (2004).
34. Sakuraba, Y. *et al.* Molecular characterization of ENU mouse mutagenesis and archives. *Biochem. Biophys. Res. Commun.* **336**, 609–616 (2005).
35. Sakuraba, Y. *et al.* Identification and characterization of new long conserved noncoding sequences in vertebrates. *Mamm. Genome* **19**, 703–712 (2008).
36. Takahasi, K. R., Sakuraba, Y. & Gondo, Y. Mutational pattern and frequency of induced nucleotide changes in mouse ENU mutagenesis. *BMC Mol. Biol.* **8**, 52 (2007).
37. Herault, Y., Rassoulzadegan, M., Cuzin, F. & Duboule, D. Engineering chromosomes in mice through targeted meiotic recombination (TAMERE). *Nature Genet.* **20**, 381–384 (1998).
38. Olson, L. E., Richtsmeier, J. T., Leszl, J. & Reeves, R. H. A chromosome 21 critical region does not cause specific Down syndrome phenotypes. *Science* **306**, 687–690 (2004).
39. Kmita, M., Fraudeau, N., Herault, Y. & Duboule, D. Serial deletions and duplications suggest a mechanism for the collinearity of *Hoxd* genes in limbs. *Nature* **420**, 145–150 (2002).
40. Couzin, J. RNA interference. Mini RNA molecules shield mouse liver from hepatitis. *Science* **299**, 995 (2003).
41. Kunath, T. Transgenic RNA interference to investigate gene function in the mouse. *Methods Mol. Biol.* **461**, 165–186 (2008).
42. Raoul, C. *et al.* Lentiviral-mediated silencing of SOD1 through RNA interference retards disease onset and progression in a mouse model of ALS. *Nature Med.* **11**, 423–428 (2005).
43. Bibikova, M., Golic, M., Golic, K. G. & Carroll, D. Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. *Genetics* **161**, 1169–1175 (2002).
44. Lloyd, A., Plaisier, C. L., Carroll, D. & Drews, G. N. Targeted mutagenesis using zinc-finger nucleases in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **102**, 2232–2237 (2005).
45. Zeevi, V., Tovkach, A. & Tzfira, T. Increasing cloning possibilities using artificial zinc finger nucleases. *Proc. Natl Acad. Sci. USA* **105**, 12785–12790 (2008).
46. Mani, M., Kandavelou, K., Dy, F. J., Durai, S. & Chandrasegaran, S. Design, engineering, and characterization of zinc finger nucleases. *Biochem. Biophys. Res. Commun.* **335**, 447–457 (2005).
47. Steuber-Buchberger, P., Wurst, W. & Kuhn, R. Simultaneous Cre-mediated conditional knockdown of two genes in mice. *Genesis* **46**, 144–151 (2008).
48. Zender, L. *et al.* An oncogenomics-based *in vivo* RNAi screen identifies tumor suppressors in liver cancer. *Cell* **135**, 852–864 (2008).
49. Hitz, C., Steuber-Buchberger, P., Delic, S., Wurst, W. & Kuhn, R. Generation of shRNA transgenic mice. *Methods Mol. Biol.* **530**, 1–29 (2009).
50. Echeverri, C. J. *et al.* Minimizing the risk of reporting false positives in large-scale RNAi screens. *Nature Methods* **3**, 777–779 (2006).
51. Meng, X., Noyes, M. B., Zhu, L. J., Lawson, N. D. & Wolfe, S. A. Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. *Nature Biotechnol.* **26**, 695–701 (2008).
52. Yan, Z., Sun, X. & Engelhardt, J. F. Progress and prospects: techniques for site-directed mutagenesis in animal models. Gene *Ther.* 19 Feb 2009 (doi:10.1038/gt.2009.16).
53. Matera, I. *et al.* A sensitized mutagenesis screen identifies *Gli3* as a modifier of *Sox10* neurocristopathy. *Hum. Mol. Genet.* **17**, 2118–2131 (2008).

54. Mohan, S., Baylink, D. J. & Srivastava, A. K. A chemical mutagenesis screen to identify modifier genes that interact with growth hormone and TGF-beta signaling pathways. *Bone* **42**, 388–395 (2008).
55. Rubio-Aliaga, I. *et al.* A genetic screen for modifiers of the delta1-dependent Notch signaling function in the mouse. *Genetics* **175**, 1451–1463 (2007).
56. Dietrich, W. F. *et al.* Genetic identification of Mom-1, a major modifier locus affecting Min-induced intestinal neoplasia in the mouse. *Cell* **75**, 631–639 (1993). **A hallmark article on the genetic mapping of the first quantitative trait gene or modifier affecting a mouse model for human disease.**
57. Erickson, R. P. Mouse models of human genetic disease: which mouse is more like a man? *Bioessays* **18**, 993–998 (1996).
58. Gregorova, S. *et al.* Mouse consomic strains: exploiting genetic divergence between *Mus m. musculus* and *Mus m. domesticus* subspecies. *Genome Res.* **18**, 509–515 (2008).
59. Rogner, U. C. & Avner, P. Congenic mice: cutting tools for complex immune disorders. *Nature Rev. Immunol.* **3**, 243–252 (2003).
60. Matin, A., Collin, G. B., Asada, Y., Varnum, D. & Nadeau, J. H. Susceptibility to testicular germ-cell tumours in a 129.MOLF-Chr 19 chromosome substitution strain. *Nature Genet.* **23**, 237–240 (1999).
61. Nadeau, J. H., Singer, J. B., Matin, A. & Lander, E. S. Analysing complex genetic traits with chromosome substitution strains. *Nature Genet.* **24**, 221–225 (2000).
62. Singer, J. B. *et al.* Genetic dissection of complex traits with chromosome substitution strains of mice. *Science* **304**, 445–448 (2004).
63. Singer, J. B., Hill, A. E., Nadeau, J. H. & Lander, E. S. Mapping quantitative trait loci for anxiety in chromosome substitution strains of mice. *Genetics* **169**, 855–862 (2005).
64. Grattan, M., Mi, Q. S., Meagher, C. & Delovitch, T. L. Congenic mapping of the diabetogenic locus Idd4 to a 5.2-cM region of chromosome 11 in NOD mice: identification of two potential candidate subloci. *Diabetes* **51**, 215–223 (2002).
65. Hill, N. J. *et al.* NOD Idd5 locus controls insulitis and diabetes and overlaps the orthologous CTLA4/IDDM12 and NRAMP1 loci in humans. *Diabetes* **49**, 1744–1747 (2000).
66. Lamhamedi-Cherradi, S. E. *et al.* Further mapping of the Idd5.1 locus for autoimmune diabetes in NOD mice. *Diabetes* **50**, 2874–2878 (2001).
67. Shao, H. *et al.* Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. *Proc. Natl Acad. Sci. USA* **105**, 19910–19914 (2008). **An insightful article on the frequent occurrence of QTLs in the rodent genome and how the interaction of QTLs is neither simple nor additive.**
68. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429–435 (2008).
69. Churchill, G. A. *et al.* The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genet.* **36**, 1133–1137 (2004).
70. Iraqi, F. A., Churchill, G. & Mott, R. The Collaborative Cross, developing a resource for mammalian systems genetics: a status report of the Wellcome Trust cohort. *Mamm. Genome* **19**, 379–381 (2008).
71. Chesler, E. J. *et al.* The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. *Mamm. Genome* **19**, 382–389 (2008).
72. Davisson, M. FIMRe: Federation of International Mouse Resources: global networking of resource centers. *Mamm. Genome* **17**, 363–364 (2006).
73. Davis, M. M. A Prescription for human immunology. *Immunity* **29**, 835–838 (2008).
74. von Herrath, M. G. & Nepom, G. T. Lost in translation: barriers to implementing clinical immunotherapeutics for autoimmunity. *J. Exp. Med.* **202**, 1159–1162 (2005).
75. Nishie, W. *et al.* Humanization of autoantigen. *Nature Med.* **13**, 378–383 (2007).
76. Traggiai, E. *et al.* Development of a human adaptive immune system in cord blood cell-transplanted mice. *Science* **304**, 104–107 (2004).
77. Cocco, M. *et al.* CD34$^+$ cord blood cell-transplanted Rag2$^{-/-}\gamma_c^{-/-}$ mice as a model for Epstein–Barr virus infection. *Am. J. Pathol.* **173**, 1369–1378 (2008).
78. Gorantla, S. *et al.* Human immunodeficiency virus type 1 pathobiology studied in humanized BALB/c-Rag2$^{-/-}\gamma_c^{-/-}$ mice. *J. Virol.* **81**, 2700–2712 (2007).
79. Proia, D. A. & Kuperwasser, C. Reconstruction of human mammary tissues in a mouse model. *Nature Protoc.* **1**, 206–214 (2006).
80. Gailus-Durner, V. *et al.* Introducing the German Mouse Clinic: open access platform for standardized phenotyping. *Nature Methods* **2**, 403–404 (2005).
81. Brown, S. D., Chambon, P. & de Angelis, M. H. EMPReSS: standardized phenotype screens for functional annotation of the mouse genome. *Nature Genet.* **37**, 1155 (2005).
82. Green, E. C. *et al.* EMPReSS: European mouse phenotyping resource for standardized screens. *Bioinformatics* **21**, 2930–2931 (2005).
83. Mallon, A. M., Blake, A. & Hancock, J. M. EuroPhenome and EMPReSS: online mouse phenotyping resource. *Nucleic Acids Res.* **36**, D715–D718 (2008).
84. Woychik, R. P., Klebig, M. L., Justice, M. J., Magnuson, T. R. & Avner, E. D. Functional genomics in the post-genome era. *Mutat. Res.* **400**, 3–14 (1998).
85. Horsch, M. *et al.* Systematic gene expression profiling of mouse model series reveals coexpressed genes. *Proteomics* **8**, 1248–1256 (2008).
86. Frey, I. M. *et al.* Profiling at mRNA, protein, and metabolite levels reveals alterations in renal amino acid handling and glutathione metabolism in kidney tissue of Pept2$^{-/-}$ mice. *Physiol. Genomics* **28**, 301–310 (2007).
87. Ntziachristos, V., Culver, J. P. & Rice, B. W. Small-animal optical imaging. *J. Biomed. Opt.* **13**, 011001 (2008).
88. Niedre, M. J. *et al.* Early photon tomography allows fluorescence detection of lung carcinomas and disease progression in mice *in vivo*. *Proc. Natl Acad. Sci. USA* **105**, 19126–19131 (2008).
89. Ahting, U. *et al.* Neurological phenotype and reduced lifespan in heterozygous *Tim23* knockout mice, the first mouse model of defective mitochondrial import. *Biochim. Biophys. Acta* 9 Dec 2008 (doi:10.1016/j.bbabio.2008.12.001).
90. Soker, T. *et al.* Pleiotropic effects in *Eya3* knockout mice. *BMC Dev. Biol.* **8**, 118 (2008).
91. Hoelter, S. M. *et al.* "Sighted C3H" mice — a tool for analysing the influence of vision on mouse behaviour? *Front. Biosci.* **13**, 5810–5823 (2008).
92. Schmidt, S. *et al.* Deletion of glucose transporter GLUT8 in mice increases locomotor activity. *Behav. Genet.* **38**, 396–406 (2008).
93. Fuchs, H. *et al.* Phenotypic characterization of mouse models for bone-related diseases in the German Mouse Clinic. *J. Musculoskelet. Neuronal Interact.* **8**, 13–14 (2008).
94. Bender, A. *et al.* Creatine improves health and survival of mice. *Neurobiol. Aging* **29**, 1404–1411 (2008).
95. Vauti, F. *et al.* The mouse *Trm1-like* gene is expressed in neural tissues and plays a role in motor coordination and exploratory behaviour. *Gene* **389**, 174–185 (2007).
96. Barrantes Idel, B. *et al.* Generation and characterization of *dickkopf3* mutant mice. *Mol. Cell Biol.* **26**, 2317–2326 (2006).
97. Colucci-Guyon, E., Gimenez, Y. R. M., Maurice, T., Babinet, C. & Privat, A. Cerebellar defect and impaired motor coordination in mice lacking vimentin. *Glia* **25**, 33–43 (1999).
98. Colucci-Guyon, E. *et al.* Mice lacking vimentin develop and reproduce without an obvious phenotype. *Cell* **79**, 679–694 (1994).
99. Eckes, B. *et al.* Impaired wound healing in embryonic and adult mice lacking vimentin. *J. Cell Sci.* **113**, 2455–2462 (2000).
100. Henrion, D. *et al.* Impaired flow-induced dilation in mesenteric resistance arteries from mice lacking vimentin. *J. Clin. Invest.* **100**, 2909–2914 (1997).
101. Schiffers, P. M. *et al.* Altered flow-induced arterial remodeling in vimentin-deficient mice. *Arterioscler. Thromb. Vasc. Biol.* **20**, 611–616 (2000).
102. Terzi, F. *et al.* Reduction of renal mass is lethal in mice lacking vimentin. Role of endothelin-nitric oxide imbalance. *J. Clin. Invest.* **100**, 1520–1528 (1997).
103. Welsh, E., Jirotka, M. & Gavaghan, D. Post-genomic science: cross-disciplinary and large-scale collaborative research and its organizational and technological challenges for the scientific research process. *Philos. Transact. A Math. Phys. Eng. Sci.* **364**, 1533–1549 (2006). **A sociological study on the far-reaching impact that the advent of 'big science' in life science research is beginning to have, for example, in the areas of organizational cultures, working practice, rewarding systems, education and communication technology.**
104. Paigen, K. & Eppig, J. T. A mouse phenome project. *Mamm. Genome* **11**, 715–717 (2000).
105. Bogue, M. Mouse Phenome Project: understanding human biology through mouse genetics and genomics. *J. Appl. Physiol.* **95**, 1335–1337 (2003).
106. Bogue, M. A. & Grubb, S. C. The Mouse Phenome Project. *Genetica* **122**, 71–74 (2004).
107. Bogue, M. A., Grubb, S. C., Maddatu, T. P. & Bult, C. J. Mouse Phenome Database (MPD). *Nucleic Acids Res.* **35**, D643–D649 (2007).
108. Hancock, J. M. & Mouse Phenotype Database Integration Consortium. Integration of mouse phenome data resources. *Mamm. Genome* **18**, 157–163 (2007).
109. Brown, S. D., Hancock, J. M. & Gates, H. Understanding mammalian genetic systems: the challenge of phenotyping in the mouse. *PLoS Genet.* **2**, e118 (2006).
110. Richter, S. H., Garner, J. P. & Wurbel, H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? *Nature Methods* **6**, 257–261 (2009).
111. Valérie Gailus-Durner *et al.* in *Gene Knockout Protocols* 2nd edn Vol. 530 (eds Wurst, W. & Kühn, R.) 436–509 (Humana Press, New Jersey, 2009).

Acknowledgements
The authors are funded through the German Ministry of Science and Education and the European Commission (grant numbers: 01GS0850, LSHG-2006-037188, 211414, LSHG-CT-2006-518240, MRTN-CT-2006-035468).

FURTHER INFORMATION

Australian Phenomics Facility (APF): http://www.apf.edu.au
CASIMIR: http://www.casimir.org.uk
Charles River's phenotyping screens, Massachusetts: http://www.criver.com/en-US/ProdServ/ByType/Discovery/Pages/PhenotypingServices.aspx
Comparative Pathology Laboratory (CPL): http://www.vetmed.ucdavis.edu/ars/cpl.htm
EMPReSS: http://empress.har.mrc.ac.uk
EUCOMM: http://www.eucomm.org
EUMODIC: http://www.eumodic.org
EUMORPHIA: http://www.eumorphia.org
EuroPhenome: http://www.europhenome.org
FIMRe: http://www.fimre.org
Frimorfo, Switzerland: http://www.frimorfo.com
German Mouse Clinic (GMC): http://www.mouseclinic.de
Helmholtz Centre Munich: http://www.helmholtz-muenchen.de/en
IMGS: http://imgs.org
Infrafrontier: http://www.infrafrontier.eu
Institut Clinique de la Souris (ICS): http://www-mci.u-strasbg.fr/index.html
InterPhenome: http://www.interphenome.org
Jackson Laboratory Phenotyping Services: http://jaxservices.jax.org/phenotyping/index.html
KOMP: http://www.nih.gov/science/models/mouse/knockout
Laboratory Animal Sciences Program (LASP): http://web.ncifcrf.gov/rtp/lasp/phl
Mammalian Genetics Phenotyping: http://www.gnf.org/technology/organismal/mammalian-genetics-phenotyping.htm
Mary Lyon Centre: http://www.har.mrc.ac.uk
MGD: http://www.informatics.jax.org
Mouse Genetics Programme — Phenotyping: http://www.sanger.ac.uk/Teams/Team109/phenotyping.shtml
Mouse Phenome Database: http://phenome.jax.org/pub-cgi/phenome/mpdcgi?rtn=docs/home
Mouse Phenotyping Shared Resource (MPSR): http://www.vet.ohio-state.edu/255.htm
NorCOMM: http://norcomm.phenogenomics.ca
Phenotyping Core: http://www.hopkinsmedicine.org/mcp/PHENOCORE
Research Animal Diagnostic Laboratory (RADIL): http://www.radil.missouri.edu
RIKEN BioResource Center: http://www.brc.riken.go.jp/inf/en/index.shtml
Taconic Farms, Inc.: http://www.taconic.com/RAS/phenotyping.htm
TIGM: http://www.tigm.org
Toronto Centre for Phenogenomics (TCP): http://www.phenogenomics.ca
Unit for Laboratory Animal Medicine: http://www.ulam.umich.edu/services/pathcons.htm
Yale University Mouse Research Pathology (YMRP): http://mrp.yale.edu/index.html

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Mapping genes for complex traits in domestic animals and their use in breeding programmes

Michael E. Goddard[*‡] *and Ben J. Hayes*[‡]

Abstract | Genome-wide panels of SNPs have recently been used in domestic animal species to map and identify genes for many traits and to select genetically desirable livestock. This has led to the discovery of the causal genes and mutations for several single-gene traits but not for complex traits. However, the genetic merit of animals can still be estimated by genomic selection, which uses genome-wide SNP panels as markers and statistical methods that capture the effects of large numbers of SNPs simultaneously. This approach is expected to double the rate of genetic improvement per year in many livestock systems.

Quantitative trait
A measurable trait that depends on the cumulative action of many genes and the environment, and that can vary among individuals over a given range to produce a continuous distribution of phenotypes.

Estimated breeding value
An estimate of the additive genetic merit for a particular trait that an individual will pass on to its descendents.

Heritability
The proportion of phenotypic variance caused by additive genetic variation.

*Department of Agriculture and Food Systems, University of Melbourne, Royal Parade, Parkville 3010, Australia. ‡Biosciences Research Division, Department of Primary Industries, Victoria, 1 Park Drive, Bundoora 3083, Australia. Correspondence to M.E.G. e-mail: mike.goddard@dpi.vic.gov.au
doi:10.1038/nrg2575

Traits that are controlled by a single gene, such as flower colour in pea plants, have been important in elucidating the mechanisms of heredity, yet most traits that are important in agriculture, medicine and evolution are complex or quantitative traits. These traits include: susceptibility to many diseases, such as diabetes in humans; agriculturally important traits, such as the milk yield of dairy cows; and traits that affect fitness in the wild, such as the clutch size of birds.

Identifying genes for complex traits would greatly enhance our understanding of these traits, but in domestic animals there would also be a practical benefit to agriculture. Traditionally, the genetics of complex traits in these species has been studied without identifying the genes involved. Selection has been based on estimated breeding values calculated from phenotypic records and pedigrees, and on knowledge of the heritability of each trait. This has been successful, but the process is slow if the trait can only be measured in one sex (for example, milk yield), after death (for example, meat quality) or late in life (for example, longevity), or if measuring the trait is expensive (for example, methane production, feed requirement or disease resistance). Therefore, to improve on these traits, it would be advantageous to identify genes for them and select animals carrying the desirable alleles[1]. Thus, compared with research on complex traits in humans, there is greater emphasis in domesticated animals on predicting genetic merit and phenotype and less emphasis on discovering genes and pathways. However, both aims are important and are covered in this Review.

Over the past 20 years, two approaches have been used to discover the genes and polymorphisms contributing to variation in complex traits. In one approach candidate genes have been targeted based on their role in the physiology of the trait (for example, the expression level of milk proteins), and in the other approach the genes that affect a trait of interest have been mapped to a chromosomal location using genetic markers[2–4]. However, progress in identifying the causal genes for complex traits has been slow as linkage mapping results in large confidence intervals.

The recent availability of large panels of SNPs in domestic species has given new momentum to the search for the mutations underlying variation in complex traits through the use of genome-wide association (GWA) studies. This Review concentrates on domesticated species, especially those for which the possibility of dissecting the architecture of important quantitative traits is enhanced by the availability of genome-wide SNP panels — these species include cattle, dogs and chickens. Genome-wide SNP panels for sheep, pigs and horses have also recently become available.

We first discuss the experimental designs, statistical analyses and results of this research. GWA studies have successfully identified genes causing simple Mendelian traits, but these studies have not yet identified genes for complex traits. Animal breeders can nevertheless use the results of GWA studies for genetic improvement of domestic animals by using a technique called genomic selection, which potentially leads to large increases in the rate of genetic improvement[5]. Genomic selection is

already being used successfully to select dairy cattle carrying desirable alleles for milk yield and other traits, and is likely to be applied to all livestock in the near future as well as to crops and aquaculture. We describe the science behind this revolution and, finally, we discuss the directions for future research into complex traits in domesticated animals.

The history of QTL mapping

In the decades before the advent of GWA studies, genes affecting production and fitness traits in domestic animals were mapped to chromosomal regions using linkage analysis and linkage disequilibrium (LD) between markers and QTLs. As we explain below, differences between animals and humans in family structure and in the extent of LD affected the power and precision of mapping in animal studies compared with observations in human studies.

Linkage analysis. Early attempts to map QTLs used blood groups as genetic markers[6–8]. However, the power of this approach was markedly enhanced by the identification of numerous highly variable microsatellite markers. Typically, a linkage analysis was performed using microsatellite markers, often with large half-sibling families[9]. The ability to generate hundreds of offspring per sire made this approach more powerful in livestock than in humans. However, linkage analysis usually mapped the QTLs to a large interval of 20 centimorgans (cM) or more[4,9], which made it difficult to identify the underlying mutation and to use the marker information in animal breeding programmes.

LD and the effect of effective population size. More precise mapping is possible using LD between markers and QTLs because LD decays quickly as the distance between marker and QTL increases. A linkage analysis uses recombination events in the recorded pedigree and traces chromosome segments to a common ancestor in the pedigree. By contrast, LD mapping relies on chromosome segments inherited from a common ancestor before the recorded pedigree — this is because it is the inheritance of identical chromosome segments by multiple descendents from a common ancestor that causes LD.

The pattern of LD observed in a population depends on the history of the population, especially the history of its effective population size (N_e)[10–12]. A small N_e means that alleles in the current population coalesce in a common ancestor in a small number of generations. This means that there are few generations of recombination; the chromosome segments that are identical by descent are large, and so LD extends for a long distance. This explanation assumes a constant N_e but, in practice, the N_e of a population can change over time. For instance, in *Bos taurus* cattle N_e was large before domestication (>50,000), declined to 1,000–2,000 after domestication and, in many breeds, declined to approximately 100 after breed formation[13,14]. This causes some LD to exist at long distances (>1 cM) but not to increase markedly until very short distances are reached[13]. However, the long-range LD does not apply across breeds because animals from different breeds do not share a recent common ancestor. This N_e history is similar to that experienced by dogs, which show a similar pattern of LD to cattle[15], but is the opposite of that experienced by humans. The European human N_e was only ~3,000 but then increased enormously in the last 10,000 years[12] (FIG. 1). Consequently, humans have similar LD to cattle at short distances but almost no LD at long distances (FIG. 2a).

This pattern of LD in domestic animals means that a marker may be in LD with a QTL some distance away and hence show an association with the trait affected by the QTL. Consequently, one does not need as dense a panel of SNPs for a GWA study in many domestic animal species as in humans. Conversely, markers that are located several centimorgans from the QTL can show an association to the trait, making precise mapping more difficult. This problem can be overcome by using multiple breeds: markers that show a consistent pattern of LD with a QTL across breeds must be close to that QTL (FIG. 2b).

The traditional mapping strategy was to use linkage to map a QTL to a large region and then use LD to map it more precisely. However, discovering and typing a panel of dense markers across the confidence interval of a single QTL was a major undertaking until the arrival of genome-wide panels of SNPs.

GWA studies

Principles and tools. The basic design of a GWA study is that a sample of animals are recorded for a trait of interest and assayed for a genome-wide panel of markers to detect statistical associations between the trait and any of the markers. Design parameters include the choice and number of animals and markers. Most commonly, the data from a GWA study are analysed one SNP at a time using a simple linear model that includes: the effect of a SNP; fixed effects, such as the cohort or group to which the animal belongs; and the polygenic breeding value of each animal, which is due to all other genes affecting the trait.

The genomic sequence is available for several domestic species, including cattle, horses, chickens and dogs, and large numbers of SNPs were discovered as a by-product of the sequencing or in subsequent resequencing. These SNPs can be typed using the same technology as in humans, and commercial 'SNP chips' exist for cattle (50,000 SNPs), dogs (22,362 SNPs; Illumina CanineSNP20 BeadChip), sheep (56,000 SNPs), pigs (60,000 SNPs; Illumina PorcineSNP60 BeadChip), horses (54,602 SNPs; Illimina EquineSNP50 BeadChip) and chickens. However, these chips contain less SNPs than the latest human SNP chips with over 1,000,000 SNPs.

Sources of bias. An important source of false positive associations is admixture in the sample of individuals used. The most obvious case of this problem would be a sample consisting of a mixture of breeds. Fortunately this problem is easily avoided by including breed (if known) in the statistical model used to analyse the data.

Genetic improvement
Deliberate genetic change in a population of domestic animals or plants brought about by human control of their selection and breeding that makes them more suitable for the purpose for which they are kept.

Genomic selection
Selection of animals for breeding based on estimated breeding values calculated from the joint effects of genetic markers covering the whole genome.

Linkage disequilibrium
The absence of linkage equilibrium so that the allele at one locus is correlated with the allele at another locus.

Effective population size
The number of individuals in an idealized population with random mating and no selection that would lead to the same rate of inbreeding as observed in the real population. The effective population size can be much less than the actual population size owing to the unequal genetic contribution of individuals to the next generation.

Linear model
A statistical model that assumes that the observed phenotypic value can be explained by the sum of the effects of independent variables and a random error, which is usually assumed to be normally distributed.

Polygenic breeding value
The additive genetic merit an individual passes on to its descendents owing to the combined contribution of many genes of small effect, but possibly excluding some specified genes.

Admixture
A population or sample of individuals derived from more than one race or breed and that have not undergone random mating.

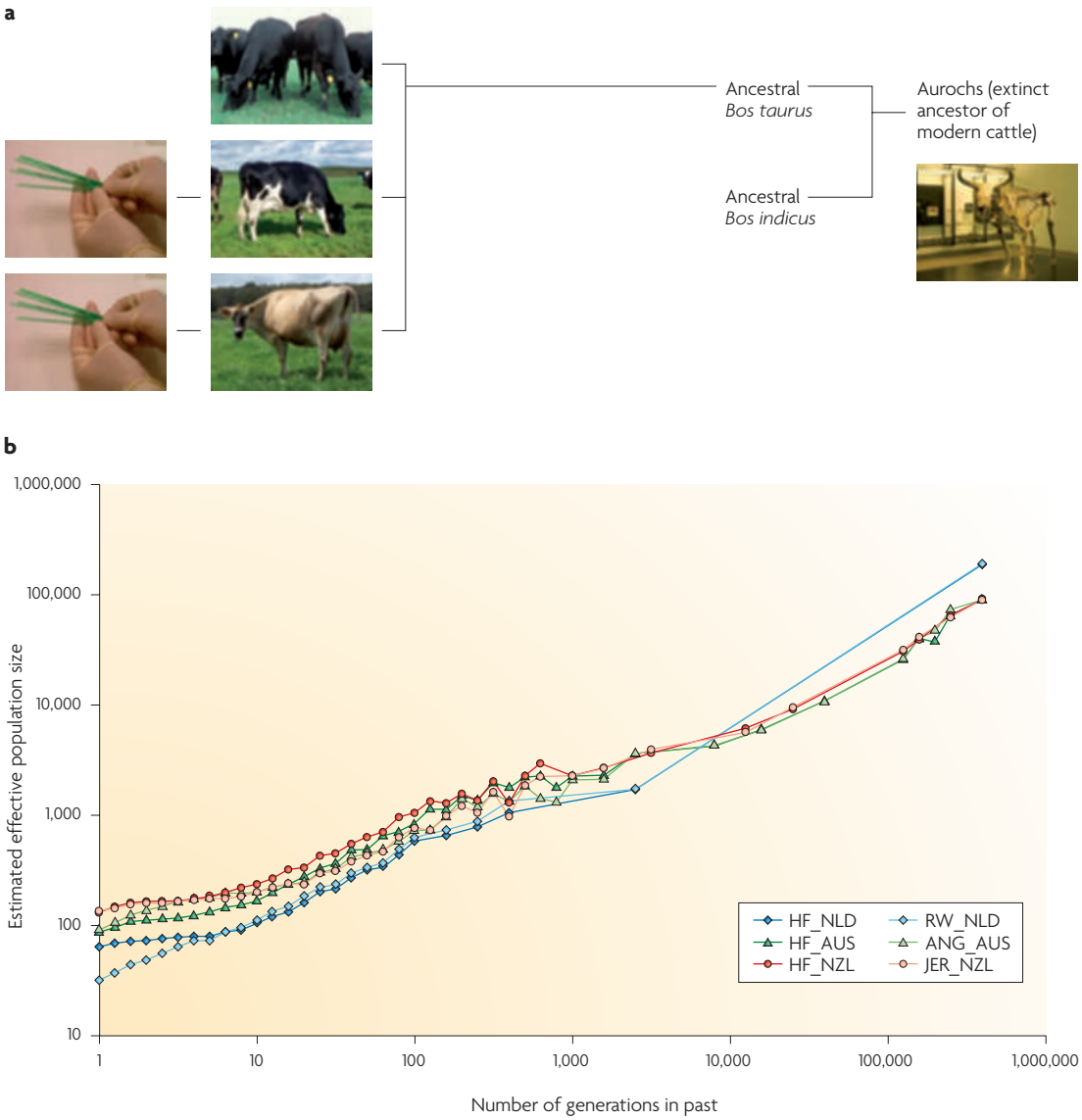


Figure 1 | **Key events in the history of cattle. a** | Approximately 1 million years ago, the *Bos* genus diverged from other *Bovidae*. Between 1,000,000 and 500,000 years ago, the *Bos taurus* species and *Bos indicus* species diverged. Approximately 10,000 years ago, both species were domesticated. 400 to 100 years ago, deliberate breed formation began. Recently, the widespread use of artificial insemination has further reduced effective population size in some breeds. **b** | The graph depicts effective population size along the population history, estimated from the average linkage disequilibrium at different marker distances, for Dutch black and white Holstein–Friesian bulls (HF_NLD), Dutch red and white Holstein–Friesian bulls (RW_NLD), Australian Holstein–Friesian bulls (HF_AUS), Australian Angus cattle (ANG_AUS), New Zealand Friesian cows (HF_NZL) and New Zealand Jersey cows (JER_NZL)[13]. Effective population size was large before domestication (>50,000) and declined to 1,000–2,000 after domestication, and then declined again to ~100 owing to breed formation and modern breeding programmes. Aurochs image in part **a** is courtesy of Roberto Fortuna, National Museum of Denmark. Part **b** is modified, with permission, from REF. 13 © Genetics Society of America (2008).

A more subtle form of admixture is the existence of relationships among the animals. The implication of an 'association' between a marker and a trait is that it exists across the whole population, and so specially designed mapping families are not needed. In fact, a sample of unrelated individuals would be ideal. However, livestock are usually bred in half-sibling families (for example, cattle) or full-sibling families (for example, pigs). Relationships among the animals in the sample cause LD between loci even if they are unlinked. For instance, if the sire of a family carries rare alleles at two unlinked loci, his offspring will be more likely than other animals to carry both rare alleles. If one of these loci is a QTL, this generates an association between the other locus and the trait. This problem can be overcome by including in the statistical model a term for the effect of all other genes affecting the trait (the polygenic term). In our experience, omitting the polygenic term in the

model approximately doubles the number of false positive associations (I. Macleod, personal communication). Fortunately, the inclusion of a polygenic term in the statistical analysis is becoming more common.

Although the inclusion of a polygenic term in the statistical model eliminates associations between a QTL and markers that are not linked to it, it does not eliminate associations between QTLs and linked markers. To

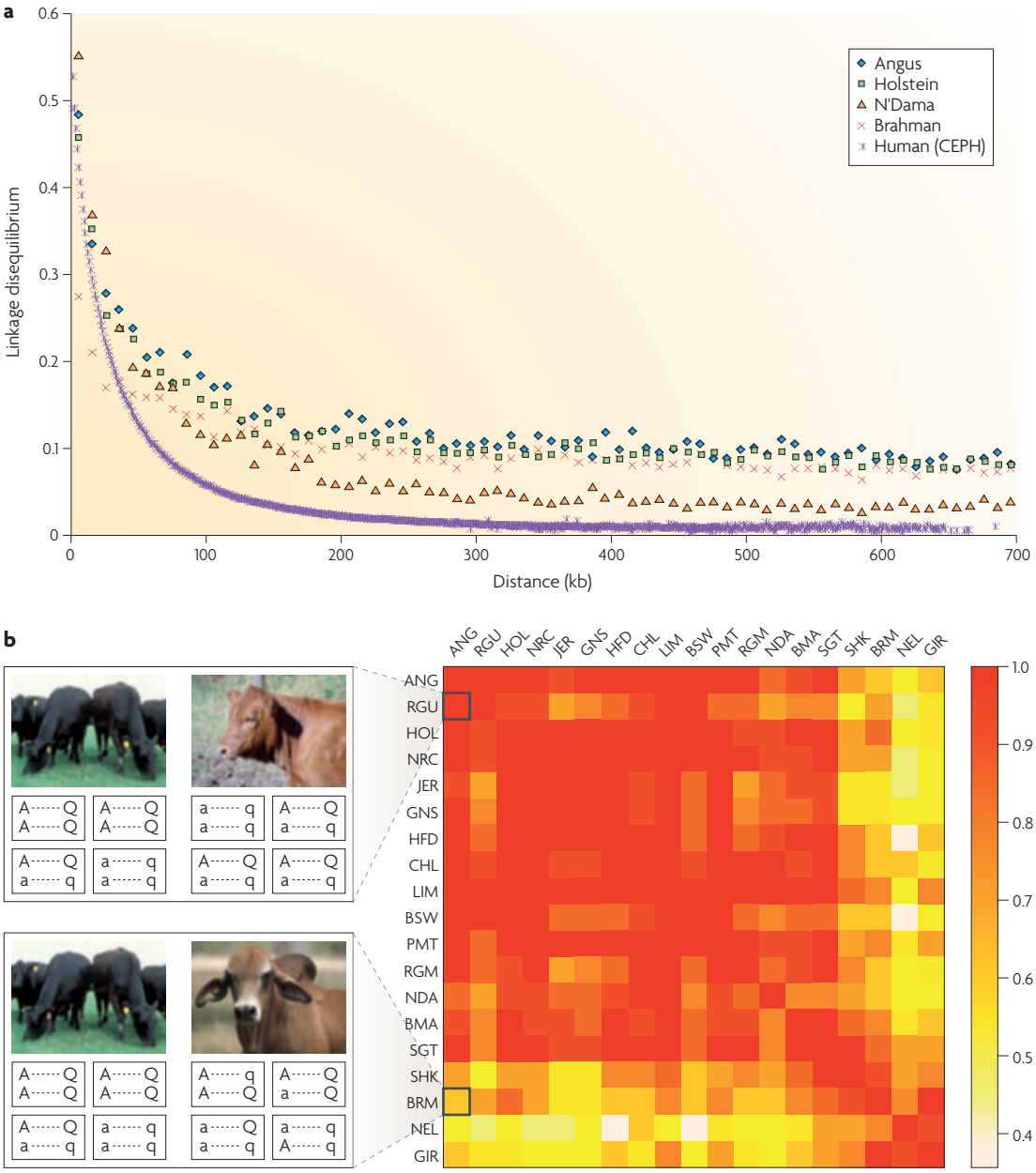


Figure 2 | **Linkage disequilibrium (LD) in cattle breeds. a** | Decline of LD with distance between pairs of SNPs as measured by LD within breeds of cattle (derived from approximately 35,000 SNPs, and a human population with northern and western European ancestry (CEPH cohort))[12]. **b** | LD between breeds of cattle. The heat map shows the correlation between LD in different breeds for SNPs within 10 kb of each other. For two closely related breeds (Angus and Red Angus) the correlation is high, as shown in a hypothetical example in which a–q and A–Q chromosomes are common in both breeds (upper box). However, when Angus is compared with Brahman (a distantly related breed) the correlation is low and, in the hypothetical example, Brahman chromosomes often carry a–Q, which is a rare haplotype in Angus (lower box). In fact, the correlation is low for any combination of a *Bos indicus* breed and a *Bos taurus* breed[65]. ANG, Angus; BMA, Beefmaster; BRM, Brahman; BSW, Brown Swiss; CHL, Charolais; GIR, Gir; GNS, Guernsey; HFD, Hereford; HOL, Holstein; JER, Jersey; LIM, Limousin; NDA, N'Dama; NEL, Nelore; NRC, Norwegian Red; PMT, Piedmontese; RGM, Romagnola; RGU, Red Angus; SGT, Santa Getrudis; SHK, Sheko. Data for part **a** is taken from REFS 12,65. Data for part **b** are courtesy of the Bovine HapMap Consortium. Part **a** and the heat map in part **b** are modified, with permission, from REF. 65 © (2009) American Association for the Advancement of Science.

extend the previous example, if a sire carries two rare alleles at linked loci, there will be an association between them in the offspring. That is, if there are relationships among the animals used for the GWA study, the associations discovered represent a mixture of associations caused by LD and associations caused by linkage. By linkage, we mean that the associations exist within a family or families but not across the whole population. This has two effects. First, SNPs some distance from a QTL may show an association to the trait, exacerbating the problem caused by long-range LD in many species of domestic animals. Second, these associations may be specific to the sample of families studied and may not be replicated in another sample of families from the same population.

Meuwissen *et al.*[16] published results that were produced by accurately incorporating linkage disequilibrium and linkage analysis (LDLA), and this technique has been used to fine map QTLs in chickens, dairy cattle and pigs[17–19]. However, the volume of data generated by GWA studies has caused scientists to revert to simpler linear models that consume less computer time than LDLA. One feature of LDLA is that it can be used to estimate the confidence interval for the position of the QTL[16]. By contrast, single SNP analysis in livestock often yields a collection of significant SNPs spread over many centimorgans, reflecting the extensive range of low-level LD, and it is difficult to define the most likely QTL position and its confidence interval.

Calculating the number of SNPs to be analysed. The number of SNPs needed depends on the distance over which LD operates. If SNPs are too far apart a QTL may not be in sufficient LD with any of the markers, and so will be undetected. Increasing the SNP density will increase the power to detect QTLs and, to some extent, increase the precision of mapping. However, if LD is high over a chromosome segment, increasing SNP density may still not allow one to position the QTLs precisely within this segment.

Sutter *et al.*[15] stated that only 10,000 SNPs are needed for within-breed analyses in dogs, but that 30,000 SNPs are needed for between-breed analyses. In cattle, significant associations were found within a breed using only 10,000 SNPs, but we estimate that 300,000 SNPs would be needed for between-breed analyses in *B. taurus* cattle. This number was calculated using the data in FIG. 2b, which shows that the SNPs need to be spaced less than 10 kb apart to show consistent LD phase across breeds. Between-breed analyses would need a SNP density similar to the widely used 375,000 SNP chip in humans; this is not surprising as short-range LD is similar in humans and cattle. The common ancestor of *B. taurus* and *Bos indicus* cattle dates to >500,000 years ago, so several million SNPs would be needed for SNPs to be close enough to each QTL that the same LD phase is consistently found in both subspecies. This marker density across the two subspecies could lead to very accurate mapping of QTLs if they segregate in both subspecies, but it is not known how often this is the case. Among the small number of known QTLs, a mutation in the *DGAT1* gene, which affects fat percentage in milk, segregates only in *B.taurus*[20], whereas mutations in calpastatin and calpain, which affect meat tenderness, segregate in both *B. taurus* and *B. indicus*[21].

Estimating the number of study animals. The number of animals needed for a GWA study depends on the size of the effects that one wishes to detect. The crucial parameter is the proportion of the variance explained by the SNP. This parameter combines the allele frequency with the mean difference between the SNP genotypes. An approximate idea of the number needed can be gained from the following simple calculation. The correlation (r) between the marker and the trait, $r(t,m)$, is equal to $r(m,q) \times r(q,g) \times r(g,t)$, in which m is the marker genotype (usually scored 0, 1 or 2), q is the QTL genotype, g is the genetic value of the animal and t is the phenotypic value of the animal. $r^2(m,q)$ is the conventional r^2 measure of LD, $r^2(q,g)$ is the proportion of genetic variance explained by the QTL, and $r^2(g,t)$ is the heritability of the trait.

For instance, if $r^2(m,q) = 0.50$, $r^2(q,g) = 0.04$ and $r^2(g,t) = 0.25$, then $r(t,m) = 0.07$. If we require a standard error equal to 0.33, then the expected correlation and the number of animals required is 1,800 (as the standard error of a correlation coefficient is the square root of the number of animals). In practice, some SNPs explain more than 4% of the genetic variance assumed above, and so a smaller experiment would suffice but, in fact, most SNPs associated with human complex traits explain less than 4% of the genetic variance, and so over 1,800 animals would be needed[22]. This assumes that the sizes of QTL effects are similar in domestic animals to those reported in humans; our preliminary findings support this assumption (M.E.G and B.J.H., unpublished observations).

The number of animals referred to above is the number for which both genotypes and phenotypes have been directly measured. The number can be reduced by using animals that have been progeny tested so that the mean of their progeny can be used instead of their own phenotypic value. This is advantageous as long as the mean phenotype of the progeny is more highly correlated with the animal's breeding value than is its own phenotype. The formulae above still apply but now $r^2(g,t)$ is the reliability of the progeny test and t is the progeny mean.

Results of GWA studies

Monogenic trait mapping. As has been the case for older experimental methods of gene mapping, the greatest successes in identifying genes using GWA have been for monogenic traits (TABLE 1). For instance, Karlsson *et al.*[23] discovered that mutations in the pigmentation-related gene *MITF* cause white spotting in dogs. To map this gene they used two breeds, boxers and bull terriers, in which the white spotting trait was segregating, and then used common haplotypes between the two breeds to increase the precision of mapping. In cattle, Charlier *et al.*[24] used GWA to identify three genes harbouring mutations causing three recessive abnormalities. This

LD phase
If linkage disequilibrium (LD) exists between genes A and B, each with two alleles (A or a and B or b), then gametes that carry allele A can carry B or b. Thus, LD can exist in one of two phases: gametes that are more commonly AB and ab, or gametes that are more commonly Ab and aB.

Table 1 | **Genes harbouring mutations affecting monogenic traits in dogs and cattle discovered by genome-wide association**

Phenotype	Breed	Number of samples used	Gene harbouring causative mutation	Gene description	Refs
Dog					
Hairless	Chinese crested, Peruvian hairless, Mexican hairless	195	*FOXI3*	Forkhead box transcription factor family, expressed in developing hair and teeth	66
Degenerative myelopathy	Pembroke Welsh corgi	38 cases, 17 controls	*SOD1*	Superoxide dismutase 1, soluble	67
Cone–rod dystrophy	Wire-haired dachshund	13 discordant sibling pairs	*NPHP4*	Nephronophthisis 4	68
White spotting	Boxers, bull terriers	146	*MITF*	Microphthalmia-associated transcription factor	23
Size and weight	148 breeds	2,801	*IGF1*	Insulin-like growth factor 1	35
Hair ridge	Rhodesian and Thai ridgeback, other ridgeless breeds	21 in genome-wide association, 91 in subsequent fine mapping	Duplication of *FGF3*, *FGF4*, *FGF19* and *ORAOV1*	133-kb duplication involving three fibroblast growth factor genes	69
Cattle					
Congenital muscular dystonia 1	Belgian Blue	12 cases, 14 controls	*ATP2A1*	ATPase, Ca^{2+} transporting, cardiac muscle, fast twitch 1	24
Congenital muscular dystonia 2	Belgian Blue	7 cases, 24 controls	*SLC6A5*	Solute carrier family 6 (neurotransmitter transporter, glycine), member 5	24
Ichthyosis fetalis	Italian Chianina	3 cases, 9 controls	*ABCA12*	ATP-binding cassette, sub-family A (ABC1), member 12	24

The genes here were selected as they have been validated in additional studies and, in many cases, have led to the discovery of the causative mutation.

study was particularly efficient because the low N_e in cattle means that, typically, calves suffering from a fatal recessive disorder are homozygous for a large chromosome segment containing the causative gene, allowing this segment to be detected using moderately dense markers. For complex traits much larger numbers of animals would be needed, but the basic strategy seems applicable.

Complex trait mapping. For complex traits the results generally indicate many mutations, suggesting that the individual mutations each have a small effect. For example, Kolbehdari *et al.*[25] reported 196 SNPs in Canadian Holstein bulls with significant associations with traits describing the size and shape of cows. In the same breed, Daetwyler *et al.*[26] found 144 significant SNPs for milk protein yield, Barendse *et al.*[27] found significant SNPs for feed conversion efficiency, and Lillehammer *et al.*[28] found significant SNPs for a genotype × environment interaction for milk yield at the level of herd production. In chickens, SNPs with significant associations with mortality in broilers have been reported, both associations across environments and associations that differentially affect mortality in two different hygiene environments[29]. Another study in chickens reported 21 SNPs linked to 19 genes associated with resistance to *Salmonella enterica* colonization[30].

Validating SNP associations. However, in none of these cases were the significant SNPs confirmed in an independent sample. There are three reasons for this. First, the effect size of each association is small, even in the original GWA study. However, when you account for the so-called Beavis effect[31], the true effects are even smaller and so a very large confirmation experiment is needed to have the required power to confirm the effect. Second, the LD between the SNP and the QTL may be present in the original sample of animals but not in other samples from a different breed or even from different families within the same breed (FIG. 2b). Third, the false discovery rate is often high, and so most of the significant associations are just those expected by chance when so many SNPs are tested.

In our own unpublished GWA studies, we find that SNP associations are most likely to be confirmed when the original GWA study used a large number of animals (>1,000) that were widely sampled from one breed, when the SNPs were highly significant, and when the confirmation was carried out in a large sample of the same breed (M.E.G and B.J.H., unpublished observations).

By contrast, GWA studies in humans have found many confirmed associations with complex traits, such as height[22,32] and susceptibility to disease[33]. However, recently published papers have used consortia of scientists to amass GWA studies with tens of thousands of subjects and confirmation samples of approximately equal or greater size (the Wellcome Trust Case Control Consortium[34] is one such consortium). Hopefully, experiments in domestic animals will soon be published that reflect the lessons learned from human studies. For instance, researchers in the Animal Improvement Programs Laboratory (AIPL) at the United States Department of Agriculture have genotyped 7,000 Holstein bulls that could be in used in a powerful GWA study (see their Genomic Comparison of Young Bulls).

Beavis effect
The tendency for statistically significant effects to be overestimated when many effects are tested for significance.

Box 1 | **Genetic architecture of complex traits**

One of the surprising results in human genome-wide association (GWA) studies has been the small size of the observed effects[22,32,34], which implies that many SNPs have effects on complex traits. This conclusion is supported by the results in domestic animals. However, the number of QTLs could be less than the number of SNPs with significant effects. Each QTL could be tracked by many SNPs because no individual SNP is in complete linkage disequilibrium (LD) with the QTL, especially in domestic animals in which LD extends over a wide distance and SNP density is not high.

A second surprising result from human and domestic animal GWA studies is that the SNPs with validated effects only explain a small proportion of the genetic variance, leading to a question — where are the missing genes?[58] Most QTLs might explain such a small proportion of the variance that even large human GWA studies lack the power to find them. This is consistent with the recent findings that combining data sets from different GWA studies leads to the discovery of additional genes[59]. Also, it may be that many QTLs are not in high LD with any one SNP. Although LD between common SNPs is high, QTLs may have different properties to common SNPs. For instance, most QTLs may be subject to selection so that polymorphisms are typically young and the minor allele frequency is low. Such polymorphisms show less LD with markers than common SNPs do, and so may not be detected by GWA studies that rely on LD. This is the case for mutations that are fatal and hence never reach high frequency, but one might expect that even QTLs of small effect are subject to some selection. If this were not true then genetic variance, which is caused each generation by mutation, would accumulate in large populations to reach high levels, and hence heritabilities would be high for all traits. This is not the case; therefore, selection must be eliminating variance in complex traits.

Some QTLs of larger effect have been discovered, such as the *DGAT1* polymorphism that explains about 40% of the genetic variation in fat content in the milk of Holstein cattle[60]. Therefore, the distribution of effects of QTLs must have many small effects but a tail with larger but rare effects. An exponential distribution has been suggested[61,62]. It is likely that a mutation with a large effect on a complex trait will also be subject to larger selection pressure. In many cases this will tend to eliminate the mutant allele, but in some cases artificial selection by humans may select for the mutation. For instance, mutations in the myostatin gene that increase muscling would be detrimental in the wild but have been selected for by cattle and sheep breeders[63,64].

Do genes affecting complex traits typically have multiple alleles segregating? If so, this could be one reason why QTLs are hard to pinpoint but, unfortunately, we do not know the answer because not enough QTLs have been found. It is clear that multi-allelic series exist at major genes — such as for the myostatin gene in cattle for which many double muscling mutations exist, as well as mutations of lesser effect[63]. However, these allelic series may only be discovered when the mutations are positively selected for by humans owing to their novelty or practical value. Therefore, if nature selects against most QTL mutations, it may be rare for a QTL to have more than two alleles segregating, especially within a breed whose effective population size is small.

Another design has also been employed successfully in dogs. Jones *et al.*[35] used the extensive variation between 148 dog breeds as the basis for mapping genes affecting size and behaviour. This approach identified insulin-like growth factor 1 (*IGF1*) as a gene that affects size or weight, and found SNPs that are possibly associated with pointing and herding behaviour. The validity of this approach rests on the assumption that the breeds used are a random sample of unrelated breeds. However, breeds tend to come in related groups (such as gun dogs) and so a SNP and a trait might seem to be associated because they both occur in a group of related breeds. This problem might be overcome by a very wide sampling of breeds so that related breeds make up a small part of the sample. This design assumes that the same mutations are polymorphic in different breeds. This is true for some well-characterized mutations, such as the K232A mutation in *DGAT1*, which is polymorphic in Holstein, Jersey and Ayrshire cattle[36]. Other mutations, such as some functional mutations in the myostatin gene, seem to be breed specific[37].

The ideal design of a GWA study depends on the genetic architecture of complex traits, and the results of GWA studies in humans and in livestock are providing information about this architecture. In BOX 1 we argue that many mutations of different types affect a typical complex trait and that most of their effects are small but, despite this, they are often subject to weak natural selection and consequently have a low minor allele frequency (MAF). This architecture means that large numbers of individuals are needed for a GWA study to have the necessary power to find the genes explaining most of the genetic variance in a complex trait.

Minor allele frequency
The frequency of the less frequent allele in a two-allele polymorphism.

Marker-associated selection

One justification for conducting GWA studies in livestock is to use the validated markers to select better livestock through marker-assisted selection (MAS)[38]. There are two types of MAS. The first makes use of a causative mutation that has been identified in a gene or regulatory region — such mutations typically have a major effect, as in a monogenic trait. Examples include: the halothane gene in pigs that increases muscle growth but makes the pigs susceptible to stress and halothane anaesthesia[39]; the booroola gene in sheep[40] that increases the number of lambs born to ewes that carry the gene; numerous recessive abnormalities, such as bovine leukocyte adhesion deficiency in cattle[41]; and a mutation in the *PRNP* locus that alters the resistance of sheep to scrapie[42]. In most cases the purpose of selection is to eliminate the abnormal allele from the population, although it can also be used to increase the frequency of a rare allele, for instance, by introgressing the booroola gene from Merino sheep into Border Leicester sheep[43].

Genomic breeding value
An estimate of an animal's genetic merit, including genomic information

The second type of MAS makes direct use of SNPs that are in LD with QTLs. First, the effect associated with each allele of the significant marker or markers is estimated; to avoid bias the effect is ideally estimated in a population that is independent from the one in which the significant markers were discovered. Breeding values for selection candidates can then be estimated by combining pedigree, marker and phenotype information[44,45]. This type of MAS has been applied to improve reproduction rate, feed intake, growth rate and body composition in various livestock species, meat quality in commercial lines of pigs, muscle development in sheep, and milk yield in dairy cattle[45,46]. The key criticism of MAS applied in this way is that its ability to predict breeding values is limited. This is because a low number of markers with validated associations typically explain a small proportion of the genetic variance in the trait.

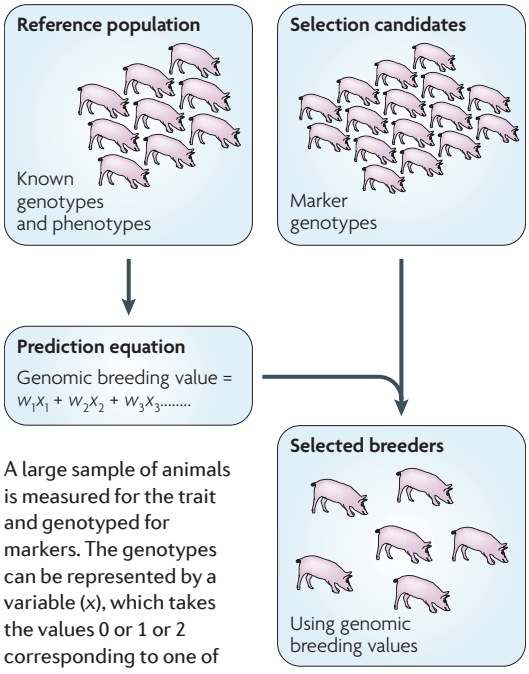
Genomic selection

Experimental design. To overcome the deficiencies of MAS, Meuwissen *et al.*[5] suggested a different approach, known as genomic selection. The key difference between the two approaches is that MAS concentrates on a small number of QTLs that are tagged by markers with well-verified associations, whereas genomic selection uses a genome-wide panel of dense markers so that all QTLs are in LD with at least one marker. Genomic selection has two advantages. First, all the genetic variance for a trait can be tracked by the marker panel. This is true even if the experiment lacks the power needed to detect all significant QTLs, as a marker effect does not need to exceed a stringent significance threshold to be used to predict breeding value or phenotype. Increasing the power does, however, increase the accuracy with which the marker effects are estimated. Second, the effect of the marker alleles can be estimated on a population basis rather than within each family, because the markers and the QTLs are in LD.

For genomic selection one needs a sample of animals that have been assayed for the markers and recorded for the trait — this is the reference population. This sample is analysed to derive a prediction equation that predicts breeding value from marker genotypes — the genomic breeding value — such that the effect of each marker is predicted simultaneously with the other markers. This formula can then be applied to predict the breeding value of selection candidates that have marker genotypes but no trait record (BOX 2). Thus, as for other forms of MAS[1], genomic selection is particularly advantageous for traits that are difficult to record at a young age. For instance, dairy bulls are 5 years old by the time they can be assessed on the basis of their daughters' milk yields. Genomic selection of dairy bulls at 1 year of age could greatly reduce the generation interval and hence speed up the rate of genetic improvement[47]. Of course, the success of genomic selection depends on the accuracy with which breeding value can be predicted in the selection candidates (FIG. 3).

Results of genomic selection. In simulated data Meuwissen *et al.*[5] found the accuracy of the genomic breeding value — that is, the correlation between the genomic breeding value and the true breeding value — to be 0.85. Results from real data have not reached this level of accuracy, but VanRaden *et al.*[48] reported a correlation of 0.71 in Holstein–Friesian dairy cattle, averaged across a number of traits. They used a reference population of 3,576 bulls genotyped for 38,416 SNPs. Phenotypes for the bulls were the averages of their daughters' production records. For comparison, the accuracy of estimated breeding values for calves at birth, based on the average of their parents' breeding value, is only about 0.5. Harris *et al.*[49] reported similar accuracies of genomic breeding value in New Zealand Holstein–Friesian and Jersey dairy cattle, and Hayes *et al.*[50] reported somewhat lower accuracies for the genomic breeding value from a much smaller reference population in Australian Holstein–Friesians. In mice, using genomic predictions — including additive SNP effects or both additive and dominance SNP effects — instead of using pedigree information alone can give a higher accuracy of phenotype prediction for various traits, including weight, growth slope, body mass index, body length, coat colour, percentage of $CD8^+$ cells present and mean cellular haemoglobin[51,52]. In chickens, González-Recio *et al.*[53] were able to show an almost fourfold increase in the accuracy of prediction of yet-to-be observed phenotypes for

Box 2 | **Genomic selection**



A large sample of animals is measured for the trait and genotyped for markers. The genotypes can be represented by a variable (*x*), which takes the values 0 or 1 or 2 corresponding to one of the homozygotes, the heterozygote or the other homozygote. The statistical analysis of the reference population estimates effects for each marker (*w*), and hence a prediction equation can be generated that combines all the marker genotypes with their effects to predict the breeding value of each animal. This prediction equation can then be applied to a group of animals that have genotypes but not phenotypes, and the estimated breeding values calculated from this can be used to select the best animals for breeding.

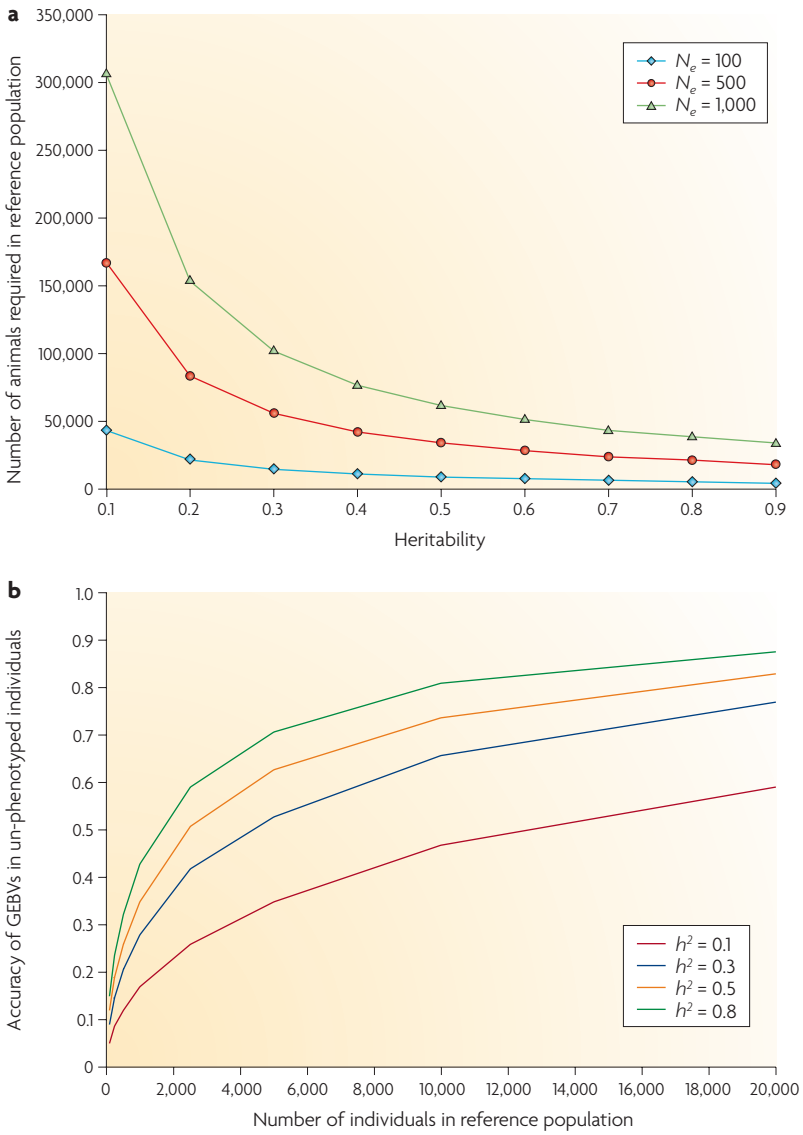


Figure 3 | **Calculation of number of animals in a reference population and accuracy of breeding values. a** | Number of animals needed in a reference population. To achieve an accuracy of 0.7 for estimated genomic breeding values (GEBVs) calculated from SNPs requires an increasing number of animals in the reference population as the heritability declines or the N_e of the population increases. **b** | Accuracy of GEBVs of un-phenotyped individuals with increasing number of phenotype records in the reference population used to estimated SNP effects, for different heritabilities (h^2). N_e was 100.

food conversion rate in broilers when genomic predictions of phenotype were used compared with pedigree predictions of phenotype.

Some of the statistical methods for genomic selection have been reviewed elsewhere[54,55]. The various methods make assumptions about the distribution of SNP effects on the trait, such as the proportion of the SNPs that have any effect on the trait. The best results have been obtained by methods that assume that many thousands of SNPs have an effect on traits such as milk yield[48], which is consistent with the results of GWA studies[26] (M.E.G. and B.J.H., unpublished observations). If many SNPs have an effect, these effects on average must be small. To estimate small effects accurately requires a large sample size and, not surprisingly, the accuracy of genomic selection increases as sample size increases, at least up to a reference population size of 3,500 (REF. 48).

We have developed an analytical method for predicting the accuracy of genomic selection[54,56] assuming that all SNPs have an effect and these effects are normally distributed. The size of the reference population that is needed to achieve a given accuracy is shown in FIG. 3. Unless the N_e is small, a large sample of animals is needed in the reference population if accurate prediction of breeding value is desired. This theory predicts the upper limit of the number of animals required. If the SNP effects are not normally distributed, with some large effects and many SNPs with no effect, the number of animals needed is reduced[54].

Challenges for genomic selection. The major challenge is assembling the large reference population that is required to accurately estimate SNP effects. In some cases this has been achieved; for example, a project run by the US Department of Agriculture has assembled a reference population of approximately 6,700 dairy bulls, leading to an accuracy of genomic breeding values for young dairy bulls of greater than 0.8 (REF. 57). These accuracies are sufficiently high that some US breeding companies are marketing semen from young bulls on the basis of their DNA and pedigree information alone. Smaller reference populations of dairy bulls have been assembled in Australia, New Zealand and the Netherlands, resulting in impressive but lower accuracies of genomic breeding values[50]. Another major challenge, particularly in the beef cattle and sheep industries, is the involvement of multiple breeds. Given the limited extent of LD across breeds, large multi-breed reference populations must be assembled and genotyped for many (>300,000) SNPs before genomic selection can be applied.

There are still several unknowns in the implementation of genomic selection. For instance, how often will the marker effects have to be re-estimated and new markers discovered? The cost of genotyping may delay implementation in species such as sheep and chickens, in which individual animals are less valuable than in cattle. However, even in these species, selection in the top layers of the stud pyramid should prove profitable because the benefits can be recouped from a large population descended from the genotyped and selected animals.

The future

The benefits from the study of complex traits in domestic species are an increase in scientific knowledge and practical improvements in breeding programmes. Large populations with recorded phenotypes exist and, in some cases, there are males with accurate estimates of breeding value for traits that are based on a progeny test, allowing designed mating programmes to be implemented. The breeds within a species show a large amount of genetic variation owing to deliberate selection and genetic drift in populations of small N_e. Long-range LD within a breed, but not between breeds, allows

rapid mapping to a large region and then more precise mapping to a single gene or a few genes. The small N_e should also lead to greater homogeneity within a breed so that there are fewer genes causing variation within a breed.

These advantages are likely to be translated into many newly discovered QTLs from GWA studies in the near future as dense SNP arrays for nearly all the major domestic species have recently become available. However, to capture the benefits of GWA studies, experiments on domestic animals should learn from those reported on humans. This means an increase in the number of animals and the number of SNPs, and routine validation of significant associations in an independent sample of animals.

QTL effects are typically small, and so many animals are needed to estimate them accurately. One way to achieve this would be by collaboration between different scientists, but at present this is inhibited by commercial use of the SNP genotypes. A larger number of animals is needed to detect associations with traits of low heritability (for example, fertility) and, unless these large numbers can be achieved, genomic selection for these traits will be less accurate and fewer QTLs will be discovered. Genomic selection would be especially useful for traits that are expensive to measure (for example, methane production), but unfortunately this also makes it expensive to carry out the experiments needed to find markers.

Increasing the number of SNPs above 50,000 may not be necessary if one worked entirely within a breed, such as Holstein. However, mapping QTLs is much more accurate if advantage is taken of multiple breeds — this requires denser SNPs because LD extends for only a short distance between breeds. Also, genomic selection would be more accurate if SNP effects could be estimated across breeds, but this will also require denser SNPs.

To support this research, the genomes of some species (such as the goat) need to be sequenced and SNP chips need to be made available for many more species. Information from SNP genotyping will soon be supplemented with genome resequencing. This should increase the power to detect QTLs and simplify the discovery of causative mutations. However, if neutral mutations exist in high LD with a causative mutation across all breeds, it will remain difficult to identify the causal mutation. Although the use of these resources will initially be directed to practical outcomes, such as genomic selection, we expect that many causal mutations underlying QTLs will also be discovered in domestic animals.

1. Meuwissen, T. H. E. & Goddard, M. E. The use of marker haplotypes in animal breeding schemes. *Genet. Sel. Evol.* **28**, 161–176 (1996). **Quantifies the benefits of MAS.**
2. Falconer, D. S. & McKay, T. F. X. *Introduction to Quantitative Genetics* 4th edn (Longmans Green, UK, 1996).
3. Andersson, L. & Georges, M. Domestic animal genomics: deciphering the genetics of complex traits. *Nature Rev. Genet.* **5**, 202–212 (2004).
3. Dekkers, J. C. M. & Hospital, F. Multifactorial genetics: the use of molecular genetics in the improvement of agricultural populations. *Nature Rev. Genet.* **3**, 22–32 (2002).
4. Van Laere, A. S. *et al.* A regulatory mutation in *IGF2* causes a major QTL effect on muscle growth in the pig. *Nature* **425**, 832–836 (2003).
5. Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001). **Introduced the concept and statistical methods for genomic selection.**
6. Tolle, A. in *Rep. VIth Int. Bloodgroup Congr.* 40–52 (Inst. Blutgruppenforshung, Munich, Germany, 1959).
7. Neimann-Sorensen, A. & Robertson, A. The association between blood groups and several production characteristics in three Danish cattle breeds. *Acta Agric. Scand.* **11**, 163–196 (1961).
8. Rendel, J. Relationships between blood groups and the fat percentage of the milk in cattle. *Nature* **189**, 408–409 (1961).
9. Georges, M. *et al.* Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139**, 907–920 (1995).
10. Sved, J. A. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.* **2**, 125–141 (1971).
11. Hayes, B. J. Visscher, P. M., McPartlan, H. & Goddard, M. E. A novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* **13**, 635–643 (2003).
12. Tenesa, A. *et al.* Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* **17**, 520 – 526 (2007).
13. De Roos, A. P. W., Hayes, B. J., Spelman, R. & Goddard, M. E. Linkage disequilibrium and persistence of phase in Holstein Friesian, Jersey and Angus cattle. *Genetics* **179**, 1503–1512 (2008).
14. MacEachern, S., Hayes, B. J., McEwan, J. & Goddard, M. E. An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (Bos taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. *BMC Genomics* **10**, 181 (2009).
15. Sutter, N. B. *et al.* Extensive and breed-specific linkage disequilibrium in *Canis familiaris*. *Genome Res.* **14**, 2388–2396 (2004).
16. Meuwissen, T. H. E. & Goddard, M. E. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet. Sel. Evol.* **36**, 261–279 (2004).
17. Uleberg, E. *et al.* Fine mapping of a QTL for intramuscular fat on porcine chromosome 6 using combined linkage and linkage disequilibrium mapping. *J. Anim. Breed Genet.* **122**, 1–6 (2005).
18. Gautier, M. *et al.*Fine mapping and physical characterization of two linked quantitative trait loci affecting milk fat yield in dairy cattle on BTA26. *Genetics* **172**, 425–436 (2006).
19. Olsen, H. G., Meuwissen, T. H., Nilsen, H., Svendsen, M & Lien, S. Fine mapping of quantitative trait loci on bovine chromosome 6 affecting calving difficulty. *J. Dairy Sci.* **91**, 4312–4322 (2008).
20. Tantia, M. S. *et al.* DGAT1 and ABCG2 polymorphism in Indian cattle (*Bos indicus*) and buffalo (*Bubalus bubalis*) breeds. *BMC Vet. Res.* **7**, 32 (2006).
21. Barendse, W., Harrison, B. E., Bunch, R. J. & Thomas, M. B. Variation at the calpain 3 gene is associated with meat tenderness in zebu and composite breeds of cattle. *BMC Genet.* **9**, 41 (2008).
22. Visscher, P. M. Sizing up human height variation. *Nature Genet.* **40**, 489–490 (2008). **Uses published results to demonstrate the small effect size of most QTLs.**
23. Karlsson, E. K. *et al.* Efficient mapping of mendelian traits in dogs through genome-wide association. *Nature Genet.* **39**, 1321–1328 (2007). **Shows how mapping the same locus within two different breeds of dog can lead to discovery of a causal mutation.**
24. Charlier, C. *et al.* Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nature Genet.* **40**, 449–454 (2008).
25. Kolbehdari, D. *et al.* A whole-genome scan to map quantitative trait loci for conformation and functional traits in Canadian Holstein bulls. *J. Dairy Sci.* **91**, 2844–2856 (2008).
26. Daetwyler, H. D., Schenkel, F. S., Sargolzaei, M. & Robinson, J. A. A genome scan to detect quantitative trait loci for economically important traits in Holstein cattle using two methods and a dense single nucleotide polymorphism map. *J. Dairy Sci.* **91**, 3225–3236 (2008).
27. Barendse, W. *et al.* A validated whole-genome association study of efficient food conversion in cattle. *Genetics*. **176**, 1893–1905 (2007).
28. Lillehammer, M., Hayes, B. J., Meuwissen, T. H. E. & Goddard, M. E. Gene by environment interactions for production traits in Australian dairy cattle. *J. Dairy Sci.* (in the press).
29. Long, N., Gianola, D., Rosa, G. J., Weigel, K. A. & Avendaño, S. Marker-assisted assessment of genotype by environment interaction: a case study of single nucleotide polymorphism-mortality association in broilers in two hygiene environments. *J. Anim. Sci.* **86**, 3358–3366 (2008).
30. Hasenstein, J. R., Hassen, A. T., Dekkers, J. C. & Lamont, S. J. High resolution, advanced intercross mapping of host resistance to Salmonella colonization. *Dev. Biol.* **132**, 213–218 (2008).
31. Beavis, W. D. in *Molecular Dissection of Complex Traits* (ed. Patterson, A. H.) 145–162 (CRC, New York, 1998).
32. Sanna, S. *et al.* Common variants in the *GDF5-UQCC* region are associated with variation in human height. *Nature Genet.* **40**, 198–203 (2008).
33. Franke, A. *et al.* Replication of signals from recent studies of Crohn's disease identifies previously unknown disease loci for ulcerative colitis. *Nature Genet.* **40**, 713–715 (2008).
34. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
35. Jones, P. *et al.* Single-nucleotide-polymorphism-based association mapping of dog stereotypes. *Genetics* **179**, 1033–1044 (2008).
36. Spelman, R. J. Ford, C. A., McElhinney, P., Gregory, G. C. & Snell, R. G. Characterization of the DGAT1 gene in the New Zealand dairy population. *J. Dairy Sci.* **85**, 3514–3517 (2002).
37. Dunner, S. *et al.* Haplotype diversity of the myostatin gene among beef cattle breeds. *Genet. Sel. Evol.* **35**, 103–118 (2003).
38. Smith, C. Improvement of metric traits through specific genetic loci. *Anim. Prod.* **9**, 349–358 (1967).
39. Fujii, J. *et al.* Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. *Science* **253**, 448–451 (1991).

40. Piper, L., R., Bindon, B. M. & Davis, G. H. in *Genetics of Reproduction in Sheep* (eds Land, R. B. & Robinson D. W) 115–125 (Butterworths, London, 1985).
41. Shuster, D. E., Kehrli, M. E. Jr, Ackermann, M. R. & Gilbert, R. O. Identification and prevalence of a genetic defect that causes leukocyte adhesion deficiency in Holstein cattle. *Proc. Natl Acad. Sci. USA* **89**, 9225–9229 (1982).
42. Goldman, W. N. *et al.* Two alleles of a neural protein gene linked to scrapie in sheep. *Proc. Natl Acad. Sci. USA* **87**, 2476–2480 (1990).
43. Davis, G. H. Major genes affecting ovulation rate in sheep. *Genet. Sel. Evol.* **37** (Suppl. 1), S11–S23 (2005).
44. Van Arendonk, J. A. M. *et al.* in *From Jay L. Lush to Genomics: Visions for Animal Breeding and Genetics* (eds Dekkers, J. C. M., Lamont, S. J. & Rothschild, M. F.) 60–69 (Iowa State Univ., Ames, 1999).
45. Plastow, G. S. *et al.* in *Proceedings 28th Annual Meeting National Swine Improvement Federation* 151–154 (Iowa State Univ., Ames, 2003).
46. Dekkers, J. C. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *J. Anim. Sci.* **82**, E313–E328 (2004).
47. Schaeffer, L. R. Strategy for applying genome-wide selection in dairy cattle. *J. Anim. Breed. Genet.* **123**, 218–223 (2006).
Calculates the gain in selection response from genomic selection in dairy cattle.
48. VanRaden, P. M. *et al.* Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* **92**, 16–24 (2009).
49. Harris, B. L., Johnson, D. L. & Spelman, R. J. in *Proc. Interbull Meeting, Bulletin 39* (Niagara Falls, Canada, 2008).
50. Hayes, B. J., Bowman, P. J., Chamberlain, A. C. & Goddard, M. E. Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* **92**,433–443 (2008).
51. Legarra A., Robert-Granié, C., Manfredi, E. & Elsen, J. M. Performance of genomic selection in mice. *Genetics*. **180**, 611–618 (2008).
52. Lee, S. H., van der Werf, J. H., Hayes, B. J., Goddard, M. E. & Visscher, P. M. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet.* **4**, e1000231 (2008).
53. González-Recio, O., Gianola, D., Rosa, G. J., Weigel, K. A., Kranis, A. Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. *Genet. Sel. Evol.* **41**, 3 (2009).
54. Goddard, M. E. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 14 Aug 2008 (doi: 10.1007/s10709-008-9308-0).
Presents formulae for the accuracy of genomic selection and the optimization of long-term selection response.
55. Goddard, M. E. & Hayes, B. J. Genomic selection. *J. Anim. Breed. Genet.* **124**, 323–330 (2007).
56. Hayes, B. J., Visscher, P. M. & Goddard, M. E. Increased accuracy of selection by using the realised relationship matrix. *Genet. Res.* **91**, 47–60 (2009).
57. Dalton, R. No bull: genes for better milk. *Nature* **457**, 369 (2009).
58. Maher, B. The case of the missing heritability. *Nature* **456**, 18–21 (2008).
59. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genet.* **41**, 25–34 (2009).
60. Grisart, B. *et al.* Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine *DGAT1* gene with major effect on milk yield and composition. *Genome Res.* **12**, 222–231 (2002).
61. Hayes, B. J. & Goddard, M. E. The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* **33**, 209–229 (2001).
62. Weller, J. I. Shlezinger, M. & Ron, M. Correcting for bias in estimation of quantitative trait loci effects. *Genet. Sel. Evol.* **37**, 501–522 (2005).
63. Bellinge, R. H., Liberles, D. A., Iaschi, S. P., O'Brien, P. A. & Tay, G. K. Myostatin and its implications on animal breeding: a review. *Anim. Genet.* **36**, 1–6 (2005).
64. Clop, A. *et al.* A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nature Genet.* **38**, 813–818 (2006).
65. The Bovine HapMap Consortium. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **324**, 528–532 (2009).
66. Drögemüller, C. *et al.*A mutation in hairless dogs implicates *FOXI3* in ectodermal development. *Science* **321**, 1462 (2008).
67. Awano, T. *et al.* Genome-wide association analysis reveals a *SOD1* mutation in canine degenerative myelopathy that resembles amyotrophic lateral sclerosis. *Proc. Natl Acad. Sci. USA* **106**, 2794–2799 (2009).
68. Wiik, A. C. *et al.* A deletion in nephronophthisis 4 (*NPHP4*) is associated with recessive cone-rod dystrophy in standard wire-haired dachshund. *Genome Res.* **18**, 1415–1421 (2008).
69. Salmon Hillbertz, N. H. *et al.* Duplication of *FGF3*, *FGF4*, *FGF19* and *ORAOV1* causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. *Nature Genet.* **39**, 1318–1320 (2007).

Acknowledgements
The authors would like to thank H. Campbell and H. Burrow for cattle pictures used in this Review.

FURTHER INFORMATION
APIL's Genomic Comparison of Young Bulls: http://aipl.arsusda.gov/reference/genomic_comparison_yng_0901.htm

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

GENOME-WIDE ASSOCIATION STUDIES

Detecting gene–gene interactions that underlie human diseases

Heather J. Cordell

Abstract | Following the identification of several disease-associated polymorphisms by genome-wide association (GWA) analysis, interest is now focusing on the detection of effects that, owing to their interaction with other genetic or environmental factors, might not be identified by using standard single-locus tests. In addition to increasing the power to detect associations, it is hoped that detecting interactions between loci will allow us to elucidate the biological and biochemical pathways that underpin disease. Here I provide a critical survey of the methods and related software packages currently used to detect the interactions between genetic loci that contribute to human genetic disease. I also discuss the difficulties in determining the biological relevance of statistical interactions.

Data mining
The process of extracting hidden patterns and potentially useful information from large amounts of data.

Machine learning
The ability of a program to learn from experience, that is, to modify its execution on the basis of newly acquired information. A major focus of machine-learning research is to automatically produce models (rules and patterns) from data.

Bayesian model selection
A statistical approach for selecting models by incorporating both prior distributions for parameters of the models and the observed experimental data.

Institute of Human Genetics, Newcastle University, International Centre for Life, Central Parkway, Newcastle upon Tyne NE1 3BZ, UK.
e-mail: heather.cordell@ncl.ac.uk
doi:10.1038/nrg2579
Published online 12 May 2009

The search for genetic factors that influence common complex traits and the characterization of the effects of those factors is both a goal and a challenge for modern geneticists. In recent years, the field has been revolutionized by the success of genome-wide association (GWA) studies[1–5]. Most of these studies have used a single-locus analysis strategy, in which each variant is tested individually for association with a specific phenotype. However, a reason that is often cited for the lack of success in genetic studies of complex disease[6,7] is the existence of interactions between loci. If a genetic factor functions primarily through a complex mechanism that involves multiple other genes and, possibly, environmental factors, the effect might be missed if the gene is examined in isolation without allowing for its potential interactions with these other unknown factors. For this reason, several methods and software packages[8–15] have been developed that consider the statistical interactions between loci when analysing the data from genetic association studies. Although in some cases the motivation for such analyses is to increase the power to detect effects[16], in other cases the motivation has been to detect statistical interactions between loci that are informative about the biological and biochemical pathways that underpin the disease[7]. We return to this complex issue of biological interpretation of statistical interaction later in the article.

The purpose of this Review is to provide a survey of the methods and related software packages that are currently being used to detect the interactions between the genetic loci that contribute to human genetic disease. Although the focus is on human genetics, many of the concepts and approaches are strongly related to methods used in animal and plant genetics. I begin by describing what is meant by statistical interaction and by setting up the definitions and notation for the following sections. I then explain how one might test for interaction between two or more known genetic factors and how one might address the slightly different question of testing for association with a single factor while allowing for interaction with other factors. In practice, one rarely wishes to test for interactions that occur only between known factors, unless perhaps to replicate a previous finding or to test a specific biological hypothesis. It is more common to search for interactions or for loci that might interact, given genotype data at potentially many sites (for example, from a GWA analysis or from a more focused candidate gene study). I continue the article by outlining different methods and software packages that search for such interactions, ranging from simple exhaustive searches to data-mining and machine-learning approaches to Bayesian model selection approaches. Throughout these sections I use the analysis of a publicly available genome-wide data set on Crohn's disease from the Wellcome Trust Case Control Consortium (WTCCC) as a recurring example[1]. I conclude the article with a section discussing the biological interpretation of results found from such statistical interaction analyses.

There is a long history of the investigation of interactions in genetics, ranging from classical quantitative genetic studies of inbred plant and animal populations[17–19] to evolutionary genetic studies[20] and, finally, to linkage and association studies in outbred human populations. In this article, I focus primarily on human genetic association studies; readers are referred to REFS 20–25 for a

Box 1 | **Statistical models of interaction**

Linear, multiple and logistic regression

Statistical interaction can best be described in relation to a linear model that describes the relationship between an outcome variable and some predictor variable or variables. In linear regression, we model a quantitative outcome y as a function of a predictor variable x using the regression equation $y = mx + c$. Here the regression coefficient m corresponds to the slope of the best-fit line and the regression coefficient c corresponds to the intercept. We use the values of pairs of data points (x, y) (for example, if x and y are, respectively, measurements of height and weight in different individuals) to estimate m and c, such that the line $y = mx + c$ fits the observed data as closely as possible.

In multiple regression, we extend this idea to include several different predictor variables using an equation such as $y = m_1x_1 + m_2x_2 + m_3x_3 + c$. Here we are implicitly assuming that there is a linear relationship between each of the predictor variables x_1, x_2 and x_3 and the outcome variable y, so that for each unit increase in x_1, y is expected to increase by m_1 (and similarly for x_2 and x_3).

In logistic regression, rather than modelling a quantitative outcome y, we model the log odds $\ln(p/(1-p))$ (in which p is the probability of having a disease). For example, we might propose the model $\ln(p/(1-p)) = \alpha + \beta x_B + \gamma x_C + i x_B x_C$, in which x_B and x_C are measured binary indicator variables that represent the presence or absence of genetic exposures at loci B and C respectively, β and γ are regression coefficients that represent the main effects of exposures at B and C, and the coefficient i represents an interaction term[16] (a term that is required in addition to the linear terms for B and C).

Testing for interaction

Tests of interaction correspond to testing whether the regression coefficients that represent interaction terms in the above mathematical formula equal zero or not. In the logistic regression example above, this would correspond to a one degree of freedom test of $i = 0$. In the saturated genotype model described in Supplementary information S1 (box), it would correspond to a four degrees of freedom test of $i_{11} = i_{12} = i_{21} = i_{22} = 0$. Tests of association (for example, at a given locus C) while allowing for interaction (for example, with another locus B) correspond to comparing a linear model in which the main effects of B, C and their interactions are included with a model in which all the terms (main or interaction) that involve locus C are removed. For example, if modelling the log odds as $\ln(p/(1-p)) = \alpha + \beta x_B + \gamma x_C + i x_B x_C$, then the test of association at C allowing for interaction with B corresponds to a two degrees of freedom test of $\gamma = i = 0$. This is in contrast to the one degree of freedom pure interaction test of $i = 0$. One could also construct a pairwise test of the joint effects at both loci, including interactions, by comparing a model in which the main effects of loci B, C and their interactions are included with a model in which only the baseline intercept α is included. This gives a three degrees of freedom test of association allowing for interaction if a binary or allelic code is used, or an eight degrees of freedom test[52] if a saturated genotype model (Supplementary information S1 (box)) is used. Tests with fewer degrees of freedom could be used by prior grouping of the two-locus genotypes according to certain prespecified classification schemes[15,29].

discussion of interactions in the context of evolutionary genetics or in human genetic linkage analysis.

Definition of statistical interaction

Interaction as departure from a linear model. The most common statistical definition of interaction relies on the concept of a linear model that describes the relationship between an outcome variable and a predictor variable or variables. We propose a particular model for how we believe the predictors might relate to the outcome and we use data (measurements of the relevant variables from a number of individuals) to determine how well the model fits our observed data and to compare the fit of different models. Arguably the most well-known form of this type of analysis is simple linear or least squares regression[26], in which we relate an observed quantitative outcome y (for example, weight) to a predictor variable x (for example, height) using a 'best fit' line or regression equation $y = mx + c$. More generally, we might use multiple regression[26] to include several different predictor variables (for example, x_1, x_2 and x_3, to represent height, age and gender).

From a statistical point of view, interaction represents departure from a linear model that describes how two or more predictors predict a phenotypic outcome (BOX 1). For a disease outcome and case–control data, rather than modelling a quantitative trait y, the usual approach is to model the expected log odds of disease as a linear function of the relevant predictor variables[26,27]. Using genotype data, we can evaluate the likelihood of the data under this model and use maximum likelihood or other methods to estimate the regression coefficients and test hypotheses, such as the hypothesis that the interaction term (i in the mathematical formula in BOX 1) equals zero.

Supplementary information S1 (box) describes some specific models that follow this general formula, including the saturated genotype model. Although this model provides the best possible fit to the data, it includes many parameters. We can make parameter restrictions to generate fewer degrees of freedom and thus increase power. Although written in terms of nine or fewer regression parameters, the models in Supplementary information S1 (box) represent an infinite number of different models, depending on the values taken by the regression parameters. There has been some interest in categorizing these models[28–30] to aid mathematical or biological interpretation. As discussed below, biological interpretation is usually easiest when the penetrance values all equal either zero or one, leading to a clear relationship between the genotype and phenotype; however, this situation is unlikely for complex genetic diseases.

Marginal effects. An important issue in genetic studies is whether there are factors that display interaction effects without displaying marginal effects[6,31]. Factors that display interaction effects without displaying marginal effects will be missed in a single-locus analysis, as they do not lead to any marginal correlation between the genotype and phenotype when each locus is considered individually. It is not clear in practice how often this might occur, as many models that include an interaction term even in the absence of main effects (α and β in the mathematical formula in BOX 1) lead to substantial marginal effects, that is, they show correlations between the genotype and phenotype that are detectable in a single-locus analysis. Thus, although one may derive mathematical models (sets of specific values for the regression coefficients) that lead to single-locus models without marginal effects[6], it remains to be seen whether such models represent common underlying scenarios — and thus a potentially serious problem — in complex genetic diseases.

For simplicity, I have concentrated here on defining interaction in relation to two genetic factors (two-locus interactions). In practice, however, for complex diseases we might also expect three-locus, four-locus and even higher-level interactions. Mathematically, such higher-level interactions are simple extensions to the two-locus models described earlier. The problem with these models

Maximum likelihood
A statistical approach that is used to make inferences about the combination of parameter values that gives the greatest probability of obtaining the observed data.

Saturated
A term for a statistical model that is as full as possible (saturated) with parameters. Such a model is sometimes useful as it serves as a benchmark to quantify how well a simpler model (one with fewer parameters) fits the data.

is that they contain many parameters, and extremely large data sets would be required to accurately estimate these parameters. Interpreting the resulting parameter estimates is also complicated, except perhaps in some simple cases; for example, when risk alleles at all loci are required to alter disease risk (that is, when only the full multi-locus interaction term differs from zero).

Testing for interaction between known factors

Regression models. For two or more known or hypothetical genetic factors that influence disease risk, arguably the most natural way to test for statistical interaction on the log odds scale is to fit a logistic regression model that includes the main effects and relevant interaction terms and then to test whether the interaction terms equal zero. A similar approach can be used for quantitative phenotypes, in which case linear rather than logistic regression is used. These analyses can be performed in almost any statistical analysis package after construction of the required genotype variables. Alternatively, the '--epistasis' option in the whole-genome analysis package PLINK[12] provides a logistic regression test for interaction that assumes an allelic model for both the main effects and the interactions.

A more powerful approach in case–control studies is to use a case-only analysis[32–34]. Case-only analysis exploits the fact that, under certain conditions, an interaction term in the logistic regression equation corresponds to the dependency or the correlation between the relevant predictor variables within the population of cases. A case-only test of interaction can therefore be performed by testing the null hypothesis that there is no correlation between alleles or genotypes at the two loci in a sample that is restricted to cases alone. This test can easily be performed using a simple χ^2 test of independence between genotypes (a four degrees of freedom test) or alleles (a one degree of freedom test), or using logistic or multinomial regression in any statistical analysis package.

The main problem with the case-only test is its requirement that the genotype variables are not correlated in the general population. It is this assumption, rather than the design *per se*, that provides the increased power compared with case–control analysis. The case-only test is therefore unsuitable for loci that are either closely linked or show correlation for another reason (for example, if certain genotype combinations are related to viability). In contrast to epidemiological studies of environmental factors, in which correlation and confounding between variables is common, in genetic studies the assumption of independence between unlinked genetic factors seems reasonable. One could use a two-stage procedure to test first for correlation between the loci in the general population and then use the outcome to determine whether to perform a case-only or case–control interaction test. However, this procedure has potential bias[35].

A preferable approach is to incorporate the case-only and case–control estimators into a single test. Zhao *et al.*[36] proposed a test based on the difference in inter-locus allelic association between cases and controls, an idea originally suggested by Hoh and Ott[37]. The '--fast-epistasis' option in PLINK[12] performs a similar test. Zhao *et al.*[36] found that their test had greater power than a four degrees of freedom logistic regression test of gene–gene interaction. However, this increase in power might be largely due to the lower number of degrees of freedom in their allelic test compared with a genotypic test. Mukherjee and Chatterjee[35,38] proposed an empirical Bayes procedure that uses a weighted average of the case–control and case-only estimators of the interaction. This approach exploits the gene–gene independence assumption and thus the power of case-only analysis, and additionally incorporates controls, allowing the estimation of main effects. Routines that implement this procedure are available for Microsoft Office Excel and MATLAB.

Other approaches. Although regression-based tests of interaction seem the most natural approach, given the definition of interaction as departure from a linear regression model, alternative approaches have been proposed. Yang *et al.*[39] proposed a method based on partitioning of χ^2 values that, similarly to REF. 36, compares inter-locus association between cases and controls. Their method was more powerful than logistic regression when the loci had no marginal effects. Recently, there has been interest in information theory or entropy-based approaches for modelling genetic interactions[40–43]. It is unclear whether this framework offers any advantage over more standard statistical methods of modelling of the same predictor variables as, in most cases, the conditional probability statements that are implied by the two approaches are equivalent[44].

Family-based studies. Here I focus on testing for interaction in the context of case–control or population-based studies. Several related methods have been proposed to test for interaction in the context of family-based association studies[45–49]. The case–pseudocontrol approach[46] offers a regression-based framework that allows interaction tests that are similar to those described here. Given the larger sample sizes that are required when testing for interaction rather than main effects[50,51], it is unclear whether investigators will have family-based cohorts of a sufficient size to provide high power to detect interactions. However, such cohorts might provide a useful resource for the replication and characterization of interaction effects that have been found using alternative methods.

Tests for association allowing for interaction

Rather than testing for interaction *per se,* many researchers are interested in allowing for interaction with other genetic or environmental factors when testing for association at a given genetic locus. The rationale is that, if the test locus influences the disease or phenotypic outcome by interacting with another factor, then allowing for this interaction should increase the power to detect the effect at the test locus. From a mathematical point of view, a test for association at a given locus C while allowing for interaction with another locus B (a joint test[16]) corresponds to comparing the fit to the observed data

Penetrance
The probability of displaying a particular phenotype (for example, succumbing to a disease) given that one has a specific genotype.

Marginal effects
The average effects (for example, penetrances) of a single variable, averaged over the possible values taken by other variables. These could be calculated for one locus of a two-locus system as the average of the two-locus penetrances, averaged over the three possible genotypes at the other locus.

Logistic regression model
A statistical model that is used when the outcome is binary. It relates the log odds of the probability of an event to a linear combination of the predictor variables.

Multinomial regression
A statistical approach, similar to logistic regression, which is used when the outcome takes one of several possible categorical values.

Confounding
A phenomenon whereby the measure of association between two variables is distorted because other variables, associated with both variables of interest, are not controlled for in the calculation.

Empirical Bayes procedure
A hierarchical model in which the hyperparameter is not a random variable but is estimated by another (often classical) method.

Information theory
A branch of applied mathematics involving the quantification of information.

Entropy
A key measure used in information theory that quantifies the uncertainty associated with a random variable. For example, a variable indicating the outcome from a toss of a coin will have less entropy than a variable indicating the outcome from a roll of a die (two versus six equally likely outcomes).

of a linear model in which the main effects of B, C and their interactions are included with a model in which all the terms (main or interaction) involving locus C are removed (BOX 1).

Theoretically, if no interaction effects exist, these joint tests will be less powerful than marginal single-locus association tests. However, if interaction effects exist, then the power of joint tests can be higher than that of single-locus approaches[52]. Kraft *et al.*[16] showed that the joint test of a genetic effect while allowing for interaction with a known environmental factor had a near optimal performance over a wide range of plausible underlying models. This test uses case–control data to test the combination of a main effect at locus C and an interaction effect. As case-only analysis provides a more powerful test for the interaction effect[32–34], Chapman and Clayton[53] proposed using a version of the joint test that combines a case–control main effect component with a case-only interaction component.

The joint test of association while allowing for interaction assumes that there is some known or hypothetical measured factor that might interact with the test locus. In the absence of a specific factor of this type, a natural approach is to average over all other potentially interacting genetic factors when performing a test at a locus. A Bayesian method for this approach in the context of GWA studies is in development[14] and a beta version of the associated Bayesian Interaction Analysis software is available in limited release from its authors on request. Rather than averaging over all possible interacting loci, Chapman and Clayton[53] proposed using the maximum value of the joint test evaluated over a predefined set of potentially modifying loci and assessing significance using a permutation argument.

I have concentrated on the issue of testing either for interaction or for association while allowing for interaction at one or two specific genetic variants of interest. Rather than testing a single variant, it is now common to have genotype data for many variants that might or might not have any prior evidence for involvement with disease. Given such data, various model selection approaches have been proposed that allow one to step through a sequence of regression models searching for significant effects, including both main effects and interactions[8–10,13,37,54–56]. These approaches are described in more detail in subsequent sections. First, I describe an approach that is feasible provided the number of main and interaction effects to be examined is not too large, namely, a simple exhaustive search.

Exhaustive search

Two-locus interactions. Given genotype data at several different loci, arguably the simplest way to search for interactions between these loci is by an exhaustive search. For example, to test all two-locus interactions, one could analyse all possible pairs of loci and perform the desired interaction test for each pair. Similarly, if testing for association while allowing for interaction, one could perform the relevant three or eight degrees of freedom test[52] (BOX 1, Supplementary information S1 (box)). Clearly, an exhaustive search of this type raises a multiple testing issue analogous to the multiple testing issue encountered in single-locus analysis of GWA studies[1]. If all the tests are independent, a Bonferroni correction is appropriate[52]; however, linkage disequilibrium between loci can induce correlation between many of the tests. When testing for association while allowing for interaction, additional correlation occurs owing to the fact that the main effect of a locus will be a component of all tests that involve that locus. Theoretically, one can use permutation[53] to assess significance while allowing for the multiplicity of and correlation between the tests performed, but, for several loci, this approach might be computationally prohibitive.

A pragmatic approach to the multiple testing issue in single-locus analysis of GWA studies is to use a stringent significance threshold (for example, $p = 5 \times 10^{-7}$) coupled with replication in an independent data set to avoid generating large numbers of false positives. Stringent significance thresholds can also be motivated by Bayesian arguments concerning the low prior probability of any given variant being associated with disease[1]. In practice, the Q–Q plot[1] has emerged as the tool of choice for visualizing the results from an entire-genome scan.

An exhaustive search of all two-locus interactions from a genome scan is time consuming but computationally feasible. Marchini *et al.*[52] quote a time of 33 hours on a 10-node cluster to perform all pairwise tests of association allowing for interaction at 300,000 loci in 1,000 cases and 1,000 controls. The PLINK[12] website quotes 24 hours to test (using the '--fast-epistasis' option) all pairwise interactions at 100,000 loci typed in 500 individuals. Given that genome-wide studies now routinely generate between 500,000 and 1,000,000 markers in 5,000 or more individuals, these times will need to be scaled upwards by several weeks or even months, but an exhaustive search of all two-locus interactions still remains feasible. In addition, as each test can be computed independently of all other tests, the entire search can be split up into several separate jobs and analysed by parallel processing facilities, if they are available.

Higher-order interactions. The problem with an exhaustive search is that it does not scale up to analyse higher-order interactions. Because the number of tests and therefore the time taken to perform the analysis increases exponentially with the order of interaction analysed, an exhaustive search of all three-way, four-way or higher-level interactions seems impractical in a genome-wide setting. For this reason, two-stage procedures have been proposed[52,57,58], in which a subset of loci that pass some single-locus significance threshold are chosen, and an exhaustive search of all two-locus interactions (or a higher order if required, perhaps conditional on significant lower-order effects[58]) is carried out on this 'filtered' subset. The obvious drawback with this approach is that loci will only be filtered into the second or subsequent stages of the testing procedure if they show a marginal association with the phenotype. Therefore, this procedure would not be expected to be useful for detecting interactions that genuinely occur in the absence of marginal effects.

Permutation
This method is often used in hypothesis testing. An empirical distribution of a test statistic is obtained by permuting the original sample many times and recalculating the value of the test statistic in each permuted data set. Each permuted sample is considered to be a sample of the population under the null hypothesis.

Multiple testing
An analysis in which multiple independent hypotheses are tested. If a large number of tests are performed, the significance level (*p* value) of any particular test must be interpreted in light of this fact, as the overall combined probability of making a type I error will increase.

Bonferroni correction
The simplest correction of individual *p* values for multiple hypothesis testing can be calculated using $p_{corrected} = 1 - (1 - p_{uncorrected})^n$, in which n is the number of hypotheses tested. This formula assumes that the hypotheses are all independent, and simplifies to $p_{corrected} = np_{uncorrected}$ when $np_{uncorrected} << 1$.

Q–Q plot
A quantile–quantile plot is a diagnostic plot that can be used to compare the distribution of observed test statistics with the distribution expected under the null hypothesis. Those tests that lie significantly above the line of equality between observed and expected quantiles are considered significant in the context of the number of tests performed.

Use of a single-locus significance threshold is not the only way to reduce the number of markers for testing. Several of the machine-learning approaches described in the next section (in particular ReliefF and random forests) could be used, as they do not require a locus to have a significant marginal effect. Biological plausibility offers an alternative strategy. Bochanovits *et al.*[59] used evidence of co-adaptation between loci in the mammalian genome to select genes for interaction testing in a human study. Emily *et al.*[60] used experimental knowledge of biological networks to reduce the number of interaction tests from 1.25×10^{11} to 7.1×10^{4} when analysing genotype data from the WTCCC[1]. In their analysis of seven disease cohorts, they found four significant interaction effects, including one of $p = 1 \times 10^{-9}$ between rs6496669 on chromosome 15 and rs434157 on chromosome 5 in Crohn's disease. An example of applying semi-exhaustive testing to this same data set using the '--fast-epistasis' and '--case-only' options in PLINK[12] is shown in FIG. 1.

Data-mining methods and related approaches

Traditional regression-based methods are often criticized[8,31,61] for their inability to deal with nonlinear models and with high-dimensional data that contain many potentially interacting predictor variables, leading to sparse contingency tables that have many empty cells. For this reason, machine-learning or data-mining methods developed in the field of computer science are sometimes preferred. The selection of predictor variables and the interactions between them that predict an outcome variable is a well-known problem in the fields of machine learning and data mining. Data-mining approaches do not fit a single prespecified model, nor do they attempt an exhaustive search, but rather they attempt to step through the space of possible models, including potentially large numbers of main effects and multiway interactions, in a computationally efficient way. Many data-mining approaches are equivalent to stepping through a particular sequence of regression models and attempting to find the model that best fits the data; the distinction that is often made between data-mining and regression models is therefore, to some extent, false. Nonlinearity is not an issue when fitting a saturated model, although it might be an issue for more restricted models. One common theme in data mining is the use of cross-validation[62] to avoid overfitting problems.

Data-mining methods typically have problems dealing with incomplete or unbalanced data sets; for example, when the number of cases and controls are unequal[63]. They also do not always deal well with correlated predictors that show colinearity. This has been addressed in the mainstream statistics literature by the introduction of penalized regression approaches[64,65] that allow large numbers of predictor variables to be included in a regression model but with many estimated regression coefficients reduced towards zero. In genetics, the use of such techniques is just starting to emerge, including penalized logistic regression[66,67] and least-angle regression[68] for identifying gene–gene interactions[69,70] in binary traits.

A good overview of several machine-learning approaches for detecting gene–gene interactions is given by McKinney *et al.*[31]. For the remainder of this section, I focus on several methods that have become popular or seem to show promise for detection of gene–gene interactions or, more precisely, for detection of genes that might interact.

Recursive partitioning approaches. Recursive partitioning approaches (BOX 2) have been used as an alternative to traditional regression methods for detecting the genetic loci and their interactions that influence a phenotypic outcome[71–73]. These approaches produce a graphical structure that resembles an upside-down tree that maps the possible values of certain predictor variables (for example, SNP genotypes) to a final expected outcome (for example, disease status). Each vertex or node of the tree represents a predictor variable and there are arcs or edges from each node leading down to 'child' nodes, in which each edge corresponds to a different possible value that could be taken by the variable in the 'parent' node. A path through the tree represents a particular combination of values taken by the predictor variables that are present within that path. Recursive partitioning approaches do not include interaction variables *per se* in the model. Rather, the trees constructed allow for interaction in the sense that each path through a tree corresponds to a particular combination of values taken by certain predictor variables, thus including the potential interactions between them. The aim of tree-based approaches therefore corresponds most closely to testing for association while allowing for interaction rather than testing for interaction *per se*. One limitation of recursive partitioning is that, because it conditions on the main effects of variables at the first stage and on the main effects conditional on previously selected variables at subsequent stages, pure interactions in the absence of main effects can be missed[74].

Rather than using a single tree, substantial improvements in classification accuracy can result from growing an ensemble of trees. A popular ensemble tree approach is the random forests approach[75] (BOX 2), which has been used in several genetic studies[76,77]. Apart from the classification of future observations (which is not our focus of interest), the main result of a random forests analysis is a list of variable importance measures. These measure the effect of each predictor variable both individually and through multiway interactions with other predictor variables, and therefore have an advantage over a list of significance values from single-locus association testing.

Random forests provide a fast algorithm that can be applied in parallel for measuring variable importance partly because, at each split, only a small random subset of predictors is used. To allow each predictor the opportunity to enter the model and to make an accurate prediction, one must carefully choose important parameters, such as the number of trees in the forest, the number of randomly chosen SNPs analysed at each node and the number of permutations used to assess variable importance. Ideally, one would repeat the analysis several times to assess the sensitivity to the choice of these

High-dimensional data
Data that contain information on a large number of variables, albeit possibly measured in a small number of subjects or replicates.

Cross-validation
This approach involves partitioning a data set into smaller subsamples, performing an analysis in one subsample and using the other subsample to measure or validate how well the analysis has performed. To reduce variability, multiple rounds of cross-validation are often performed using different partitions of the data and the validation results are averaged over the rounds.

Overfitting
The phenomenon in which a complex model might provide a good fit to the current data set but is overfitted to the random quirks present in that particular data set and therefore cannot be generalized to future data sets in the way that a simpler model might be.

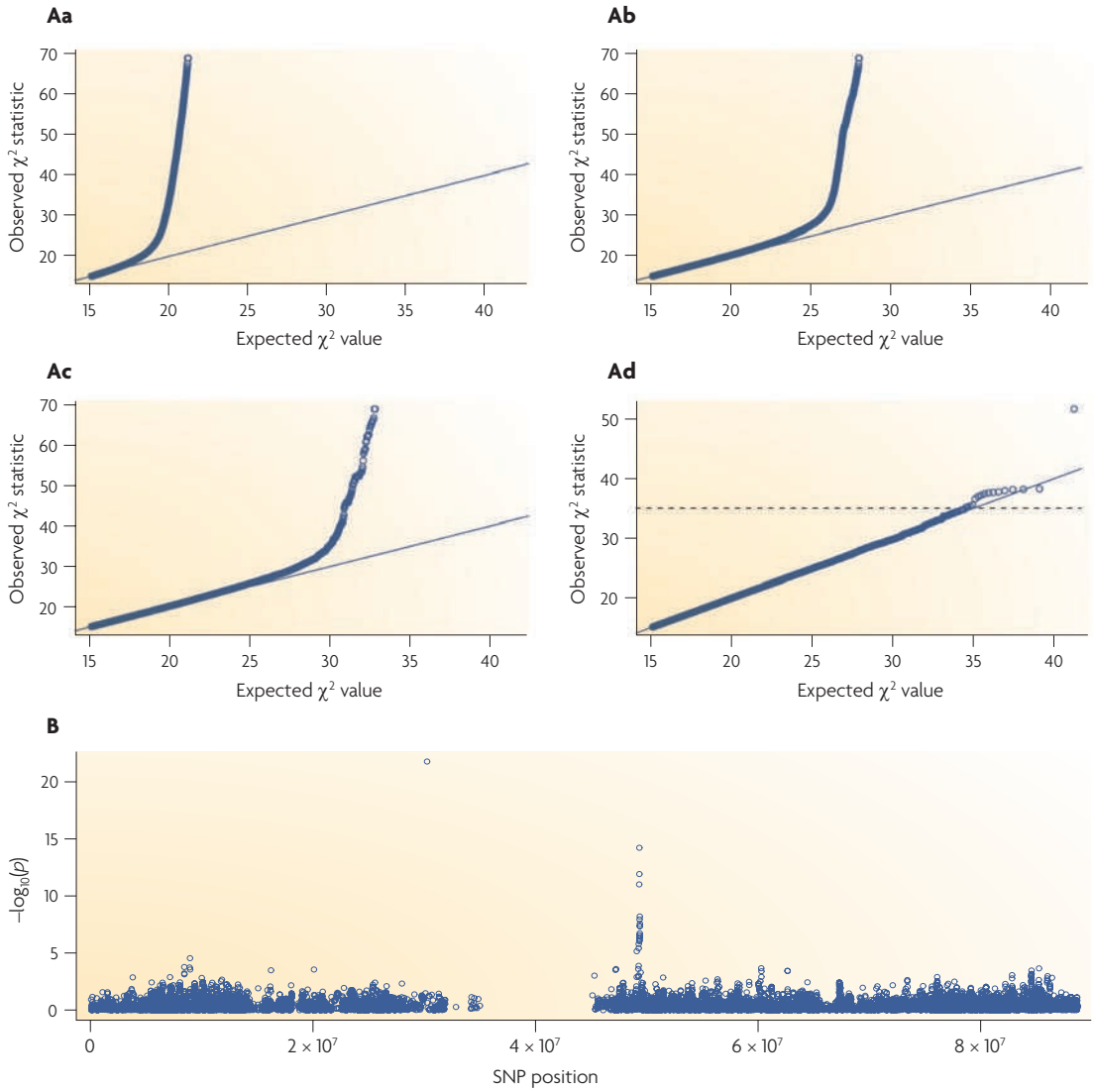
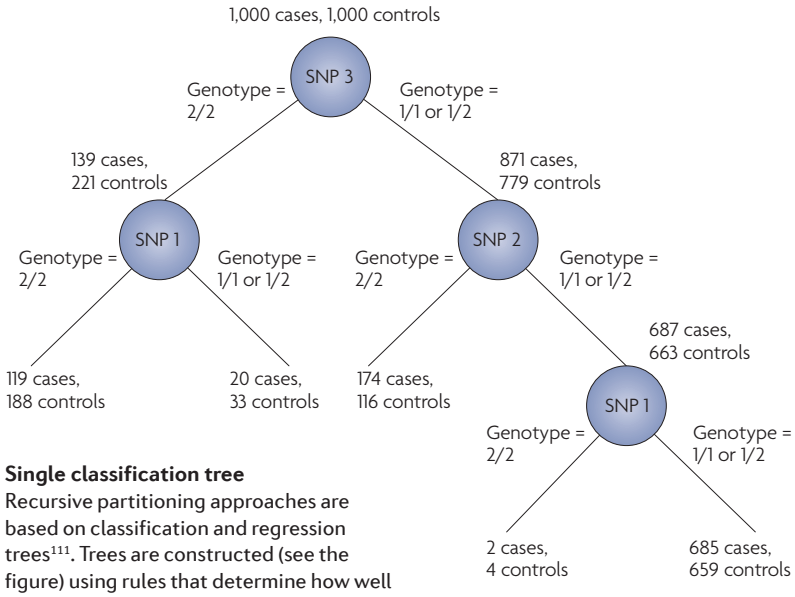


Figure 1 | **Semi-exhaustive search of pairwise interactions between 89,294 SNPs.** I used the '--fast-epistasis' and '--case-only' options in PLINK to analyse the Wellcome Trust Case Control Consortium (WTCCC) Crohn's disease and control samples. I used the same quality control procedures as the WTCCC to remove poor quality SNPs and samples before analysis. I additionally discarded 561 SNPs that had been analysed by WTCCC but were subsequently discarded on the basis of visual inspection of the SNP intensity cluster plots (J. Barrett, personal communication). To reduce the number of interaction tests to be performed, I selected a set of 89,294 SNPs that passed a single-locus *p* value threshold of 0.2. Analysis of the 89,294 SNPs on a single node of a computer cluster took 14 days. Unfortunately, neither SNP in the interaction detected by Emily *et al.*[60] were included in my analysis, as neither had a single-locus $p \leq 0.2$. **A** | Results from '--case-only' analysis, in which SNP pairs were discarded if they were <1 Mb apart (panel **a**), <5 Mb apart (panel **b**), and <50 Mb apart (panel **c**). The default in PLINK is to exclude tests of pairs of SNPs that are less than 1 Mb apart. Even when extreme separations of 5 Mb or 50 Mb are enforced (panels **b** and **c**), we find a large number of apparently significant results. A closer inspection showed that in many cases, these significant results are due to correlation within the sample of cases between alleles at loci on different chromosomes. Given the general departure from the expected distribution, it seems likely that these significant case-only results are artefacts rather than genuine interaction effects. Panel **d** shows a Q–Q plot of all results from the '--fast-epistasis' option with $p < -0.0001$. These results lie much closer to the expected line; only one result seems to show strong departure from the expected significance. The top-ranking results (those with $\chi^2 > 35$, as indicated by the dashed line on panel **d** are shown in Supplementary information S3 (table). Interestingly, most of the SNPs involved in the putative interactions show little single-locus significance, apart from rs4471699 on chromosome 16. This SNP was not reported as significantly associated by WTCCC[1]. **B** | Single-locus association results across chromosome 16. rs4471699 at position 30,227,808 shows the highest significance but is far removed from most of the significant results, which are situated close to nucleotide-binding oligomerization domain containing 2 (*NOD2*) (approximate position 49,297,083). Further investigation showed that this SNP had been excluded from the WTCCC analysis owing to poor genotype clustering (J. Barrett, personal communication), even though it passed the stated WTCCC exclusion criteria and was not present in the original list of additional exclusions I was given. It therefore seems likely that both the single-locus and interaction results at rs4471699 are false positives.

Box 2 | **Recursive partitioning approach**



Single classification tree

Recursive partitioning approaches are based on classification and regression trees[111]. Trees are constructed (see the figure) using rules that determine how well a split at a node (based on the values of a predictor variable such as a SNP) can differentiate observations with respect to the outcome variable (such as case–control status). A popular splitting rule is to use the variable that maximizes the reduction in a quantity known as the Gini impurity[111,112] at each node. In the figure, SNP 3 maximizes the reduction in the Gini impurity at the first node and is therefore chosen for splitting (according to the genotype at SNP 3) the original data set of 1,000 cases and 1,000 controls into two smaller data sets. Once a node is split, the same logic is applied to each child node (hence the recursive nature of the procedure). The splitting procedure stops when no further gain can be made (for example, when all terminal nodes contain only cases or only controls, or when all possible SNPs have been included in a branch) or when some preset stopping rules are met. At this stage, it is usual to prune the tree back (that is, to remove some of the later splits or branches) according to certain rules[111] to avoid overfitting and to produce a final more parsimonious model.

Ensemble approaches: random forests

Rather than using a single classification tree, substantial improvements in classification accuracy can result from growing an ensemble of trees and letting them 'vote' for the most popular outcome class, given a set of input variable values. Such ensemble approaches can be used to provide measures of variable importance, a feature that is of great interest in genetic studies and that is often lacking in machine-learning approaches. The most widely used ensemble tree approach is probably the random forests method[75]. A random forest is constructed by drawing with replacement several bootstrap samples of the same size (for example, the same number of cases and controls) from the original sample. An unpruned classification tree is grown for each bootstrap sample, but with the restriction that at each node, rather than considering all possible predictor variables, only a random subset of the possible predictor variables is considered. This procedure results in a 'forest' of trees, each of which will have been trained on a particular bootstrap sample of observations. The observations that were not used for growing a particular tree can be used as 'out-of-bag' instances to estimate the prediction error. The out-of-bag observations can also be used to estimate variable importance in different ways including through use of a permutation procedure[31,77,113].

The true model in which the important predictor variables act or interact to influence phenotype is somewhat obscured because it results from the predictions of many different classification trees, and so one might wish to follow a random forests analysis with another approach. For example, one might choose the top-ranking variables from a random forests analysis as input variables for a simple regression-based search, a standard classification and regression trees analysis or for analysis using an alternative data-mining procedure.

See REFS 31,74,113 for a good summary of the approach, the available R software (the 'randomForest', 'cforest' and 'party' libraries) and a discussion of some of the limitations of the method.

parameters. An example of applying random forests to the WTCCC Crohn's disease and control data using the Random Jungle software package[78] is shown in FIG. 2.

Multifactor Dimensionality Reduction method. A range of data-mining approaches have been used for the detection of interactions or potentially interacting variables in genetic association studies, including logic regression[79,80], genetic programming[81], neural networks[54,55] and pattern mining[82,83]. One particularly popular method is Multifactor Dimensionality Reduction (MDR)[8–10]. MDR has been used to identify potential interacting loci in several phenotypes, including breast cancer[8], type 2 diabetes[84], rheumatoid arthritis[85] and coronary artery disease[86], although to date it is unclear whether any of these identified interactions have been replicated in larger samples.

The MDR algorithm is described in BOX 3 and in detail elsewhere[8–11,49]. Rather than testing for interaction *per se*, MDR seeks to identify combinations of loci that influence a disease outcome, possibly by interactions rather than — or in addition to — by main effects. MDR reduces the number of dimensions by converting a high-dimensional multilocus model to a one-dimensional model, thus avoiding the issues of sparse data cells and models with too many parameters that can cause problems for traditional regression-based methods. MDR classifies genotypical classes as either high risk or low risk according to the ratio of cases and controls in each class. This approach could be considered overly simplistic, and improvements that embed a more traditional regression-based approach into the cell classification step, allowing application of the method to continuous as well as binary traits and adjustment for covariates, have been proposed[87,88].

The main problem with MDR, as with other exhaustive search techniques, is that it does not scale up to allow analysis of large numbers of predictor variables (for example, many loci from a GWA study)[8,9]. If an exhaustive search for the best *n*-locus combination (within each of ten cross-validation replicates) is performed, anything more than a two-locus screen on more than a few hundred variables will be computationally prohibitive. An additional problem with early versions of the widely used Java implementation of the MDR software (but note that other software implementations exist[11,88]) is that it was not designed with genome-wide data sets in mind and thus could fail owing to memory and disc usage issues. However, these problems seem to have been addressed in the most recent version of the software.

For investigation of higher-order interactions, MDR is therefore perhaps best suited for use with small numbers of loci (up to a few hundred), which have perhaps been discovered from a candidate gene study or selected from a larger set of potential predictors using a prior processing or filtering step[40]. This step could be as simple as using a single-locus significance threshold, but that seems counter-intuitive if the goal is to detect interactions in the absence of marginal effects. Perhaps a more appealing approach would be to use a measure of variable importance that allows for possible interactions,

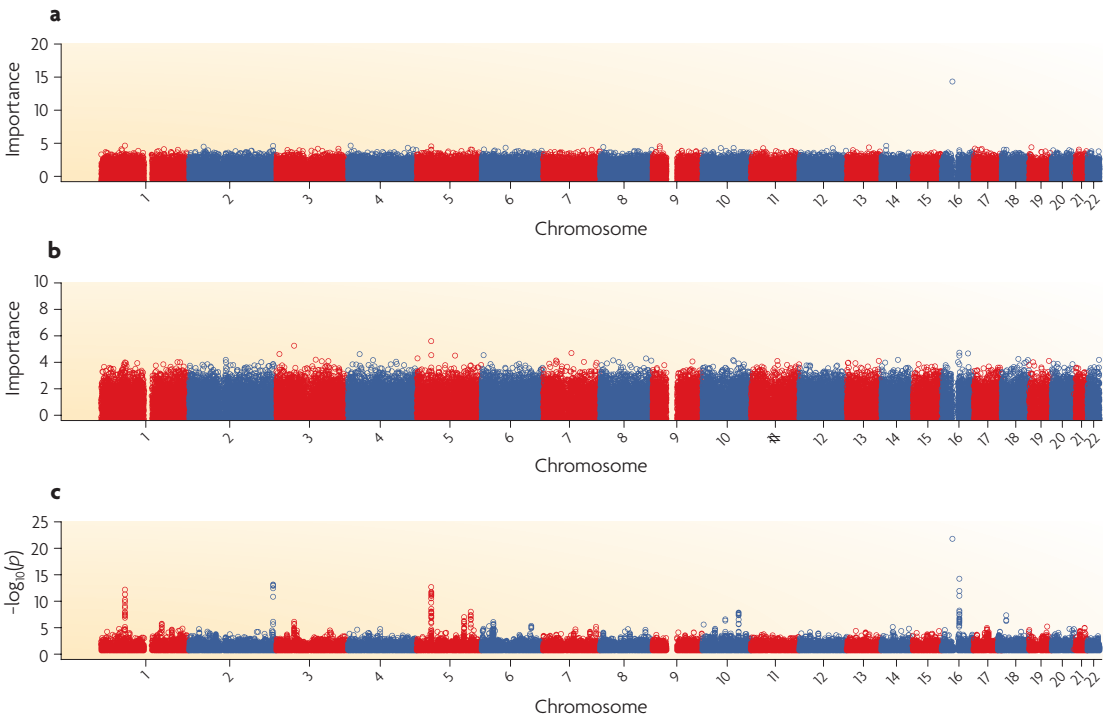


Figure 2 | **Random Jungle analysis of 89,294 SNPs.** I used the software package Random Jungle[78] to perform a random forests analysis of the 89,294 SNPs that passed a single-locus *p* value threshold of 0.2 in the Wellcome Trust Case Control Consortium (WTCCC) Crohn's disease and control data. As Random Jungle, in common with many other machine-learning approaches, prefers not to have missing genotype data, the missing genotypes were imputed as the single most likely values on the basis of the genotype frequencies in the case–control data set. Analysis of the 89,294 SNP set took approximately 5 hours, using 6,000 trees in the forest and $\sqrt{n} = \sqrt{89{,}294}$ randomly chosen variables at each node. **a** | Importance values from the Random Jungle analysis. These are clearly dominated by the result at rs4471699 on chromosome 16, which is likely to be a false positive. **b** | Results from Random Jungle analysis with SNP rs4471699 removed. Once this SNP is removed, the remaining SNPs are better distinguished, but it is unclear whether this analysis offers any greater insight than the single-locus analysis. **c** | Results from single-locus association analysis of all 6,113 SNPs using the trend test implemented in PLINK. In many cases, the highest ranking SNPs are in similar locations to (**b**), but with clearer significance in (**c**).

such as the variable importance measure from a random forests analysis or from one of the alternative filtering methods described below.

ReliefF, Tuned ReliefF and evaporative cooling. One promising filtering algorithm that has been proposed[40] is ReliefF[89] or its modified version, Tuned ReliefF (TuRF)[90]. This approach uses a measure of proximity between observations (individuals) — which is calculated, for example, on the basis of the genome-wide genetic similarity between individuals — to determine the nearest neighbours of each individual from within their own phenotype class and from within the opposite phenotype class. The difference in the value of each predictor variable between the pairs of neighbouring individuals, weighted negatively or positively according to whether the individuals come from the same or different phenotype classes, can be used to construct an importance measure for that variable[90]. The algorithm is simple and scalable, and should be applicable to large numbers of predictor variables and observations; an in-house C++ implementation was able to analyse 1 million loci in 200 individuals in approximately 4 minutes[90].

ReliefF and TuRF have both been implemented in the Java version of the MDR software. One problem with ReliefF is that it can be affected by large backgrounds of genetic variants that do not contribute to the phenotype[74]. This has motivated the development of an alternative approach, evaporative cooling[74,91], which can be used to combine the strengths of ReliefF with those of random forests methods[74].

An example of analysis using the Java implementation of TuRF and MDR applied to the WTCCC Crohn's disease data is shown in FIG. 3.

Bayesian model selection approaches

Bayesian model selection techniques[92] offer an alternative approach for selecting predictor variables and the interactions between them that are the best predictors of phenotype. The key difference between Bayesian model selection and simple comparisons of nested regression models using frequentist (non-Bayesian) procedures is the specification of prior distributions for the unknown regression parameters as well as for a dimension parameter in a Bayesian approach. This dimension parameter specifies how many non-zero predictors are included

Bootstrap samples
These are data sets obtained by taking a random sample of the original data, usually with replacement. One then applies the same analysis as was applied to the real data. This is repeated many times, allowing one to assess the variability in results incurred owing to random sampling.

Frequentist
A statistical approach for testing hypotheses by assessing the strength of evidence for the hypothesis provided by the data.

Box 3 | **Multifactor Dimensionality Reduction**

The Multifactor Dimensionality Reduction (MDR) method is a constructive induction algorithm[40] that proceeds as follows: the observed data is divided into ten equal parts and a model is fit to each nine-tenths of the data (the training data), and the remaining one-tenth (the test data) is used to assess model fit, thus using ten-fold cross-validation. Within each nine-tenths of the data, a set of *n* genetic factors is selected and their possible multifactor classes or cells are represented in *n* dimensional space. For example, for $n = 2$ diallelic loci, there are nine possible genotype classes or cells (Supplementary information S1 (box)). The ratio of the number of cases to the number of controls is estimated in each cell and the cell is labelled as either high risk if the case–control ratio reaches or exceeds a predetermined threshold (for example, ≥1) and low risk if it does not reach this threshold. This reduces the original *n*-dimensional model to a one-dimensional model (that is, one variable with two classes: high risk and low risk). The procedure is repeated for each possible *n*-factor combination and the combination that maximizes the case–control ratio of the high-risk group (that is, the combination that fits the current nine-tenths of the data best, giving minimum classification error among all *n*-locus models) is selected. The testing accuracy (which is equal to 1 – prediction error) of this best *n*-locus model can be estimated using the remaining test data portion of the data. The whole procedure is repeated for each of the nine-tenth-one-tenth partitions of the data, and the final best *n*-locus model is the model that maximizes the testing accuracy or, equivalently, minimizes the prediction error. The cross-validation consistency is defined as the number of cross-validation replicates (partitions) in which that same *n*-locus model was chosen as the best model (that is, the number of replicates in which it minimized classification error). The average prediction error is defined as the average of the prediction errors over the ten cross-validation test data sets. Note that the prediction error of each individual cross-validation replicate refers to the prediction error of the *n*-locus model chosen as the best model in that replicate, which will not always correspond to the final best *n*-locus model.

In practice, rather than selecting a single value of *n* in each cross-validation replicate, one might consider all possible values of *n* up to a certain maximum; for example, all single-locus genotype combinations ($n = 1$), all two-locus combinations ($n = 2$) or all three-locus combinations ($n = 3$). One thus generates a best model within each cross-validation replicate as well as a final best model (with the associated cross-validation consistency and average prediction error) for each different value of *n*. The cross-validation consistencies and average prediction errors can be used to determine the best value of *n* that gives the highest cross-validation consistency or lowest average prediction error, and thus the resulting overall best model.

Burn-in period
In Markov chain Monte Carlo analysis, a period at the start of the computation in which the values taken by the parameters are ignored when constructing the posterior distribution.

Compositional epistasis
The blocking of one allelic effect by an allele at another locus.

Statistical epistasis
The average effect of substitution of alleles at combinations of loci, with respect to the average genetic background of the population.

Functional epistasis
The molecular interactions that proteins and other genetic elements have with one another.

in the regression equation. A posterior distribution for these parameters, given the observed data, can then be calculated using Markov chain Monte Carlo (MCMC)[93] simulation techniques, in which one traverses the space of the possible models (sets of parameter values), sampling the outputs of the simulation run at intervals. Although MCMC is a flexible approach, it can require some care with respect to the choice of prior distributions, proposal schemes (determining how one moves between models) and the number of iterations required to achieve convergence.

Lunn *et al.*[56] proposed a Bayesian version of stepwise regression implemented in the software WinBUGS. This method focuses on the main effects of loci rather than interactions, but the inclusion of interaction effects is a straightforward extension. The main problem with this method is that it can deal with only a few hundred variables at most[56] and does not scale to the large numbers of predictor variables that might be encountered in a genome-wide study. However, related approaches that can deal with data sets with more dimensions have been proposed[94].

Bayesian Epistasis Association Mapping. A recently proposed MCMC approach that is specifically designed to detect interacting, as well as non-interacting, loci is Bayesian Epistasis Association Mapping[13], which is implemented in the software package BEAM. In BEAM, predictors in the form of genetic marker loci are divided into three groups: group 0 contains markers that are not associated with disease, group 1 contains markers that contribute to disease risk only by main effects and group 2 contains markers that interact to cause disease by a saturated model. Given prior distributions that describe the membership of each marker in each of the three groups and prior distributions for the values of the relevant regression coefficients given group membership, a posterior distribution for all relevant parameters can be generated using MCMC simulation. In addition to making inferences in a fully Bayesian inferential framework, one can use the results from BEAM in a frequentist hypothesis-testing framework by calculating a 'B-statistic'[13] that tests each marker or set of markers for significant association with a disease phenotype.

BEAM can handle large numbers of markers (for example, 100,000 SNPs typed in 500 cases and 500 controls[13]) although, in practice, some modification to the default parameters (namely the burn-in period, number of starting points and number of MCMC iterations) might be required to apply the method in a reasonable period of time. BEAM cannot currently handle the 500,000–1,000,000 markers that are now routinely being genotyped in genome scans of 5,000 or more individuals. In theory, BEAM can account for linkage disequilibrium between adjacent markers[13]. However, it is unclear whether linkage disequilibrium between non-adjacent markers is fully accounted for, suggesting that reducing the number of markers in the marker set might be required, not only for computational reasons, but also to ensure that the markers are in low linkage disequilibrium. An example of applying BEAM to the WTCCC Crohn's data is shown in FIG. 4.

Biological interpretation

The extent to which statistical interaction implies biological or functional interaction has been extensively debated in both the genetics[19,21,95–99] and epidemiological[100–102] literature. One problem has been the inherently different nature of definitions of interaction and the use of a common term, epistasis, to encapsulate these definitions[21,95] (Supplementary information S2 (box)). In a recent review, Phillips[20] defines three different forms of epistasis — compositional epistasis, statistical epistasis and functional epistasis — that capture different concepts that are often grouped together under this single term. A unified framework, the natural and orthogonal interactions (NOIA) model, was proposed by Alvarez-Castro and Carlborg[98] for modelling both statistical and functional epistasis. However, Alvarez-Castro and Carlborg's definition of functional differs from that of Phillips. The NOIA model is actually a mathematical reparameterization of classical quantitative genetics models[19] (Supplementary information S2 (box)). The NOIA model allows the main effects to be defined with

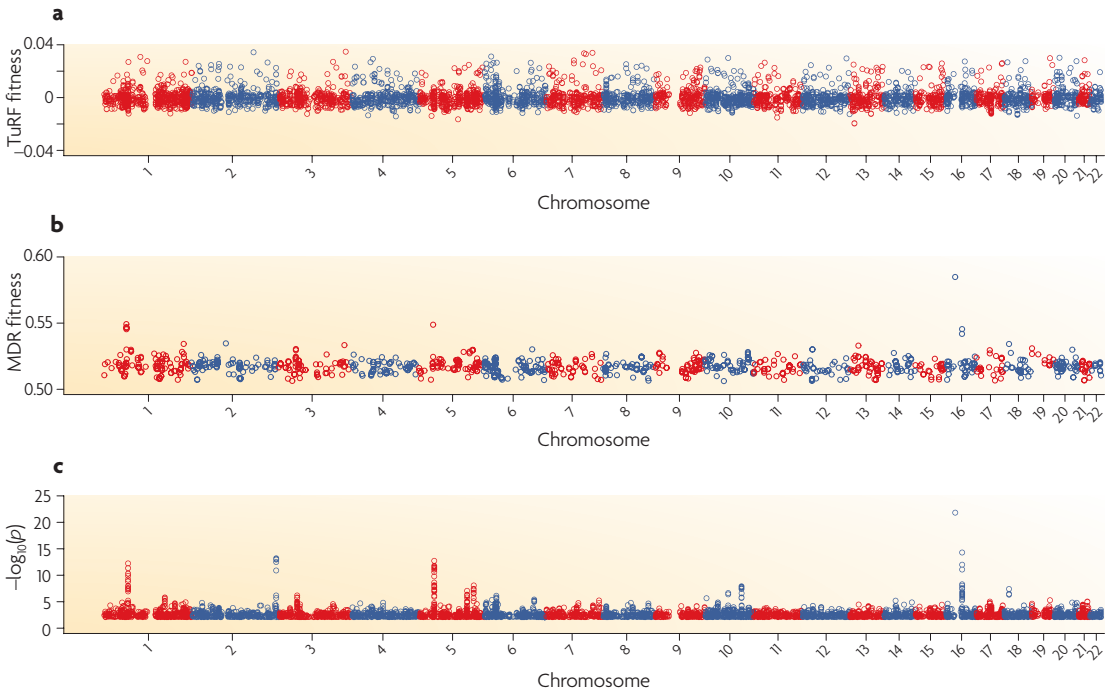


Figure 3 | **Multifactor Dimensionality Reduction (MDR) and Tuned ReliefF (TuRF) analysis of 6,113 SNPs.** I used the Java implementation of MDR to analyse 6,113 SNPs that passed a single-locus *p* value threshold of 0.01 in the Wellcome Trust Case Control Consortium (WTCCC) Crohn's disease and control data, with missing genotypes imputed as the single most likely values on the basis of the genotype frequencies in the case–control data set. Examination of all pairwise combinations in the entire 6,113 SNP set was computationally prohibitive but analysis using a prior filtering step with ReliefF or TuRF, which reduced the data set for MDR analysis to 1,000 SNPs, was achievable. The best single-locus model identified was rs4471699, providing a testing accuracy of 0.5852 and cross-validation consistency of 10 out of 10. The best two-locus model identified was rs4471699 and rs2076756, providing a testing accuracy of 0.5879 and cross-validation consistency of 4 out of 10. MDR, in common with the other methods investigated, has clearly been dominated by the false positive result at rs4471699. Interestingly, however, this SNP is not selected by TuRF when filtering down the set of SNPs for MDR analysis to include only 100 SNPs. Using the 100 SNP set, the best single-locus model identified was rs931058, providing a testing accuracy of 0.5114 and cross-validation consistency of 5 out of 10. The best two-locus model identified was rs931058 and rs10824773, providing a testing accuracy of 0.5205 but cross-validation consistency of only 2 out of 10. Using the 100 SNP set, it was computationally feasible to fit three-locus and four-locus models; however, the resulting best models had cross-validation consistencies as low as for the two-locus model. I also found extreme sensitivity in both TuRF and MDR to the choice of the random number seed (data not shown), suggesting that, overall, these results should be interpreted with caution. A problem with MDR is that it outputs only the best model rather than a measure of significance for all of the models or variables considered. An idea of the importance of the variables can be determined by examining the 'fitness landscape' output from the program, shown here. **a** | Fitness landscape scores from TuRF analysis of all 6,113 SNPs. **b** | Fitness landscape scores from MDR analysis using the top 1,000 out of 6,113 SNPs filtered using TuRF. **c** | Results from single-locus association analysis of all 6,113 SNPs using the trend test implemented in PLINK. It is unclear whether the fitness landscape results from TuRF (**a**) or MDR (**b**) offer any great advantage over standard single-locus analysis (**c**) with respect to determining the importance of variables.

respect to a different reference point and interaction effects to be defined with respect to different definitions of the independence of the main effects, thus allowing mapping of models between different experimental populations. As the whole issue in interaction modelling is how one defines the effect of a variable and, therefore, how one measures departure from the independence of effects (Supplementary information S2 (box)), this reparameterization does not seem to be biologically enlightening.

It may seem reasonable to assume that functional epistasis in the form of biomolecular or protein–protein interaction is a ubiquitous component of the underlying biological pathways that determine disease progression[7,103]. However, this does not mean that epistasis will be detected as a mathematical or statistical interaction[102,104], particularly if the variables that are being examined are, as in many cases, simply surrogates for the true underlying causal variants that are correlated with the causal variants because of linkage disequilibrium. The historical lack of success in genetic studies of complex disease can largely be attributed, not to ignored biological interactions[7,61,67], but to underpowered studies that surveyed only a fraction of genetic variation. The recent success of GWA studies[1–5] has shown that single-locus association analysis in sufficiently large sample collections can reliably detect modest genetic effects that are robustly replicated[105,106].

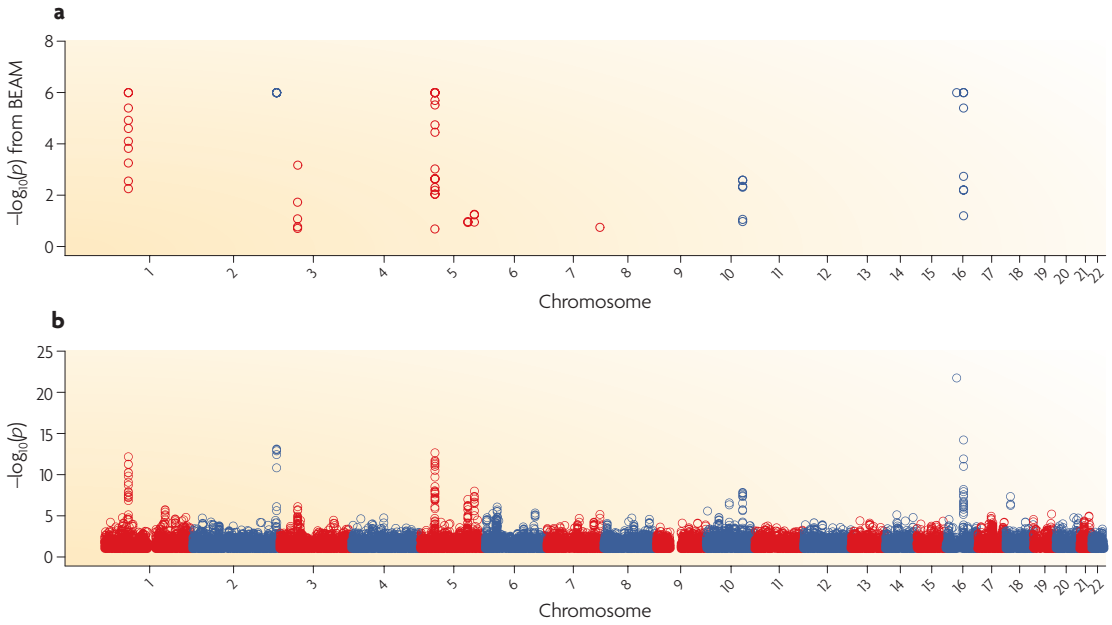


Figure 4 | **Bayesian Epistasis Association Mapping (BEAM) analysis of 47,727 SNPs.** I used BEAM to analyse a set of 47,724 SNPs that passed a single-locus *p* value threshold of 0.1 in the Wellcome Trust Case Control Consortium (WTCCC) Crohn's disease and control samples. Analysis of the 47,724 SNPs took 8 days (with some modification to the default settings, most notably imposing a maximum of 5×10^{-7} Markov chain Monte Carlo (MCMC) iterations[13] rather than using the default value of n^2, in which n is the number of loci). I estimated that analysis of the 89,294 SNP set passing a single-locus *p* value threshold of 0.2 with a similar number of MCMC iterations would have taken more than 5 weeks. **a** | 'B-statistic' *p* values for the 1,321 single-locus associations detected by BEAM. **b** | Results from single-locus association analysis of all 47,727 SNPs using the trend test implemented in PLINK. BEAM detects the same loci as are detected by single-locus analysis. BEAM additionally detects (with a quoted *p* value of 0.000000) four two-locus interactions, each involving an interaction of rs2532292 on chromosome 17 with a nearby SNP (either rs12150547, rs17689882, rs17650381 or rs17574824) within the same cluster. None of these SNPs shows particularly strong single-locus associations and so this putative interaction is intriguing. However, none of these pairs of SNPs showed significant (defined as a $p < 0.0001$) interaction in the PLINK '--fast-epistasis' analysis. Closer inspection of these SNPs in the control sample indicated that they are in strong linkage disequilibrium ($D' > 0.99$) with one another, suggesting that the detected interactions might correspond to marker dependencies owing to linkage disequilibrium, rather than to genuine interaction effects.

Although the extent to which biological interaction can be inferred from statistical interaction might be limited[102], some interesting recent studies[107–109] have focused on whether, given a strong prior biological model or set of models, one can use genetic or genomic data from outbred populations or inbred strains to assess the fit of the model and compare the fits of competing models. This is a more modest goal because it relies on a prior understanding or at least a strong biological hypothesis with respect to the action of the relevant predictors.

Conclusions

As we have seen, there are numerous methods and an even larger number of software implementations that allow investigators to examine or test for interaction between loci, using data that is currently generated from large-scale genotyping projects. Although the precise details of the methods differ, in many cases there are close conceptual links between the different approaches. The best way to understand these links might be provided by understanding the difference between testing for interaction versus testing for association while allowing for interaction.

From a practical point of view, probably the main difference between the methods I have described is the computational time required to implement the analysis. As data sets become larger, the development of efficient computational algorithms that can be implemented in parallel will become more important. On this note, the use of filtering approaches that allow one to preselect a subset of potentially interesting loci to input into a more computer-intensive exhaustive or stochastic search algorithm might hold promise. In my application of various methods to the WTCCC Crohn's disease data, I found that a semi-exhaustive search of two-locus interactions implemented in PLINK[12] and a random forests analysis implemented in Random Jungle[78] were the most computationally feasible of the methods examined. Bayesian Epistasis Association Mapping implemented in BEAM[13] was feasible only for a filtered data set and with some modification to the default recommended input parameter settings; it is unclear what effect, if any, this will have had on the reliability of the results. MDR was feasible for examining two-locus interactions in a filtered data set or for examining higher-level interactions in an even further reduced data set.

To date, few publications have incorporated interaction testing of GWA data. This is perhaps unsurprising as GWA studies have naturally focused on single-locus testing in the first instance. Curtis[110] performed pairwise tests of association at 396,591 markers using 541 subjects (cases and controls) from a genome-wide study of Parkinson's disease. He found no significant epistatic interactions, possibly because of the small sample size or because of the interaction test that was used, which might have been more powerful if it was restricted to cases alone. Gayan *et al.*[15] used the same data set to perform two-locus interaction testing using their interaction detection approach, hypothesis-free clinical cloning. This approach involves testing for association while allowing for interaction under a set of prespecified fully penetrant disease models, and the tests are performed in several different subgroups of the data, which are considered as replication groups. For the Parkinson's disease analysis, each subgroup consisted of approximately 90 cases and 90 controls, which seems a very small sample size for this kind of analysis. Unsurprisingly, little consistency between results was found when the analysis was repeated using different partitions of the data. Emily *et al.*[60] reported four significant cases of epistasis in the WTCCC data using an approach that narrows the search space on the basis of experimental knowledge of biological networks.

Given the large number of GWA studies that have recently been or are currently being performed, it is clear that, for many, genome-wide interaction testing will be the natural next step following single-locus testing. We await with interest the results of these analyses.

1. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
In this study of 17,000 individuals, many new complex trait loci were identified and key methodological and technical issues related to GWA studies were explored.
2. Easton, D. F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
3. Frayling, T. M. *et al.* A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894 (2007).
4. Plenge, R. M. *et al. TRAF1-C5* as a risk locus for rheumatoid arthritis — a genome-wide study. *N. Engl. J. Med.* **357**, 1199–1209 (2007).
5. Fellay, J. *et al.* A whole-genome association study of major determinants for host control of HIV-1. *Science* **317**, 944–947 (2007).
6. Culverhouse, R., Suarez, B. K., Lin, J. & Reich, T. A perspective on epistasis: limits of models displaying no main effect. *Am. J. Hum. Genet.* **70**, 461–471 (2002).
7. Moore, J. H. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.* **56**, 73–82 (2003).
8. Ritchie, M. D. *et al.* Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* **69**, 138–147 (2001).
This was the original paper describing the popular MDR method.
9. Hahn, L. W., Ritchie, M. D. & Moore, J. H. Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics* **19**, 376–382 (2003).
10. Moore, J. H. Computational analysis of gene–gene interactions using multifactor dimensionality reduction. *Expert Rev. Mol. Diagn.* **4**, 795–803 (2004).
11. Chung, Y., Lee, S. Y., Elston, R. C. & Park, T. Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions. *Bioinformatics* **23**, 71–76 (2007).
12. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
13. Zhang, Y. & Liu, J. S. Bayesian inference of epistatic interactions in case–control studies. *Nature Genet.* **39**, 1167–1173 (2007).
This paper proposed a new Bayesian approach for the detection of loci that might interact in the context of GWA studies. The related BEAM software package provides a computationally efficient implementation of the proposed algorithm.
14. Ferreira, T., Donnelly, P. & Marchini, J. Powerful Bayesian gene–gene interaction analysis. *Am. J. Hum. Genet.* **81** (Suppl.), 32 (2007).
15. Gayan, J. *et al.* A method for detecting epistasis in genome-wide studies using case–control multi-locus association analysis. *BMC Genomics* **9**, 360 (2008).
16. Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J. & Gauderman, W. J. Exploiting gene–environment interaction to detect genetic associations. *Hum. Hered.* **63**, 111–119 (2007).
17. Fisher, R. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edin.* **52**, 399–433 (1918).
18. Hayman, B. I. & Mather, K. The description of genetic interactions in continuous variation. *Biometrics* **11**, 69–82 (1955).
19. Zeng, Z. B., Wang, T. & Zou, W. Modeling quantitative trait loci and interpretation of models. *Genetics* **169**, 1711–1725 (2005).
This paper includes an excellent discussion of issues in the definition and interpretation of interaction in quantitative genetic studies of derived populations (inbred lines).
20. Phillips, P. C. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Rev. Genet.* **9**, 855–867 (2008).
An excellent review describing the differing definitions and interpretations of epistasis.
21. Cordell, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **11**, 2463–2468 (2002).
22. Cordell, H. J., Todd, J. A., Bennett, S. T., Kawaguchi, Y. & Farrall, M. Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of *IDDM2* and *IDDM4* with *IDDM1* in type 1 diabetes. *Am. J. Hum. Genet.* **57**, 920–934 (1995).
23. Cox, N. J. *et al.* Loci on chromosomes 2 (*NIDDM1*) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nature Genet.* **21**, 213–215 (1999).
24. Cordell, H. J., Wedig, G. C., Jacobs, K. B. & Elston, R. C. Multilocus linkage tests based on affected relative pairs. *Am. J. Hum. Genet.* **66**, 1273–1286 (2000).
25. Strauch, K., Fimmers, R., Baur, M. & Wienker, T. F. How to model a complex trait 2. Analysis with two disease loci. *Hum. Hered.* **56**, 200–211 (2003).
26. Armitage, P., Berry, G. & Matthews, J. N. S. *Statistical Methods in Medical Research* 4th edn (Blackwell Science, Chichester, 2002).
27. McCullagh, P. & Nelder, J. A. *Generalized Linear Models* (Chapman & Hall, London, 1989).
28. Neuman, R. J. & Rice, J. P. Two-locus models of disease. *Genet. Epidemiol.* **9**, 347–365 (1992).
29. Li, W. & Reich, J. A complete enumeration and classification of two-locus disease models. *Hum. Hered.* **50**, 334–349 (2000).
30. Hallgrimsdottir, I. B. & Yuster, D. S. A complete classification of epistatic two-locus models. *BMC Genet.* **9**, 17 (2008).
31. McKinney, B. A., Reif, D. M., Ritchie, M. D. & Moore, J. H. Machine learning for detecting gene–gene interactions: a review. *Appl. Bioinformatics* **5**, 77–88 (2006).
32. Piegorsch, W. W., Weinberg, C. R. & Taylor, J. A. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case–control studies. *Stat. Med.* **13**, 153–162 (1994).
An important paper showing the use of case-only designs for detection of gene–environment interactions in epidemiological studies.
33. Yang, Q., Khoury, M. J., Sun, F. & Flanders, W. D. Case-only design to measure gene–gene interaction. *Epidemiology* **10**, 167–170 (1999).
34. Weinberg, C. R. & Umbach, D. M. Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *Am. J. Epidemiol.* **152**, 197–203 (2000).
35. Mukherjee, B. *et al.* Tests for gene–environment interaction from case–control data: a novel study of type I error, power and designs. *Genet. Epidemiol.* **32**, 615–626 (2008).
36. Zhao, J., Jin, L. & Xiong, M. Test for interaction between two unlinked loci. *Am. J. Hum. Genet.* **79**, 831–845 (2006).
37. Hoh, J. & Ott, J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Rev. Genet.* **4**, 701–709 (2003).
38. Mukherjee, B. & Chatterjee, N. Exploiting gene–environment independence for analysis of case–control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* **64**, 685–694 (2008).
39. Yang, Y., Houle, A. M., Letendre, J. & Richter, A. *RET* Gly691Ser mutation is associated with primary vesicoureteral reflux in the French-Canadian population from Quebec. *Hum. Mutat.* **29**, 695–702 (2008).
40. Moore, J. H. *et al.* A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J. Theor. Biol.* **241**, 252–261 (2006).
41. Chanda, P. *et al.* Information-theoretic metrics for visualizing gene–environment interactions. *Am. J. Hum. Genet.* **81**, 939–963 (2007).
42. Kang, G. *et al.* An entropy-based approach for testing genetic epistasis underlying complex diseases. *J. Theor. Biol.* **250**, 362–374 (2008).
43. Dong, C. *et al.* Exploration of gene–gene interaction effects using entropy-based methods. *Eur. J. Hum. Genet.* **16**, 229–235 (2008).
44. Zwick, M. An overview of reconstructability analysis. *Kybernetes* **33**, 877–905 (2004).
An excellent overview of some of the principles and techniques used in information-theory modelling of frequency and probability distributions.
45. Cordell, H. J. & Clayton, D. G. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to *HLA* in type 1 diabetes. *Am. J. Hum. Genet.* **70**, 124–141 (2002).
46. Cordell, H. J., Barratt, B. J. & Clayton, D. G. Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene–gene and gene–environment interactions and parent-of-origin effects. *Genet. Epidemiol.* **26**, 167–185 (2004).
This paper describes a regression-based framework for the analysis of family-based data that allows tests of interaction that are similar to the tests often used in case–control studies to be performed.
47. Martin, E. R., Ritchie, M. D., Hahn, L., Kang, S. & Moore, J. H. A novel method to identify gene–gene effects in nuclear families: the MDR-PDT. *Genet. Epidemiol.* **30**, 111–123 (2006).
48. Kotti, S., Bickeboller, H. & Clerget-Darpoux, F. Strategy for detecting susceptibility genes with weak or no marginal effect. *Hum. Hered.* **63**, 85–92 (2007).

49. Lou, X. Y. *et al.* A combinatorial approach to detecting gene–gene and gene–environment interactions in family studies. *Am. J. Hum. Genet.* **83**, 457–467 (2008).
50. Gauderman, W. J. Sample size requirements for association studies of gene–gene interaction. *Am. J. Epidemiol.* **155**, 478–484 (2002).
51. Hein, R., Beckmann, L. & Chang-Claude, J. Sample size requirements for indirect association studies of gene–environment interactions (G x E). *Genet. Epidemiol.* **32**, 235–245 (2008).
52. Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genet.* **37**, 413–417 (2005).
This paper highlights the importance and feasibility of fitting interaction models using GWA data.
53. Chapman, J. & Clayton, D. Detecting association using epistatic information. *Genet. Epidemiol.* **31**, 894–909 (2007).
54. Motsinger, A., Lee, S., Mellick, G. & Ritchie, M. GPNN: power studies and applications of a neural network method for detecting gene–gene interactions in studies of human disease. *BMC Bioinformatics* **7**, 39 (2006).
55. Motsinger-Reif, A. A., Dudek, S. M., Hahn, L. W. & Ritchie, M. D. Comparison of approaches for machine-learning optimization of neural networks for detecting gene–gene interactions in genetic epidemiology. *Genet. Epidemiol.* **32**, 325–340 (2008).
56. Lunn, D. J., Whittaker, J. C. & Best, N. A Bayesian toolkit for genetic association studies. *Genet. Epidemiol.* **30**, 231–247 (2006).
57. Hoh, J. *et al.* Selecting SNPs in two-stage analysis of disease association data: a model-free approach. *Ann. Hum. Genet.* **64**, 413–417 (2000).
58. Millstein, J., Conti, D. V., Gilliland, F. D. & Gauderman, W. J. A testing framework for identifying susceptibility genes in the presence of epistasis. *Am. J. Hum. Genet.* **78**, 15–27 (2006).
59. Bochdanovits, Z. *et al.* Genome-wide prediction of functional gene–gene interactions inferred from patterns of genetic differentiation in mice and men. *PLoS ONE* **3**, e1593 (2008).
60. Emily, M., Mailund, T., Schauser, L. & Schierup, M. H. Using biological networks to search for interacting loci in genomewide association studies. *Eur. J. Hum. Genet.* 11 Mar 2009 (doi: 10.1038/ejhg.2009.15).
61. Moore, J. H. & Williams, S. M. New strategies for identifying gene–gene interactions in hypertension. *Ann. Med.* **34**, 88–95 (2002).
62. Golub, G., Heath, M. & Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215–224 (1979).
63. Velez, D. R. *et al.* A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet. Epidemiol.* **31**, 306–315 (2007).
64. Copas, J. B. Regression, prediction and shrinkage. *J. Roy. Stat. Soc., Series B* **45**, 311–354 (1983).
65. Hastie, T., Tibshirani, R., & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (Springer, New York, 2001).
66. Lee, A. & Silvapulle, M. Ridge estimation in logistic regression. *Comm. Stat. Simul. Comput.* **17**, 1231–1257 (1988).
67. Le Cessie, S. & Van Houwelingen, J. Ridge estimators in logistic regression. *Appl. Stat.* **41**, 191–201 (1992).
68. Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least angle regression. *Ann. Statist.* **32**, 407–499 (2004).
69. Park, M. Y. & Hastie, T. Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**, 30–50 (2008).
70. Zhang, Z., Zhang, S., Wong, M. Y., Wareham, N. H. & Sha, Q. An ensemble learning approach jointly modelling main and interaction effects in genetic association studies. *Genet. Epidemiol.* **32**, 285–300 (2008).
71. Zhang, H. & Bonney, G. Use of classification trees for association studies. *Genet. Epidemiol.* **19**, 323–332 (2000).
72. Nelson, M. R., Kardia, S. L., Ferrell, R. E. & Sing, C. F. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.* **11**, 458–470 (2001).
73. Culverhouse, R., Klein, T. & Shannon, W. Detecting epistatic interactions contributing to quantitative traits. *Genet. Epidemiol.* **27**, 141–152 (2004).
74. McKinney, B. A., Crowe, J. E., Guo, J. & Tian, D. Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS Genet.* **5**, e1000432 (2009).
75. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
76. Lunetta, K. L., Hayward, L. B., Segal, J. & Van Eerdewegh, P. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet.* **5**, 32 (2004).
77. Bureau, A. *et al.* Identifying SNPs predictive of phenotype using random forests. *Genet. Epidemiol.* **28**, 171–182 (2005).
78. Schwartz, D. F., Ziegler, A. & König, I. R. Beyond the results of genome-wide association studies. *Genet. Epidemiol.* **32**, 671 (2008).
79. Kooperberg, C., Ruczinski, I., LeBlanc, M. & Hsu, L. Sequence analysis using logic regression. *Genet. Epidemiol.* **21**, S626–S631 (2001).
80. Kooperberg, C. & Ruczinski, I. Identifying interacting SNPs using Monte Carlo logic regression. *Genet. Epidemiol.* **28**, 157–170 (2005).
81. Nunkesser, R., Bernholt, T., Schwender, H., Ickstadt, K. & Wegener, I. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics* **23**, 3280–3288 (2007).
82. Li, Z., Zheng, T., Califano, A. & Floratos, A. Pattern-based mining strategy to detect multi-locus association and gene × environment interaction. *BMC Proc.* **1** (Suppl. 1), S16 (2007).
83. Long, Q., Zhang, Q. & Ott, J. Detecting disease-associated genotype patterns. *BMC Bioinform.* **10** (Suppl. 1), S75 (2009).
84. Cho, Y. M. *et al.* Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus. *Diabetologia* **47**, 549–554 (2004).
85. Julia, A. *et al.* Identification of a two-loci epistatic interaction associated with susceptibility to rheumatoid arthritis through reverse engineering and multifactor dimensionality reduction. *Genomics* **90**, 6–13 (2007).
86. Tsai, C. T. *et al.* Renin–angiotensin system gene polymorphisms and coronary artery disease in a large angiographic cohort: detection of high order gene–gene interaction. *Atherosclerosis* **195**, 172–180 (2007).
87. Lee, S. Y., Chung, Y., Elston, R. C., Kim, Y. & Park, T. Log-linear model based multifactor-dimensionality reduction method to detect gene–gene interactions. *Bioinformatics* **23**, 2589–2595 (2007).
88. Lou, X. Y. *et al.* A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am. J. Hum. Genet.* **80**, 1125–1137 (2007).
89. Robnik-Sikonja, M. & Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* **53**, 23–69 (2003).
90. Moore, J. H. & White, B. C. Tuning ReliefF for genome-wide genetic analysis. *Lect. Notes Comp. Sci.* **4447**, 166–175 (2007).
91. McKinney, B. A., Reif, D. M., White, B. C., Crowe, J. & Moore, J. H. Evaporative cooling feature selection for genotypic data involving interactions. *Bioinformatics* **23**, 2113–2120 (2007).
92. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis* (Chapman and Hall, London, 1995).
93. Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. *Markov Chain Monte Carlo in Practice* (Chapman and Hall, London, 1996).
94. Hoggart, C. J., Whittaker, J. C., De Iorio, M. & Balding, D. J. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* **4**, e1000130 (2008).
95. Phillips, P. C. The language of gene interaction. *Genetics* **149**, 1167–1171 (1998).
An important paper that describes the differing definitions and interpretations of epistasis used in different fields and the lack of equivalence between these definitions.
96. Moore, J. H. & Williams, S. M. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays* **27**, 637–646 (2005).
97. Cheverud, J. M. & Routman, E. J. Epistasis and its contribution to genetic variance components. *Genetics* **139**, 1455–1461 (1995).
98. Alvarez-Castro, J. M. & Carlborg, O. A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics* **176**, 1151–1167 (2007).
99. McClay, J. L. & van den Oord, E. J. Variance component analysis of polymorphic metabolic systems. *J. Theor. Biol.* **240**, 149–159 (2006).
100. Thompson, W. D. Effect modification and the limits of biological inference from epidemiologic data. *J. Clin. Epidemiol.* **44**, 221–232 (1991).
101. Siemiatycki, J. & Thomas, D. C. Biological models and statistical interactions: an example from multistage carcinogenesis. *Int. J. Epidemiol.* **10**, 383–387 (1981).
102. Greenland, S. Interactions in epidemiology: relevance, identification, and estimation. *Epidemiology* **20**, 14–17 (2009).
A useful commentary on the relationship between statistical and biological interaction assessed from epidemiological studies.
103. Gibson, G. Epistasis and pleiotropy as natural properties of transcriptional regulation. *Theor. Popul. Biol.* **49**, 58–89 (1996).
104. Vanderweele, T. J. Sufficient cause interactions and statistical interactions. *Epidemiology* **20**, 6–13 (2009).
105. Todd, J. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genet.* **39**, 857–864 (2007).
106. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
107. Sepulveda, N., Paulino, C. D., Carneiro, J. & Penha-Goncalves, C. Allelic penetrance approach as a tool to model two-locus interaction in complex binary traits. *Heredity* **99**, 173–184 (2007).
108. Sepulveda, N., Paulino, C. D. & Penha-Goncalves, C. Bayesian analysis of allelic penetrance models for complex binary traits. *Comp. Stat. Data Anal.* **53**, 1271–1283 (2009).
109. Aylor, D. L. & Zeng, Z. B. From classical genetics to quantitative genetics to systems biology: modeling epistasis. *PLoS Genet.* **4**, e1000029 (2008).
110. Curtis, D. Allelic association studies of genome wide association data can reveal errors in marker position assignments. *BMC Genet.* **8**, 30 (2007).
111. Breiman, L., Freidman, J. H., Olshen, R. A. & Stone, C. J. *Classification and Regression Trees* (Chapman and Hall/CRC, New York, 1984).
112. Bastone, L., Reilly, M., Rader, D. J. & Foulkes, A. S. MDR and PRP: a comparison of methods for high-order genotype–phenotype associations. *Hum. Hered.* **58**, 82–92 (2004).
113. Strobl, C., Boulesteix, A. L., Zeileis, A. & Hothorn, T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25 (2007).
This paper gives an overview of some of the strengths and limitations of random forests analysis for measuring variable importance.

Acknowledgements
Support for this work was provided by the Wellcome Trust (Grant reference 074524). I thank J. Barrett for assistance with interpretation of the WTCCC Crohn's results, and the WTCCC for making their data freely available. I also thank J. Moore for useful discussions of data-mining methods in general and MDR in particular, and K. Keen for pointing out the origins of the term epistasis.

DATABASES
OMIM: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
Crohn's disease

FURTHER INFORMATION
Heather J. Cordell's homepage: http://www.staff.ncl.ac.uk/heather.cordell
BEAM: http://www.people.fas.harvard.edu/~junliu/BEAM
MDR: http://sourceforge.net/projects/mdr
Nature Reviews Genetics Series on Genome-wide association studies: http://www.nature.com/nrg/series/gwas/index.html
Nature Reviews Genetics Series on Modelling: http://www.nature.com/nrg/series/modelling/index.html
PLINK: http://pngu.mgh.harvard.edu/~purcell/plink
Random Jungle: http://randomjungle.com

SUPPLEMENTARY INFORMATION
See online article: S1 (box) | S2 (box) | S3 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Human language as a culturally transmitted replicator

Mark Pagel

Abstract | Human languages form a distinct and largely independent class of cultural replicators with behaviour and fidelity that can rival that of genes. Parallels between biological and linguistic evolution mean that statistical methods inspired by phylogenetics and comparative biology are being increasingly applied to study language. Phylogenetic trees constructed from linguistic elements chart the history of human cultures, and comparative studies reveal surprising and general features of how languages evolve, including patterns in the rates of evolution of language elements and social factors that influence temporal trends of language evolution. For many comparative questions of anthropology and human behavioural ecology, historical processes estimated from linguistic phylogenies may be more relevant than those estimated from genes.

Languages
Linguists identify two languages as distinct when, according to various criteria, they become mutually unintelligible.

School of Biological Sciences, University of Reading, Reading, Berkshire RG6 6AH, UK; and Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico, USA.
e-mail: m.pagel@reading.ac.uk
doi:10.1038/nrg2560
Published online 7 May 2009

Here is a remarkable fact about humans: we speak approximately 7,000 mutually unintelligible languages around the world[1]. This means that a person plucked from one corner of the Earth is not able to communicate with another human in a different corner of the Earth, or often from next door. Apart from a few songbird species, and possibly some whales that learn their songs locally and show dialectical differences, this is unique among animals. For example, a chimpanzee or an elephant removed from its range and placed among any other chimpanzees or elephants will know what to do and how to communicate.

Large as the number of extant human languages is, it has probably reduced from a maximum of perhaps 12,000 to 20,000 different languages before the spread of agriculture[2], and it pales in comparison with the possibly hundreds of thousands of different languages humans have ever spoken[2]. Elsewhere I have pondered the question of why humans would evolve a system of communication that prevents them from communicating with other members of its species, and have suggested that human societies come to behave in ways that are not so different from that of biological species[3]. Whether or not that explanation is correct, the human tendency to separate into distinct societies has given human language a geographical mosaic on which to play out its evolution. My interest here is to use the phenomenon of language diversity to understand the evolution of what turns out to be a remarkable culturally transmitted replicator, one with many of the properties we have come to expect of genes, but also with many of its own.

In this Review I shall first describe how a new and expanding field of phylogenetic and comparative studies of language evolution has made use of concepts, data and statistical modelling approaches that draw inspiration from genetics to exploit the genetic-like properties of language. I shall then move on to describe recent work in four areas of language evolution in which statistical modelling approaches have begun to return results. These include the reconstruction of language phylogenies and their relationship to genetic trees; investigations of the rate, tempo and time-depth of language evolution; social influences on language; and studies of the structure of language.

My coverage of these topics will be selective, but is designed to give a flavour of what language evolution is like and of what is possible. I will not discuss the tricky and very large literatures on language origins or how we acquire it, whether our language skills are innate, or possible genetic influences on language abilities. Instead, I will treat language as evolving against what I will regard for sake of discussion as a more or less homogeneous genetic background in its human hosts.

Descent with modification

One of the best-known theories for the diversity of human languages is a creation myth. According to the bible story of the Tower of Babel, humans developed the conceit that they could construct a tower that would take them all the way to heaven. Angered at the attempt to usurp his control, God destroyed the tower. To ensure

Table 1 | **Some analogies between biological and linguistic evolution**

Biological evolution	Language evolution
Discrete heritable units (for example, nucleotides, amino acids and genes)	Discrete heritable units (for example, words, phonemes and syntax)
Mechanisms of replication	Teaching, learning and imitation
Mutation (for example, many mechanisms yielding genetic alterations)	Innovation (for example, formant variation, mistakes, sound changes, and introduced sounds and words)
Homology	Cognates
Natural selection	Social selection and trends
Drift	Drift
Cladogenesis* (for example, allopatric speciation (geographic separation) and sympatric speciation (ecological or reproductive separation))	Lineage splits (for example, geographical separation and social separation)
Anagenesis‡	Linguistic change without split
Horizontal gene transfer	Borrowing
Hybridization (for example, horse with zebra and wheat with strawberry)	Language Creoles§ (for example, Surinamese)
Correlated genotypes and phenotypes (for example, allometry‖ and pleiotropy¶)	Correlated cultural terms (for example, 'hasta' and 'spear')
Geographic clines#	Dialects and dialect chains
Fossils	Ancient texts
Extinction	Language death

Darwin noted many of these parallels in *The Descent of Man*[4]. Table is modified, with permission, from *Nature* REF. 26 © (2007) Macmillan Publishers Ltd. All rights reserved. *Cladogenesis: the formation of separate groups by evolutionary splitting. ‡Anagenesis: the evolutionary process whereby one species evolves into another without any splitting of the lineage into separate groups or species. §Creole: a language that emerges in the second or later generations of the speakers of pidgins (which are the rudimentary languages that form when two language communities mix and seek a common basis for simple communication). Creoles are typically more complex than pidgins, although less so than fully developed languages. ‖Allometry: the relationship between size and shape. ¶Pleiotropy: the action of a single gene on two or more distinct phenotypic characters. #Clines: a gradual change in phenotype in a species over a given area.

that it could not be rebuilt, God confused the workers by giving them different languages, leading to the irony that language exists to stop us from communicating.

Delightful as the Babel story is, ideas taken from the theory of evolution give us the conceit that we can improve on it. Darwin[4] asserted that languages, like biological species, evolve by a process of descent with modification. If correct, we can expect human languages to form into family trees, known as phylogenies, which chart the history of their evolution in a manner analogous to that for biological species. It also means that the diversity of extant languages reflects the actions of various shared historical evolutionary processes, including features of the rate and tempo of linguistic evolution, timings and correlations, as well as the starting points or ancestral languages. This raises the possibility that, far from settling for each language being a distinct object of creation, we can use the combination of phylogenetic trees of language along with statistical models of how languages evolve to detect and characterize the signature of these historical processes. In effect, we wish to discover what the past must have been like and how it evolved given what we now see.

TABLE 1 records analogies between the ways that genes and languages evolve, giving hope that the use of phylogenetic methods will succeed. Key among these analogous features is that both systems of replicators are digital, comprising discrete heritable units: the four nucleotides in the case of genes, and words in the case of language. Without this property neither system would retain fidelity through repeated bouts of transmission from parents to offspring (genes) or from teachers to learners (language), and historical signals would quickly be lost. Other features of language evolution that might be thought to vitiate its historical signature have analogies in genetic systems. For example, languages can acquire new unrelated words by borrowing, and genes can arrive from bouts of lateral transfer. These influences often occur at lower rates in genetic systems, but do not represent a qualitative difference between genes and language.

Data and statistical modelling

The starting point for most comparative statistical investigations of language evolution is a set of discrete characters that can be scored in each of the languages. These might include features of the syntax or structure of a language, other aspects of grammar, phonemes and, most obviously, lexical items or words. I shall confine my remarks here to the lexicon, although most of what I have to say applies to these other classes of discrete traits. Owing to pioneering work by Morris Swadesh[5] in the 1950s a common list of 200 words known as the fundamental vocabulary is available for a large number of the world's languages (see the further information box for a link to an example of a Swadesh list). This type of list contains words for things that are expected to be found in all languages, such as names for body parts, pronouns, common verbs and numerals, but excludes technological words and words related to specific ecologies or habitats. It can be thought of as like a list of universal genes. Other lists are possible, but Swadesh's has simply proven to be well chosen and widely available. His words tend to evolve slowly and are largely resistant to outside influences and borrowing[6].

Phylogeny
A branching diagram describing the set of ancestral–descendant relationships among a group of species or languages.

Borrowing
The acquisition of a new non-cognate word from another language.

Phoneme
Characteristically thought of as the smallest units of speech-sounds that are distinguished by the speakers of a particular language. Phonemes are not universal, but act as the fundamental building blocks to produce all of the words of a given language.

Box 1 | **A linguistic alignment and a statistical model of evolution**

Matrix of cognates

Whereas a gene sequence alignment identifies homologous sites in genes, a lexical 'alignment' identifies sets of cognate words, or words that descend from a common ancestral word. Let a matrix of these lexical alignments be denoted M (see also REFS. 7,9,19) to signify that it is a matrix of meanings (for example, hand, who and ear), and write M as:

$$M = \begin{array}{c} \\ \text{language 1} \\ \text{language 2} \\ \vdots \\ \vdots \\ \text{language n} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{meanings} & & & & \\ 1 & 2 & 3 & & \text{m} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & . & 0 \\ 1 & 1 & 1 & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 3 & 1 & k & . & 0 \end{pmatrix} \end{array}$$

The columns of numbers designate cognate classes, or words for a given meaning that have been identified as deriving from a common ancestral word (see text). The first column of M denotes a meaning for which four distinct cognate classes of words exist (0, 1, 2 and 3), the second column shows a meaning represented by two cognate classes, the third has $k + 1$ cognate classes, and the last column shows a meaning with a single cognate class — that is, all of the words for that particular meaning among the n languages derive from a common ancestral word. This matrix is the analogue to an aligned set of gene sequences, although all gene sequences have the same four states (twenty states if considering amino acids).

A statistical model

The data in M can be used to infer phylogenetic trees of languages, or perhaps to investigate some feature of lexical replacement or the change from one cognate class to another. A statistical model that is widely used in phylogenetic inference from gene sequences is written as:

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ . \\ . \\ k \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & . & . & k \end{array} \\ \begin{pmatrix} - & q_{01} & . & . & q_{0k} \\ q_{10} & - & . & . & q_{1k} \\ . & . & . & . & . \\ . & . & . & . & . \\ q_{k0} & q_{k1} & . & . & - \end{pmatrix} \end{array}$$

The matrix Q is the central element of the finite-state continuous time Markov transition model (see text). Each q_{ij} term in this matrix describes the instantaneous rate of change from state i to state j over the short interval dt. In gene sequence data the states are the bases A, C, G or T, and Q is always a 4×4 matrix. If protein sequences are used, the states are amino acids and Q becomes a 20×20 matrix. Using lexical data, the states are the cognate classes, and a different Q needs to be estimated for meanings with different numbers of cognate classes. For phylogenetic inference with lexical data it is convenient to rewrite M such that all of the columns have the same number of cognate classes and thus a common Q can be estimated for the entire matrix, as is common for gene sequences. This is achieved by converting M to a binary form such that the $k + 1$ cognate classes for each meaning are written as k + 1 binary vectors, each one of which identifies a different cognate class as '1' with the remainder designated '0'. Then, Q becomes a 2×2 matrix estimating a common rate at which new cognate classes occur.

The elements of the Q matrix are presumed to apply equally well to each site in a gene or protein sequence or, in the case of lexical data, to different words. To accommodate variation in the rates of evolution among sites or among words, the well-known gamma rate variability correction can be applied[40]. This correction amounts to multiplying each of the q_{ij} in Q by carefully chosen constants that are either less than or greater than one to achieve an overall slower or faster rate of evolution.

Outside the linguistic context M could contain any set of cross-cultural or other comparative data, and Q could then be used to estimate their evolutionary transitions (for example, REFS 8,9).

Given a list of m meanings (such as hand, tree, I, walk, run) in n languages, a data set (M) can be written as a matrix, analogous to an alignment of gene sequences (BOX 1), but recording sets of cognate words. Linguists, using careful rules of sound correspondences within language families, can assign words from different languages into classes denoting words that derive from a common ancestral word, analogous to identifying homologous genes. The word 'two' in English and *dos* in Spanish are cognate. The French *fleur* and Dutch *blumen* are not. Cognacy data, by recording evolved similarities and differences among languages, can be used to infer linguistic phylogenetic trees or to study features of lexical evolution itself[7]. Phylogenies are of interest in their own right as descriptions of historical relationships, and they form the backbone of comparative studies that seek to understand the evolution of linguistic traits, and how other cultural traits have evolved and co-evolved[7–9]. For these wider cross-cultural studies, M can be broadened to include the cultural data.

Statistical approaches apply models of evolution to characterize the probability of observing the data in M under various evolutionary scenarios[10,11]. A model that will be familiar to geneticists is the finite-state continuous time Markov transition model (BOX 1). Often designated Q, it was introduced into studies of phylogenetic inference from gene sequences by Felsenstein[12] along with the sum over histories logic. Applied to lexical data, this model estimates the instantaneous rates at which the words of one cognate class evolve into another unrelated set of words[7]. Other statistical approaches to analysing M include a 'stochastic Dollo' model[13] that allows each new cognate class to arise only once on a tree of languages. The name is a conscious nod to the Belgian palaeontologist Louis Dollo (1857–1931), who suggested that identical complex forms do not arise more than once in nature. A different stochastic treatment of M (described in REF. 14) allows for borrowing while estimating the underlying tree.

Parsimony or distance-based methods can be used instead of statistical approaches. However, statistical approaches, unlike the other methods, allow one to estimate directly parameters of the models of evolution (such as the q_{ij} transition rates in Q, see BOX 1) and to test among different models for the same data[15]. Whether inferring trees or studying the evolution of traits on trees, the common currency for testing models is a quantity known as the likelihood or L, defined as an amount proportional to the probability of the data given the model[16]. It is conventionally written as $L \propto P(\text{M} \mid \text{Q}, T)$, where M and Q are as defined here, and T refers to the phylogenetic tree on which the data in M are presumed to have evolved. Likelihood methods regard the observed data as a fixed observation. This makes them particularly suited to historical inference problems such as those in linguistics, in which the observed data arise only once. Thus, the likelihood does not describe the probability that the events under study happened (they did) or that the model is true; it merely describes the 'fit' between the observed data and an inferred tree, or model (such as Q) of how a trait evolved on a tree.

The likelihood can be found by maximum likelihood methods — in this case many different trees or models are tried and the one that gives the largest value of *L* is preferred. Alternatively, Markov chain Monte Carlo (MCMC) methods[17] are increasingly being employed to infer models and trees[9,18]. Rather than seeking a single 'best' solution, MCMC methods attempt to derive a distribution of outcomes consistent with the data, called the posterior distribution. Posterior distributions can be formed for trees, for their likelihoods and for the parameters of the model of evolution. Their attraction is in providing a measure of the uncertainty in the estimates of these various components. The posterior distribution of trees also provides the logical background against which to estimate models of evolution for other traits, this being a tidy way to account for the effects of uncertainty about the past.

Language trees and gene trees

Features of language trees. An early attempt to apply a likelihood sum over histories approach to languages made use of 7 Indo-European languages, 18 meanings and the finite-state Markov transition model Q described above[19]. The analysis yielded a phylogeny with the expected monophyletic groupings of Romance languages (Spanish, French and Romanian) and Germanic languages (German, Dutch and English), and Welsh as an out-group. In the same year a tree of 77 Austronesian languages appeared, which was derived from parsimony methods[20]. Holden[21], also using parsimony and a 100-word Swadesh list, inferred a tree of 93 Bantu languages.

Later analyses of the Bantu data with likelihood models returned more or less the same tree[22]. Gray and Atkinson[23] applied the Markov transition model in a MCMC context to analyse 87 Indo-European languages using the entire Swadesh 200-word list, estimating an ancestral age for Indo-European languages of between 7,800 and 9,800 years. Trees of Papuan languages have been inferred from both typological and lexical features of language[24]. Recently, Gray and colleagues have expanded their Austronesian sample to include over 400 languages, inferring the tree using MCMC approaches and the Markov transition model[25]. Their tree supports a scenario for the origin of this group in Formosa, beginning approximately 6,000 years ago.

My interest here is less in the trees *per se* than in their characteristics. FIGURE 1 shows a consensus tree derived from the Bayesian posterior distribution of trees for 87 Indo-European languages[23,26]. The tree recovers the expected clades of Romance, Germanic, Slavic, Indo-Iranian and Celtic languages, and suggests their deeper relationships. But what is remarkable about this tree is how tree-like it is, given all of the ways that a linguistic signal can be corrupted — most obviously by borrowing. The numbers near to the nodes of this tree record the posterior support for that node, defined as the proportion of trees in the posterior distribution in which that node was found. These posterior support values rival those found for many gene trees of a similar size[27]. Comparable degrees of posterior support are reported for the Bantu and Austronesian trees[20,22], if not for the Papuan languages[24]. Techniques designed to reveal conflicting phylogenetic signals (for example, SplitsTree and Neighbour-net analyses)[28], such as would arise from borrowing, typically reveal a healthy pattern of tree-like data[29].

The Indo-European and Bantu trees reflect population expansions or radiations into new areas, riding on the back of agriculture[21,23,30], whereas the Austronesian tree records an expansion that may have been propelled in fits and starts linked to developments in sea-going boat technologies[20,25]. These population processes might contribute to the elegance of the three phylogenies by reducing the opportunities for borrowing of lexical items among the speakers of differing languages. Trees must always be carefully checked for borrowing, but unless it regularly occurs among distantly related languages, the broad structure of language phylogenies should be relatively unaffected[31]. Owing to a battle lost at Hastings, England, in 1066, English was bombarded by words of Romance origin and now approximately 50% of its vocabulary derives from such stock. Still, despite its history, English correctly appears among the Germanic languages in the I-E tree (FIG. 1), although linguists often place it closer to Frisian than the basal position it occupies in its portion of the Germanic clade.

Comparison of gene trees and language trees. If languages are not the 'closed shop' to outside influences that we have come to expect of eukaryotic organisms with sequestered germ lines, the strength of descent with modification in language trees shows that the cultural processes of language teaching and learning that transmit language from one generation to the next can have a surprisingly high fidelity and can show resistance to outside effects. Although genes may only be replicated once or a few times between generations, vocabulary items are replicated by producing a sound that is copied by a listener and then produced anew in a cyclical process that may occur many tens of thousands of times (or more) per word per speaker. The opportunities for mutation and corruption of this signal, not to mention for innovation and borrowing, are great and yet these simple lists of words can reconstruct the cultural history of groups of speakers spanning thousands of years.

Trees derived from language bear a range of relationships to gene trees for the same population, and this is as we should expect. Cavalli-Sforza[32] demonstrated in the late 1980s that the major genetic groupings of people around the world conform, with few exceptions, to their language groupings. This reveals, unsurprisingly, that people divided by large geographical distances drift apart genetically and linguistically. More fine-grained analyses reveal a different picture. Sometimes language groups conform closely to genetic groups even on a small geographical scale[33] and other times they do not[34]. This does not invalidate one kind of tree or elevate another. It tells us that some trees are good for tracking the movements of genes, and others for tracking the movement of cultures.

Cognate
Two words are deemed cognate if they derive by a process of descent with modification from a common ancestral word.

Sum over histories
A mathematical technique that accounts for all possible ancestral states (that is, all possible histories) when finding the likelihood of observing the gene sequence or other data among extant species.

Parsimony
When applied to phylogenetic inference in a linguistic context, parsimony is a method that seeks the phylogenetic tree that implies the fewest number of changes among cognate classes.

Distance
As applied to phylogenetic inference in a linguistic context, distance is a set of methods that infer an underlying phylogenetic tree from a matrix of the pair-wise differences among all languages.

Likelihood
A statistical quantity defined as an amount that is proportional to the probability of observing some set of data given a particular model of how those data arose. In linguistic phylogenetic applications one finds the likelihood of the lexical data on the proposed tree given some model of how words evolve.

Maximum likelihood method
A statistical technique for finding the parameters of a model that make the observed data most likely or probable under that model.

Markov chain Monte Carlo
(MCMC). A statistical method for searching a complex high-dimensional space. As applied to phylogenetic inference in a linguistic context, MCMC methods return a sample of trees that are statistically representative of the trees that might arise from a given model of how words evolve.

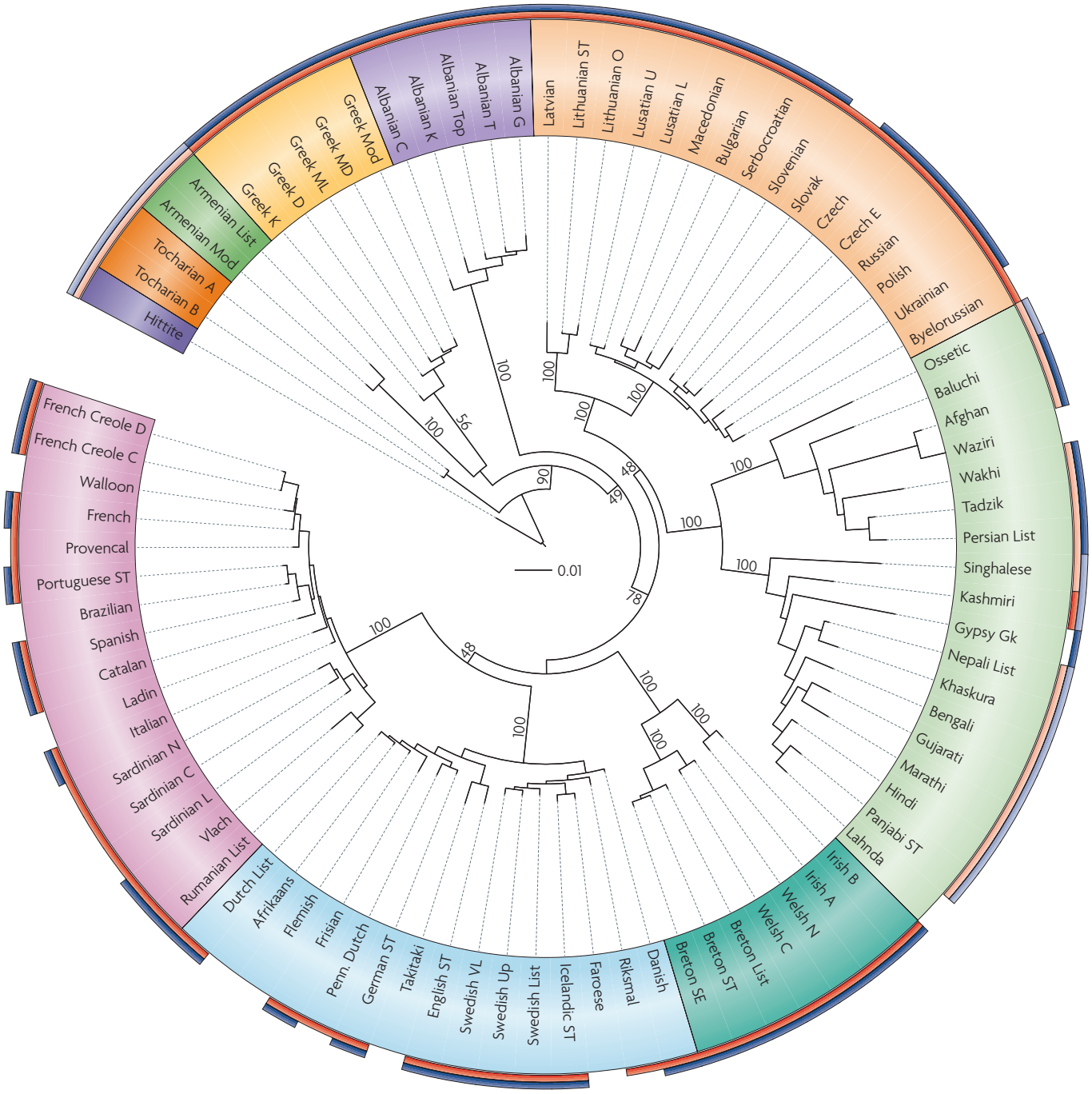


Figure 1 | **Tree of Indo-European languages.** Consensus tree of 87 Indo-European languages derived from the Swadesh list of 200 words[5,23,26]. Inner coloured ring identifies major clades as shown in the legend. The tree is rooted using ancient Hittite and Tocharian languages[23,26]. Branch lengths measure the expected number of lexical replacements (word changes) between two points on the tree. Numbers along branches are the Bayesian posterior probabilities of selected deep nodes of the tree, showing that words can resolve old relationships. Many of those nodes not labelled have high posterior support, although some are low and suffer from conflicting signals[26]. The outer colours identify a language's sentence word order in terms of subject (S) verb (V) and object (O) (red bars), and whether it employs pre or postpositional modification of sentence objects (blue bars) (see text, data from REF. 69). Red, SVO or VSO; light red, SOV; blue, prepositional; light blue, postpositional. Celtic languages are VSO; Greek, German, Dutch, Byelorussian and others are sometimes classified as no dominant word order (NDO) (here coded red). Blue–red pairs and light blue–light red pairs conform to Greenberg's[58] prediction (see text and FIG. 5)

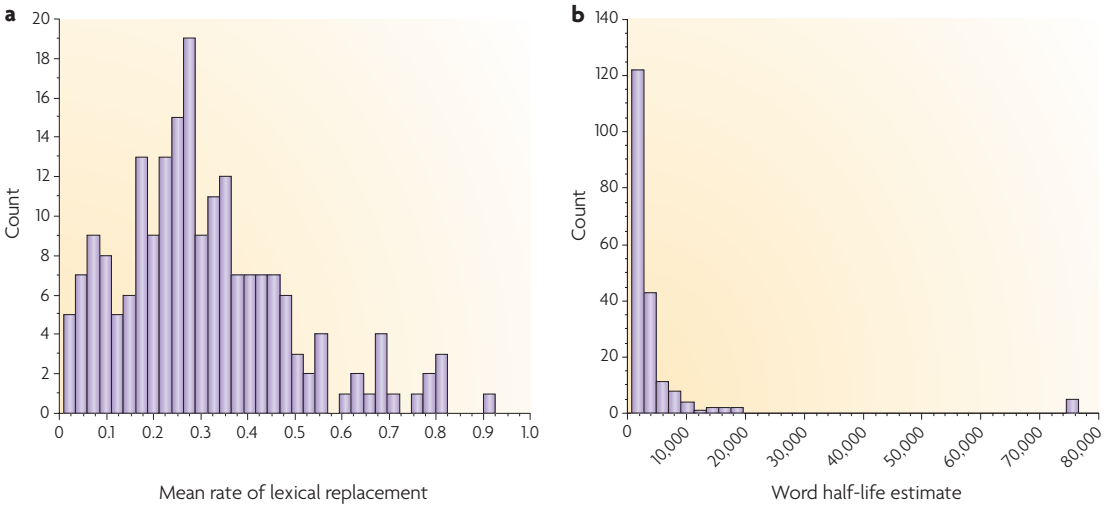


Figure 2 | **Rates of lexical replacement. a** | Histogram of mean rates of lexical replacement in Indo-European languages for the 200 words in the Swadesh list, measured in units of numbers of new cognates per 1,000 years of evolution. Fastest to slowest rate represents over 100-fold difference. Values were found as the mean of the posterior distribution of the elements of Q matrices (see text) integrated over a Bayesian posterior distribution of trees[26]. Mean = 0.3 ± 0.18 new cognates per 1,000 years, median = 0.27, range = 0.009 to 0.93. **b** | Histogram of the word half-life estimates as derived from the rates of lexical replacement, measuring the expected amount of time before a word has a 50% chance of being replaced by a new non-cognate word. A half-life of >70,000 years is indicative of a transition rate that is compatible with observing a single cognate class (that is, no changes) over the entire ~130,000 'language years' of the Indo-European tree[26] (calculated as the total obtained by adding the number of years of evolution represented by the sum of the branches of the tree in FIG. 1). Existence of at least five such classes in Indo-European lexicon lends support to this estimate. Mean = 5,300 years, median = 2,500, range = 750 to 76,000.

Indo-European languages
A family of related languages that derive from a common ancestral language that probably arose in Anatolia around 8,000 years ago and then spread throughout Europe, India, and what is now Afghanistan, Pakistan and Iran.

Monophyletic
In a phylogenetic context, a group of species (or languages) is monophyletic if they derive from a common ancestor not shared with any other species (or languages). The Germanic languages are monophyletic and are distinct from the monophyletic group of Romance languages. Monophyly implies that the group has just one origin.

Bantu languages
A group of approximately 500 languages that is part of the larger Niger-Congo language family. Bantu languages probably arose 3,000 years ago in West Africa, possibly close to present day Cameroon, and then spread east and then south eventually reaching to present day South Africa.

Clade
In the context of languages, a clade is a group of related languages.

Lexical replacement
The rate of lexical replacement is the rate at which a word is replaced by a new non-cognate word.

To understand why this is true, consider a thought experiment in which human genes flow among populations or even around the world largely invisible to the human phenotypes they inhabit (although there are some hints of genes and languages co-evolving[35]). Culture can, in principle, rest easily above this flow, as migrants adopt the local traditions, such that cultural variants and changes are independent of genetic changes. A situation similar to this has recently been reported for some Melanesian islanders[34]. Accordingly, phylogenetic trees derived from language may be preferable to gene trees in cross-cultural studies whenever the variables of interest are culturally transmitted[8]. These studies must separate the influence of common ancestry on a trait's representation among cultures from independent instances of the acquisition or evolution of that trait[8]. For cultural data, such as bride wealth and dowry, matriliny, patriliny, modes of subsistence and even sex ratio[36–38], the cultural phylogenetic tree provides the description of common ancestry that makes this separation possible. Linguists and anthropologists need not suffer from gene envy when it comes to building and using phylogenies.

In the next three sections I move away from inferring and interpreting language trees to discuss examples of how they have been used as the backbones of investigations into how features of language evolve, including rates of word evolution and the structure of languages, and to investigate social influences on the rates of lexical evolution.

Rates of evolution and time depth

Differing rates of word evolution. What English speakers call a bird, the Italians call *ucello*, the French *oiseau*, the Spanish *pajaro*, the Germans *vogel*, the Greeks *pouli*, and Caesar would have said *avis*. There are approximately 15 different cognate classes for 'bird' among the 90 or so Indo-European languages. By comparison, all Indo-European language speakers use a related form of the word 'two' (*dos*, *deux*, *due*, *zwei*; the Latin is *duo*) to describe two objects. Just as some sites in a gene sequence alignment evolve slowly and others rapidly, words in the Indo-European languages show ~100-fold variation in their rates of lexical replacement or in the acquisition of a new non-cognate form[7,26] (FIG. 2a). These rates were found from estimating the q_{ij} in Q separately for each word in the Swadesh list, integrating over a Bayesian posterior sample of Indo-European trees. Slowly evolving words include 'two', 'three', 'I', 'five' and 'who', each of which has just a single cognate class among the Indo-European languages. By comparison, words such as 'bird', 'tail', 'sand', and 'belly' evolve more rapidly, with the word 'dirty' having, at 46, the largest number of cognate classes.

The rates can be expressed as word half-lives[7,19,26] (FIG. 2b) corresponding to the expected amount of time before a word has a 50% chance of being replaced by a new non-cognate word. The median half-life is 2,000–2,500 years. This may seem fast, especially when compared to genes, but it is slower than the average rate at which new languages appear (which is approximately every 500–1,000 years for Indo-European languages).

Even the most rapidly evolving words have fewer cognate classes than the number of languages. This means that, in general, words can achieve a measure of immortality by escaping into a new language before a new form replaces them.

Some words, like highly conserved genes, evolve at very slow rates. For each of the 5 most slowly evolving words there is a single cognate class in Indo-European languages. This is consistent with a half-life of over 70,000 years[26], a rate of evolution as slow as some genes[39], and shows that a culturally transmitted replicator can achieve a surprising fidelity. The sound an Indo-European language speaker makes to describe two objects is ancient, a related sound having been used by every speaker of an Indo-European language. If the unique time that each language has evolved is summed over languages this amounts to 130,000 or so language years that 'two' has remained stable. The same is true for the other words with a single cognate class. Even a word with a 6,000-year lexical half-life has a 25% chance of not changing in 11,500 years. Putting all these figures together, comparative linguists who seek evidence of very old linguistic signals are not simply chasing unicorns: there is every reason to expect that a linguistic signal exists that can identify relationships among distantly related language families.

Reasons for rate heterogeneity. Heterogeneity in the rates of evolution of words can be accommodated when inferring language phylogenies in the same way as correction for differing rates of substitution in genetics (using the gamma correction)[11,40]. But at a deeper level we want to understand why this rate variation exists. We sought a general explanation for variation in rates of replacement by studying the 'expression level' of a word, that is, the frequency with which it is used in everyday speech[26]. Speech is dominated by a small number of frequently used words, the remainder being used infrequently[41,42]. We found that slowly evolving words in Indo-European are those with higher expression levels; they are used more frequently in everyday speech[26]. Within English, frequently used words are more likely to be of Old English origin[43]. For example, irregular English verbs retain their ancestral morphology[44,45] and are the more commonly used verbs.

Speakers of different Indo-European languages use the various words in the Swadesh list at similar frequencies in their everyday speech[26]. It might be that the way we use language and its structure means that some words inevitably will be used more than others; it is, for example, difficult to avoid verbs and pronouns. If so, then frequency of use has potentially been a general historical influence in the world's language families. FIGURE 3 plots the rates of lexical replacement we have reported for the Indo-European languages[26] against a list of 110 words that the late Russian comparative linguist Sergei Starostin identified as among the most stable in 14 language families from around the world[46]. The figure shows that slowly evolving words in Indo-European languages are also slowly evolving in the world's other language families, and vice versa; remarkably, this suggests that rates of evolution have been conserved throughout human history. This result attests to the generality and historical influence of the frequency effect, and gives additional support to the search for deep language relationships.

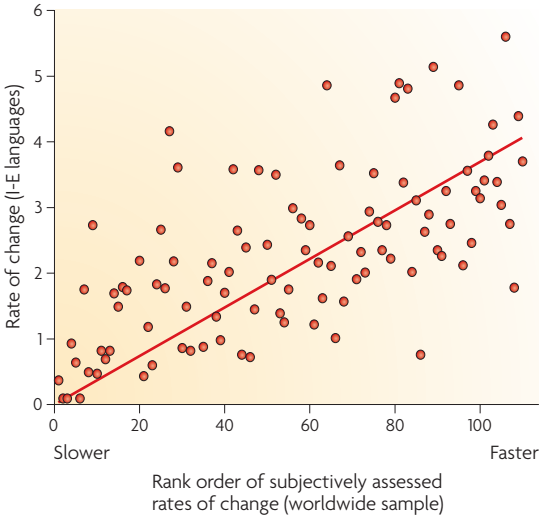


Figure 3 | **Rates of lexical replacement are stable among language families.** Statistically estimated rates of lexical replacement for 110 words from the Swadesh list in the Indo-European (I-E) languages (data from REF. 26 and FIG. 2a) correlated with rank ordering of subjectively assessed rates of change for the same words in a worldwide sample of 14 languages families, correlation (r) = 0.65. The 14 language families assessed were Sino-Tibetan, Austroasiatic, Altaic, Austronesian, Australian, Khoisan, North Caucasian, Dravidian, Indo-European, Kartvelian, Afroasiatic, Tai, Uralic and Yenisean. The rank order list was taken from REF. 46.

Frequency of use might affect rates of lexical replacement by altering 'production errors' — akin to the mutation rate in genetics — or by altering the rate at which a new form is adopted in a speech community (akin to selection), or both[26,47]. Word use may be under strong purifying selection within populations of speakers, if only through the rule 'speak as most others do'. It is difficult to understand how entire populations of speakers could otherwise agree on a single or small number of mostly arbitrary sounds to represent a given meaning. Such a rule would have been advantageous in our history if speakers who make mistakes are disadvantaged. If I say that the war-like tribe coming over the hill numbers two when in fact I meant two hundred, there may be consequences. Some words may acquire connections in the cognitive or semantic space[48], connections the strength or size of which may influence how rapidly words evolve. For example, *hasta* is the Sanskrit word for hand, but among Latin speakers it became the word for spear. The sound 'hasta' may have been saved by the cognitive connection between hand and spear. Questions surrounding why different words evolve at different rates are areas rich for discovery and are only just beginning to be investigated — they are likely to unlock fundamental aspects of how languages evolve.

Language year
In a phylogenetic context, each of the branches of a phylogeny represents some amount of evolution that occurs independently of the evolution in other branches. If the times in years that these branches represent are added together, the result records the total number of years of evolution that the tree represents; that is, the total number of language years.

Gamma correction
An elegant mathematical technique developed for characterizing the evolution of gene sequences that allows the nucleotides at different sites in the gene to evolve or be replaced at varying rates. The same technique can be applied to characterize the differing rates of evolution among lexical items.

Linguistic universals
A set of features of language and relationships among those features that the great comparative linguist Joseph Greenberg proposed would be found in all or nearly all languages, or which would at least show statistical evidence for being linked.

Word order
The typical order of subjects, verbs and objects in a sentence.

Social effects and bursts of linguistic change

Are there external forces that affect linguistic change independently of the ways we use language in everyday speech? Here I briefly discuss one way in which languages, by acting as markers of social identity, may influence the rate of linguistic evolution.

Languages are not evenly distributed geographically. Cultural groups are more densely packed in coastal than inland regions[49,50]. Similarly, the density or number of different indigenous languages spoken in a given area of North America before European contact sharply increases in the more southerly regions of that continent, and is startling in its similarity to a plot of the density of different biological species from the same area[3,51] (FIG. 4). Human cultural–linguistic groups seem to partition the landscape in a manner similar to species, and perhaps for similar reasons: where in the southerly regions the landscape is richer and more ecologically diverse, a greater variety of species seems able to coexist. The puzzle is that humans are all the same species, and so their higher densities in tropical regions may suggest a tendency for cultural groups to fission whenever the environment will support it[3,51].

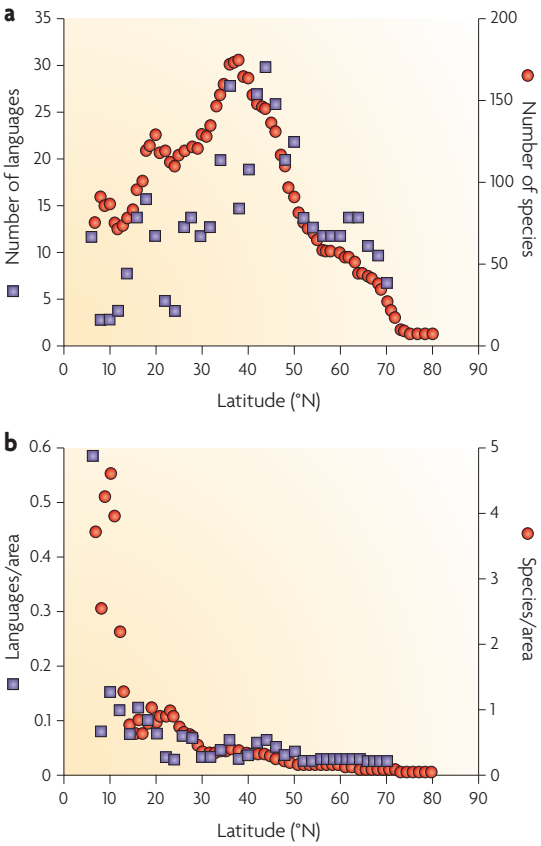


Figure 4 | **Relationships between language and species distribution.** North American human language–cultural groups (before European contact) and mammal species are distributed similarly across degrees of latitude. **a** | Numbers of languages and numbers of mammal species at each degree of north latitude in North America. The trends reflect the shape of the continent, being narrow in the south regions and growing wider at higher latitudes. Both trends peak at approximately 40°N, where North America is ~3,000 miles wide. **b** | Densities of languages and mammal species, calculated as the number of each found at the specific latitude divided by the area of the continent for a 1° latitudinal slice at each latitude. Figure and data is reproduced, with permission, from *Nature* REF. 3 © (2004) Macmillan Publishers Ltd. All rights reserved.

Gene flow is often reduced across linguistic boundaries[52], and anthropologists speculate that language may be used to advertise affinity to particular social groups[53,54]. The eighteenth century American educator Noah Webster put this view trenchantly at the time of American independence from Britain saying that "as an independent nation, our honor [*sic*] requires us to have a system of our own, in language as well as government"[55,56]. Phylogenetic trees of languages for Austronesian, Bantu and Indo-European languages all suggest that Webster was stating a general phenomenon[56]. Extant languages with a rich history of language splitting events, such as that between the speakers of American and British English, have diverged more from their ancestral languages than extant languages with fewer splitting events in their pasts. A similar pattern is observed for genetic evolution among biological species[57]. Humans seem to adjust languages at crucial times of cultural evolution, such as during the emergence of new and rival groups. Maybe there is some truth to the Babel myth after all.

Language structure

The late eminent American comparative linguist Joseph Greenberg pioneered the study of the structural properties of languages, most famously seeking properties he called linguistic universals that could be found in all or nearly all known languages[58,59]. Languages can be classified according to structural properties of syntax, grammar and other features. Associations among these features reveal the internal structure of language and what combinations are possible. I give only the briefest treatment of this very large area, which is ripe for quantitative approaches[60,61], confining my remarks to showing how language phylogenies are fundamental to understanding how language structures evolve.

One structural feature of language is the word order of its sentences. Of the six possible orderings of subjects (S), verbs (V) and objects (O) in a sentence, two — SVO and SOV — dominate the world's languages, two others — VSO and VOS — account for ~10% of languages, and the remaining two — OSV and OVS — are rare[58,62]. From analyses of the relative frequencies of these differing orders among the world's major language families, it has been suggested that the ancestral human language was SOV[62] (M. Gell-Mann and M. Ruhlen, personal communication). One of Greenberg's best-known universals was his proposal that VSO and SVO languages use prepositional phrases to modify sentence objects, whereas SOV languages tend to use postpositional phrasing. Counts of languages support Greenberg's proposal[63].

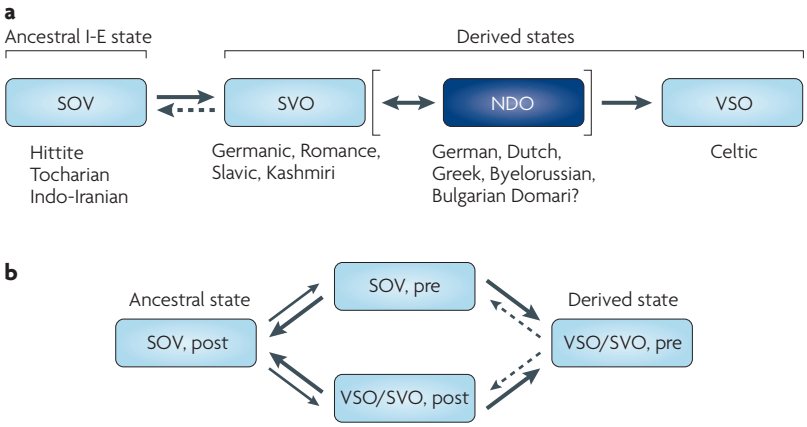


Figure 5 | **Evolution of word order changes. a** | Suggested evolution of word order changes in Indo-European (I-E) languages. Only well-supported transitions are shown. SOV, SVO and VSO refer to orderings of subject (S), verb (V) and object (O) in sentences (see FIG. 1), NDO is for languages categorized as having no dominant word order. The brackets indicate that the NDO category is questioned by some linguists, and the evolutionary relationship excluding it is represented by the light blue boxes. Statistical modelling[9,67] reconstructs ancestral Indo-European word order as SOV. SVO later evolves from SOV (dashed arrow indicates transitions back to SOV that occur in Indo-Iranian languages), and SVO gives rise to VSO. SVO and (possibly SOV) may switch to NDO in case-marked languages such as German, Latin or Greek. This broad result is predicted in REF. 62, although NDO transitions need further study. There are many possible transitions between SVO and NDO within the Indo-European tree. Uncertainty about the true Indo-European phylogeny and the models of evolution is taken into account by integrating the estimates of the model's parameters over a Bayesian sample of Indo-European trees. **b** | A diagram showing correlated evolution of word order — SOV versus VSO or SVO (VSO/SVO) including NDO — and pre versus postpositional phrasing ('pre' and 'post') to modify sentence objects (log Bayes factor test of association ~12, which indicates very strong support)[9,67]. This model reconstructs the ancestral state as SOV, post. Solid arrows indicate statistically supported evolutionary transitions, dashed arrows are not supported. Thickness conveys relative strength of the effect. Here, these arrows indicate that the derived state of VSO/SVO and prepositional phrasing evolves from the ancestral state either by adopting a different word order first, becoming SOV, pre, or by adopting a different positional phrasing first, becoming VSO/SVO, post. Examples of each evolutionary process occur in Indo-European languages (FIG. 1). These intermediate states violate Greenberg's[58] predictions but are short lived, as indicated by the thick arrows pointing to the ancestral and derived states. The derived state seems to be stable in Indo-European languages. Data is taken from the World Atlas of Linguistic Structures[71]. The evolutionary relationships shown here might change owing to a lack of consensus on the classification of some languages on these two traits (see also FIG. 1).

Do phrases such as 'I built a house' and 'I a house built' (SVO and SOV, respectively) owe their dominance to inherent properties of those systems or are they accidental winners, having ridden on the backs of people who came to dominate the globe for some other reason. Why do English speakers use the prepositional phrase in 'I built a house for you' rather than the postpositional 'I built a house you for'? Is the pairing of word order and pre versus postpositioning a chance association or does it represent co-evolution of these two structural traits? If it represents co-evolution, which feature of language changes first or can either change? These kinds of questions have direct parallels in cross-cultural studies[8,64] and in comparative biology[65], and must be studied using phylogenies. A co-evolutionary explanation would be favoured if the relationship arose independently many times in unrelated languages.

Pre versus postpositioning
Whether a language places the phrase that modifies a sentence object before (preposition) or after (postposition) that object in the sentence.

Phylogenetic statistical approaches. I illustrate a phylogenetic comparative approach to the word order and positional phrasing predictions with data for the Indo-European languages (FIG. 1). Germanic, Romance and Slavic languages are mostly SVO, many Indo-Iranian languages and the ancient Tocharian and Hittite languages are SOV, and the Celtic languages are VSO. German, Greek, Bulgarian and the Indo-Iranian language Domari are among a handful of Indo-European languages sometimes regarded as having no dominant word order (NDO). The historical evolution of these four states can be studied for the Indo-European tree using the finite-state Markov model in Q, and implemented using a technique called reversible jump MCMC[66] that allows one to explore the space of possible models[9,67].

The approach outlined above reconstructs the ancestral or proto-Indo-European language as SOV (FIG. 5a). Early in Indo-European language evolution SOV gave way to SVO (or NDO, which then later resolved to SVO) before reverting to SOV in the Indo-Iranian languages. The Celtic VSO evolved from SVO in the common ancestor to the Celtic, Romance and Germanic languages. There is an intriguing hint that languages can rapidly switch between a fixed or NDO word order, perhaps using case marking in place of order. The same methodology can then be used to test Greenberg's proposal for a relationship between word order and pre versus postpositioning. Languages are scored as VSO or SVO (VSO/SVO) or as SOV, and also as prepositional or postpositional — yielding four possible combinations of paired states. The analysis finds the correlation Greenberg predicted (FIG. 5b). The analysis also shows that languages can evolve from one of Greenberg's preferred states to the other, by changing either of the individual traits first, but suggests these 'intermediate' states are unstable. These results could only have emerged from a phylogenetic analysis and should be replicated in additional families. It should be straightforward to repeat this exercise for Bantu and Austronesian languages (R. Gray and M. Dunn, personal communication).

Once many of these structural features of language have been analysed for their correlations across languages it will be possible to construct network diagrams like those used to display protein or metabolic interaction networks, in which links between pairs of features correspond to significant evolutionary correlations across species[68]. These have the potential to reveal the structural hubs and satellites of language — that is, features that are highly connected and those that are not — and the traits that are most likely to be gained or lost over time.

Discussion

Phylogenetic and statistical methods have only begun to be used to study language evolution, but they have already returned important insights into its evolution. Much remains to be done. Models for phylogenetic inference could be improved by allowing words to alter their rates of change throughout the tree, and it should also be possible to automate cognacy judgements in a

manner analogous to automated gene sequence alignment. Greenberg's[58] proposals for linguistic universals describe dozens of associations among pairs of structural features, and these are suitable candidates for phylogenetic testing. At the level of lexicon, rather than structure, little is known about whether some words tend to change together, and whether these potential co-evolutionary linkages affect how these words evolve.

Swadesh's original vocabulary list comprises words that are used at a higher than average frequency, and so could be expanded to include less frequently used and consequently more rapidly evolving words. In a similar vein, although historically much emphasis has been placed on how words come to be replaced by new non-cognate words, there is room to study how words come to acquire new meanings, or how words gradually change their sounds while retaining their meanings and while remaining cognate. An important aspect of this process, in turn, relates the ways that languages are learned and transmitted within communities to the rates at which existing words (or other features of language) change or new words emerge and replace old ones[69]. This is analogous to attempts in evolutionary biology to describe how within-population processes give rise to differences between populations or species[70].

Language trees provide the logical backbone on which to test these and many other anthropological questions. There is no doing comparative linguistics or comparative anthropology without them, and new linguistic or anthropological research programmes should routinely make their construction a priority. Already projects such as the World Atlas of Linguistic Structures[71] or the Austronesian Basic Vocabulary Database[72] document hundreds of thousands of observations on language and these databases need to be developed in a similar way to GenBank and other genetic databases.

For geneticists, or for anyone interested in molecular evolution, the parallels between linguistic and genetic evolution should be striking, and all the more so because language is a cultural rather than a physical replicator, without built-in error correction mechanisms and potentially subject to far greater effects of borrowing and other influences that could corrupt its signal. Like genomes, the languages we observe today are the survivors of a long process of being tried out and tested by their speakers. Like genomes, we can speculate that we have retained those languages that adapted best to our minds[73], and this may be the most obvious reason why we find them easy to learn and use.

For a language system to survive it must adapt as a coherent whole, and this governs the likely combinations of language elements, be they words, grammar, syntax or morphology. These functional restraints on languages coupled with the observation of high fidelity in the transmission of linguistic elements means that there are far fewer languages and less linguistic diversity than might otherwise be possible. Are some of these languages somehow better than others or somehow better suited to their own speakers, or do existing languages represent alternative and equally functional outcomes of the linguistic evolutionary process? It is the many differences between what we see and what is possible that reveal the ways that languages adapt. How they do it and why is an area that holds great promise for furthering our understanding of this uniquely human trait as the complex and adaptively evolving system that it is[74].

1. Gordon, R. G. *Ethnologue: Languages of the World* 15th edn (SIL International, Dallas, 2005).
2. Pagel, M. in *The Evolutionary Emergence of Language* (eds Knight, C., Studdert-Kennedy, M. & Hurford, J.) 391–416 (Cambridge Univ. Press, Cambridge 2000). **An overview of linguistic diversity and how it can be studied phylogenetically and statistically.**
3. Pagel, M. & Mace, R. The cultural wealth of nations. *Nature* **428**, 275–278 (2004).
4. Darwin, C. *The Descent of Man* (Murray, London, 1871).
5. Swadesh, M. Lexico-statistic dating of prehistoric ethnic contacts. *Proc. Am. Phil. Soc.* **96**, 453–463 (1952).
6. Embleton, Sheila M. *Statistics in Historical Linguistics. Quantitative Linguistics* Vol. 30 (Bochum, Brockmeyer, 1986).
7. Pagel, M. & A. Meade. in *Phylogenetic Methods and the Prehistory of Languages* (eds Forster, P. & Renfrew, C.) 173–182 (McDonald Institute for Archaeological Research, Cambridge, 2006).
8. Mace, R. & Pagel, M. The comparative method in anthropology. *Curr. Anthropol.* **35**, 549–564 (1994). **This paper formally introduced use of phylogenetic trees into comparative anthropology.**
9. Pagel M, Meade A. in *The Evolution of Cultural Diversity: a Phylogenetic Approach* (eds Mace R., Holden C. J. & Shennan S.) 235–256 (UCL Press, London, 2005).
10. Kruskal, J., Dyen, I. & Black, P. in *Mathematics in the Archeological and Historical Sciences* (eds Hodson, F. R., Kendall, D. G. & Tautu, P.) 361–380 (Edinburgh Univ. Press, Edinburgh, 1971).
11. Sankoff, D. in *Current Trends in Linguistics 11: Diachronic, Areal and Typological Linguistics* (ed. Sebeok, T. A.) 93–112 (Mouton, The Hague, 1973).
12. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
13. Nicholls, G. K. & Gray, R. D. in *Phylogenetic Methods and the Prehistory of Languages* (eds Forster, P. & Renfrew, C.) 161–171 (McDonald Institute for Archaeological Research, Cambridge, 2006).
14. Warnow, T., Evans, S. N., Ringe, D. & Nakhleh, L. in *Phylogenetic Methods and the Prehistory of Languages* (eds Forster, P. & Renfrew, C.) 75–87 (McDonald Institute for Archaeological Research, Cambridge, 2006).
15. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).
16. Edwards, A. W. E. *Likelihood* (Cambridge Univ. Press, Cambridge, 1972).
17. Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. in *Markov Chain Monte Carlo in Practice* (eds Gilks, W. R., Richardson, S. & Spiegelhalter, D. J.) 1–19 (Chapman and Hall, 1996).
18. Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314 (2001).
19. Pagel, M. in *Time-Depth in Historical Linguistics* (eds Renfrew, C., MacMahon, A. & Trask L.) 189–207 (The McDonald Institute of Archaeology, Cambridge, 2000).
20. Gray, R. & Jordan, F. Language trees support the express-train sequence of Austronesian expansion. *Nature* **405**, 1052–1055 (2000).
21. Holden, C. J. Bantu language trees reflect the spread of farming across Sub-Saharan Africa: a maximum-parsimony analysis. *Proc. R. Soc. Lond., B* **269**, 793–799 (2002). **This paper describes an early application of phylogenetic methods in linguistics.**
22. Holden, C. J., Meade, A. & Pagel, M. in *The Evolution of Cultural Diversity: a Phylogenetic Approach* (eds Mace R., Holden C. J. & Shennan S.) 53–65 (UCL Press, London, 2005).
23. Gray, R. D. & Atkinson, Q. D. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439 (2003). **This study used language phylogeny to test a historical hypothesis for the timing of the origin of Indo-European languages.**
24. Dunn, M., Terrill, A., Reesink, G., Foley, R. A. & Levinson, S. C. Structural phylogenetics and the reconstruction of ancient language history. *Science* **309**, 2072–2075 (2005).
25. Gray, R. D., Drummond, A. J. & Greenhill, S. J. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483 (2009). **This paper describes the use of a language phylogeny to test a historical hypothesis for the timing of the origin of Austronesian languages.**
26. Pagel, M., Atkinson, Q. D. & Meade, A. Frequency of word use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717–719 (2007). **A statistical phylogenetic study that proposed a general explanation for variation in rates of lexical replacement.**
27. Sanderson, M. J. & Donoghue, M. J. Patterns of variation in levels of homoplasy. *Evolution* **43**, 1781–1795 (1989).
28. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).
29. Bryant, D., Filimon, F. & Gray, R. D. in *The Evolution of Cultural Diversity: Phylogenetic Approaches* (eds Mace, R., Holden, C. & Shennan, S.) 69–85 (UCL Press, London, 2005).
30. Renfrew, C. *Archaeology and Language: the Puzzle of Indo-European Origins* (Cape, London, 1987). **Classic text on the origin of the Indo-European language family.**
31. Greenhill, S., Currie, T. & Gray. R. Does horizontal transmission invalidate cultural phylogenies? *Proc. R. Soc. Lond., B* 18 Mar 2009 (doi:rspb.2008.1944).

32. Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. & Mountain, J. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc. Natl Acad. Sci. USA* **85**, 6002–6006 (1988).
This paper is a widely cited early attempt to link genetic and linguistic diversity.
33. Lansing, J. S. *et al.* Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc. Natl Acad. Sci. USA* **104**,16022–16026 (2007).
34. Hunley, K. *et al.* Genetic and linguistic coevolution in Northern Island Melanesia. *PLoS Genet*. **4**, 1–14 (2008).
35. Dediu, D. & Ladd, D. R. Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and microcephalin. *Proc. Natl Acad. Sci. USA* **104**, 10944–10949 (2007).
36. Holden, C. J. & Mace, R. Spread of cattle led to the loss of matriliny in Africa: a co-evolutionary analysis. *Proc. R. Soc. Lond., B* **270**, 2425–2433 (2003).
A good example of the use of language trees to study cultural evolution.
37. Fortunato, L., Holden, C. J. & Mace, R. From bridewealth to dowry? A Bayesian estimation of ancestral states of marriage transfers in Indo-European groups. *Human Nature* **17**, 355–376 (2006).
38. Mace, R. & Jordan, F. in *The Evolution of Cultural Diversity: a Phylogenetic Approach* (eds Mace, R., Holden, C. & Shennan, S.) 207–216 (UCL Press, London, 2005).
39. Burger, J., Kirchner, M., Bramanti, B., Haak, W. & Thomas, M. G. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proc. Natl Acad. Sci. USA* **104**, 3736–3741 (2007).
40. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994).
41. Zipf, G. K. *Human Behaviour and the Principle of Least Effort* (Addison-Wesley, Reading, Massachusetts, 1949).
42. Leech, G., Rayson, P. & Wilson, A. *Word Frequencies in Written and Spoken English: Based on the British National Corpus* (Longman, London, 2001).
43. Zipf, G. K. Prehistoric 'cultural strata' in the evolution of Germanic: the case of Gothic. *Mod. Lang. Notes* **62**, 522–530 (1947).
44. Francis, W. N., Kuçera, H. & Mackie, A. W. *Frequency Analysis of English Usage: Lexicon and Grammar* (Houghton Mifflin, Boston, 1982).
45. Lieberman, E., Michel, J.-B., Jackson, J., Tang, T. & Nowak, M. A. Quantifying the evolutionary dynamics of language. *Nature* **449**, 713–716 (2007).
46. Starostin, S. A. in *Works on Linguistics* (ed. Starostin, S. A.) 827–839 (Languages of the Slavic Culture, Moscow, 2007).
47. Ellis, N. C. Frequency effects in language processing: a review with implications for theories of implicit and explicit language acquisition. *Stud. Second Lang. Acquisit.* **24**, 143–188 (2002).
48. Huettig, F., Quinlan, P. T., McDonald, S. A. & Altmann, G. T. M. Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychol.* **121**, 65–80 (2006).
49. Birdsell, J. B. Some environmental and cultural factors influencing the structuring of Australian Aboriginal populations. *Am. Nat.* **87**, 171–207 (1953).
50. Nichols, J. *Linguistic Diversity in Space and Time* (Univ. of Chicago Press, Chicago, 1992).
51. Mace, R. & Pagel, M. A latitudinal gradient in the density of human languages in North America. *Proc. Roy. Soc. Lond., B* **261**, 117–121 (1995).
52. Barbujani, G. & Sokal, R. R. Zones of sharp genetic change in Europe are also linguistic boundaries. *Proc. Natl Acad. Sci. USA* **87**, 1816–1819 (1990).
53. Labov, W. *Principles of Linguistic Change: Social Factors* (Blackwell, Oxford, 2001).
54. Milroy, J. & Milroy, L. Linguistic change, social network and speaker innovation. *J. Linguist* **21**, 229–284 (1985).
55. Webster, N. *Dissertations on the English Language* (Isaiah Thomas, Boston, 1789).
56. Atkinson, Q., Meade, A., Venditti, C., Greenhill, S. & Pagel, M. Languages evolve in punctuational bursts. *Science* **319**, 588 (2008).
57. Pagel, M., Venditti, C. & Meade, A. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* **314**, 119–121 (2006).
58. Greenberg, J. H. (ed.) *Universals of Languages* (MIT Press, Cambridge, Massachusetts, 1963).
59. Kirby, S. *Function, Selection, and Innateness: the Emergence of Language Universals* (Oxford Univ. Press, Oxford, 1999).
60. Cysouw, M. in *Quantitative Linguistics: An International Handbook* (eds Altmann, G., Köhler, R. & Piotrowski, R.) 554–578 (Mouton de Gruyter, Berlin, 2005).
61. Croft, W. *Explaining Language Change: an Evolutionary Approach* (Longman, Harlow, 2000).
This text provides a good overview of evolutionary thinking about language.
62. Newmeyer, F. J. in *The Evolutionary Emergence of Language* (eds Knight, C., Studdert-Kennedy, M. & Hurford, J.) 372–388 (Cambridge Univ. Press, Cambridge, 2000).
63. Haspelmath, M. & Siegmund, S. Simulating the replication of some of Greenberg's word order predictions. *Linguistic Typology* **10**, 74–82 (2006).
64. Mace, R. *The Evolution of Cultural Diversity: a Phylogenetic Approach* (eds Mace R., Holden C. J. & Shennan S.) 1–10 (UCL Press, London, 2005).
65. Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, Oxford, 1991).
66. Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732 (1995).
67. Pagel, M. & Meade, A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am. Nat.* **167**, 808–825 (2006).
68. Pagel, M., Meade, A. & Scott, D. Assembly rules for protein interaction networks. *BMC Evol. Biol.* **7** (Suppl. 1), S16 (2007).
69. Niyogi, P. *The Computational Nature of Language Learning and Evolution* (MIT Press, Cambridge, Massachusetts, 2006).
70. Mangel, M. & Clark, C. W. *Dynamic Modeling in Behavioral Ecology* (Princeton Univ. Press, Princeton, New Jersey, 1988).
71. Haspelmath, M., Dryer M. S., Gil, D. & Comrie, B. (eds) The World Atlas of Linguistic Structures *Max Planck Digital Library* [online], www.wals.info (2008).
72. Greenhill, S. J., Blust, R. & Gray, R. D. The Austronesian Basic Vocabulary Database: from bioinformatics to lexomics. *Evol. Bioinform. Online* **4**, 271–283 (2008).
73. Christiansen, M. H. & Chater, N. Language as shaped by the brain. *Behav. Brain Sci.* **31**, 489–558 (2008).
74. Gell-Mann, M. *The Quark and the Jaguar: Adventures in the Simple and Complex* (W.H. Freeman New York, 1994).

Acknowledgements
I thank C. Venditti, A. Calude, I. Peiros, A. Meade, Q. Atkinson, M. Ruhlen, M. Cysouw and M. Haspelmath for help, comments and suggestions. Grants to M.P. from the Leverhulme Trust and the Natural Environment Research Council supported this work.

FURTHER INFORMATION
Mark Pagel's homepage: www.evolution.reading.ac.uk
Austronesian Basic Vocabulary Database: http://language.psy.auckland.ac.nz/austronesian
SplitsTree: http://www.splitstree.org
Swadesh list: http://en.wiktionary.org/wiki/Appendix:Swadesh_list
World Atlas of Linguistic Structures: http://www.wals.info

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

OPINION

The future of evo–devo: model systems and evolutionary theory

Ralf J. Sommer

Darwin200

Abstract | There has been a recent trend in evolutionary developmental biology (evo–devo) towards using increasing numbers of model species. I argue that, to understand phenotypic change and novelty, researchers who investigate evo–devo in animals should choose a limited number of model organisms in which to develop a sophisticated methodological tool kit for functional investigations. Furthermore, a synthesis of evo–devo with population genetics and evolutionary ecology is needed to meet future challenges.

Evolutionary developmental biology (evo–devo) investigates the evolution of developmental processes, aiming for a mechanistic understanding of phenotypic change[1,2]. Building on the analysis of model organisms in developmental biology, evo–devo has seen a fruitful expansion in the last two decades and has successfully integrated various comparative research strategies[3–7]. The investigation of several concepts, including modularity, redundancy, developmental constraints, evolutionary novelties and phenotypic plasticity, forms a framework for evo–devo. However, evo–devo suffers from a sometimes misguided selection of model organisms, often with a limited availability of technical tools[8,9] and, most importantly, poor integration with other areas of evolutionary biology[10]. In this Opinion article, I argue that the future success of evo–devo in animals depends on two major technical and conceptual aspects: first, evo–devo has to concentrate on a few well-selected model organisms to allow the development of a sophisticated analytical tool kit for functional investigations; and second, evo–devo has to enhance its connections to other areas of evolutionary biology. Specifically, synthesis with population genetics can reveal how phenotypic evolution is initiated at the microevolutionary level, and synthesis with evolutionary ecology can add an ecological perspective to these evolutionary processes.

Limiting the number of models

The principle that focusing on a few organisms can be effective is demonstrated by the fact that the initial rise of developmental genetics was largely based on two invertebrate model systems, *Drosophila melanogaster* and *Caenorhabditis elegans*. The mechanistic understanding of development in these model organisms was also one of the important starting points for 'modern' evo–devo. Initial evo–devo work, which focused mainly on the cloning and expression pattern analysis of genes homologous to *D. melanogaster* developmental control genes[11], pointed towards an unexpected conservation of developmental genes. This work was, however, largely descriptive.

In some new evo–devo model organisms, such as the insects *Tribolium castaneum*[12] and *Nasonia vitripennis*[13] and the nematode *Pristionchus pacificus*[14], researchers started to build a more sophisticated tool kit to investigate the mechanisms of evolutionary change in developmental processes (TABLE 1). However, the development of these methods — including forward genetics to allow gene knockout or knockdown, and transgenesis to allow experimental manipulation — proved challenging. Method development depends mostly on empirical optimizations, which are largely species specific, so protocols cannot be transferred from one organism to another. Large research communities can overcome these challenges, but in evo–devo, with its relatively small research communities, method development is much harder.

One reverse genetics technology that has been used extensively in evo–devo in recent years to overcome technical limitations is RNAi. Although RNAi is becoming increasingly accessible, it is not easily transferable to every organism, and even in *C. elegans*, in which it was originally described, it does not work in all cells and tissues. By definition, RNAi is biased towards candidate genes identified in model organisms and is a transient method. Both of these features influence the type of questions that can be addressed by RNAi and the accuracy of the conclusions. Two of the strongest applications of RNAi in model organisms are genome-wide RNAi screens and the generation of double mutants by performing RNAi in a mutant background, but these are not yet realistic in evo–devo systems.

Owing to the technical limitations discussed above, evo–devo has largely followed the classical strategy of comparative morphology by analysing more organisms to provide unbiased phylogenetic sampling[8]. Particularly in the animal kingdom, with its deep branches and vast diversity of form and species, one can always look at new taxa and investigate their molecular inventory. If species are selected from a phylogenetic perspective, such studies can increase our understanding of the molecular evolution of developmental control genes; this research strategy provides important insight into evolutionary patterns. However, this strategy also has a serious trade-off: because of the limited resources and small number of researchers, large phylogenetic sampling will often result in few studies per organism and a superficial understanding of each system. In addition, it has been argued that analysing species because of their phylogenetic position rather than their conceptual value could leave the discovery of law-like generalities to chance[8].

I argue that the analysis of the central concepts of evo–devo can best be achieved by the selection of a limited number of model organisms and the development of sophisticated made-to-measure tool kits: this principle has been highly successful

in developmental genetics and its application in evo–devo seems equally promising. One reason for this optimism is that most conceptual themes in evo–devo arose from developmental genetics. Phenomena such as redundancy might be observed as widespread[15], and yet their significance in developmental processes and their contribution to evolution cannot be identified by the analysis of a single species; their role in evo–devo requires comparative studies between related species of the same taxa. Classical model organisms are a valuable starting point for such studies; by comparing *D. melanogaster* with other insects, or *C. elegans* with other nematodes, one can use the mechanistic insights provided by classical models to investigate evo–devo themes.

Considerations for comparative studies. To ensure that the comparative studies introduced above will be valuable for elucidating changes in development and the influence of these changes on evolution, two factors must be considered.

First, the species that are compared should be related in such a way that distinct, but still homologous, developmental patterns can be studied. Changes in developmental processes and mechanisms can then be identified as the cause of morphological diversity and novelty. By contrast, if organisms are completely unrelated, comparisons often result in a descriptive list of their molecular inventories, thus not going much beyond the information that genome projects provide. The intellectual merit of comparative studies in unrelated organisms often rests with providing evidence for the co-option of conserved transcription factor modules and signalling networks in independent evolutionary lineages[3].

Second, comparative studies should concentrate on mechanisms rather than, for example, gene conservation and gene expression. For transcription factors and cell–cell signalling molecules this is of particular importance because studies in model organisms constantly reveal that protein function is context dependent. One well-known example is Wnt signalling, which has both β-catenin-dependent and β-catenin-independent functions[16]. Therefore, studies that rest on the analysis of expression patterns of shared components of such pathways can easily be misleading. Only functional investigations and comparisons between a developmental model system and an evo–devo 'model system' can reveal how mechanisms change during evolution to create phenotypic diversity or novelty (discussed further in the following section). Furthermore, such studies can indicate the importance of evo–devo concepts for studying the evolution of developmental processes.

Taking these two considerations together, I argue that restricting the number of model organisms would help the field of evo–devo in its search for a theory. Developing a theory is of utmost importance for any discipline. This is clearly shown in evolutionary genetics, which builds on the framework of population genetics. In the context of developing a theory, it has been argued that signalling pathways and transcription factor modules could serve as a theoretical framework for elucidating developmental changes in evolution[1]. As functional investigations of development require the generation of sophisticated methods (TABLE 1), the limitation of the number of evo–devo model organisms is a logical consequence, and is a prerequisite for the long-term success of evo–devo.

Table 1 | **Several central criteria for evo–devo model species**

Methodology or approach	Scientific aim
Forward genetics	Unbiased identification of developmental mechanisms
Reverse genetics (RNAi, small interfering RNA morpholinos)	Functional studies from gene predictions
Genome projects	Evolution of genome architecture
Transgenesis	Experimental manipulation of gene function
Phylogenetic reconstructions	Directionality of evolutionary changes
Microevolutionary comparison of different isolates of the same species	Natural variation in developmental control genes
Genome-wide association studies	
Recombinant inbred line analysis	
Evo–devo in relation to ecology	Environmental influence on developmental control genes

The need for sophisticated tools

The importance of in-depth functional studies for achieving the aims of evo–devo, and by consequence limiting the number of organisms used, can be illustrated by case studies from nematodes and insects. These two cases indicate how the use of forward and reverse genetics can provide mechanistic insights into the evolution of development.

The nematode vulva. The nematode *P. pacificus* has been developed as a model system in evo–devo for comparison with *C. elegans*[14] (TABLE 2). *P. pacificus* shares many technical features with *C. elegans*, such as a 3–4 day life cycle, simple culture and self-fertilization as mode of reproduction. Its hermaphroditic mode of reproduction makes forward genetics feasible, the *P. pacificus* genome has recently been sequenced[17] and a DNA-mediated transformation method allows genetic manipulation[18].

Although *P. pacificus* shares technical features with *C. elegans*, many aspects of its development are strikingly different. Particular attention has been given to the development of the vulva, the nematode egg-laying structure. *C. elegans* vulva formation is one of the best studied developmental processes in animals[19], providing a platform for mechanistic studies in evo–devo[20]. Two hallmarks of *C. elegans* vulva formation are the generation of a vulva equivalence group and the induction of the vulva by the gonadal anchor cell. *P. pacificus* reveals striking differences with respect to both aspects of vulva development (BOX 1). Vulva induction requires different signalling pathways, and the reduction of the size of the vulva equivalence group in *P. pacificus* involves a transcriptional module that is absent from *C. elegans*, although it is otherwise conserved among metazoans[21,22]. Recent genetic studies in just these two species have allowed the molecular and mechanistic basis for these evolutionary changes in pattern formation and induction to be identified.

Insect dorso–ventral patterning. The red flour beetle *T. castaneum* is one of a few insects that have been developed as a model organism for mechanistic investigation in evo–devo[12]. This beetle can be easily cultured, has a short life cycle and is amenable to forward genetics analysis. The genome of *T. castaneum* has been sequenced, and an RNAi technique has been developed[23]. RNAi has proved particularly powerful and efficient in this organism, providing a tool for the large-scale elucidation of gene function[23].

Table 2 | **A selection of emerging evo–devo model systems with genetic tools in the vicinity of classical model organisms**

Classical model organism	Evo–devo model	Evo–devo themes	Refs
Drosophila melanogaster (arthropod)	*Tribolium castaneum*	Segmentation, appendix formation	12,26,28
	Nasonia vitripennis	Segmentation	13
	Daphnia pulex	Response to environmental variation	58
Caenorhabditis elegans (nematode)	*Caenorhabditis briggsae*	Sex determination, convergent evolution	63
	Pristionchus pacificus	Pattern formation, induction	14,20–22
Zebrafish	*Astyanax mexicanus*	Developmental and morphological response to environmental variation	54
	Sticklebacks	Developmental and morphological response to environmental variation	64
Hydra (cnidarian)	*Nematostella vectensis*	Evolution of body plan, ecological evo–devo	53
Arabidopsis thaliana (higher plant)	*Antirrhinum* (snapdragon)	Flowering	65

In *T. castaneum* embryogenesis, posterior segments develop successively and two extra-embryonic membranes cover the egg. By contrast, in *D. melanogaster* all segments form simultaneously and extra-embryonic membranes are fused to the amnioserosa[24] (BOX 2). RNAi studies of known dorso–ventral patterning genes have shown striking differences between *T. castaneum* and *D. melanogaster* in the function of individual genes and of genetic networks (BOX 2). In particular, gene duplications and subfunctionalization are crucial for extra-embryonic membrane formation and dorso–ventral patterning[25–28].

Structure–function dualism. The genetic experiments in *T. castaneum* and *P. pacificus* described above, and others like them[13], indicate that the exact mechanisms by which developmental control genes work can change rapidly during the course of evolution. For example, homologous genes can assume different functions in different species so that elimination of these genes results in different phenotypes[22,28]. Also, some developmental control genes are present in one organism but not in another[21], and genes that are duplicated during the course of evolution can undergo subfunctionalization in individual evolutionary lineages[26]. Therefore, comparative studies between phylogenetically related species can reveal how induction, pattern formation and segmentation evolve and contribute to the generation of evolutionary novelty. The examples of the nematode vulva and insect embryogenesis also show how homologous characteristics — characteristics that are shared because of a common ancestry — can be uncoupled at different levels: although the cells that form the nematode vulva and the organ itself are homologous, the genes regulating the underlying molecular processes are not necessarily homologous[29,30]. This allows deBeer's proposal, that homologous structures can be built by different genes[31,32], to be tested at a molecular level[29].

Genetic experiments give insights into how the function of a homologous gene can change during evolution. Isolation of a known gene in a new species or expression studies do not allow us to identify function and potential functional alterations during the course of evolution; this requires specific tools, such as forward and reverse genetics. The genes *zerknüllt* and *Toll*, for example, are both expressed during dorso–ventral patterning in *D. melanogaster* and *T. castaneum*, but their differing functions were only revealed by genetic manipulation experiments[28]. Although this conclusion is worthy in itself, it also provides an additional argument for the selection of a limited number of evo–devo model systems and the development of functional tools in these species.

The future of evo–devo models

The *T. castaneum* and *P. pacificus* case studies show how the use of new models can give novel insights into evo–devo. Therefore, going beyond the classical model systems can be of value. *T. castaneum* and *P. pacificus* are two evo–devo models that have a sophisticated tool kit — but how many species should there be? The number of species worked on in evo–devo is constantly changing, with species being added and being removed: a recent monograph provides a detailed list of 'emerging model organisms'[33]. In some cases these organisms have received special attention because they offer the analysis of themes that have not received particular attention in classical models, such as regeneration, which can be efficiently studied in planarians and ascidians[34,35]. Similarly, some themes in evo–devo can only be studied in a particular species, or group of species. Under such circumstances, alternative models should also be used. But in the more general evo–devo context most concepts are based on widespread phenomena. For example, redundancy, phenotypic plasticity and developmental constraints are found in most organisms, and their role in evo–devo can therefore be studied in several systems if the appropriate tools are available. Thus, broad phylogenetic sampling is not a necessary prerequisite for studying the mechanisms behind important evo–devo concepts. With the two criteria identified above, namely the technical considerations and the need to compare the phylogenetic relationship of the evo–devo and the classical model organism, a realistic starting number of evo–devo model species should not be much higher than a dozen because the long-term value of a species depends on its conceptual merit (TABLE 2).

Implications for the funding of evo–devo research. Another sensitive issue for evo–devo studies is research funding. Relative to comparative morphology, one of its intellectual forerunners, research in evo–devo requires substantially more investment. An emerging consequence is, therefore, the problem of securing funding for evo–devo in the modern life sciences, which largely aim to address applied research questions. This difficulty arises when evo–devo studies are compared with mechanistically driven applied research projects. A second significant problem is obtaining the initial funding for technology development in new model organisms. I argue that evo–devo projects that focus on functional studies are the most likely to be successful in competition with other research fields. In addition, allocation of research funds for technology development, as has been seen for comparative

genomics, could further help evo–devo to succeed in a world of limited funds. Specific funding allocation could, for example, target the exploration of new species to extend the number of model systems over a longer time period. Together, seeking funding for functional studies and technology development might even result in a gain of funding for evo–devo overall.

Integration with evolutionary theory

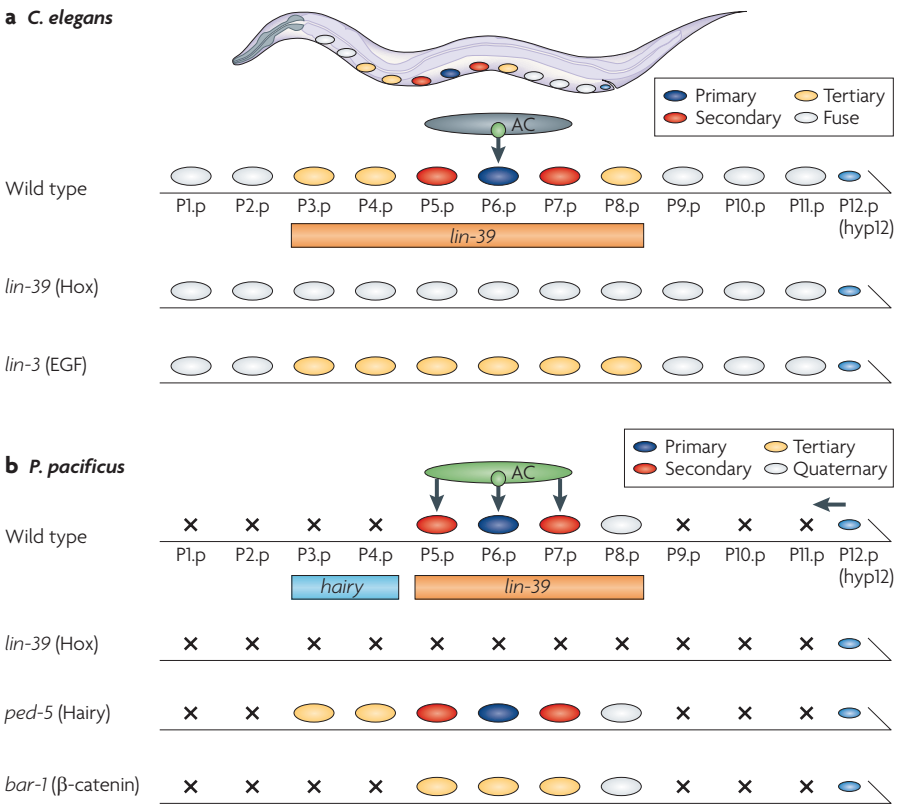
In addition to practical considerations regarding the number of model organisms and the development of appropriate analytical tools, the interaction of evo–devo with other research areas needs to be re-considered to ensure future successes in the field. Specifically, I argue that more integration with evolutionary biology would be mutually beneficial (TABLE 1). The relationship between development and evolution has changed several times in the past 150 years (discussed in REF. 36). Currently, there is growing consensus that development has to be integrated into evolutionary theory, because the evolution of form and the generation of morphological novelty are of utmost importance in a general philosophical framework of biology. However, working solely within the conceptual framework of evo–devo results in a gene-centred and development-centred perspective that lacks interrelationships with other areas of evolutionary biology. If evo–devo wants to establish itself as a part of evolutionary theory, it has to find a suitable way of incorporating evolutionary thinking and recent advances, such as genomics[10]. Specifically, I argue that a synthesis with population genetics and evolutionary ecology is required.

A synthesis with population genetics. Why are developmental control genes conserved at the sequence level, when their functions can change? This question and the original observations that led to it are important because they help to distinguish, in the evo–devo context, between the contrasting theories of neo-Darwinism and neutral evolution. In neo-Darwinism, positive (that is, directional) selection is thought to be the major mechanism driving the change of allele frequencies and it predicts that genes would not be conserved among species[37,38]. By contrast, Kimura's neutral theory of molecular evolution proposes that the majority of mutations in non-coding areas of the genome are selectively neutral or nearly neutral, whereas most mutations in genes are selectively deleterious[39]. The neutral theory predicts that in coding regions

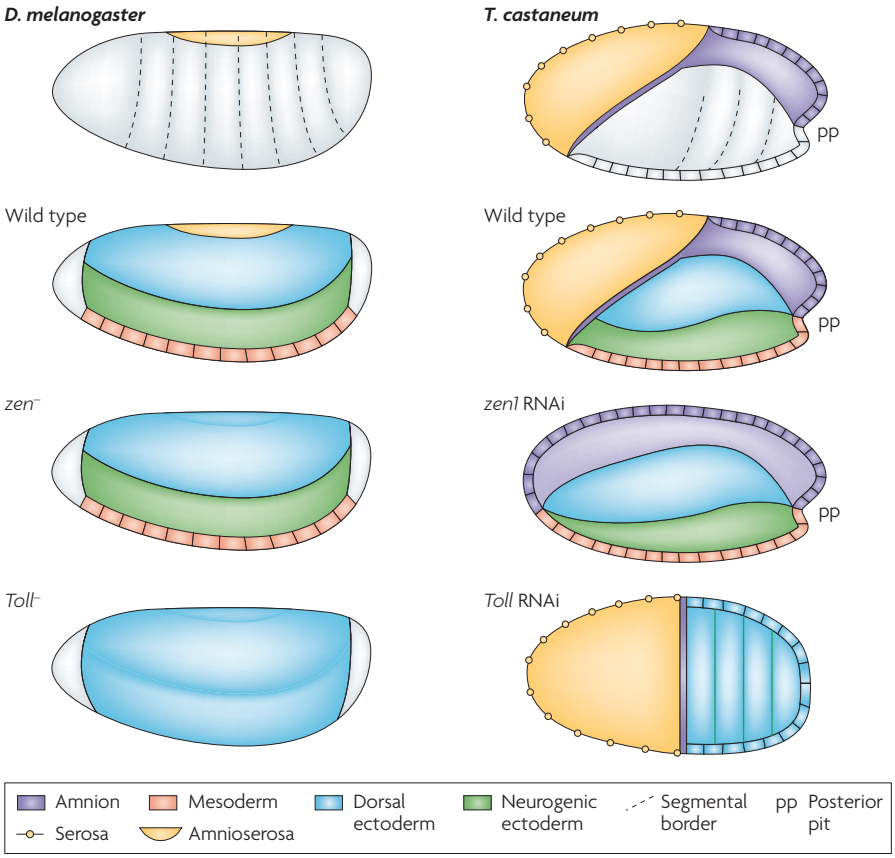
Box 1 | Vulva induction in *C. elegans* and *P. pacificus*

In *Caenorhabditis elegans* the vulva is a derivative of the ventral epidermis, which consists of 12 ectoblasts, named P1.p–P12.p according to their antero–posterior position[19] (see the figure, part **a**). In wild-type animals, the vulva is formed from the progeny of P5.p–P7.p. P6.p has the primary fate and generates eight progeny (represented by a blue oval) and P5.p and P7.p have the secondary fate and form seven progeny each (represented by red ovals). P3.p, P4.p and P8.p have the tertiary fate (represented by yellow ovals). These cells are competent to form vulval tissue, but remain epidermal under wild-type conditions. The remaining ectoblasts (light grey ovals) fuse with the hypodermis and are not competent to form part of the vulva. P12.p is a special cell called hyp12, and forms part of the rectum. The vulva equivalence group, consisting of P3.p–P8.p, is located in the central body region and is specified by the homeobox (Hox) gene *lin-39*. In *C. elegans lin-39* mutants, positional information for the formation of the vulva equivalence group is missing, and P3.p–P8.p fuse with the hypodermis. *C. elegans* vulva induction depends on a signal from the anchor cell (AC, green circle) of the somatic gonad (dark grey oval). Ablation of the AC at birth is sufficient to prevent vulva induction and mutations in the epidermal growth factor (EGF) family member *lin-3* result in a vulvaless phenotype.

As in *C. elegans*, the *Pristionchus pacificus* vulva forms from the ventral epidermis, which is generated by homologous precursor cells, P1.p–P12.p (see the figure, part **b**). In *P. pacificus*, however, P1.p–P4.p and P9.p–P11.p die of programmed cell death and reduce the size of the vulva equivalence group to four cells[20]. In contrast to *C. elegans*, P3.p and P4.p are unable to form part of the vulva in *P. pacificus* because they die early in development. P5.p–P7.p have a secondary–primary–secondary pattern, as in *C. elegans*, and P8.p is a special epidermal cell (light grey oval), which is designated a quaternary cell fate. The vulva equivalence group, although reduced in size, is also formed by positional information of the Hox gene *lin-39*. In *P. pacificus lin-39* mutants, the vulva equivalence group is not formed and P5.p–P8.p die of programmed cell death. The reduction of the size of the vulva equivalence group in *P. pacificus* involves the transcription factor *hairy*[21]. In *hairy* mutants, P3.p and P4.p survive and form a vulva equivalence group with a pattern that is reminiscent of the pattern in *C. elegans*. Genetic and biochemical studies showed that, in *P. pacificus*, HAIRY and GROUCHO form a heterodimer that downregulates the activity of *lin-39* in P3.p and P4.p. Surprisingly, there is no 1:1 orthologue of *hairy* in the *C. elegans* genome. Moreover, vulva induction in *P. pacificus* requires multiple cells of the somatic gonad instead of only one, as is the case in *C. elegans*. Mutations in the β-catenin-like gene *bar-1* in *P. pacificus* result in a vulvaless phenotype, indicating that Wnt signalling controls vulva induction. Indeed, genetic studies showed a redundant role of several Wnt ligands, which are expressed in the somatic gonad and the posterior region of the animal (arrows)[22].



Box 2 | **Dorso–ventral patterning in *D. melanogaster* and *T. castaneum***



Drosophila melanogaster is a long germ band insect that forms all body segments simultaneously during the blastoderm stage[24] (see the figure, left panel). By contrast, *Tribolium castaneum* is a short germ band insect in which posterior segments develop successively[24] (see the figure, right panel). As a result, the extra-embryonic membranes differ between *D. melanogaster* and *T. castaneum*. *T. castaneum* has two extra-embryonic membranes: the serosa, surrounding the complete embryo, and the amnion, covering the embryo proper on the ventral side. In *D. melanogaster*, both membranes are fused to an amnioserosa, which covers the embryo only at the dorsal side. Dorso–ventral patterning and extra-embryonic membrane formation require homologous genes that have divergent functions. Mutations in the homeobox transcription factor *zerknüllt* (*zen*) in *D. melanogaster* result in the replacement of the amnioserosa by ectodermal tissue[25]. *T. castaneum* contains two *zen* genes, *zen1* and *zen2*, and RNAi experiments revealed sub-functionalization of these genes[26]. RNAi against *zen1* results in the absence of the serosa and an expansion of the germ rudiment towards the anterior, indicating that *zen1* acts in antero–posterior development and specifies the border between the embryonic and extra-embryonic tissue[26]. In *D. melanogaster*, the loss of the transmembrane receptor Toll results in completely dorsalized embryos, whereas RNAi against *T. castaneum Toll* results in the absence of the central nervous system and the amnion. These differences reflect the different regulatory linkage of signalling networks in *D. melanogaster* and *T. castaneum*[28].

purifying selection dominates over positive selection and, as a result, genes should be conserved over large evolutionary time spans[39]. The evolutionary conservation of developmental control genes — as indicated by studies in evo–devo — strongly supports Kimura's neutral theory.

Recent advances in population genetics have come through comparative genomics, with genome sequencing projects revealing an enormous amount of natural variation[10]. But is natural variation also seen in developmental control genes? How do developmental control genes change in microevolution? More generally, are non-adaptive forces important for developmental evolution? Work at the interface between population genetics and evo–devo will indicate the contribution of natural variation to the evolution of development. This requires the research portfolio of population genetics to be added to evo–devo[10,40] (TABLE 1).

The comparison of very closely related species and independent isolates of the same species can indicate to what extent developmental processes evolve at the microevolutionary level. High-resolution mapping, through genome-wide association studies or through recombinant inbred lines, combined with next-generation sequencing can identify the molecular changes that cause a particular effect. Such studies can easily be performed in any species, as long as enough natural isolates have been or can be obtained. A few inroads into the microevolution of development have been taken; for example, studies in *P. pacificus* and *C. elegans* indicate that vulva development is subject to microevolutionary change[41,42]. In *C. elegans*, several recent studies show the power of QTL analysis for other developmental and life history traits, such as copulatory plug formation and pathogen susceptibility[43,44]. Therefore, 'next-generation genetics', as recently proposed for plants[45], can be a powerful new tool when applied to evo–devo. Ultimately, such studies might indicate how natural variation contributes to macroevolutionary alterations. Neo-Darwinism assumes that macroevolutionary change results from repeated microevolutionary alterations, but there is no substantial proof for this assumption. Current population genetics lacks an in-depth consideration of developmental control genes in the same way as evo–devo lacks a serious consideration of microevolutionary processes. Therefore, a synthesis of evo–devo and population genetics would provide a substantial contribution to evolutionary theory.

A synthesis with evolutionary ecology. All processes required for phenotypic change — natural variation, selection, genetic drift and developmental change — occur in populations that live in a specific ecological context. As the environmental conditions that organisms are exposed to change, it is crucial to ask whether the environment influences development. But are the developmental response to the environment and the ecological interactions of the organism important for the evolution of new phenotypes? How do developmental processes evolve under changing environmental conditions?

Research programmes in 'ecological developmental biology' are now actively propagated[40,46]. For some evo–devo models the ecological niche is well described. For example, *P. pacificus* lives on a scarab beetle[47,48] and *T. castaneum* in dry environments, such as wheat[49]. Both species are now the subject of 'ecological evo–devo'

research[50–52]. Other evo–devo models, such as the cnidarian *Nematostella vectensis* and some of its close relatives, differ from each other in their ecological niche and tolerance, and research programmes that involve ecology-oriented studies are well underway[53]. Other models have been established largely owing to ecological considerations. For example, studies in the cavefish *Astyanax mexicanus* can indicate how the developmental networks regulating eye development have been altered in response to the dark environment in caves[54]. Phenotypic plasticity is a central concept of evo–devo and is, by definition, at the interface between evo–devo and ecology[55,56]. However, although it is a widespread phenomenon[57–59], further studies are required to reveal whether phenotypic plasticity is a common route for the generation of developmental novelty. One advocate of this idea was van Valen, who was ahead of his time when he proposed that "evolution is the control of development by ecology"[60] — a statement that is now being transferred to a highly interdisciplinary research agenda.

Conclusions

I argue that the attempt of evo–devo to understand phenotypic change and novelty requires functional investigations. This is best achieved by choosing a limited number of model organisms and by developing a sophisticated methodological tool kit in those organisms. Although such a research strategy is constrained by unbiased phylogenetic sampling, it can help evo–devo to develop its own theory and to secure funding as part of the modern life sciences. Insight into the change of developmental mechanisms provides a platform for the integration of evo–devo into evolutionary theory — the single most important requirement for the long-term success of this young discipline. The partial ignorance of evo–devo with respect to the complexity of evolutionary theory[61], and the naive assumption that all developmental patterns observed in nature are adaptive[62], is an important threat to evo–devo. A synthesis with population genetics and evolutionary ecology can help evo–devo meet these challenges, but requires new research strategies and intense consideration of evolutionary theory.

Ralf J. Sommer is at the Max Planck Institute for Developmental Biology, Department for Evolutionary Biology, Spemannstrasse 37, D-72076 Tübingen, Germany.

e-mail: ralf.sommer@tuebingen.mpg.de

doi:10.1038/nrg2567
Published online 16 April 2009

1. Wilkins, A. *The Evolution of Developmental Pathways* (Sinauer Associates, Sunderland, Massachusetts, 2002).
2. Rudel, D. & Sommer, R. J. The evolution of developmental mechanisms. *Dev. Biol.* **264**, 15–37 (2003).
3. Raff, R. *The Shape of Life* (Chicago Univ. Press, Chicago, 1996).
4. Gerhard, J. & Kirschner, M. *Cells, Embryos and Evolution* (Blackwell Science, Oxford, 1997).
5. Minelli, A. *The Development of Animal Form* (Cambridge Univ. Press, Cambridge, 2003).
6. Carroll, S. B. *Endless Forms Most Beautiful* (Norton & Comp., New York, 2005).
7. Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, Oxford, 1991).
8. Jenner, R. J. & Wills, M. A. The choice of model organisms in evo–devo. *Nature Rev. Genet.* **8**, 311–319 (2007).
9. Rieppel, O. Development, essentialism, and population thinking. *Evol. Dev.* **10**, 504–507 (2008).
10. Lynch, M. *The Origin of Genome Architecture* (Sinauer Associates, Sunderland Massachusetts, 2007).
11. Akam, M., Holland, P., Ingham, P. & Wray, G. (eds) *The Evolution of Developmental Mechanisms. Development Supplement* (The Company of Biologists, Cambridge, 1994).
12. Roth, S. & Hartenstein, V. Development of *Tribolium castaneum*. *Dev. Genes Evol.* **218**, 115–118 (2008).
13. Lynch, J. A., Brent, A. E., Leaf, D. S., Pultz, M. A. & Desplan, C. Localized maternal *orthodenticle* patterns anterior and posterior in the long germ wasp *Nasonia*. *Nature* **439**, 728–732 (2006).
14. Hong, R. L. & Sommer, R. J. *Pristionchus pacificus*: a well-rounded nematode. *BioEssays*, **28**, 651–659 (2006).
15. Cooke, J., Nowak, M. A., Boerlijst, M. & Maynard-Smith, J. Evolutionary origin and maintenance of redundant gene expression during metazoan development. *Trends Genet.* **13**, 360–364 (1997).
16. Veeman, M. T., Axelrod, J. D. & Moon, R. T. A second canon: functions and mechanisms of β-catenin-independent Wnt signaling. *Dev. Cell* **5**, 367–377 (2003).
17. Dieterich, C. *et al.* The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nature Genet.* **40**, 1193–1198 (2008).
18. Schlager, B. *et al.* Molecular cloning of a dominant Roller mutant and establishment of DNA-mediated transformation in the nematode model *Pristionchus pacificus*. *Genesis* (in the press).
19. Sternberg, P. W. Vulva development. *Wormbook* [online], <http://www.wormbook.org/chapters/www_vulvaldev/vulvaldev.html> 25 Jun 2005 (doi:10.1895/wormbook.1.6.1)., ed.
20. Sommer, R. J. & Sternberg, P. W. Apoptosis limits the size of the vulval equivalence group in *Pristionchus pacificus*: a genetic analysis. *Curr. Biol.* **6**, 52–59 (1996).
21. Schlager, B. *et al.* HAIRY-like transcription factors and the evolution of the nematode vulva equivalence group. *Curr. Biol.* **16**, 1386–1394 (2006).
22. Tian, H; Schlager, B., Xiao, H. & Sommer, R. J. Wnt signaling by differentially expressed Wnt ligands induces vulva development in *Pristionchus pacificus*. *Curr. Biol.* **18**, 142–146 (2008).
23. *Tribolium* Genome Sequencing Consortium. The genome of the model beetle and pest *Tribolium castaneum*. *Nature* **452**, 949–955 (2008).
24. Roth, S. in *Gastrulation: From cells to embryos* (ed. C. Stern) 105–122 (Cold Spring Harbor Laboratory Press, 2004).
25. Rushlow, C. & Levine, M. Role of the *zerknüllt* gene in dorsal-ventral pattern formation in *Drosophila*. *Adv. Genet.* **27**, 277–307 (1990).
26. van der Zee, M., Berns, N. & Roth, S. Distinct functions of the *Tribolium zerknüllt* genes in serosa specification and dorsal closure. *Curr. Biol.* **15**, 624–636 (2005).
27. Stathopoulos, A., Van Drenth, M., Erives, A., Markstein, M. & Levine, M. Whole-genome analysis of dorsal-ventral patterning in the *Drosophila* embryo. *Cell* **111**, 687–701 (2002).
28. da Fonseca, R. N. *et al.* Self-regulatory circuits in dorsoventral axis formation of the short-germ beetle *Tribolium castaneum*. *Dev. Cell* **14**, 605–615 (2008).
29. Wagner, G. The developmental genetics of homology. *Nature Rev. Genet.* **8**, 473–479 (2007).
30. Sommer, R. J. Homology and the hierarchy of biological systems. *BioEssays*, **30**, 653–658 (2008).
31. de Beer, G. R. *Embryos and Ancestors* (Claredon Press, Oxford, 1958).
32. de Beer, G. R. *Homology: An Unsolved Problem* (Oxford Univ. Press, Oxford, 1971).
33. Behringer, R. *et al.* (eds) *Emerging Model Organisms. A Laboratory Manual* (Cold Spring Harbor Laboratory Press, 2009).
34. Reddien, P. W. & Sanchez, Alvarado, A. Fundamentals of planarian regeneration. *Annu. Rev. Cell Dev. Biol.* **20**, 725–757 (2004).
35. Tiozzo, S., Brown, F. D. & De Tomaso, A. W. in *Stem Cells From Hydra to Man* (ed. Thomas Bosch) 95–112 (Springer, Heidelberg, 2008).
36. Amundson, R. *The Changing Role of the Embryo in Evolutionary Thought* (Cambridge Univ. Press, Cambridge, 2005).
37. Dobzhansky, T. *Evolution, Genetics and Man* (Wiley, New York, 1955).
38. Mayr, E. *Animal Species and Evolution* (Harvard Univ. Press, Cambridge, Massachusetts, 1966).
39. Kimura, M. *The Neutral Theory of Molecular Evolution*. (Cambridge Univ. Press, Cambridge, 1983).
40. Zauner, H. & Sommer, R. J. in *Evolving pathways: Key Themes in Evolutionary Developmental Biology* (eds Minelli, A. & Fusco, g.). 151–171 (Cambridge Univ. Press, Cambridge, 2008).
41. Zauner, H. & Sommer, R. J. Evolution of robustness in the signaling network of *Pristionchus* vulva development. *Proc. Natl Acad Sci. USA* **104**, 10086–10091 (2007).
42. Milloz, J., Duveau, F., Nuez, I. & Felix, M.-A. Intraspecific evolution of the intercellular network underlying a robust developmental system. *Genes Dev.* **22**, 3064–3075 (2008).
43. Palopoli, M. F. *et al.* Molecular basis of the copulatory plug polymorphism in *Caenorhabditis elegans*. *Nature* **454**, 1019–10222 (2008).
44. Reddy, K. C., Andersen, E. C., Kruglyak, L. & Kim, D. H. A polymorphism in *npr-1* is a behavioral determinant of pathogen susceptibility in *C. elegans*. *Science* **323**, 382–384 (2009).
45. Nordborg, M. & Weigel, D. Next-generation genetics in plants. *Nature* **456**, 720–723 (2008).
46. Gilbert, S. F. & Bolker, J. A. Ecological developmental biology: preface to the symposium. *Evol. Dev.* **5**, 3–8 (2003).
47. Herrmann, M., Mayer, E. W. & Sommer, R. J. Nematodes of the genus *Pristionchus* are closely associated with scarab beetles and the Colorado potato beetle in western Europe. *Zoology* **109**, 96–108 (2006).
48. Herrmann, M. *et al.* The nematode *Pristionchus pacificus* (Nematoda: Diplogastridae) is associated with the Oriental beetle *Exomala orientalis* (Coleoptera: Scarabaeidae) in Japan. *Zool. Sci.* **24**, 883–889 (2007).
49. Sokoloff, A. *The Biology of Tribolium* (Oxford Clarendon, Oxford, 1972).
50. Hong, R. L., Svatos, A., Herrmann, M. & Sommer, R. J. The species-specific recognition of beetle cues by *Pristionchus maupasi*. *Evol. Dev.* **10**, 273–279 (2008).
51. Hong, R. L., Witte, H. & Sommer, R. J. Natural variation in *P. pacificus* insect pheromone attraction involves the protein kinase EGL-4. *Proc. Natl Acad. Sci. USA* **105**, 7779–7784 (2008).
52. Jackowska, M. *et al.* Genomic and gene regulatory signatures of cryptozoic adaptation: loss of blue sensitive photoreceptors through expansion of long wavelength-opsin expression in the red flour beetle *Tribolium castaneum*. *Front. Zool.* **4**, 24 (2007).
53. Darling, J. *et al.* Rising starlet: the starlet sea anemone *Nematostella vectensis*. *BioEssays* **27**, 211–221 (2005).
54. Jeffery, W. R. Cavefish as model system in evolutionary developmental biology. *Dev. Biol.* **231**, 1–12 (2001).
55. Pigliucci, M. Evolution of phenotypic plasticity: where are we going now? *Trends Ecol. Evol.* **20**, 481–486 (2005).
56. West-Eberhard, M. J. *Developmental Plasticity and Evolution* (Oxford Univ. Press, Oxford, 2003).
57. Saenko, S. V., French, V., Brakefield, P. M. & Beldade, P. Conserved developmental processes and the formation of evolutionary novelties: examples from butterfly wings. *Philos. Trans. R. Soc. Lond., B* **363**, 1549–1555 (2008).
58. Laforsch, C. & Tollrian, R. Embryological aspects of inducible morphological defense in *Daphnia*. *J. Morph.* **262**, 701–707 (2004).

59. Abouheif, E. & Wray, G. Evolution of the gene network underlying wing polymorphism in ants. *Science* **297**, 249–252 (2002).
60. Van Valen, L. Festschrift. *Science* **180**, 488 (1973).
61. Hoekstra, H. E. & Coyne, J. A. The locus of evolution: evo devo and the genetics of adaptation. *Evolution* **61**, 995–1016 (2007).
62. Lynch, M. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl Acad Sci. USA* **104**, 8597–8604 (2007).
63. Hill, R. C. *et al.* Genetic flexibility in the convergent evolution of hermaphroditism in *Caenorhabditis* hermaphrodites. *Dev. Cell* **10**, 531–538 (2006).
64. Shapiro, M. D. *et al.* Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717–723 (2005).
65. Corley, S. B., Carpenter, R., Copsey, L. & Coen, E. Floral asymmetry involves an interplay between TCP and MYB transcription factors in *Antirrhinum*. *Proc. Natl Acad Sci. USA* **102**, 5068–5073 (2005).

Acknowledgements
I would like to thank S. Roth, F. Brown, M. Riebesell and three anonymous reviewers for useful comments on the manuscript.

DATABASES

Entrez Genome: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genome
Caenorhabditis elegans | *Drosophila melanogaster* | *Nasonia vitripennis* | *Nematostella vectensis* | *Pristionchus pacificus* | *Tribolium castaneum*

FURTHER INFORMATION

Ralf Sommer's homepage: http://www.eb.tuebingen.mpg.de/departments/4-evolutionary-biology/department-4-evolutionary-biology
Max Plank Institute for Developmental Biology: http://www.pristionchus.org

ALL LINKS ARE ACTIVE IN THE ONLINE PDF