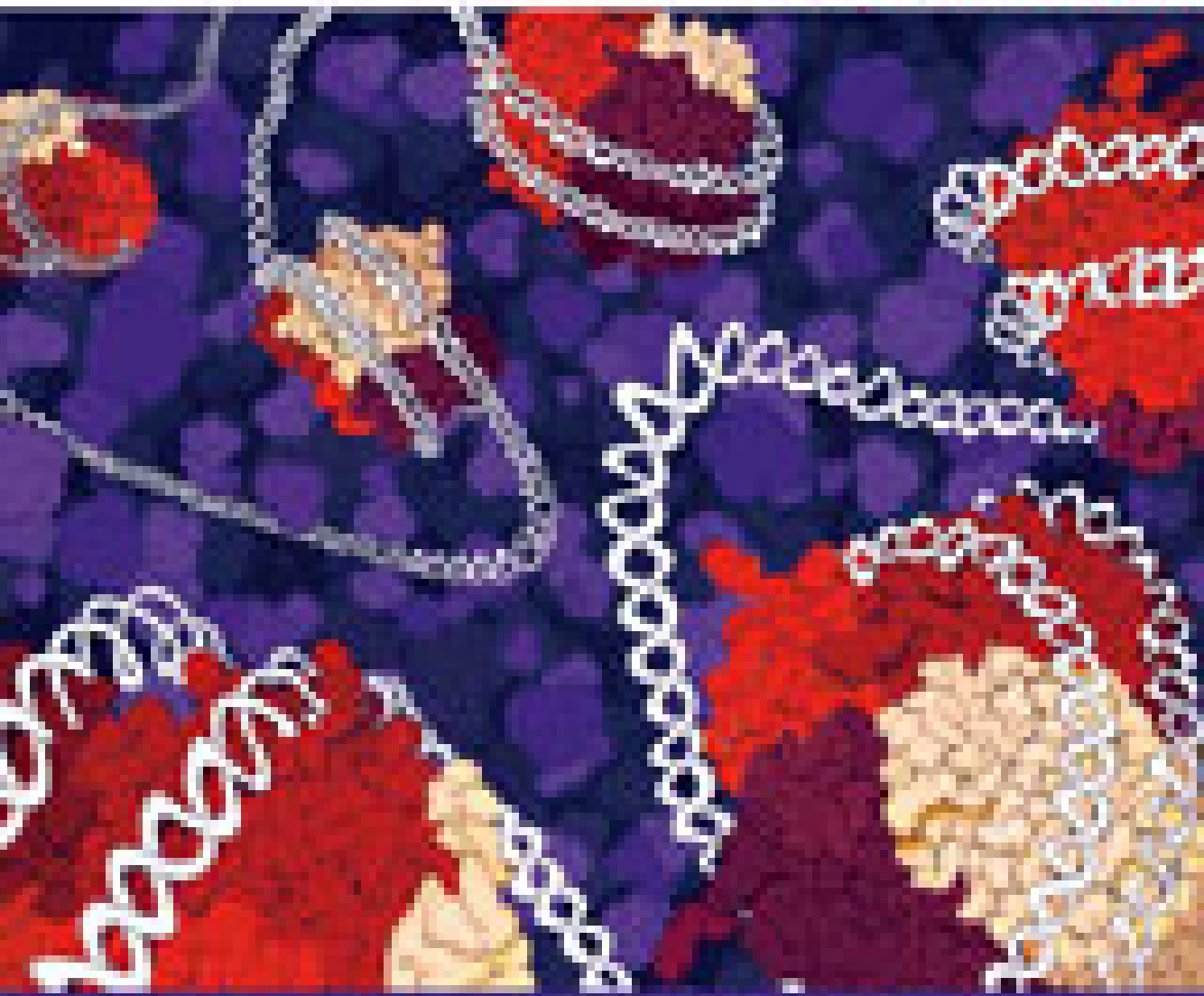
nature
REVIEWS

GENETICS



TAKE YOUR PLACES

Control of nucleosome positioning and its influence on transcription

Genetics of gene expression

Mapping diverse variants with transcript levels

nature publishing group

► **COVER:** 'Left reeling' by Patrick Morgan, inspired by the Review on p161.


LOUISA FLINTOFT


TANITA CASCI


MARY MUERS


MEERA SWAMI

How much of the genome is transcribed and what is the function of the RNA output? It used to be possible to give straightforward answers to these questions. With a few exceptions, transcription was thought to start from defined regions — promoters of protein-coding genes — and to serve as an intermediate step in the production of proteins.

Recent years have seen a major rethink about the origins and purpose of transcription. One important turning point was the discovery of RNAi and the subsequent finding that many regulatory small RNAs, notably microRNAs, lurked — previously unnoticed — in the genomes of eukaryotes. But the picture is far more complicated than this. The ability to probe the transcriptome at high resolution using microarrays and, more recently, high-throughput sequencing, has revealed a bewilderingly complex 'tangle' of transcription occurring throughout the genome. For instance, a recent flurry of papers has documented extensive bidirectional transcription. As reported in a Research Highlight this month (p154), a new study suggests another layer of complexity, as transcripts are processed to spawn abundant smaller RNAs.

How much of this mass of transcription serves a functional purpose is a matter of debate, but insights have started to emerge. Also in this issue, a Progress article by John Mattick and colleagues (p155) focuses on long non-coding RNAs, which have been shown to have regulatory effects on protein-coding genes at the levels of chromatin modification, transcription and post-transcriptional processing. The diverse biological roles of the targets of these non-coding RNAs, and the involvement of some of them in cancer and other diseases, suggest that the field of transcriptomics will have an impact across many aspects of biological research.

CANCER GENETICS

Networking on the fly

There are now many ways in which genes that are mutated in particular cancer types can be identified, including starting from less genetically complex organisms such as *Drosophila melanogaster*. Kevin White and colleagues have taken this approach and have identified SPOP as a protein that is overexpressed in renal cell carcinoma (RCC).

D. melanogaster has proved particularly useful for identifying gene networks that are conserved throughout evolution. White and colleagues started with two pair-rule genes — *eve* and *ftz* — homeobox genes that are part of a gene network often disrupted in human disease. Using a variety of information about Eve and Ftz, they built a predictive gene network model. Initially, gene expression patterns from wild-type embryos or embryos with mutated *eve* or *ftz* taken over a time course of 2–7 hours after egg laying (AEL) were compared. Chromatin immunoprecipitation analysed on DNA microarrays (ChIP–chip) was also used to map Eve and Ftz DNA binding sites 2 hours AEL. Genes that were both differentially expressed and identified by the ChIP–chip approach were considered as putative targets: 137 Ftz target genes and 98 Eve target genes. To extend the network further, the authors then added in information on the target genes obtained from automated literature-mining techniques and yeast two-hybrid protein–protein interaction data. The resulting network contained 4,084 genes and proteins and 6,648 interactions between them.

Having tested that specific links within the network behave as expected, the authors analysed the 150 genes that have validated human homologues. The top candidate (a major network hub) was *roadkill* (*rdx*), which is 79% identical to the human protein SPOP. The network model indicated that D-SPOP (Rdx) is a Ftz target 2–3 hours AEL and then becomes an Eve target 6–7 hours AEL, and that D-SPOP interacts with the Jun kinase phosphatase Puckered (Puc). Further analyses indicated that D-SPOP and Puc interact and this is important for the regulation of Eiger (tumour necrosis factor)-induced apoptosis in neurons. Indeed, the authors found that D-SPOP, like human SPOP, can induce protein ubiquitylation and degradation, and D-SPOP induced the degradation of Puc, which functions to inhibit Eiger-mediated JUN N-terminal kinase-induced apoptosis.

As homologues of Eve and Ftz targets are involved in tumorigenesis in humans, the authors investigated whether SPOP shows altered expression in tumours using a human tissue microarray. They found that SPOP is highly expressed in 85% of RCCs, whereas the protein is expressed at a low level in normal kidney tissue. Moreover, they found that SPOP expression can be used to identify different types of RCC. In clear cell RCC, which in some cases can be difficult to distinguish from other types of RCC, 99% of cases were positive for SPOP expression, as were 86% of chromophobe RCCs, whereas only 22% of papillary-type RCCs were positive.

The authors conclude that analysing gene networks on the basis of information in *D. melanogaster* is an effective method for understanding the biological function of these networks, for identifying conserved gene networks and identifying new genes within these networks that are deregulated in human disease.

Nicola McCarthy,
Chief Editor, Nature Reviews Cancer

ORIGINAL RESEARCH PAPER Liu, J. *et al.* Analysis of *Drosophila* segmentation network identifes a JNK pathway factor overexpressed in kidney cancer. *Science* 22 Jan 2009 (doi 10.1126/science1157669)

Nature Reviews Genetics | AOP, published online 3 February 2009; doi:10.1038/nrg2542

CHROMATIN

Completing the picture

Characterizing the proteins that bind to and regulate specific genes is crucial for understanding gene regulation. A new method called proteomics of isolated chromatin segments (PiCh) promises to accelerate progress in this area, by allowing the full complement of proteins that bind to a particular gene to be assayed in a single experiment.

Chromatin immunoprecipitation (ChIP) is a technique that is widely used to determine whether a protein of interest binds to a particular genomic region. But this method relies on antibodies to known DNA-binding proteins and does not provide a complete description of protein composition. To overcome these limitations PiCh uses a specific DNA probe to isolate proteins from fixed cells, which are then characterized using mass spectrometry (MS). Because MS analysis requires a relatively large quantity of protein (at least a picomole) Dejardin and Kingston, who developed PiCh, tested their approach using telomeres — the abundance of which necessitates less starting material than for single copy loci.

Using PiCh, the authors were able to identify most known telomeric factors. By comparing different cell lines that expressed mutant telomerase, they were also able to pinpoint differences in protein composition of telomeres maintained by the reverse transcriptase and alternative lengthening of telomeres (ALT) pathways. Importantly, they also discovered many novel telomeric associations; for example, between telomeres that are maintained by ALT and orphan receptors from the nuclear hormone receptor superfamily. Although further optimization will be required to study single copy genes in humans and other organisms with large genomes, PiCh could be immediately applied to other repeat sequences and to low copy sequences in organisms with small genomes.

Meera Swami

ORIGINAL RESEARCH PAPER Dejardin, J. & Kingston, R. E. Purification of proteins associated with specific genomic loci. *Cell* **136**, 175–186 (2009).
FURTHER READING Schones, D. E. & Zhao, K. Genome-wide approaches to studying chromatin modifications. *Nature Rev. Genet.* **9**, 179–191 (2008).

IN BRIEF

CHROMATIN

Role of Jhdm2a in regulating metabolic gene expression and obesity resistance

Tateishi, K. *et al. Nature* 4 Feb 2009 (doi:10.1038/nature07777)

This paper reveals a physiological role for a chromatin-modifying protein — the histone H3 lysine 9 (H3K9) demethylase JHDM2A. Disruption of *Jhdm2a* function in mice led to obesity and hyperlipidaemia. The authors showed that JHDM2A expression is induced by β-adrenergic stimulation and that this leads to direct activation of two genes that are involved in metabolic regulation — peroxisome proliferator activated receptor-α (*Ppara*) and uncoupling protein 1 (*Ucp1*) — by both decreasing levels of H3K9 dimethylation and facilitating the recruitment of transcriptional coactivators.

HUMAN GENOMICS

Population analysis of large copy number variants and hotspots of human genetic disease

Itsara, A. *et al. Am. J. Hum. Genet.* **84**, 1–14 (2009)

Using genome-wide SNP data from ~2,500 apparently normal individuals, these authors found that large (>100 kb) copy number variants are common in humans and that at least 1% of individuals carry variants longer than 1 Mb. However, individual large variants segregate at low frequencies (0.1–1%) in the general population. The authors also suggest that the anticorrelation between both the size of a variant and its gene density with allele frequency indicates that large variants are generally deleterious and so may contribute to disease phenotypes.

VIRAL GENETICS

Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration

Geuking, M. B. *et al. Science* **323**, 393–396 (2009)

This study reveals that RNA viruses, not just retroviruses, can integrate into the host genome. RNA viruses acquire this ability by recombining with an endogenous retrotransposon. The authors observed *in vitro* and *in vivo* that a mouse RNA virus (lymphocytic choriomeningitis virus) can use the reverse transcriptase and integrase functions of an endogenous retrotransposon (intracisternal A-type particle) to insert a recombinant cDNA sequence into the host genome. This finding raises the need to look closely at endogenous, largely inactive retroviral elements before attempting viral-based gene therapy.

FUNCTIONAL GENOMICS

Integrating computational biology and forward genetics in *Drosophila*

Aerts, S. *et al. PLoS Genet.* **5**, e1000351 (2009)

A new study shows that integrating genome-wide computational gene prioritization with large-scale *in vivo* genetic screening — termed systems genetics — increases the efficiency of identifying functional genes. The authors combined the ENDEAVOUR-HIGHFLY web resource for gene prioritization with genetic screening of *Drosophila* species mutant, deficiency and RNAi collections, identifying a novel gene interaction network for the *Drosophila* proneural transcription factor Atonal. They then used systems genetics to prioritize the entire *Drosophila melanogaster* genome for 10 canonical biological pathways, creating a publicly available database of prioritized pathway candidates.

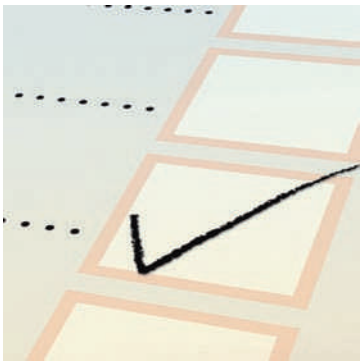
Nature Reviews Genetics | AOP, published online 3 February 2009; doi:10.1038/nrg2541

EPIGENETICS

Transcription makes a mark

How specific maternal and paternal patterns of DNA methylation are established in the germ line is a crucial aspect of genomic imprinting that is poorly understood. A new paper has now added transcription to the mix of factors needed for methylation in mammalian oocytes.

Differentially methylated regions (DMRs) are CpG islands that are methylated in either maternal or paternal gametes and they lead to monoallelic expression of imprinted genes in the offspring. In maturing oocytes the DNA methyltransferase DNMT3a and its co-factor DNMT3L are responsible for setting up the maternal imprint, but why DMRs are specifically targeted for methylation is unclear. The periodicity of CpG dinucleotides in DMRs and the absence of histone H3 lysine 4 methylation seem to be important and yet are insufficient to explain why DMRs but not other CpG islands become methylated.



To try to fill in the picture, Chotalia and colleagues explored whether transcription had an impact on maternal methylation at the *Gnas* locus in mouse oocytes — three DMRs at this locus control monoallelic expression of five imprinted transcripts. The protein-coding *Nesp* gene starts upstream of the DMRs and traverses the entire locus. When the authors truncated the *Nesp* transcript by inserting a termination cassette, they found that germ line methylation of the DMRs was lost and this disrupted imprinted expression of the *Gnas* cluster in offspring.

The authors also observed transcription in oocytes across DMRs at several other maternally marked imprinted domains, but transcription in oocytes was not a common feature of similar non-imprinted or paternally methylated CpG islands. Coupling these findings to the known link between histone methylation and DNA methylation they suggest two hypotheses: that transcription through DMRs in oocytes might create a chromatin environment that allows access of the DNA methylation machinery to the DMRs; or that the transcripts could help to recruit proteins involved in DNA methylation or histone demethylation.

As well as offering potential insights into the evolution of imprinting, this new-found role for transcription in establishing imprinting could also help to explain the basis of some human imprinting disorders. For example, some individuals with Angelman's syndrome or pseudohypoparathyroidism type 1b have maternally transmitted microdeletions upstream of DMRs that result in loss of methylation at the DMRs. These deletions might disrupt an upstream transcript that is required for methylation in the female germ line, so loss of transcription could be the molecular defect in these cases.

Mary Muers

ORIGINAL RESEARCH PAPER Chotalia, M. *et al.* Transcription is required for establishment of germline methylation marks at imprinted genes. *Genes Dev.* **23**, 105–117 (2009)
FURTHER READING Sasaki, H. & Matsui, Y. Epigenetic events in mammalian germ-cell development: reprogramming and beyond. *Nature Rev. Genet.* **9**, 129–140 (2008)

Nature Reviews Genetics | AOP, published online 3 February 2009; doi:10.1038/nrg2543

DISEASE GENETICS

The importance of networking

Remember cloning a gene the old-school way? Or studying only one gene at a time? A recent paper by Quigley *et al.* is the latest in a series of studies to apply the construction of genetic networks to linking genotypes to phenotypes, specifically in mice. The results highlight the potential of this approach for identifying genes with key roles in phenotypes that are related to disease.

To identify genetic motifs linked to skin cancer and inflammation, Quigley and colleagues crossed mice of two species, *Mus spretus* and *Mus musculus*, as they are respectively resistant and susceptible to tumour development. After inducing skin tumours in a number of the mice, they extracted mRNA from uninvolved tail skin for gene expression analysis. Combining mRNA profiling with linkage analysis allowed the authors to construct a 'susceptibility network' of gene expression and regulation in the normal skin.

The normal skin network identified at least 62 genes involved in hair follicle biology, which are regulated by numerous expression QTLs (eQTLs; genetic loci that influence other genes in *cis* or in *trans*). A candidate regulatory gene in the hair follicle network was the G-protein-coupled receptor *Lgr5*, agreeing with its previously described role as a stem cell marker in hair follicles and intestinal cells. Further work is needed to determine whether *Lgr5* is the 'master regulator' of hair follicles; presumably this would be of great interest to industries concerned with hair regrowth.

Construction of the gene expression networks from skin tumour-susceptible and resistant mice suggested that normal skin from susceptible mice showed enriched expression of genes involved in inflammation as well as in cell growth and its regulation. One locus on mouse chromosome 15 seems to be linked to expression of numerous genes in a network for inflammation and barrier function. The researchers focused on *Vdr* as the best candidate master regulator of this network. *Vdr* encodes the mouse vitamin D receptor, and vitamin D levels in humans have been linked to cancer susceptibility in previous studies. The authors speculate that several sequence changes between the *M. spretus* and *M. musculus* alleles affect the protein's function and resulting tumour susceptibility. These results support the notion that adequate levels of vitamin D are important for health, as well as confirming the power of the genetic network approach for identifying genotype–phenotype relationships.

Chris Gunter,
HudsonAlpha Institute for Biotechnology

ORIGINAL RESEARCH PAPER Quigley, D. A. *et al.* Genetic architecture of mouse skin inflammation and tumour susceptibility. *Nature* 11 Jan 2009 (doi:10.1038/nature07683)

WEB SITE

Balmain laboratory homepage:
http://cancer.ucsf.edu/balmain/index.php

Nature Reviews Genetics | AOP, published online 10 February 2009; doi:10.1038/nrg2545

EVO–DEVO

Failsafe flowers

The design of regulatory circuits determines their function — for example, whether they behave like developmental switches — but the specific details of the wiring can also be influenced by evolutionary pressures. A study of flower organ development shows that the interaction between two key regulatory proteins arose not because it is indispensable for function, but because it provides robustness against incorrect cell fate decisions.

Class B floral homeotic genes encode transcription factors that specify petals and male reproductive organs. The authors specifically examined two such genes, *DEFICIENS* and *GLOBOSA*, the products of which heterodimerize and maintain their own expression via a positive autoregulatory loop. Given that a single gene is capable of maintaining an autoregulatory loop, why would it be necessary to retain heterodimerization? The hypothesis tested here is that obligate heterodimerization between these two ancient paralogues is necessary to filter out noise, and hence to canalize flower development.

The authors used a stochastic modelling approach to examine alternative evolutionary scenarios — homodimerization, heterodimerization or a mixture of the two. These alternatives were compared with respect to the activity (ON/OFF) of the self-regulating circuit once the initial activator is withdrawn. The robustness of the switch function is clearly favoured by heterodimerization, which allows the establishment of the autoregulatory loop (when it is most labile) to occur more accurately.

Given the generality of the mathematical model, the conclusions might apply to equivalent genetic relationships in other contexts.

Tanita Casci

ORIGINAL RESEARCH PAPER Lenser, T. *et al.* Developmental robustness by obligate interaction of class B floral homeotic genes and proteins. *PLoS Comp. Biol.* **5**, e1000264 (2009)
FURTHER READING Wilkinson, D. J. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Rev. Genet.* **10**, 122–133 (2009)

IN BRIEF

RNA INTERFERENCE

Antiviral immunity in *Drosophila* requires systemic RNA interference spread

Saleh, M.-C. *et al. Nature* 8 Feb 2009 (doi:10.1038/nature07712)

Insects can mount a local antiviral RNAi defence; however, it has now been shown that *Drosophila melanogaster* can also generate a systemic RNAi response. Inoculation of flies with dsRNA corresponding to regions of the Sindbis and *Drosophila* C viral genomes led to a sequence-specific systemic immune response, which required a recently defined dsRNA uptake pathway. This study suggests that immunity in vertebrates and invertebrates may be more highly conserved than previously believed.

DEVELOPMENT

Nodal points and complexity of Notch–Ras signal integration

Hurlbut, G. D. *et al. Proc. Natl Acad. Sci. USA* 26 Jan 2009 (doi:10.1073/pnas.0812024106)

Although it is known that signalling pathways interact during development, how signals are integrated remains unexplored. By examining the genome-wide, common transcriptional targets of two pathways — Notch and receptor tyrosine kinase (RTK) — in transgenic fly embryos, the authors reveal extensive crosstalk between the two pathways, identify the integration points (which were validated through genetic interaction analysis) and suggest that Notch increases the output specificity of RTK signalling.

GENOME EVOLUTION

Cryptic variation in the human mutation rate

Hodgkinson, A. *et al. PLoS Biol.* **7**, e1000027 (2009)

Hotspots of biased nucleostide substitution in human genes

Berglund, J. *et al. PLoS Biol.* **7**, e1000026 (2009)

These two papers examine factors affecting the rate of mutation in the human genome. Hodgkinson and colleagues identified nucleotide positions with rapid mutation rates as those that have SNPs in both humans and chimpanzees. Substantially increased mutation rate at these sites was not due to CpG dinucleotides or neighbouring nucleotides, but was influenced by sequence context in a complex and previously undetected way. A bias towards AT-to-GC substitutions in genes with accelerated substitutions in humans, detected by comparison with primate genomes, suggested to Berglund and colleagues that increased mutation in these genes is influenced by recombination rather than positive selection.

CANCER GENETICS

The dynamic DNA methylomes of double-stranded DNA viruses associated with human cancer

Fernandez, A. F. *et al. Genome Res.* 10 Feb 2009 (doi:10.1101/gr.083550.108)

This paper suggests that the progression of viral-linked cancers might be caused by epigenetic changes in the viral DNA. The authors created a methylation map of three oncogenic viruses — Epstein–Barr virus, human papilloma virus and hepatitis B virus — and show that these genomes become progressively methylated during disease progression. Methylation, which might shield viruses from the immune system, could therefore be used as a biomarker for disease.

Nature Reviews Genetics | AOP, published online 10 February 2009; doi:10.1038/nrg2546

DEVELOPMENT

Deciphering the Wingless gradient

A key principle of developmental biology is that during pattern formation a cell detects the local concentration of a morphogen within a gradient, which guides the cell down the appropriate differentiation pathway. But how can such simple information reliably activate the range of intricate signalling pathways required to pattern a complex structure? A recent study by Piddini and Vincent shows that, in *Drosophila melanogaster*, the Wingless (Wg) morphogen does more than just trigger gene expression in a dose-dependant fashion. It also activates two nonautonomous inhibitory pathways that modulate the reaction of surrounding cells to the Wg signal, thereby enhancing their ability to recognize their position within the Wg gradient.

The authors used models in which the Wg signal was disrupted in all or parts of the *D. melanogaster* imaginal disc and looked at the expression of Wg target genes. These targets included *dll* (which is induced by mid to low Wg levels at the first to second instar) and *sens* (which is induced by high Wg levels at the third instar).

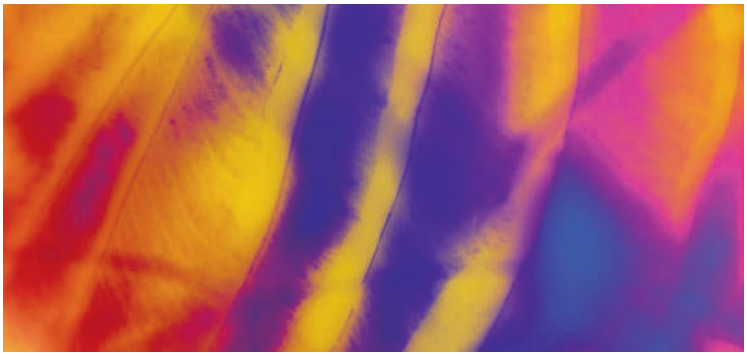
When the Wg signal was gradually depleted across the whole imaginal disc, so that it was absent by the end of the third larval instar, *dll* expression was largely unaffected. This shows that, once it is established, *dll* expression is maintained in the absence of Wg. However, expression of *sens* was fully depleted in these models, showing that *sens* expression requires continuous Wg signalling.

The authors also looked at what happened when Wg signalling was overactivated in a mosaic fashion. In discs with small patches of high signalling cells among wild-type (WT) cells, *dll* expression was expressed normally in most WT cells; however, in WT cells directly adjacent to cells undergoing high Wg signalling activity, *dll* expression was reduced. In mosaic discs with large patches of high signalling cells, *dll* expression was reduced in most WT cells and *sens* expression was completely absent in all WT cells. Excess Wg signalling inhibited the expression of *dll* and *sens* in neighbouring WT cells, indicating that these Wg target genes are sensitive to a nonautonomous negative feedback mechanism that is activated as a result of Wg signal transduction.

One candidate for this negative signal was Notum, a phospholipase that is an inhibitor of Wg signalling. The authors used the same mosaic disc model described above, but the Wg-overexpressing cells were also depleted of Notum through RNAi. *sens* expression was restored in WT cells, showing that Notum is required to suppress *sens* expression. However, *dll* expression was not restored in the WT cells: an additional signal must therefore be responsible for the observed *dll* suppression. The authors suggest that this second signal could be mediated by a secreted protein or by a more subtle influence, such as mechanical tension.

This study highlights one case in which the final cellular response to a patterning signal is modulated by secondary cell interactions. Such mechanisms may operate in other developmental scenarios to achieve the striking precision observed in patterning processes.

Elizabeth Neame



ORIGINAL PAPER Piddini, E. & Vincent, J.-P. Interpretation of the Wingless gradient requires signaling-induced self-inhibition. *Cell* **136**, 396–307 (2009)

FURTHER READING Affolter, M. & Basler, K. The Decapentaplegic morphogen gradient: from pattern formation to growth regulation. *Nature Rev. Genet.* **8**, 663–674 (2007)

TRANSCRIPTOMICS

Processing adds to the complexity

A recent study reveals how some of the increasingly apparent complexity of mammalian transcriptomes arises. It identifies an unanticipated abundance of stable, capped small RNAs in mammalian cells, which seem to be processed from both protein-coding and non-coding RNAs.

Deep sequencing identified 102,159 new small RNAs less than 200 nucleotides long from two human cell lines. Many of these RNAs can be categorized as members of a previously identified class of small RNAs that lie no more than 500 nucleotides from a transcriptional start site (TSS) — so-called promoter-associated small RNAs (PASRs). The authors showed by enzymatic treatment and immunoprecipitation that PASRs have 5′ cap structures. There was also a correlation between many long RNA 5′ ends that have been previously identified by CAGE tagging (which relies on the presence of a 5′ cap) and the 5′ ends of the new small RNAs.

PHOTOALTO

Both small RNAs and CAGE tags are particularly abundant in first exons, and the authors showed that many CAGE tags actually extend across splice sites, which suggests that they arise from the processing of spliced mRNAs, and then acquire a cap — a previously unrecognized fate for mRNAs. These long RNAs might then be further processed to generate smaller capped transcripts. The more detailed mapping of CAGE tags and small RNAs in one human gene strongly supported this conclusion.

These results suggest the regulated generation of abundant short RNAs in many mammalian protein-coding genes — but do these RNAs have any function? The authors made a series of synthetic single-stranded RNAs, 30–35 nucleotides in length, to mimic PASRs that lie near the annotated TSS of the human *MYC* gene. When these RNAs were transfected into cells, the result was a reduction in levels of *MYC* mRNA. This study has therefore not only revealed a new pathway of mRNA processing, but has also highlighted a potentially widespread transcriptional regulatory mechanism.

Louisa Flintoft

ORIGINAL RESEARCH PAPER Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* 25 Jan 2009 (doi:10.1038/nature07759)

FURTHER READING Wang, Z. *et al.* RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009) | Kapranov, P. *et al.* Genome-wide transcription and the implications for genomic organization. *Nature Rev. Genet.* **8**, 413–423 (2007)

Long non-coding RNAs: insights into functions

Tim R. Mercer, Marcel E. Dinger and John S. Mattick

Abstract | In mammals and other eukaryotes most of the genome is transcribed in a developmentally regulated manner to produce large numbers of long non-coding RNAs (ncRNAs). Here we review the rapidly advancing field of long ncRNAs, describing their conservation, their organization in the genome and their roles in gene regulation. We also consider the medical implications, and the emerging recognition that any transcript, regardless of coding potential, can have an intrinsic function as an RNA.

The RNA world hypothesis proposes that early life was based on RNA, which subsequently devolved the storage of information to more stable DNA, and catalytic functions to more versatile proteins. Consequently, despite crucial roles in the ancient processes of translation and splicing, RNA is assumed to have been largely relegated to an intermediate between gene and protein, encapsulated in the central dogma 'DNA makes RNA makes protein'. However, the finding that most of the genome in complex organisms is transcribed, apparently in a developmentally regulated fashion[1–6], and the discovery of new classes of regulatory non-coding RNAs (ncRNAs), challenges this assumption and suggests that RNA has continued to evolve and expand alongside proteins and DNA.

Although the current literature is dominated by short RNAs, there are an increasing number of reports describing long transcripts that, rather than encoding protein, act functionally as RNAs. Although we currently lack satisfactory classifications for these transcripts, long ncRNAs are arbitrarily considered to be longer than ~200 nucleotides, on the basis of a convenient practical cut-off in RNA purification protocols that excludes small RNAs[4].

Identification of long ncRNAs

As a transcriptional class, long ncRNAs were first described during the large-scale sequencing of full-length cDNA libraries in the mouse[6]. Although distinguishing long ncRNAs from other protein-coding mRNAs is not a trivial process (BOX 1), it has nevertheless become apparent that a significant portion of the transcriptome has little or no protein-coding capacity. The increased sensitivity of genome tiling arrays provides an even more detailed view, revealing that the extent of non-coding sequence transcription is at least four times greater than coding sequence, and that an abundance of non-polyadenylated non-coding transcripts had been previously overlooked[4].

These studies also showed that the transcriptome is surprisingly complex, with long ncRNAs often overlapping with, or interspersed between, multiple coding and non-coding transcripts[1,5] (FIG. 1). This complexity has prompted a shift in our understanding of gene organization from a linear to a modular model, in which it is possible for a sequence to be transcribed into a range of sense and antisense, coding and non-coding transcripts. Attempts to untangle this complexity have led to crude classifications of ncRNAs based on their genomic proximity to protein-coding genes, including overlapping, *cis*-antisense, bidirectional or intronic ncRNAs. In reality many transcripts resist classification into any particular category, and instead exhibit a combination of these qualities. Other unusual species of long ncRNAs, such as *trans*-spliced transcripts, macroRNAs that encompass huge genomic distances and multigene transcripts that encompass several genes or even the whole chromosome, further confound efforts for systematic classification[2,3].

Widespread functionality of long ncRNAs

Given their unexpected abundance, long ncRNAs were initially thought to be spurious transcriptional noise resulting from low RNA polymerase fidelity[7]. However, the expression of many long ncRNAs is restricted to particular developmental contexts[8], and large numbers of mouse ncRNAs are specifically expressed during embryonic stem cell differentiation[9] and in the brain, often exhibiting precise subcellular localization[10]. The binding of transcription factors to non-coding loci, together with evidence of purifying selection acting on ncRNA promoters, suggests that this type of expression is explicitly regulated[1,11,12].

Nevertheless, despite such signatures of functionality, the generally low sequence conservation of long ncRNAs has fuelled the assertion that they are not functional. However, this conclusion needs to be carefully considered. First, it ignores many examples that are conserved, and a recent study ascribes functional roles to a high proportion of such ncRNAs[13]. Second, long ncRNAs are likely to exhibit different patterns of conservation to protein-coding genes, which are subject to strict functional constraints and must preserve an ORF. By contrast, long ncRNAs can exhibit shorter stretches of sequence that are conserved to maintain functional domains and structures. Indeed, many long ncRNAs with a known function, such as *Xist*, only exhibit high conservation over short sections of their length[14]. Third, rather than being indicative of non-functionality, low sequence conservation can also be explained by high rates of primary sequence evolution if long ncRNAs have, like promoters and other regulatory elements, more plastic structure–function constraints than proteins[14]. Many conserved regions of the human genome that have been subject to recent and rapid evolutionary change are transcribed into long ncRNAs, including *HAR1*, a ncRNA expressed in Cajal–Retzius neurons in the developing neocortex[15]. Moreover, the adaptive radiation

of non-coding (that is, regulatory) sequences is likely to specify most of the phenotypic differences between, and within, species[16].

Low sequence conservation of long ncRNAs also prompted the alternative suggestion that it is the process rather than the product of transcription that is functional. For example, the cascading transcription of ncRNAs across the fructose bisphosphate *fbp1*$^{+}$ promoter in yeast is associated with the progressive opening of chromatin, thereby increasing access to transcriptional activators and RNA polymerase[17]. However, genome-wide evidence of conserved secondary structure[18], splicing patterns[12] and subcellular localization[10] suggest that a significant portion of ncRNAs fulfil functional roles beyond transcriptional remodelling. The diverse selection pressures acting on long ncRNAs probably reflect the wide range of their functions and their relative importance.

Functions of long ncRNAs

Unlike microRNAs or proteins, ncRNA function cannot currently be inferred from sequence or structure, with the diversity of long ncRNAs described to date precluding simple generalizations. The broad functional repertoire of long ncRNAs includes roles in high-order chromosomal dynamics, telomere biology and subcellular structural organization[8]. One major emergent theme is the involvement of these ncRNAs in regulating the expression of neighbouring protein-coding genes. The importance of this localized regulation was foreshadowed by the phenomenon of 'transvection', in which non-coding loci affect the expression of nearby protein-coding genes in *trans*[19]. Additionally, the recent observation that human chromosome 21 largely recapitulates its native expression profile in mouse cells, despite interspecies differences in epigenetic machinery, cellular environment and transcription factors, suggests that most of the information required for gene regulation is embedded in the chromosome sequence[20]. In the following sections we focus on the ability of long ncRNAs to regulate gene expression at the level of chromatin modification, transcription and post-transcriptional processing.

Chromatin modification. Long ncRNAs can mediate epigenetic changes by recruiting chromatin remodelling complexes to specific genomic loci. For example, hundreds of long ncRNAs are sequentially expressed along the temporal and spatial developmental axes of the human homeobox (Hox) loci, where they define chromatin domains of differential histone methylation and RNA polymerase accessibility[21]. One of these ncRNAs, Hox transcript antisense RNA (*HOTAIR*), originates from the *HOXC* locus and silences transcription across 40 kb of the *HOXD* locus in *trans* by inducing a repressive chromatin state, which is proposed to occur by recruitment of the Polycomb chromatin remodelling complex PRC2 by *HOTAIR*[21] (FIG. 2a). This model fits other chromatin modifying complexes, such as MLL, PcG, and G9a methyltransferase, which can be similarly directed by their associated ncRNAs[9,22–24]. Such a mechanism might resolve the paradox of how a small repertoire of chromatin remodelling complexes, which often have RNA binding domains but little DNA sequence specificity, are able to specify the complex array of chromatin modifications that are apparent throughout development.

Although the model of recruitment of chromatin modifying complexes by ncRNAs has been informative in our understanding of epigenetic phenomenon such as imprinting, it is probably only part of the story. X chromosome inactivation is mediated by the iconic long ncRNA, *Xist*. A small internal non-coding transcript from the *Xist* locus, *RepA*, recruits PRC2 to silence one X chromosome[25], whereas PRC2 is titrated from the remaining active X chromosome by the antisense transcript *Tsix*. However, another study describes an alternative mechanism whereby *Xist* and *Tsix* anneal to form an RNA duplex that is processed by Dicer to generate small interfering RNAs (siRNAs), which are required for the repressive chromatin modifications on the inactive X chromosome[26]. The contribution of these two different pathways to coordinate long and small RNAs in chromatin remodelling infers the existence of a global, integrated regulatory network based in RNA.

Transcriptional regulation. The pervasive transcription of enhancers[27] and promoters[28] anticipates a core role for long ncRNA in regulating the process of transcription. The means by which such ncRNAs regulate transcription are expanding to encompass a diversity of mechanisms, as shown by the following examples.

Proximal promoters can be transcribed into long ncRNAs that recruit and integrate the functions of RNA binding proteins into the transcriptional programme, as exemplified by the repression of cyclin D1 transcription in human cell lines[29]. DNA damage signals induce the expression of long ncRNAs associated with the cyclin D1 gene promoter, where they act cooperatively to modulate the activities of the RNA binding protein TLS. TLS subsequently inhibits the histone acetyltransferase activities of CREB binding protein and p300 to silence cyclin D1 expression (FIG. 2b). The ability of

Box 1 | Parsing coding and non-coding transcripts

Coding and non-coding RNAs (ncRNAs) can be difficult to distinguish. In eukaryotes, a protein-coding transcript is commonly defined by the presence of an ORF greater than 100 amino acids. However, a long ncRNA might contain such an ORF by chance alone, and many well-characterized long ncRNAs do indeed contain long ORFs. Reciprocally, proteins smaller than 100 amino acids might also be translated, with functional peptides as small as 11 amino acids being reported in *Drosophila* species[42]. The observation that selection favours synonymous over non-synonymous mutations to preserve codon usage has been exploited to help distinguish between transcripts with true rather than spurious ORFs[43]. Nevertheless, despite such improvements in the annotation of transcripts in recent years, we still lack a satisfactory definition of ncRNAs and there remain many ambiguous transcripts that exhibit both coding and non-coding traits. This might ultimately reflect the likelihood that the genome has evolved to encode a continuous spectrum of transcripts and information with little regard for our arbitrary definitions of coding and non-coding transcripts[44].

The extensive overlapping of alternatively spliced coding and non-coding isoforms further confounds the problem of distinguishing coding and non-coding transcripts, and indeed there might be a false dichotomy between them. One of the first and best characterized long ncRNAs, *SRA*, was later found to also encode a protein that acts antagonistically to the function of the ncRNA[45]. Reciprocally, many mRNAs can also function at an RNA level. For example, the *p53* mRNA acts intrinsically as an RNA regulator by binding the Mdm2 protein, which in turn induces *p53* expression and function[46]. A synonymous 'silent' mutation interferes with this process, and other similar silent mutations that affect protein translation are prevalent throughout the genome[47]. Moreover, 3′ UTRs of mRNAs can be expressed separately from the associated mRNA[48] and can impart functional information independently of the encoded protein[49,50]. Together these studies indicate that transcripts can potentially function both at an RNA level and to encode protein.

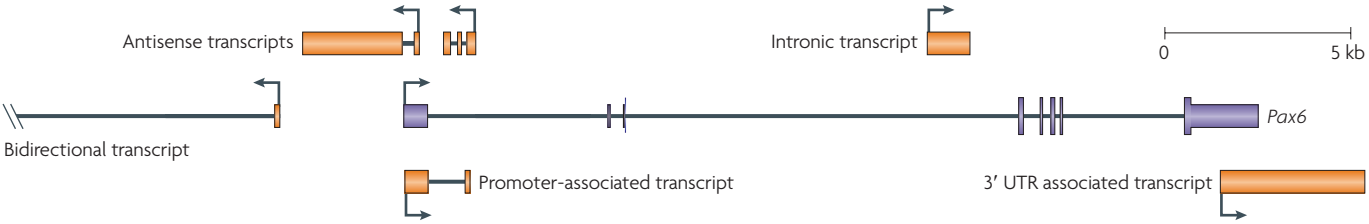


Figure 1 | **Genomic organization of coding and non-coding transcripts.** Schematic diagram illustrating the complexity of the interleaved networks of long non-coding transcripts (orange) that are associated with paired box gene 6 (*Pax6*; purple).

ncRNAs to recruit RNA binding proteins, one of the largest protein classes in the mammalian proteome, to gene promoters hugely expands the regulatory repertoire available to the transcriptional programme[29].

Long ncRNAs also act as co-factors to modulate transcription factor activity. For example, in mice, the ncRNA *Evf2* is transcribed from an ultraconserved distal enhancer and recruits the binding and action of the transcription factor DLX2 to this same enhancer to induce expression of adjacent protein-coding genes[30] (FIG. 2c). Many similar enhancers are transcribed in cells in which they are active — this could be a general strategy for regulating the expression of key developmental genes[27].

Long ncRNAs can regulate RNA polymerase (RNAP) II activity through other mechanisms, including by interaction with the initiation complex to influence promoter choice. For example, in humans, a ncRNA transcribed from an upstream region of the dihydrofolate reductase (*DHFR*) locus forms a triplex in the major promoter of *DHFR* to prevent the binding of the transcriptional co-factor TFIID[31] (FIG. 2d). This could be a widespread mechanism for controlling promoter usage as thousands of triplex structures exist in eukaryotic chromosomes[32].

Long ncRNAs can also effect global changes by interacting with basal components of the RNAP II-dependent transcription machinery. ncRNAs that interact with RNAP II machinery are typically transcribed by RNAP III, thereby decoupling their expression from the RNAP II-dependent transcription reaction they regulate. For example, *Alu* elements that are transcribed in response to heat shock bind tightly to RNAP II to preclude the formation of active preinitiation complexes[33]. *Alu* elements contain modular domains that can independently mediate polymerase binding and repression. In light of their abundance and distribution in the mammalian genome, these functional domains might have been co-opted into other ncRNAs during evolution; an observation supported by the finding that functional repeat sequence domains are a common characteristic of several known long ncRNAs[8].

Post-transcriptional regulation. The ability of ncRNAs to recognize complementary sequences also allows highly specific interactions that are amenable to regulating various steps in the post-transcriptional processing of mRNAs, including their splicing, editing, transport, translation and degradation. Most mammalian genes express antisense transcripts, which might constitute a class of ncRNA that is particularly adept at regulating mRNA dynamics[34].

Antisense ncRNAs can mask key *cis*-elements in mRNA by the formation of RNA duplexes, as in the case of the *Zeb2* (also called *Sip1*) antisense RNA, which complements the 5′ splice site of an intron in the 5′ UTR of the zinc finger Hox mRNA *Zeb2* (REF. 35). Expression of the ncRNA prevents the splicing of an intron that contains an internal ribosome entry site required for efficient translation and expression of the ZEB2 protein (FIG. 2e). This sets a precedent for ncRNAs in directing the alternative splicing of mRNA isoforms. Indeed, a number of studies have noted the prevalence of ncRNAs antisense to introns, and they could similarly regulate splicing[34].

Alternatively, the annealing of ncRNA can target protein effector complexes to the sense mRNA transcript in a manner analogous to the targeting of the RNA-induced silencing complex (RISC) to mRNAs by siRNAs. RNA duplexes resulting from the annealing of complementary transcripts or even of long ncRNAs with extended internal hairpins can be processed into endogenous siRNAs to silence gene expression, raising the possibility that many long ncRNAs feed into RNA silencing pathways[26].

There are probably many other functions of long ncRNAs awaiting discovery. For example, the ncRNA *NRON* has been shown to regulate the nuclear trafficking of the transcription factor NFAT[36], and the observation that many long ncRNAs are located in the cytoplasm[4] suggests that they might have undiscovered roles in cell biology.

Medical significance

There is increasing interest in the potential involvement of ncRNAs in disease aetiology, owing to aberrant function of ncRNAs in differentiation and developmental processes. The ability of ncRNAs to regulate associated protein-coding genes might contribute

Glossary

Adaptive radiation
Evolution of new morphological or functional characteristics in lineages that diversify in response to environmental changes or to enable colonization of new ecological niches.

Epigenetic
Heritable changes in phenotype caused by mechanisms outside of the genomic sequence. Such changes might remain through cell divisions during, for example, cellular differentiation, or they might persist through subsequent generations. Epigenetic changes include chromatin modifications, such as histone acetylation, or chemical alterations to the DNA itself, such as DNA methylation.

Long ncRNA
Transcripts longer than 200 nucleotides that have little or no protein-coding capacity. Long ncRNAs can regulate gene expression through a diversity of mechanisms.

MicroRNA
Single-stranded RNAs of approximately 21–23 nucleotides that regulate gene expression by partial complementary base pairing to specific mRNAs. This annealing inhibits protein translation and can also facilitate degradation of the target mRNA.

Transvection
Apparent cross-talk between alleles on homologous chromosomes, in which complementation is observed between promoter mutations in one allele and structural mutations in the other. Transvection can cause either gene activation or repression.

X chromosome inactivation
A process in which one of the two copies of the X chromosomes in female mammals is inactivated. X inactivation occurs so that females produce the same dosage of gene products from the X chromosome as males.

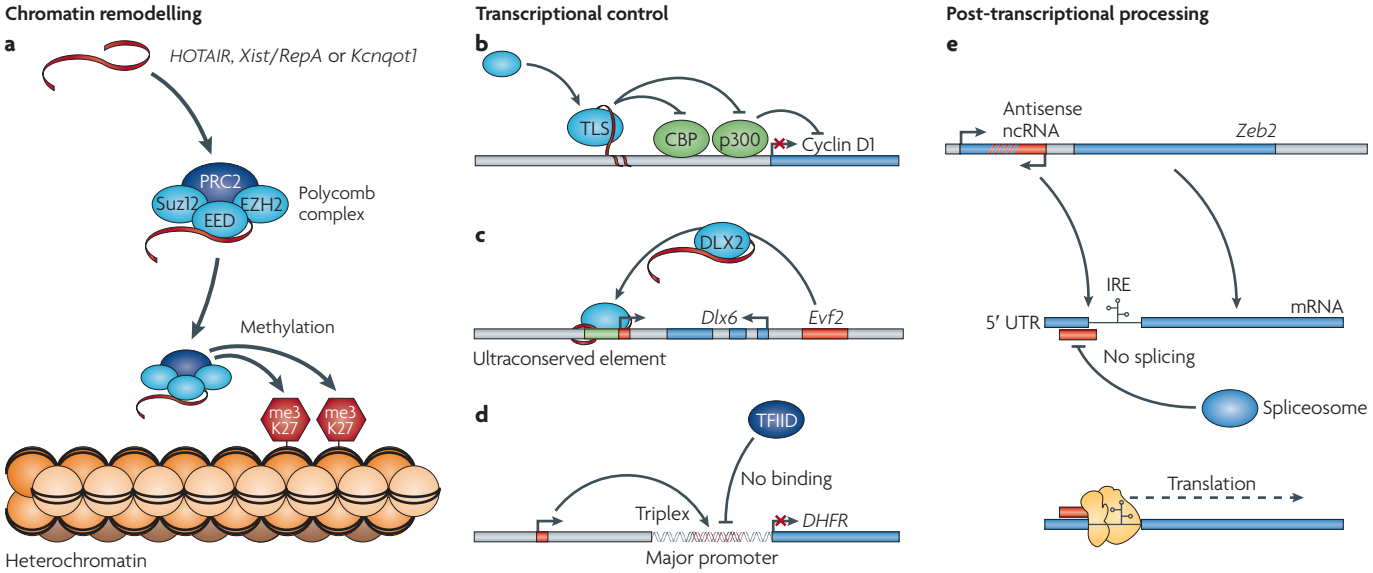


Figure 2 | **Functions of long non-coding RNAs (ncRNAs).** Illustrative mechanisms by which long ncRNAs regulate local protein-coding gene expression at the level of chromatin remodelling, transcriptional control and post-transcriptional processing. **a** | ncRNAs can recruit chromatin modifying complexes to specific genomic loci to impart their catalytic activity. In this case, the ncRNAs *HOTAIR*[21], *Xist* and *RepA* (the small internal non-coding transcript from the *Xist* locus)[25], or *Kcnqot1* (REF. 24) recruit the Polycomb complex to the *HoxD* locus, the X chromosome, or the *Kcnq1* domain, respectively, where they trimethylate lysine 27 residues (me3K27) of histone H3 to induce heterochromatin formation and repress gene expression. **b** | ncRNAs can regulate the transcriptional process through a range of mechanisms. ncRNAs tethered to the cyclin D1 gene recruit the RNA binding protein TLS to modulate the histone acetyltransferase activity of CREB binding protein (CBP) and p300 to repress gene transcription[29]. **c** | An ultraconserved enhancer is transcribed as a long ncRNA, *Evf2*, which subsequently acts as a co-activator to the transcription factor DLX2, to regulate the *Dlx6* gene transcription[30]. **d** | A ncRNA transcribed from the *DHFR* minor promoter in humans can form a triplex at the major promoter to occlude the binding of the general transcription factor TFIID, and thereby silence *DHFR* gene expression[31]. **e** | An antisense ncRNA can mask the 5′ splice site of the zinc finger homeobox mRNA *Zeb2* from the spliceosome, resulting in intron retention. The translation machinery can then recognize and bind an internal ribosome entry site (IRE) in the retained intron, resulting in efficient *Zeb2* translation and expression[35].

towards disease if their misexpression deregulates a gene of clinical significance. For example, an antisense ncRNA transcribed from the *p15* tumour suppressor locus induces changes to local heterochromatin and DNA methylation status, thereby regulating *p15* expression[37], and is potentially involved in oncogenesis as the antisense ncRNA and protein have inverse expression profiles in leukaemia. Many other tumour suppressor genes that are frequently silenced by epigenetic mechanisms in cancer also have antisense partners[37].

An appreciation of long ncRNAs might inform a reinterpretation of the functional basis of many disease-associated polymorphisms and chromosomal alterations that occur in non-coding regions[38]. For example, the translocation and induced expression of an antisense long non-coding transcript causes the epigenetic silencing of an adjacent α-globin gene, resulting in α-thalassaemia[39]. The role of ncRNAs in disease is likely to have been overlooked in genetic screens because of the subtlety of their effects and the emphasis to date, both intellectually and practically, on mutation scanning of protein-coding exons. Moreover, the complexity of non-coding transcription can make understanding the specific contribution of embedded polymorphisms a bewildering exercise. For example, a SNP that occurs both in the 3′ UTR of the zinc finger gene *ZFAT* and also in the promoter of an antisense transcript increases the expression of *ZFAT* — not through increasing mRNA stability, but by repressing the expression of the antisense transcript[40].

Conclusion

Continuing advances in transcriptomics are resulting in ncRNA being recognized as an important functional expression of the genome. Rather than being reduced to a simple messenger role, it seems likely that the sophisticated structural and informational capacity of RNA has continued to be harnessed by evolution in a range of biological roles that interact with, but are distinct from, protein and DNA. The concomitant increase in non-coding content with organismal complexity supports the proposition that evolutionary innovations and expansion of regulatory RNAs were fundamental to the genetic programming of complex eukaryotes[41].

There are still huge gaps in our understanding of long ncRNAs, including the proportion that is functional and the range and mechanistic basis of their functions. What is required, and is currently developing, is the application of genome-wide techniques to reveal the full extent of ncRNA expression. Unbiased techniques, such as next-generation sequencing, have the benefit of not being constrained by current protein-centric annotations. Such data will progressively build a catalogue of ncRNAs with common characteristics that will aid in the identification and prediction of functional features, complemented by experimental analyses of individual examples to determine the mechanisms by which long ncRNAs act. This will increasingly involve the intersection of techniques from other fields, such as live cell RNA imaging, RNA proteomics (that is, the analysis of RNA-associated protein complexes) and RNA structural biology.

Long ncRNAs have the potential to rival the functional repertoire of the proteome, albeit with a different spectrum. It is already apparent that any RNA, regardless of protein-coding capacity, might have its own intrinsic messages to deliver (BOX 1),

suggesting that the genome might encompass an RNA-based information suite that is far more sophisticated than expected. Should the majority of ncRNAs prove to be functional, their characterization will have a considerable impact on our understanding of the genetic programming of complex organisms and will bring new answers to old questions in evolution, development and the understanding of disease.

Tim R. Mercer, Marcel E. Dinger and John S. Mattick are at the Australian Research Council Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland 4072, Australia.

T.R.M. and M.E.D. contributed equally to this work

Correspondence to J.S.M.
e-mail: j.mattick@imb.uq.edu.au

doi:10.1038/nrg2521
Published online 3 February 2009

1. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
2. Denoeud, F. *et al.* Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res.* **17**, 746–759 (2007).
3. Horiuchi, T. & Aigaki, T. Alternative *trans*-splicing: a novel mode of pre-mRNA processing. *Biol. Cell* **98**, 135–140 (2006).
4. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).
5. Kapranov, P. *et al.* Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* **15**, 987–997 (2005).
6. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
7. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Struct. Mol. Biol.* **14**, 103–105 (2007).
8. Amaral, P. P. & Mattick, J. S. Noncoding RNA in development. *Mamm. Genome* **19**, 454–492 (2008).
9. Dinger, M. E. *et al.* Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* **18**, 1433–1445 (2008).
10. Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F. & Mattick, J. S. Specific expression of long noncoding RNAs in the adult mouse brain. *Proc. Natl Acad. Sci. USA* **105**, 716–721 (2008).
11. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).
12. Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* **17**, 556–565 (2007).
13. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* (in the press).
14. Pang, K. C., Frith, M. C. & Mattick, J. S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* **22**, 1–5 (2006).
15. Pollard, K. S. *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**, 167–172 (2006).
16. Carroll, S. B. Evo–devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).
17. Hirota, K. *et al.* Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. *Nature* **456**, 130–134 (2008).
18. Torarinsson, E., Sawera, M., Havgaard, J. H., Fredholm, M. & Gorodkin, J. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.* **16**, 885–889 (2006).
19. Mattick, J. S. & Gagen, M. J. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.* **18**, 1611–1630 (2001).
20. Wilson, M. D. *et al.* Species-specific transcription in mice carrying human chromosome 21. *Science* **322**, 434–438 (2008).
21. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human *HOX* loci by noncoding RNAs. *Cell* **129**, 1311–1323 (2007).
22. Morris, K. V., Santoso, S., Turner, A. M., Pastori, C. & Hawkins, P. G. Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. *PLoS Genet.* **4**, e1000258 (2008).
23. Nagano, T. *et al.* The *Air* noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322**, 1717–1720 (2008).
24. Pandey, R. R. *et al. Kcnq1ot1* antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol. Cell* **32**, 232–246 (2008).
25. Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**, 750–756 (2008).
26. Ogawa, Y., Sun, B. K. & Lee, J. T. Intersection of the RNA interference and X-inactivation pathways. *Science* **320**, 1336–1341 (2008).
27. Ashe, H. L., Monks, J., Wijgerde, M., Fraser, P. & Proudfoot, N. J. Intergenic transcription and transinduction of the human β-globin locus. *Genes Dev.* **11**, 2494–2509 (1997).
28. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77–88 (2007).
29. Wang, X. *et al.* Induced ncRNAs allosterically modify RNA-binding proteins in *cis* to inhibit transcription. *Nature* **454**, 126–130 (2008).
30. Feng, J. *et al.* The *Evf-2* noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev.* **20**, 1470–1484 (2006).
31. Martianov, I., Ramadass, A., Serra Barros, A., Chow, N. & Akoulitchev, A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445**, 666–670 (2007).
32. Ohno, M., Fukagawa, T., Lee, J. S. & Ikemura, T. Triplex-forming DNAs in the human interphase nucleus visualized *in situ* by polypurine/polypyrimidine DNA probes and antitriplex antibodies. *Chromosoma* **111**, 201–213 (2002).
33. Mariner, P. D. *et al.* Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol. Cell* **29**, 499–509 (2008).
34. He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N. & Kinzler, K. W. The antisense transcriptomes of human cells. *Science* **322**, 1855–1857 (2008).
35. Beltran, M. *et al.* A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial–mesenchymal transition. *Genes Dev.* **22**, 756–769 (2008).
36. Willingham, A. T. *et al.* A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309**, 1570–1573 (2005).
37. Yu, W. *et al.* Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* **451**, 202–206 (2008).
38. Hindorff, L. A., Junkins, H. A. & Manolio, T. A. A catalog of published genome-wide association studies. *National Human Genome Research Institute* [online], <http://www.genome.gov/gwastudies> (2008).
39. Tufarelli, C. *et al.* Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nature Genet.* **34**, 157–165 (2003).
40. Shirasawa, S. *et al.* SNPs in the promoter of a B cell-specific antisense transcript, *SAS-ZFAT*, determine susceptibility to autoimmune thyroid disease. *Hum. Mol. Genet.* **13**, 2221–2231 (2004).
41. Taft, R. J., Pheasant, M. & Mattick, J. S. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* **29**, 288–299 (2007).
42. Kondo, T. *et al.* Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nature Cell Biol.* **9**, 660–665 (2007).
43. Lin, R., Maeda, S., Liu, C., Karin, M. & Edgington, T. S. A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. *Oncogene* **26**, 851–858 (2007).
44. Dinger, M. E., Pang, K. C., Mercer, T. R. & Mattick, J. S. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.* **4**, e1000176 (2008).
45. Chooniedass-Kothari, S. *et al.* The steroid receptor RNA activator is the first functional RNA encoding a protein. *FEBS Lett.* **566**, 43–47 (2004).
46. Candeias, M. M. *et al. p53* mRNA controls p53 activity by managing Mdm2 functions. *Nature Cell Biol.* **10**, 1098–1105 (2008).
47. Komar, A. A. Silent SNPs: impact on gene function and phenotype. *Pharmacogenomics* **8**, 1075–1080 (2007).
48. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38**, 626–635 (2006).
49. Jenny, A. *et al.* A translation-independent role of *oskar* RNA in early *Drosophila* oogenesis. *Development* **133**, 2827–2833 (2006).
50. Rastinejad, F. & Blau, H. M. Genetic complementation reveals a novel regulatory role for 3′ untranslated regions in growth and differentiation. *Cell* **72**, 903–917 (1993).

Acknowledgements
We thank P. Amaral and other laboratory colleagues for many discussions related to this article, and the Australian Research Council for financial support. We apologize both to readers and colleagues for references that were omitted owing to editorial constraint.

FURTHER INFORMATION
Mattick laboratory web site:
http://jsm-research.imb.uq.edu.au
CPC (coding potential calculator): http://cpc.cbi.pku.edu.cn
NRED (ncRNA expression database):
http://jsm-research.imb.uq.edu.au/NRED
RNAdb (RNA database):
http://research.imb.uq.edu.au/RNAdb

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Nucleosome positioning and gene regulation: advances through genomics

Cizhong Jiang and B. Franklin Pugh

Abstract | Knowing the precise locations of nucleosomes in a genome is key to understanding how genes are regulated. Recent 'next generation' ChIP–chip and ChIP–Seq technologies have accelerated our understanding of the basic principles of chromatin organization. Here we discuss what high-resolution genome-wide maps of nucleosome positions have taught us about how nucleosome positioning demarcates promoter regions and transcriptional start sites, and how the composition and structure of promoter nucleosomes facilitate or inhibit transcription. A detailed picture is starting to emerge of how diverse factors, including underlying DNA sequences and chromatin remodelling complexes, influence nucleosome positioning.

Chromatin remodelling complex
An ATP-dependent enzyme that is catalysed by different types of ATPase to alter nucleosome structure. The net effect of all chromatin remodelling enzymes is to modify nucleosome position or to increase accessibility of nucleosomal DNA.

Nucleosome-free region
(NFR). An ~140 bp region lacking nucleosomes that is found at the beginning and end of genes. Many regions might not be completely nucleosome free, but are depleted of nucleosomes compared with the surrounding region. Certain environmental conditions can cause nucleosomes to occupy an NFR; for example, when genes are repressed.

Center for Eukaryotic Gene Regulation, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA. Correspondence to B.F.P.
e-mail: bfp2@psu.edu
doi:10.1038/nrg2522
Published online
10 February 2009

The genetic code resides within a negatively charged DNA polymer. The resulting electrostatic repulsion from neighbouring phosphates stiffens the polymer such that it cannot fit within the small confines of a nucleus. A solution to this problem has evolved in the form of highly basic histone proteins that bind to DNA and neutralize the negative charges. The formation of chromatin through the binding of histones to DNA allows the DNA to be folded into chromosomes and compacted by as much as a factor of 10,000. The packaging of DNA creates both a problem and an opportunity: wrapping DNA around histones potentially obstructs access to the genetic code; however, the ubiquity of the histones that are bound at all regions of chromosomal DNA can be exploited so that enzymes that read, replicate and repair DNA can be directed to the appropriate entry sites. In this way, RNA polymerase (Pol) II initiates transcription at the beginning of genes rather than in the middle, DNA polymerase initiates replication at replication origins and DNA repair enzymes are directed to sites of DNA damage.

How does the cell package a helical DNA polymer in a way that is both refractory and accessible? The evolutionary solution to the packaging problem is the nucleosome[1,2] (FIG. 1a). The nucleosome is the basic unit of eukaryotic chromatin, consisting of a histone core around which DNA is wrapped. Each histone core is composed of two copies of each of the histone proteins H2A, H2B, H3 and H4 (FIG. 1b). Approximately 147 bp of DNA coils 1.65 times around the histone octamer in a left-handed toroid[2]. Amino-terminal histone 'tails' emanate from the nucleosome core, past the DNA. The polypeptide chains of the histone tails are subject to covalent modifications, including acetylation and methylation. At active genes or at genes that are poised for activation, histones H2A and H3 are replaced by the histone variants H2A.Z and H3.3 (see REFS 3,4 for reviews of histone variants). Beyond the nucleosome core is the linker histone, H1. Nucleosomes are arranged as a linear array along the DNA polymer as 'beads on a string'. This structure can be further compacted by H1 into higher-order transcriptionally inactive 30 nm fibres.

The combination of nucleosome positions and their chemical and compositional modifications are key to genome regulation. In this Review, we focus specifically on nucleosome positioning rather than on histone modifications and variants. Here we interrelate past and recent developments in our understanding of the basic organization of nucleosomes on chromosomes, and show how DNA sequences and chromatin remodelling complexes selectively position and organize nucleosomes so that they can regulate genomic function. Importantly, massively parallel DNA sequencing and microarray hybridization technologies have allowed the location of every nucleosome across a genome to be determined with unprecedented accuracy (BOX 1). We discuss how these maps reveal a common organizational theme at nearly every gene, including a nucleosome-free region (NFR) at the beginning and end of genes. We also discuss how the underlying DNA sequence and the action of chromatin remodelling complexes influence where nucleosomes are positioned. There is emerging evidence that nucleosomes regulate transcriptional initiation, and therefore understanding how nucleosomes are positioned has implications for how cells respond to

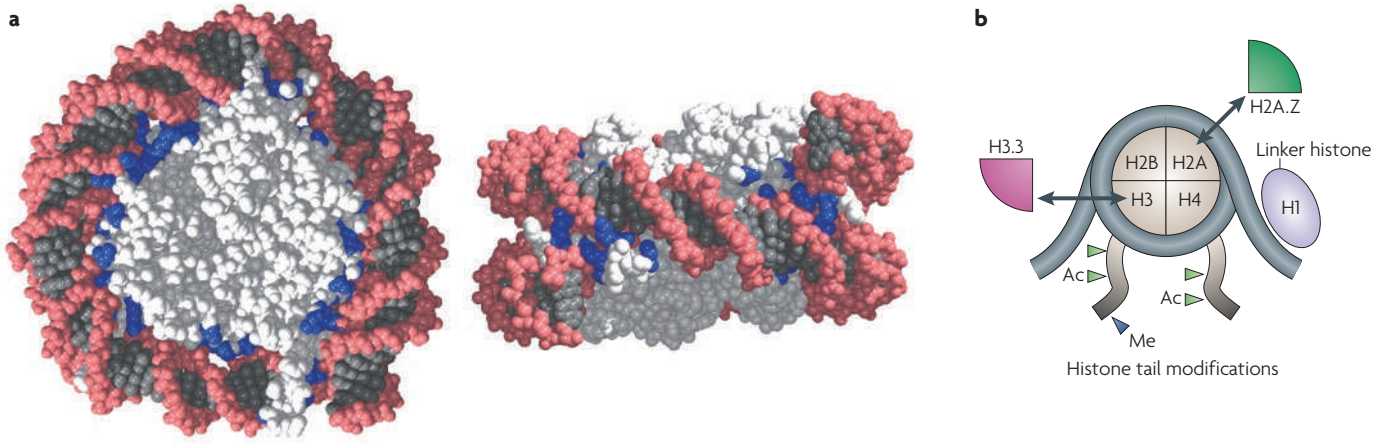


Figure 1 | **Nucleosome structure. a** | Structure of a nucleosome core particle (front and side view)[2,131]. Histones are shown in light grey, and the DNA helix is shown in dark grey with a pink backbone. Basic amino acids (lysine and arginine) within 7 Å of the DNA are shown in blue to emphasize the electrostatic contacts between the DNA phosphates and the histones. **b** | A schematic of DNA wrapped around a nucleosome. Examples of histone tail modifications (Ac, acetylation; Me, methylation) and histone variants (H2A.Z and H3.3) are shown. Arrows indicate the replacement of canonical histones with histone variants. Part **a** courtesy of S. Tan, Pennsylvania State University, USA.

external stimuli or how misregulation of nucleosome positioning leads to developmental defects and cancer.

Genomic organization of nucleosomes

Until recently, it was unclear whether deposition of histones on DNA during DNA replication occurs at random positions. Random deposition implies that nucleosomes lack positional cues and that histones are simply DNA packaging proteins that are removed and redeposited as DNA and RNA polymerases pass through them. Alternatively, individually positioned nucleosomes could take on specific physiological functions depending on where they reside in the genome. In this section, we will discuss how cells use both random deposition and specific positioning of histones to organize nucleosomes. This understanding has arisen through the development of technologies that have allowed genome-wide mapping of nucleosome positioning; we start by describing this progress then we discuss the genomic properties of nucleosomes.

A brief history of nucleosome cartography. In 2004, the exact genomic location of only a few hundred nucleosomes was known because techniques were limited to the individual interrogation of specific genomic loci. The early development of microarrays, which consisted of ~500–2,000 bp DNA probes that spanned each genic and intergenic region, provided a comprehensive view of the nucleosome landscape across the simple genome of the budding yeast *Saccharomyces cerevisiae*. The long (~1 kb) probe lengths of these early microarrays precluded the assessment of individual nucleosome states, which occur at <200 bp intervals. Nevertheless, for the first time, these pioneering studies showed a general depletion of nucleosomes in the intergenic regions where promoters are found[5–7]. It was initially unclear how the presence or absence of nucleosomes related to transcription, because the upstream intergenic regions of active and quiescent yeast genes were depleted of nucleosomes compared with other regions of the genome. However, higher-resolution nucleosome maps clarified that gene activation resulted in additional nucleosome depletion[5,7–11]. Furthermore, antibodies that were specific for individual histone post-translational modifications showed that the promoter regions of highly transcribed genes were particularly enriched with nucleosomes that contained acetylated and methylated histones[12–14]. The function of these modified histones remains an area of active research.

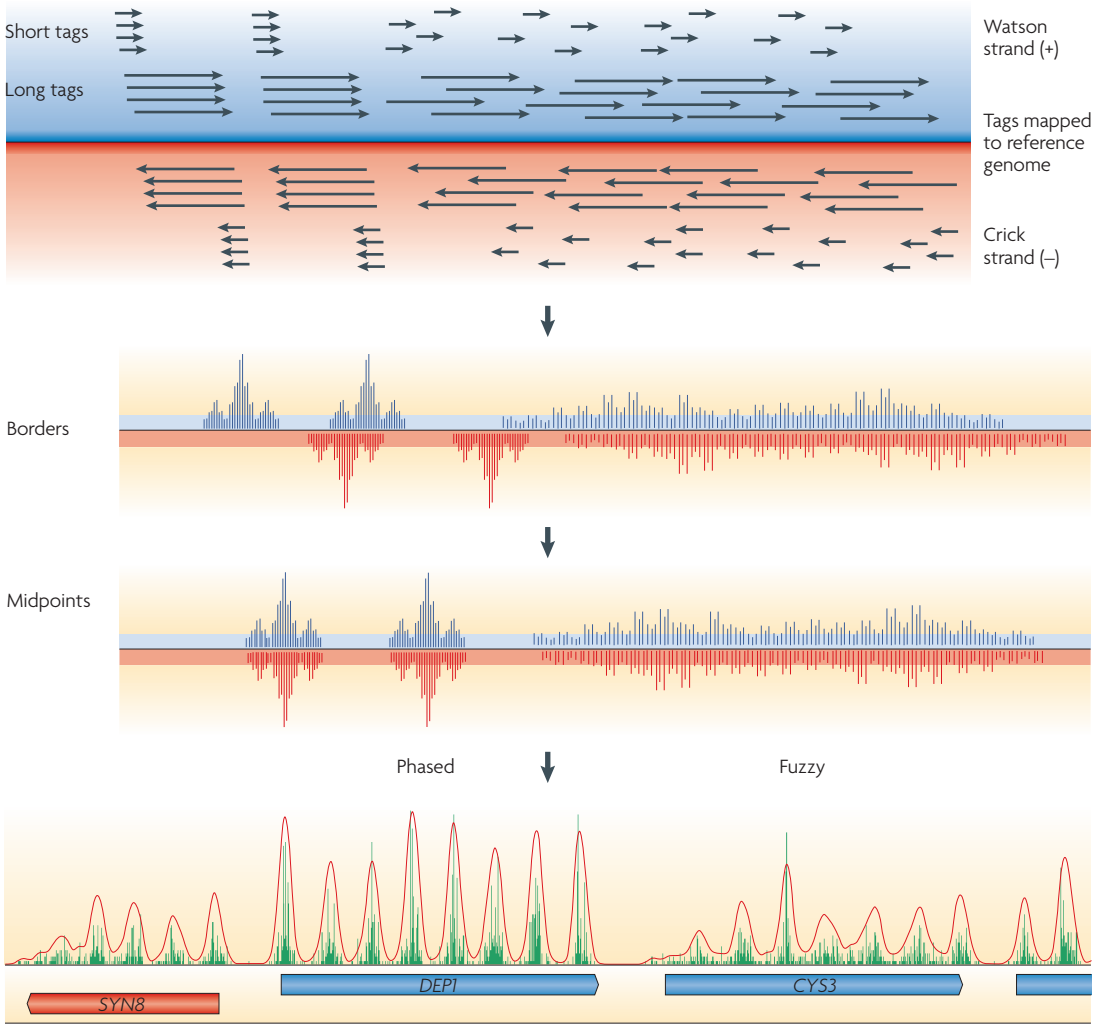
By 2005, microarrays had been developed that had shorter DNA probes and shorter probe–probe genomic distances, which provided higher resolution views of nucleosomes. However, printing technology limited the search space to small sections of small genomes. Nonetheless, these arrays showed that the nucleosomes at most genes are organized around the beginning of genes in basically the same way[15]: a NFR flanked by two well-positioned nucleosomes (the –1 and +1 nucleosomes), which is followed by a nucleosomal array that packages the gene (FIG. 2). This basic pattern has also held true for metazoans[16,17].

A complete and comprehensive map of nucleosome locations in a eukaryotic genome (*S. cerevisiae*) was completed in 2007 owing to two impressive technological advances. First, the densities of commercially printed probes on microarrays increased dramatically, allowing millions of genomic loci to be interrogated by ChIP–chip analysis in a single experiment. The genomic distances between probes were reduced to 5 bp in *S. cerevisiae* and 36 bp in *Drosophila melanogaster*[17,18]. Second, massively parallel shotgun sequencing allowed individual nucleosomal DNA molecules to be sequenced, initially at the level of hundreds of thousands of nucleosomes, and now at the level of tens of millions of nucleosomes (BOX 1). The first such ultra-high-resolution genome-wide ChIP–Seq nucleosome map was achieved for nucleosomes containing H2A.Z in *S. cerevisiae*[19]. The map showed the precise

ChIP–chip
A method for detecting the location of proteins throughout a genome using chromatin-immunoprecipitation followed by microarray analysis.

ChIP–Seq
A method for detecting the location of proteins throughout a genome using chromatin-immunoprecipitation followed by high-throughput DNA sequencing.

Box 1 | **ChIP–Seq nucleosome mapping technology**



A stringent procedure for high-resolution mapping of nucleosomes by ChIP–Seq involves an initial step to cross-link histones to nucleosomal DNA by formaldehyde treatment of living cells. In principle, cross-linking traps nucleosomes at their *in vivo* locations. Next, linker DNA is removed from isolated chromatin by digestion with high levels of micrococcal nuclease (MNase). Subsets of nucleosome particles are isolated by immunoprecipitation using antibodies directed against histones, histone variants or histone modifications. Fragments of mononucleosomal DNA that are ~150 bp long are size-selected by agarose gel electrophoresis. The 5′ ends of millions of individual DNA molecules in this library are then sequenced in parallel. Short-read technology sequences 25–35 bp fragments (called tags), whereas long-read technology produces read lengths of 100 bp or more. Sequence tags are then mapped to the reference genome using alignment algorithms (see the figure; black arrows represent the reads derived from long- and short-read technology) on either the Watson (blue) or Crick (red) strand. The 5′ ends of each tag, which correspond to nucleosome borders, are then plotted as a bar graph at each coordinate in the genome (as displayed in some publications[16,22]). Next, the tag location is adjusted to represent the nucleosome midpoint (typically +73 bp on the 'W' or '+' strand and –73 bp on the 'C' or '–' strand is added to the genomic coordinate of the 5′ end of each tag)[17,19,20]. Clusters of tags show a consensus nucleosome position. Two clusters are shown. The tighter the cluster, the more phased the corresponding nucleosome is. Randomly distributed tags reflect random ('fuzzy') positioning. A sample of tag distribution in a small section of the yeast genome is shown at the bottom of the figure, in which the red and blue tags are collapsed into a single bar graph. Each peak equates to a single consensus nucleosome position.

Three DNA sequencing technologies have been used to map nucleosomes[16,17,19,23,24]. Pyrosequencing using the Roche 454 GS20/FLX sequences nucleosomal DNA end-to-end, allowing both nucleosome borders to be linked in a single sequence. This method provides the greatest mapping accuracy, particularly in genomic regions of low complexity. By contrast, other platforms, such as those provided by the Illumina–Solexa Genome Analyzer and Applied Biosystems SOLiD, generate only 25–35 bp sequence tags, requiring both nucleosome borders to be inferred. Nonetheless, these short-read technologies produce >100 times the number of sequence tags at a similar cost as the long-read technologies, and so the short-read technology is currently the only practical technology for mapping nucleosomes in large genomes. The higher tag count of the short-read technology enhances mapping accuracy and thus provides a practical way of mapping nucleosomes.

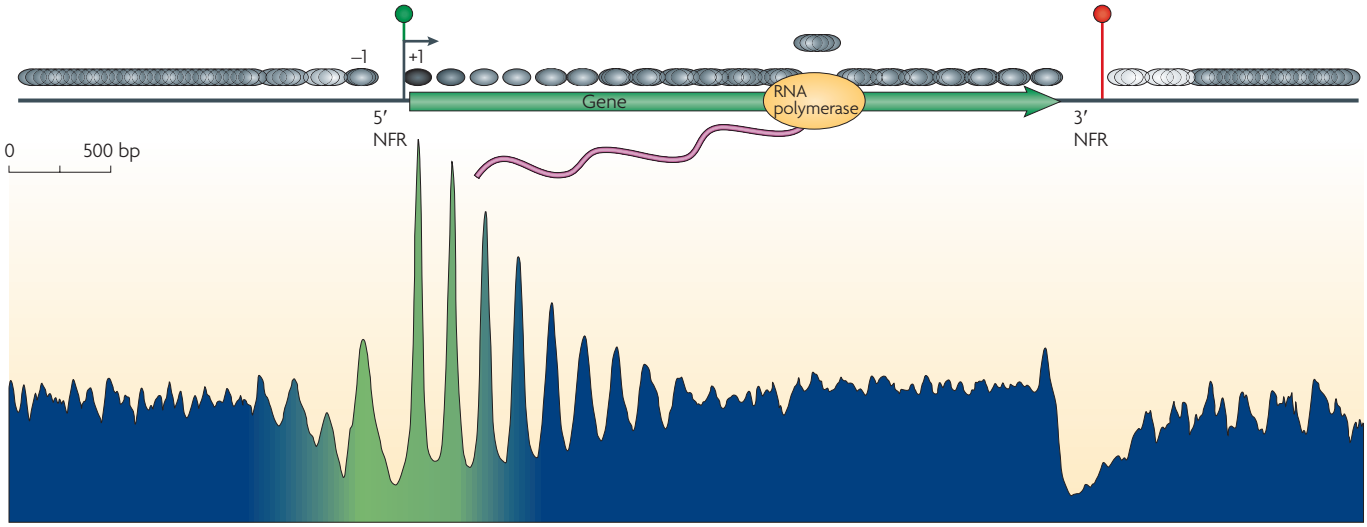


Figure 2 | **Nucleosomal landscape of yeast genes.** The consensus distribution of nucleosomes (grey ovals) around all yeast genes is shown, aligned by the beginning and end of every gene. The resulting two plots were fused in the genic region. The peaks and valleys represent similar positioning relative to the transcription start site (TSS). The arrow under the green circle near the 5′ nucleosome-free region (NFR) represents the TSS. The green –blue shading in the plot represents the transitions observed in nucleosome composition and phasing (green represents high H2A.Z levels, acetylation, H3K4 methylation and phasing, whereas blue represents low levels of these modifications). The red circle indicates transcriptional termination within the 3′ NFR. Figure is reproduced, with permission, from REF. 20 © (2008) Cold Spring Harbor Laboratory Press.

nucleosomal contexts in which gene regulatory elements function on a genomic scale. Nucleosome maps of a similar resolution in yeast, worms, flies and humans have now been published[16–18,20–24] and are likely to be produced for other model organisms soon. Future nucleosome mapping endeavours will probably focus on how nucleosome positions and histone modifications depend on cellular factors, and how they change in response to environmental signals, tissue differentiation and cellular disease states.

Lessons learned from global nucleosome maps. Genome-wide nucleosome maps allow us to explore the genomic properties of chromatin. At any given genomic locus, the preferential positioning of nucleosomes — called phasing — can be described (FIG. 3a). At most loci, there is an approximately Gaussian (normal) distribution of nucleosome positions around particular genomic coordinates, ranging from ~30 bp for highly phased nucleosomes to a random continuous distribution throughout an array. How much of this variation is due to genuine positional heterogeneity and how much is an artefact that is caused by overtrimming or undertrimming of the DNA at nucleosome borders by micrococcal nuclease during sample preparation remains to be determined.

Within each Gaussian distribution, nucleosomes have preferred positions; these positions tend to be about 10 bp apart[19] (FIG. 3b). This means that, owing to the helical nature of DNA, a DNA sequence will tend towards the same rotational setting (facing inwards or outwards) on the histone surface when a nucleosome is in alternative preferred positions (translational settings). This is important because the orientation of a DNA sequence on the histone surface determines the accessibility of its sequence and thus its activity (FIG. 3b).

Positioned nucleosomes tend to be spaced at a fixed distance from each other, with short stretches of linker DNA between them. The most common distance between adjacent nucleosome midpoints is approximately 165 bp (~18 bp linker) in *S. cerevisiae*[18,20,23], 175 bp (~28 bp linker) in *D. melanogaster*[17] and *Caenorhabditis elegans*[24], and 185 bp (~38 bp linker) in humans[16,22]. Chromatin remodelling or spacing complexes of the imitation switch (ISWI) class, such as ATP-dependent chromatin assembly and remodelling factor (ACF) and chromatin accessibility complex (CHRAC), establish nucleosome spacing[25–28]. These complexes bind nucleosomes and a finite amount of adjacent linker DNA, then use energy from ATP hydrolysis to move nucleosomes in the direction of the linker DNA[29–31]. As a result, the linker shortens until it can no longer bind the ISWI complex. Linker length is likely to be further constrained by the linker-binding histone H1 (REFS 32–34), which might reduce the amount of linker DNA that is available to the ISWI complexes. The different linker lengths in evolutionarily diverged eukaryotes might reflect the presence of evolutionarily divergent ISWI subunits or H1 proteins that have species-specific DNA length requirements for binding in these eukaryotes. Shorter linkers might result in a reduced availability of sequences for protein binding and thus these linkers might have regulatory functions. Very long linkers, or NFRs (~140 bp in length[17,20,22]), are present in the genome where a nucleosome seems to be missing or where the DNA is depleted of nucleosomes relative to the rest of the genome. As we will discuss in a later section, these NFRs are key to unlocking the mystery of how nucleosome organization and gene regulation are linked.

Phasing
The distribution of nucleosomes around a particular coordinate in a population of cells.

Rotational setting
The local orientation of the DNA helix on the histone surface.

Translational setting
The nucleosomal DNA midpoint position relative to a chromosomal locus.

Linker DNA
A short length of DNA located between nucleosomes. Long linker DNA can be considered to be a nucleosome-free region (NFR) — the DNA length cut-off for the two classes is arbitrary. However, NFRs tend to be sites of RNA and DNA polymerase loading and unloading.

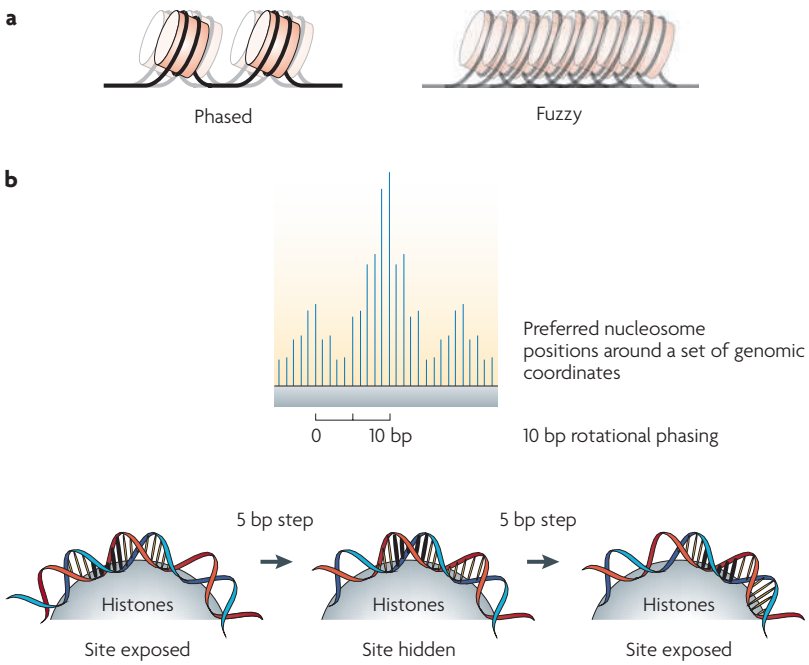


Figure 3 | **Phasing information and rotational setting. a** | In a population, individual nucleosomes are either positioned within a small range of a genomic locus (phased) or with a continuous distribution throughout an array (fuzzy). **b** | The bar graph is an idealized distribution of nucleosomal sequence tags, which form a large cluster and several subclusters, in which the subclusters are spaced about 10 bp apart and represent multiple translational settings with a single predominant rotational setting (see also BOX 1). Also shown is a schematic of alternative rotational settings of DNA and its effect on site accessibility (indicated by the black 'rungs' on the DNA helix).

The organization of nucleosomes on genes. The genome-wide maps of nucleosome location have also provided insights into the organization of nucleosomes around protein-coding genes. The *S. cerevisiae* genome provides the clearest example of a consensus pattern of organization (FIG. 2). The first predominant nucleosome located upstream of the transcription start site (TSS) (designated –1, see BOX 2) covers a region from –300 to –150 relative to the TSS, and can regulate the accessibility of promoter regulatory elements in that region. During a transcription cycle, the –1 nucleosome will experience many changes that affect its stability, including histone replacement, acetylation and methylation, as well as translational repositioning, and ultimately eviction after pre-initiation complex (PIC) formation. Whether the –1 nucleosome remains evicted during multiple rounds of transcription, or returns between each transcription cycle, remains an important unanswered question. The answer to this question would help elucidate whether reinitiation of transcription is mechanistically distinct from the initial activation event.

Downstream of the –1 nucleosome is a NFR (the 5′ NFR), then the TSS (discussed in a later section), which is followed by the +1 nucleosome. Of all the nucleosomes found in and around genes, the +1 nucleosome displays the tightest positioning (or phasing)[20]. The +1 nucleosome often contains histone variants (H2A.Z and H3.3)[35] and histone tail modifications (methylation and acetylation)[36–38], all of which might facilitate nucleosome eviction and PIC assembly. During transcription, the +1 nucleosome is likely to be evicted, but it seems to rapidly return to its original place after Pol II has passed, as it is only modestly depleted at highly transcribed genes[19]. The +2 nucleosome is found immediately downstream of the +1 nucleosome. It shares some properties with the +1 nucleosome but contains less H2A.Z, and displays less methylation, acetylation and phasing[38,39]. The +3 nucleosome and the more downstream nucleosomes each have less of these properties than the previous upstream nucleosome. The reduction in these properties might reflect a limitation in the functional distance of histone remodelling or modifying enzymes that are tethered to the 5′ end of genes.

Beyond ~1 kb from the TSS, consensus spacing from the TSS dissipates. Although phased nucleosomes are found, there is an increasing tendency for random nucleosome positions[15,20]. This might represent a loss in the functional constraints that are imposed on nucleosomes at the beginning of genes.

The array of nucleosomes that covers a gene terminates with a NFR at the 3′ end of the gene (the 3′ NFR). The 3′ NFR is the region at which Pol II terminates transcription, which is precipitated by the cleavage of the nascent RNA transcript near the 3′ end of the gene. Whether the nucleosome located at the end of the 3′ NFR contributes to termination is not known. Overall, these high-resolution genomic maps show that genes are packaged into a regular array of nucleosomes that starts at a fixed position from the TSS and are bracketed by nucleosome-free or nucleosome-depleted zones. In the next section, we discuss how this pattern might be set up.

Origins of nucleosome positions

So far, we have learned that nucleosomes adopt canonical positions around promoter regions and more random positions in the interior of genes. But how is this organization established? We describe one view using an analogy of a roulette wheel (an analogy of a parking lot is described elsewhere[40]). In a roulette wheel, the ball is allowed to land only in the designated slots (FIG. 4a). Regardless of how many balls are used, the possible positions of the balls are predetermined. Every positioned nucleosome could have an underlying DNA sequence structure (a 'slot') that favours positioning in that location. Randomly positioned nucleosomes would not be associated with any positioning sequence. This model implies that the positions of adjacent nucleosomes are independently controlled. An alternative possibility, called statistical positioning[41–44], arises from the close packing of nucleosomes into an array. The positioning of one nucleosome in the array (FIG. 4b, left side) forces the positioning of all other nucleosomes, because the tight packing restricts their lateral movement (this is termed probabilistic positioning, as indicated by the distribution trace in FIG. 4b). Thus a single genomic barrier can potentially position many nucleosomes without the need for individual positioning sequences. Below, we describe how a combination of both models might exist (FIG. 4c).

Pre-initiation complex (PIC). This assembly is found at the promoter and before the complex has initiated transcription. It includes the general transcription factors (TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH), the mediator, the RNA polymerase II complex, and activator or co-activator proteins (including SAGA).

Support vector machine classifier
A widely used method of classifying training data (for example, nucleosomal compared with non-nucleosomal genomic DNA), which can then be used to make predictions *de novo*.

Hidden Markov modelling
A method of identifying unknown or hidden states (for example, nucleosome positions) from observable states (for example, measured nucleosome positions).

Cryptic transcription
A low level of presumably unregulated transcription that originates from nucleosome-free regions. The transcripts are usually rapidly degraded.

DNA sequence patterns. The +1 nucleosome could provide the barrier for statistical positioning. So, are there DNA sequence patterns that are associated with well-positioned nucleosomes? The idea behind pattern searching is to align the 147 bp DNA sequence of thousands of well-positioned nucleosomes and determine whether particular base pair combinations are statistically enriched at particular positions along the DNA molecule. Such pattern searching began in the 1980s with a few hundred nucleosomal sequences, and showed that AA, TT and TA dinucleotides occurred at 10 bp intervals[42,45–47]. There were also 10 bp periodicities of GC dinucleotides, but their periodicity was offset by 5 bp compared with the AA, TT and TA patterns. Current alignments of thousands of nucleosomal DNAs show essentially the same pattern, including changes in nucleotide composition in linker regions[17,19,20,24,45,48–53]. Other nucleotide and DNA structural elements also exist, but they might be less universal and might be tailored for specific positioning purposes that remain to be elucidated[18].

What do the periodic AA, TT and GC patterns tell us? The 10 bp periodical presence of certain dinucleotides probably provides a rotational setting of the DNA on the histone surface because AA or TT dinucleotides tend to expand the major groove of DNA, whereas GC dinucleotides tend to contract the major groove. These alterations of the major groove might facilitate DNA wrapping around the histone core when the dinucleotides are placed in phase with the helical twist of DNA. In addition, other sequence combinations could create subtle bends in the DNA or alter the flexibility of DNA to contribute to the rotational setting of nucleosomal DNA[18,48]. Owing to rotational phasing, translational repositioning of a resident nucleosome into an adjacent linker region or NFR could obscure a DNA regulatory element in the linker without affecting the accessibility of another regulatory site that is already rotationally exposed on the surface of the nucleosome (FIG. 3b).

A key observation which showed that rotational phasing does not necessarily establish translational phasing was the inability of a 10 bp repeating pattern of AA and TT dinucleotides to predict the genomic locations of nucleosomes[20]. Instead, the nucleosome positions were more accurately predicted when the search pattern was enriched with AA dinucleotides towards the 5′ end and TT dinucleotides towards the 3′ end. Thus, partitioning of AA and TT dinucleotides towards the 5′ and 3′ ends, respectively, helps define translational positioning, whereas periodic AA and TT dinucleotides help define rotation positioning.

Box 2 | **Nucleosome numbering**

In yeast and flies, the first nucleosome upstream of the 5′ nucleosome-free region (NFR) is considered the –1 nucleosome, whereas the first nucleosome downstream of the NFR is considered the +1 nucleosome[17–21,23] (FIG. 2). In humans, the rare nucleosome that appears in the consensus NFR regions has been defined as –1, which leaves the more predominant first upstream nucleosome to be called –2 (REF. 22). As this nomenclature inconsistency between organisms could be confusing, some standardization of nucleosome numbering might be necessary, particularly as different nucleosome positions have been shown to have specific functions.

Despite the statistical enrichment of AA, TT and GC patterns associated with nucleosomes, the presence of these dinucleotide patterns in individual nucleosomes only occurs modestly above a random distribution and is largely limited to the –1 and +1 nucleosomes[20,46]. Thus, sequence-directed positioning might be subtle or diffuse, meaning that a small number of sequence determinants could be spread throughout the 147 bp nucleosomal DNA. Positioning is also likely to involve a combination of these favourable positioning sequences plus linker-enriched unfavourable sequences. It might be advantageous to have a mixture of favourable and unfavourable sequences, which results in only marginally stable nucleosome positions. An optimum mixture might strike an important balance between a state that can be disrupted to allow transcription and replication and a stable state that prevents inappropriate access to DNA. Indeed, the entire genome can be thought of as a continuous thermodynamic landscape of nucleosome occupancy, in which NFRs represent the thermodynamically least favourable regions and the +1 or –1 nucleosome positions represent the thermodynamically most favourable regions.

Predicting nucleosome positions. Many studies have attempted to computationally predict *in vivo* nucleosome locations *de novo* in yeast, flies and humans based on properties of the underlying DNA sequence[17,42,44,46,48,49,53,54], and more sophisticated strategies are now emerging. Such predictions have been successful from a statistical perspective (that is, better than random guessing), but are limited compared with the experimental determination of nucleosome positions. Two studies used a support vector machine classifier that incorporated an experimental data set of nucleosome positions to identify characteristics that could discriminate between nucleosome-forming and nucleosome-avoiding DNA sequences (AT versus GC sequences)[44,49]. Another study used a combination of favourable short distance dinucleotide periodicities and short unfavourable sequence patterns to provide a probalistic model of nucleosome positions[53]. A third, and possibly the most accurate method, involved the use of wavelet transformation sequence periodicities that were spread throughout a training set of nucleosomal DNA sequences, which were combined with nucleosomal and linker sequence differences to create discriminatory signatures that were then used to make *de novo* predictions of nucleosome positions by hidden Markov modelling[54].

It seems unlikely that a simple sequence-based algorithm will ever accurately predict all nucleosome locations. Factors other than the surrounding DNA sequence might contribute to nucleosome positioning *in vivo*. For example, nucleosome remodelling complexes, such as Isw2 in *S. cerevisiae*, override the sequence preferences of nucleosomes, causing nucleosomes to encroach into the 5′ and 3′ NFRs, thereby suppressing cryptic transcription that arises from the NFRs[55,56]. In addition, as a mechanism of gene repression, Isw2 uses the energy from ATP hydrolysis to position nucleosomes onto promoter regions that are intrinsically designed to

repel nucleosomes[56]. Such nucleosomes are said to be 'spring-loaded', because removal of Isw2 would quickly result in intrinsic nucleosome eviction or repositioning of the nucleosome away from the unfavourable sequence.

Structure and function of NFRs

Both DNA sequence and protein factors are important for establishing NFRs. It is striking that regions of the genome that possess the strongest nucleosome positioning sequences (at the +1 nucleosome) are adjacent to regions that have the strongest anti-positioning sequences (5′ NFRs). An important factor in the establishment of a 5′ NFR might be the presence of poly(dA:dT) tracts[15,20,53,57–59]. Nucleosomes tend to be excluded from these tracts owing to the rigidity imparted to the DNA by the bifurcating hydrogen bonds present between adenosine bases on one strand (at position n) and thymines located on the other strand at positions n and n + 1 (REFS 59–61). In addition, specific DNA-binding proteins, such as the Myb-related protein Reb1 in yeast, might be important in positioning nucleosomes to create NFR boundaries[62].

NFRs and transcription. The discovery of NFRs changed the way we think about how the transcription machinery assembles at promoters. We expected that promoter regions would be occluded by nucleosomes except when they were activated. This is still largely true for many genes that are repressed in specific tissues. However, the discovery of NFRs demonstrated that open promoter states are stable and common, even at genes that are transcribed so infrequently (<1% of the maximum level) that they are essentially turned off[17–19,23,63]. Thus, although a NFR is permissive for transcription, it is not sufficient to activate genes. NFRs might allow low basal levels of leaky transcription, which could be interpreted as meaningless biological noise, particularly if the transcripts are rapidly degraded[64–66]. However, low levels of genic transcription might have a general housekeeping function whereby gene products are constitutively produced at low levels.

5′ NFRs are likely to be sites for the assembly of the transcription machinery, whereas 3′ NFRs are likely to be sites for the disassembly of the transcription machinery, although in compact genomes (for example, yeast, flies and worms), the 3′ NFR of one gene could be the 5′ NFR of the next downstream gene. It is currently unclear whether the open architecture of the 5′ NFR is necessary for the initial 'pioneering' polymerase or whether transcription itself establishes the NFR from the closed state, although at heat-shock genes in *D. melanogaster*, some domain-wide chromatin reorganization occurs after heat-shock treatment but before Pol II traverses the gene[67]. Given the reasonable expectation that Pol II and nucleosomal histones cannot simultaneously occupy the same DNA sequence, any nucleosome that is present in the promoter region is likely to be evicted or substantially remodelled before Pol II binding occurs[68]. When transcription is initiated and Pol II has cleared the promoter, the resident chromatin might not return to its original closed state (at least, not immediately) but might maintain an open state in which the composition of the nucleosomes is better suited for eviction during multiple rounds of transcription. This scenario is still hypothetical and remains fertile ground for experimentation.

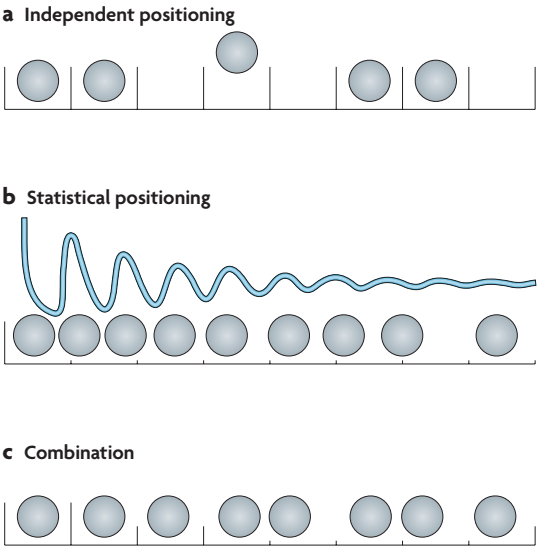


Figure 4 | **Sequence-based packing versus statistical packing. a** | Individual slots represent nucleosome positioning sequences that define where a nucleosome (grey circle) will reside on a length of DNA. **b** | In its purest form, statistical positioning relies on a single positional barrier (left side), against which nucleosomes are ordered. A probabilistic density trace of where nucleosomes would reside in a population is shown. **c** | The true cellular state is likely to be a combination of both independent and statistical positioning.

Transcription start site selection by nucleosomes?

Because many PIC components, such as the SAGA complex and TFIID, have nucleosome-binding subunits, positioned nucleosomes might define the location of the TSS by positioning the PIC. The conventional view is that most genes contain a predominant TSS, the location of which is defined by core promoter elements[69]. In PIC assembly, general transcription factors, such as TATA-binding protein (TBP) or TFIID, bind to core promoter elements and position other initiation factors, such as TFIIB and TFIIF, which then direct Pol II to initiate transcription at the initiator element (INR element) (the consensus sequence is TCAKTY in flies and YYANWYY in humans)[70,71]. The problem with this view is that, despite extensive bioinformatic searches, most promoters seem to lack core promoter elements, including a TATA box, the TFIIB recognition element (BRE), INR, downstream promoter element (DPE) or motif ten element (MTE). In the absence of core promoter elements, how does the transcription machinery establish the location of the TSS? The answer is not known. Below, we speculate that positioned nucleosomes might determine the location of the TSS.

SAGA complex
A multisubunit multifunctional complex that delivers TATA-binding protein (TBP) to promoters (by Spt3 and Spt8 subunits), acetylates nucleosomes (by the Gcn5 subunit) and is associated with activities that remodel (by Chd1) and deubiquitylate (by Ubp8) nucleosomes.

TFIID
A multisubunit general transcription factor composed of TATA-binding protein (TBP) and ~15 other subunits (TBP-associated factors).

Core promoter element
A widely used DNA sequence element that helps position the transcription initiation complex, and is typically located within 60 bp of the transcription start site.

General transcription factor
A protein that is widely considered to be required to set up a transcription initiation complex at all promoters (examples include TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH).

TATA-binding protein
(TBP). This protein is important for assembling the transcription initiation complex.

Initiator element
(INR element). A DNA sequence that specifies the transcription start site (consensus abbreviations include: K = G or T; Y = C or T; W = A or T; N = G, A, T or C).

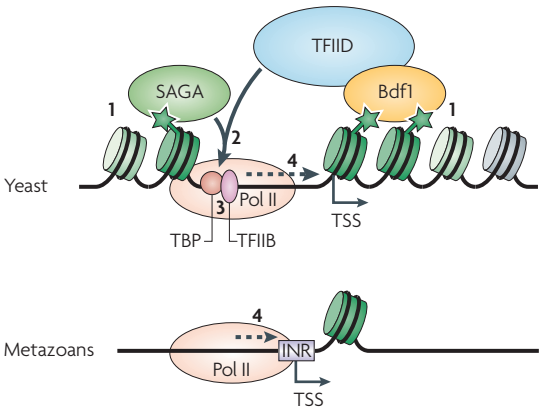


Figure 5 | **Mechanistic differences between transcription initiation in budding yeast and metazoans.** A current model of how nucleosomes might direct start site selection in yeast, compared with metazoans is shown. Each step is also described in the main text. In step 1, the acetylation marks are recognized by bromodomain modules, which are found in many chromatin regulatory complexes, including the SAGA histone acetyltransferase complex and TFIID. In step 2, SAGA and TFIID then deliver TATA binding protein (TBP) to promoters. In step 3, TBP binds TFIIB and places it immediately downstream towards the transcription start site (TSS). In step 4, TFIIB positions RNA polymerase II (Pol II) at the promoter. The diagram for metazoans is a simplified version of that shown for yeast, in which the relationship between Pol II and the initiator (INR) is emphasized. The dashed arrows in both panels indicate sliding of Pol II before transcription initiation. Acetylation marks are indicated by green stars. The green colouring represents H2A.Z enrichment in the nucleosome array.

For about 80% of the 5,700 genes in *S. cerevisiae*, there is one TSS[72]. Remarkably, this specificity is achieved without a well-defined initiator element. The minimal consensus TSS in *S. cerevisiae* is YR (a C or T followed by an initiating A or G, although other nearby sequences might influence start site selection)[73,74]. YR is predicted to occur once every 4 bp in the genome, and thus lacks selectivity. However, TSSs are tightly distributed ~10–15 bp inside of the upstream border of the +1 nucleosome[19] (FIG. 2). Given this tight linkage, it is difficult to envision how positioning of the TSS and the +1 nucleosome could have arisen independently at thousands of genes in yeast, yet maintained a fixed distance from each other.

How might the TSS be tightly linked to the position of the +1 nucleosome? First, during transcriptional activation, which is promoted by sequence-specific transcriptional activators, the –1 and +1 nucleosomes are acetylated and methylated (FIG. 5). The acetylation marks are recognized by bromodomain modules, which are found in many chromatin regulatory complexes, including the SAGA histone acetyltransferase complex[75] (step 1 in FIG. 5) and TFIID (which is contained in the TFIID-interacting protein bromodomain-containing factor 1 (Bdf1) in yeast)[76,77]. In mammals, TFIID also binds H3K4me3 (histone H3 methylated at lysine 4), which is a mark of active transcription[78]. SAGA and TFIID then deliver TBP to promoters[79,80] (step 2 in FIG. 5). Therefore, in principle, TBP positioning at promoters could be directed in part by SAGA and/or TFIID bound to nucleosomes, without the need for a positioning element.

TBP binds TFIIB and places it immediately downstream towards the TSS[81] (step 3 in FIG. 5) and TFIIB positions Pol II at the promoter[82,83] (step 4 in FIG. 5). There is experimental evidence that TFIIB controls TSS selection; for example, the TSS location can be shifted by mutations in TFIIB[84] or by replacement of TFIIB and Pol II with the same proteins from an evolutionarily diverged eukaryote that normally has a shifted TSS[74,85,86]. None of these steps invokes a need for core promoter elements. A similar scenario occurs during transcription of eukaryotic tRNA genes by Pol III. The protein complex TFIIIC binds to specific DNA sequences that are internal to the tRNA genes and positions the TBP-containing TFIIIB complex at a precise distance from the TSS without an underlying positioning element[87], and TFIIIB then positions Pol III at the TSS.

Core promoter elements might have been adopted later in evolution. In metazoans, such elements might focus the TSS[69]. At least in vertebrates, genes that lack core promoter elements tend to have many TSSs dispersed over a distance of 50–100 bp[69]. It will be interesting to learn whether such promoters also have a dispersed (fuzzy) nucleosome architecture, which might be expected if nucleosome positions define at least some of the TSSs.

Evolutionary shifts in the TSS location. Compared with *S. cerevisiae*, metazoans have a genome-wide shift in the location of the TSS with respect to the position of the +1 nucleosome (FIG. 5). The predominant metazoan TSS resides in the NFR, ~60 bp upstream of the +1 nucleosome border[16,20,24], whereas *S. cerevisiae* initiates transcription just inside the +1 nucleosome border. Part of this downstream shift in the TSS location in *S. cerevisiae* might be due to the initiation of metazoan transcription 30 bp downstream of the site at which TBP is bound[70], whereas *S. cerevisiae* initiates transcription ~60 bp downstream of TBP. Therefore TBP might reside at the same distance from the +1 nucleosome in both eukaryotic branches. Although this hypothesis remains to be tested, if it is true, this would suggest that the distance between the +1 nucleosome and the PIC is a fundamental constant in eukaryotes and that the species-specific differences in TSS location are due to species-specific differences in TFIIB and Pol II[74,85,86].

One model for how TFIIB and Pol II select a TSS is that after they are recruited to the site where TBP binds, TFIIB directs Pol II to scan downstream in a manner that does not require transcription[74,88] (step 4, dashed arrow in FIG. 5). Any mechanism that causes Pol II to dwell at a particular site might increase the probability of initiation at that site. Such a mechanism might be based in part on the nature of the TFIIB–Pol II interactions in combination with core promoter sequences and/or nucleosome positions, all of which could affect Pol II scanning efficiency. In metazoans, increased dwelling

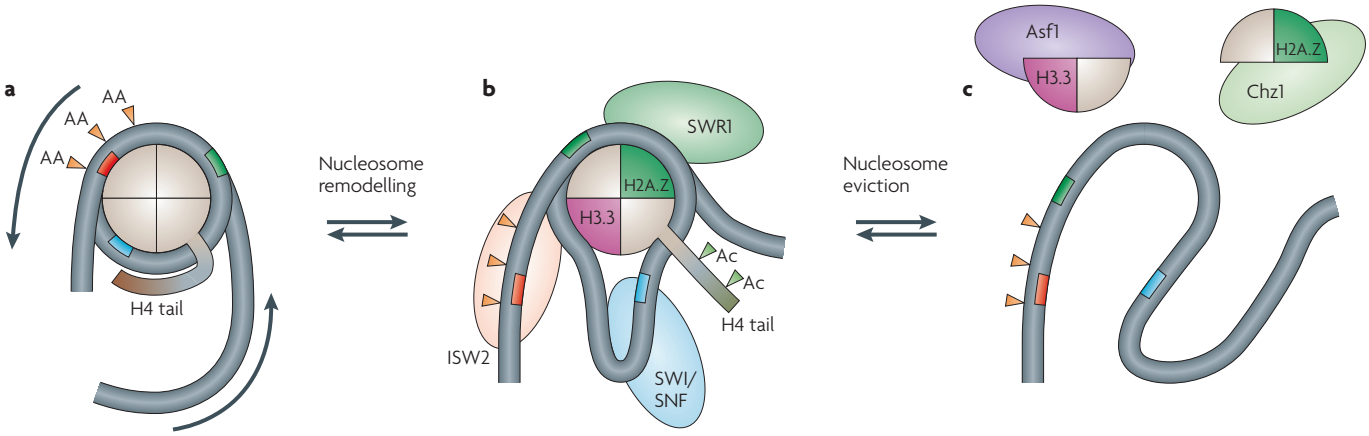


Figure 6 | **Mechanisms that allow DNA accessibility. a** | A stable nucleosome. **b** | A remodelled nucleosome. **c** | An evicted nucleosome. Three transcription factor binding sites are shown in red, green and blue, respectively. The red and blue sites become accessible only during remodelling, either by nucleosome sliding, as indicated by the arrows in **a**, or by chromatin remodelling complexes (for example, ISW2, SWR1 and SWI/SNF) that 'extract' DNA from the nucleosome surface, as shown in **b**. Owing to rotational phasing, the green site is always accessible in the various states. Nucleosome eviction (**c**) might be necessary to assemble a pre-initiation complex and to transcribe the underlying DNA. Anti-silencing function 1 (Asf1) and H2A.Z-specific chaperone (Chz1) are examples of histone chaperones. Ac, acetylation.

might be caused by core promoter elements (INR, DPE and MTE), whereas in *S. cerevisiae,* the +1 nucleosome might provide the predominant impediment because core promoter elements beyond TATA boxes might not exist. Such a mechanism remains highly speculative until tested.

What are the implications of species-specific shifts in the TSS? The location of the TSS relative to the +1 nucleosome in *S. cerevisiae* compared with its location in metazoans indicates that yeast have to displace or remodel the +1 nucleosome before initiation of transcription, whereas metazoans only need to contend with the +1 nucleosome after initiation (FIG. 5). Indeed, a large fraction of metazoan genes have an initiated, but paused, Pol II immediately upstream of, and in contact with, the +1 nucleosome[17,89–91].

Control of DNA access

We have discussed how genome-wide patterns of nucleosome positioning might influence transcription. In this final section, we turn briefly to the general question of how nucleosome disruption or displacement might allow nucleic acid polymerases to translocate along the underlying DNA or to allow transcription factor binding. There are a number of mechanisms by which the effect of a positioned nucleosome can be modulated to regulate DNA accessibility and therefore gene expression.

DNA accessibility without catalysis. Widom and colleagues have proposed a 'site exposure' model whereby thermal fluctuation of DNA on the nucleosome surface transiently exposes DNA-binding sites for transcriptional regulators[92,93]. Site exposure through thermal fluctuation posits that DNA unwrapping originates from the DNA entry and exit points of the nucleosome, and becomes energetically less favourable towards the midpoint of the nucleosome[59,94]. Consistent with this model, DNA regulatory sites tend to reside near the entry and exit sites of nucleosomes[19,55]. Binding of one factor might stabilize a partially disassembled state, allowing other transcription factors to access cognate sites that were previously buried[95]. Alternatively, certain regulatory proteins might bind to the rotationally exposed major groove of DNA that rests on the nucleosome surface (FIGS 3b, 6a)[19] or bind to sites located in the NFR, where they might be constitutively accessible[15].

DNA accessibility and remodelling complexes. Access to DNA sites that are internal to a nucleosome might require catalysed remodelling (FIG. 6b). Regulated nucleosome dynamics are driven by ATP-dependent chromatin remodelling complexes (for example, SWI/SNF[43,96–99]) in many ways: DNA 'breathing' on the nucleosome surface, in which SWI/SNF transiently exposes DNA regulatory sites by creating DNA loops on the nucleosome surface; translational repositioning (nucleosome sliding), in which complexes containing Isw2 move nucleosomes laterally to expose or cover DNA regulatory sites; nucleosome removal and deposition by the RSC complex and histone chaperones, for example, FACT (facilitates chromatin transcription), Asf1 (anti-silencing function 1) and Chz1 (H2A.Z-specific chaperone 1); and replacement of histone subunits, such as the replacement of H2A with H2A.Z by the SWR1 remodelling complex[100–102] and replacement of H3 with H3.3 by the CHD1 (chromodomain-helicase-DNA-binding 1) remodelling complex[103]. Nucleosome dynamics are important because they regulate DNA accessibility, which is key to proper gene regulation and transcription fidelity.

Nucleosome sliding might be an important way of regulating access to DNA sites that are near nucleosome borders. For example, the TATA box of the yeast *PHO5* (repressible acid phosphatase 5) gene resides near the –1 nucleosome border, and movement of this nucleosome

Histone chaperone
A member of a class of proteins that help to deposit histones onto DNA, but are not components of nucleosomes.

by as little as a few base pairs alters the accessibility of the TATA box, which ultimately alters the composition of the assembled transcription machinery[104]. In mammalian cells, induction of the interferon-β promoter by viruses involves nucleosome sliding, which is promoted by the combined action of SWI/SNF on nucleosomes and TBP binding to TATA[105]. Thus, nucleosome sliding, remodelling and/or eviction seem to be important mechanisms for promoting TSS access and gene activation.

Because the binding of sequence-specific transcription factors to their cognate sites might be largely controlled by the –1 nucleosome in *S. cerevisiae*, remodelling complexes that promote site access might be specifically targeted to such nucleosomes[68] (FIG. 6b). Consistent with this suggestion, SWI/SNF promotes the binding of the Gal4 activator to nucleosomal DNA *in vitro*[106,107] and can modulate Gal4 binding to sites near the –1 nucleosome to promote transcription *in vivo*[108].

The activity of SWI/SNF (and related complexes) can be enhanced by histone acetylation[109,110]. For example, acetylation might reduce histone–DNA electrostatic interactions by neutralizing positively charged lysines[111], which might disrupt higher-order, repressive chromatin structures[112] and also provide acetyl-lysine binding sites for SWI/SNF and other bromodomain-containing complexes[75]. A detailed discussion of this subject is beyond the scope of this Review, but it is clear from a wide range of studies that histone acetylation is an important contributor to gene activation[12,111,113–115].

Nucleosome eviction. In addition to shifting the contacts between DNA and histones, eviction of a nucleosome from a particular genomic location allows DNA-binding factors to access the DNA (FIG. 6c) and can therefore affect gene expression. Remodelling complexes remove nucleosomes, and this process is likely to be influenced by histone variants[116]. Nucleosome loss can occur as a specific response to environmental stresses or signals, leading to transcriptional reprogramming. For example, yeast genes that are activated in response to heat shock or changes in the cell cycle often lose nucleosomes in the promoter region[23,117]. Genes in which expression is turned down gain nucleosomes[5,7,10,11,23,117,118].

Nucleosome ejection during gene activation has been studied for some time at model target genes in *Saccharomyces* spp.[118–125] and other organisms[126]. For example, during erythropoiesis in humans, induction of the β-globin gene results in histone loss over DNA sequences in the locus control region (LCR), which contains enhancer elements that direct expression of the globin genes. Indeed, loss of nucleosomes allows the haematopoietic, cell-specific transcriptional activator NF-E2 (nuclear factor, erythroid-derived 2) to bind to the LCR[127]. In other cell types, nucleosomes create a closed chromatin state over the LCR, which precludes globin expression.

Conclusions and future directions

Our understanding of how nucleosome positions and dynamics regulate gene expression has increased dramatically in recent years. This is in large part due to the discovery and characterization of proteins that write, read and erase the many kinds of histone modifications, and the convergence of this knowledge on the involvement of chromatin remodelling factors, histone variants and histone chaperones in gene regulation. The ability to determine where in the genome proteins are bound and how much is bound using ChIP–chip and ChIP–Seq technologies has opened our eyes to the general mechanism or specificity of nucleosomal regulation. Because misregulation of nucleosomes can lead to developmental defects and cancer[128–130], such as mixed-lineage leukaemia, when the ability to methylate histone H3 is disrupted, understanding the extent to which nucleosome organization and modification of this organization are altered throughout a genome in normal and disease states will be an important step towards chromatin-based therapy.

Some of the next steps towards increasing our understanding of global chromatin structure will be to identify the cellular components and mechanisms that determine the canonical positioning of nucleosomes. This might be best achieved through molecular and/or genetic techniques that remove candidate nucleosome organizing factors and determine whether canonical positions are altered. The observation that many proteins contain conserved domains that interact with histone modifications raises the question of whether such proteins bind to specific nucleosomes in the genome. Addressing this question might require the development of methods to cross-link proteins to nucleosomes *in vivo* and map such interactions across the genome using ChIP–Seq technology. The development of high-resolution genome-wide mapping technologies will allow us to answer many new questions regarding the function of genomic chromatin organization and its interplay with the transcription machinery.

1. Kornberg, R. D. & Klug, A. The nucleosome. *Sci. Am.* **244**, 52–64 (1981).
2. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
3. Kamakaka, R. T. & Biggins, S. Histone variants: deviants? *Genes Dev.* **19**, 295–310 (2005).
4. Sarma, K. & Reinberg, D. Histone variants meet their match. *Nature Rev. Mol. Cell. Biol.* **6**, 139–149 (2005).
5. Lee, C. K., Shibata, Y., Rao, B., Strahl, B. D. & Lieb, J. D. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genet.* **36**, 900–905 (2004).
6. Sekinger, E. A., Moqtaderi, Z. & Struhl, K. Intrinsic histone–DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol. Cell* **18**, 735–748 (2005). **An early study that showed, using *in vitro* reconstituted nucleosomes at specific loci, that NFRs and nucleosomes positioned nearby might be dictated largely by intrinsic DNA sequence preference rather than by *trans*-acting factors.**
7. Bernstein, B. E., Liu, C. L., Humphrey, E. L., Perlstein, E. O. & Schreiber, S. L. Global nucleosome occupancy in yeast. *Genome Biol.* **5**, R62 (2004).
8. Guillemette, B. *et al.* Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. *PLoS Biol.* **3**, e384 (2005).
9. Schwabish, M. A. & Struhl, K. Evidence for eviction and rapid deposition of histones upon transcriptional elongation by RNA polymerase II. *Mol. Cell. Biol.* **24**, 10111–10117 (2004).
10. Zanton, S. J. & Pugh, B. F. Full and partial genome-wide assembly and disassembly of the yeast transcription machinery in response to heat shock. *Genes Dev.* **20**, 2250–2265 (2006).
11. Zhang, H., Roberts, D. N. & Cairns, B. R. Genome-wide dynamics of Htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss. *Cell* **123**, 219–231 (2005).
12. Kurdistani, S. K., Tavazoie, S. & Grunstein, M. Mapping global histone acetylation patterns to gene expression. *Cell* **117**, 721–733 (2004).

13. Vogelauer, M., Wu, J., Suka, N. & Grunstein, M. Global histone acetylation and deacetylation in yeast. *Nature* **408**, 495–498 (2000).
14. Bernstein, B. E. *et al.* Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc. Natl Acad. Sci. USA* **99**, 8695–8700 (2002).
15. Yuan, G. C. *et al.* Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309**, 626–630 (2005).
The first high-resolution genome-wide study to reveal a NFR and a canonical arrangement of nucleosomes, including the DNA sequences that contribute to this arrangement.
16. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
One of the most extensive catalogues of the positions of post-translationally modified nucleosomes throughout the human genome. The study used ChIP–Seq and reports on patterns associated with each nucleosome modification.
17. Mavrich, T. N. *et al.* Nucleosome organization in the *Drosophila* genome. *Nature* **453**, 358–362 (2008).
18. Lee, W. *et al.* A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genet.* **39**, 1235–1244 (2007).
19. Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446**, 572–576 (2007).
This paper provides the first report of the use of ChIP–Seq to develop high-resolution maps of nucleosome positions, which allowed the rotational and translational context of DNA regulatory elements to be determined.
20. Mavrich, T. N. *et al.* A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* **18**, 1073–1083 (2008).
This paper provides evidence that sequence-based nucleosome positioning is largely restricted to promoter regions, and that adjacent positions are dictated largely by packing principles.
21. Mito, Y., Henikoff, J. G. & Henikoff, S. Histone replacement marks the boundaries of *cis*-regulatory domains. *Science* **315**, 1408–1411 (2007).
22. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–898 (2008).
23. Shivaswamy, S. *et al.* Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.* **6**, e65 (2008).
24. Valouev, A. *et al.* A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**, 1051–1063 (2008).
25. Ito, T., Bulger, M., Pazin, M. J., Kobayashi, R. & Kadonaga, J. T. ACF, an ISWI-containing and ATP-utilizing chromatin assembly and remodeling factor. *Cell* **90**, 145–155 (1997).
26. Varga-Weisz, P. D. *et al.* Chromatin-remodelling factor CHRAC contains the ATPases ISWI and topoisomerase II. *Nature* **388**, 598–602 (1997).
27. Saha, A., Wittmeyer, J. & Cairns, B. R. Mechanisms for nucleosome movement by ATP-dependent chromatin remodeling complexes. *Results Probl. Cell Differ.* **41**, 127–148 (2006).
28. Gangaraju, V. K. & Bartholomew, B. Mechanisms of ATP-dependent chromatin remodeling. *Mutat. Res.* **618**, 3–17 (2007).
29. Kagalwala, M. N., Glaus, B. J., Dang, W., Zofall, M. & Bartholomew, B. Topography of the ISW2–nucleosome complex: insights into nucleosome spacing and chromatin remodeling. *EMBO J.* **23**, 2092–2104 (2004).
30. Ferreira, H. & Owen-Hughes, T. Lighting up nucleosome spacing. *Nature Struct. Mol. Biol.* **13**, 1047–1049 (2006).
31. Rippe, K. *et al.* DNA sequence- and conformation-directed positioning of nucleosomes by chromatin-remodeling complexes. *Proc. Natl Acad. Sci. USA* **104**, 15635–15640 (2007).
32. Blank, T. A. & Becker, P. B. Electrostatic mechanism of nucleosome spacing. *J. Mol. Biol.* **252**, 305–313 (1995).
33. Fan, Y. *et al.* H1 linker histones are essential for mouse development and affect nucleosome spacing *in vivo*. *Mol. Cell. Biol.* **23**, 4559–4572 (2003).
34. Fan, Y. *et al.* Histone H1 depletion in mammals alters global chromatin structure but causes specific changes in gene regulation. *Cell* **123**, 1199–1212 (2005).
35. Malik, H. S. & Henikoff, S. Phylogenomics of the nucleosome. *Nature Struct. Biol.* **10**, 882–891 (2003).
36. Cosgrove, M. S. & Wolberger, C. How does the histone code work? *Biochem. Cell Biol.* **83**, 468–476 (2005).
37. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).
38. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128**, 707–719 (2007).
39. Lieb, J. D. & Clarke, N. D. Control of transcription through intragenic patterns of nucleosome composition. *Cell* **123**, 1187–1190 (2005).
40. Kiyama, R. & Trifonov, E. N. What positions nucleosomes? A model. *FEBS Lett.* **523**, 7–11 (2002).
41. Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868–871 (1974).
42. Ioshikhes, I. P., Albert, I., Zanton, S. J. & Pugh, B. F. Nucleosome positions predicted through comparative genomics. *Nature Genet.* **38**, 1210–1215 (2006).
43. Rando, O. J. & Ahmad, K. Rules and regulation in the primary structure of chromatin. *Curr. Opin. Cell Biol.* **19**, 250–256 (2007).
44. Gupta, S. *et al.* Predicting human nucleosome occupancy from primary sequence. *PLoS Comput. Biol.* **4**, e1000134 (2008).
45. Satchwell, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659–675 (1986).
46. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).
Together with Reference 42, this study provides evidence that at least some genomic sequences favour nucleosome assembly, which can be used to approximately predict nucleosome positions.
47. Wang, J. P. & Widom, J. Improved alignment of nucleosome DNA sequences using a mixture model. *Nucleic Acids Res.* **33**, 6743–6755 (2005).
48. Miele, V., Vaillant, C., d'Aubenton-Carafa, Y., Thermes, C. & Grange, T. DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res.* **36**, 3746–3756 (2008).
49. Peckham, H. E. *et al.* Nucleosome positioning signals in genomic DNA. *Genome Res.* **17**, 1170–1177 (2007).
50. Trifonov, E. N. Sequence-dependent deformational anisotropy of chromatin DNA. *Nucleic Acids Res.* **8**, 4041–4053 (1980).
51. Widom, J. Role of DNA sequence in nucleosome stability and dynamics. *Q. Rev. Biophys.* **34**, 269–324 (2001).
52. Wang, J. P. *et al.* Preferentially quantized linker DNA lengths in *Saccharomyces cerevisiae*. *PLoS Comput. Biol.* **4**, e1000175 (2008).
53. Field, Y. *et al.* Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.* **4**, e1000216 (2008).
54. Yuan, G. C. & Liu, J. S. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.* **4**, e13 (2008).
55. Whitehouse, I., Rando, O. J., Delrow, J. & Tsukiyama, T. Chromatin remodelling at promoters suppresses antisense transcription. *Nature* **450**, 1031–1035 (2007).
56. Whitehouse, I. & Tsukiyama, T. Antagonistic forces that position nucleosomes *in vivo*. *Nature Struct. Mol. Biol.* **13**, 633–640 (2006).
57. Radwan, A., Younis, A., Luykx, P. & Khuri, S. Prediction and analysis of nucleosome exclusion regions in the human genome. *BMC Genomics* **9**, 186 (2008).
58. Iyer, V. & Struhl, K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.* **14**, 2570–2579 (1995).
59. Anderson, J. D. & Widom, J. Poly(dA–dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. *Mol. Cell. Biol.* **21**, 3830–3839 (2001).
60. Nelson, H. C., Finch, J. T., Luisi, B. F. & Klug, A. The structure of an oligo(dA).oligo(dT) tract and its biological implications. *Nature* **330**, 221–226 (1987).
61. Struhl, K. Naturally occurring poly(dA–dT) sequences are upstream promoter elements for constitutive transcription in yeast. *Proc. Natl Acad. Sci. USA* **82**, 8419–8423 (1985).
62. Raisner, R. M. *et al.* Histone variant H2A.Z marks the 5′ ends of both active and inactive genes in euchromatin. *Cell* **123**, 233–248 (2005).
63. Mito, Y., Henikoff, J. G. & Henikoff, S. Genome-scale profiling of histone H3.3 replacement patterns. *Nature Genet.* **37**, 1090–1097 (2005).
64. Arigo, J. T., Eyler, D. E., Carroll, K. L. & Corden, J. L. Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. *Mol. Cell* **23**, 841–851 (2006).
65. Thiebaut, M., Kisseleva-Romanova, E., Rouge-maille, M., Boulay, J. & Libri, D. Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the Nrd1–Nab3 pathway in genome surveillance. *Mol. Cell* **23**, 853–864 (2006).
66. Thompson, D. M. & Parker, R. Cytoplasmic decay of intergenic transcripts in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **27**, 92–101 (2007).
67. Petesch, S. J. & Lis, J. T. Rapid, transcription-independent loss of nucleosomes over a large chromatin domain at Hsp70 loci. *Cell* **134**, 74–84 (2008).
68. Venters, B. J. & Pugh, B. F. A canonical promoter organization of the transcription machinery and its regulators in the *Saccharomyces* genome. *Genome Res.* 5 Jan 2009 (doi:10.1101/gr.084970.108).
69. Juven-Gershon, T., Hsu, J. Y., Theisen, J. W. & Kadonaga, J. T. The RNA polymerase II core promoter — the gateway to transcription. *Curr. Opin. Cell Biol.* **20**, 253–259 (2008).
70. Smale, S. T. & Kadonaga, J. T. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72**, 449–479 (2003).
71. Thomas, M. C. & Chiang, C. M. The general transcription machinery and general cofactors. *Crit. Rev. Biochem. Mol. Biol.* **41**, 105–178 (2006).
72. David, L. *et al.* A high-resolution map of transcription in the yeast genome. *Proc. Natl Acad. Sci. USA* **103**, 5320–5325 (2006).
73. Zhang, Z. & Dietrich, F. S. Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5′ SAGE. *Nucleic Acids Res.* **33**, 2838–2851 (2005).
74. Kuehner, J. N. & Brow, D. A. Quantitative analysis of *in vivo* initiator selection by yeast RNA polymerase II supports a scanning model. *J. Biol. Chem.* **281**, 14119–14128 (2006).
75. Hassan, A. H. *et al.* Function and selectivity of bromodomains in anchoring chromatin-modifying complexes to promoter nucleosomes. *Cell* **111**, 369–379 (2002).
76. Jacobson, R. H., Ladurner, A. G., King, D. S. & Tjian, R. Structure and function of a human TAF(II)250 double bromodomain module. *Science* **288**, 1422–1425 (2000).
77. Matangkasombut, O., Buratowski, R. M., Swilling, N. W. & Buratowski, S. Bromodomain factor 1 corresponds to a missing piece of yeast TFIID. *Genes Dev.* **14**, 951–962 (2000).
78. Vermeulen, M. *et al.* Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* **131**, 58–69 (2007).
79. Pugh, B. F. & Tjian, R. Mechanism of transcriptional activation by Sp1: evidence for coactivators. *Cell* **61**, 1187–1197 (1990).
80. Sermwittayawong, D. & Tan, S. SAGA binds TBP via its Spt8 subunit in competition with DNA: implications for TBP recruitment. *EMBO J.* **25**, 3791–3800 (2006).
81. Nikolov, D. B. *et al.* Crystal structure of a TFIIB–TBP–TATA-element ternary complex. *Nature* **377**, 119–128 (1995).
82. Hausner, W., Wettach, J., Hethke, C. & Thomm, M. Two transcription factors related with the eucaryal transcription factors TATA-binding protein and transcription factor IIB direct promoter recognition by an archaeal RNA polymerase. *J. Biol. Chem.* **271**, 30144–30148 (1996).
83. Bushnell, D. A., Westover, K. D., Davis, R. E. & Kornberg, R. D. Structural basis of transcription: an RNA polymerase II–TFIIB cocrystal at 4.5 Angstroms. *Science* **303**, 983–988 (2004).
84. Pardee, T. S., Bangur, C. S. & Ponticelli, A. S. The N-terminal region of yeast TFIIB contains two adjacent functional domains involved in stable RNA polymerase II binding and transcription start site selection. *J. Biol. Chem.* **273**, 17859–17864 (1998).
85. Ghazy, M. A., Brodie, S. A., Ammerman, M. L., Ziegler, L. M. & Ponticelli, A. S. Amino acid substitutions in yeast TFIIF confer upstream shifts in transcription initiation and altered interaction with RNA polymerase II. *Mol. Cell. Biol.* **24**, 10975–10985 (2004).
86. Li, Y., Flanagan, P. M., Tschochner, H. & Kornberg, R. D. RNA polymerase II initiation factor interactions and transcription start site selection. *Science* **263**, 805–807 (1994).
87. Geiduschek, E. P. & Kassavetis, G. A. The RNA polymerase III transcription apparatus. *J. Mol. Biol.* **310**, 1–26 (2001).

88. Giardina, C. & Lis, J. T. DNA melting on yeast RNA polymerase II promoters. *Science* **261**, 759–762 (1993).
89. Muse, G. W. *et al.* RNA polymerase is poised for activation across the genome. *Nature Genet.* **39**, 1507–1511 (2007).
90. Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo. *Nature Genet.* **39**, 1512–1516 (2007).
91. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77–88 (2007).
This study showed that most genes in human embryonic stem cells seem to have a stalled RNA polymerase II at their 5′ ends (although such sites might actually have low occupancy levels).
92. Polach, K. J. & Widom, J. Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. *J. Mol. Biol.* **254**, 130–149 (1995).
93. Polach, K. J. & Widom, J. A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *J. Mol. Biol.* **258**, 800–812 (1996).
94. Anderson, J. D. & Widom, J. Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites. *J. Mol. Biol.* **296**, 979–987 (2000).
95. Adams, C. C. & Workman, J. L. Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. *Mol. Cell. Biol.* **15**, 1405–1421 (1995).
96. Smith, C. L. & Peterson, C. L. ATP-dependent chromatin remodeling. *Curr. Top. Dev. Biol.* **65**, 115–148 (2005).
97. Eisen, J. A., Sweder, K. S. & Hanawalt, P. C. Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions. *Nucleic Acids Res.* **23**, 2715–2723 (1995).
98. Cairns, B. R. Chromatin remodeling complexes: strength in diversity, precision through specialization. *Curr. Opin. Genet. Dev.* **15**, 185–190 (2005).
99. Gutierrez, J. L., Chandy, M., Carrozza, M. J. & Workman, J. L. Activation domains drive nucleosome eviction by SWI/SNF. *EMBO J.* **26**, 730–740 (2007).
100. Kobor, M. S. *et al.* A protein complex containing the conserved Swi2/Snf2-related ATPase Swr1p deposits histone variant H2A.Z into euchromatin. *PLoS Biol.* **2**, e131 (2004).
101. Mizuguchi, G. *et al.* ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* **303**, 343–348 (2004).
102. Krogan, N. J. *et al.* A Snf2 family ATPase complex required for recruitment of the histone H2A variant Htz1. *Mol. Cell* **12**, 1565–1576 (2003).
103. Konev, A. Y. *et al.* CHD1 motor protein is required for deposition of histone variant H3.3 into chromatin *in vivo*. *Science* **317**, 1087–1090 (2007).
104. Martinez-Campa, C. *et al.* Precise nucleosome positioning and the TATA box dictate requirements for the histone H4 tail and the bromodomain factor Bdf1. *Mol. Cell* **15**, 69–81 (2004).
105. Lomvardas, S. & Thanos, D. Nucleosome sliding via TBP DNA binding *in vivo*. *Cell* **106**, 685–696 (2001).
106. Kwon, H., Imbalzano, A. N., Khavari, P. A., Kingston, R. E. & Green, M. R. Nucleosome disruption and enhancement of activator binding by a human SW1/SNF complex. *Nature* **370**, 477–481 (1994).
107. Cote, J., Peterson, C. L. & Workman, J. L. Perturbation of nucleosome core structure by the SWI/SNF complex persists after its detachment, enhancing subsequent transcription factor binding. *Proc. Natl Acad. Sci. USA* **95**, 4947–4952 (1998).
108. Burns, L. G. & Peterson, C. L. The yeast SWI–SNF complex facilitates binding of a transcriptional activator to nucleosomal sites *in vivo*. *Mol. Cell. Biol.* **17**, 4811–4819 (1997).
109. Suganuma, T. *et al.* ATAC is a double histone acetyltransferase complex that stimulates nucleosome sliding. *Nature Struct. Mol. Biol.* **15**, 364–372 (2008).
110. Hassan, A. H., Neely, K. E. & Workman, J. L. Histone acetyltransferase complexes stabilize SWI/SNF binding to promoter nucleosomes. *Cell* **104**, 817–827 (2001).
111. Dion, M. F., Altschuler, S. J., Wu, L. F. & Rando, O. J. Genomic characterization reveals a simple histone H4 acetylation code. *Proc. Natl Acad. Sci. USA* **102**, 5501–5506 (2005).
112. Wang, X. & Hayes, J. J. Acetylation mimics within individual core histone tail domains indicate distinct roles in regulating the stability of higher-order chromatin structure. *Mol. Cell. Biol.* **28**, 227–236 (2008).
113. Liu, C. L. *et al.* Single nucleosome mapping of histone modifications in *S. cerevisiae*. *PLoS Biol.* **3**, e328 (2005).
This study showed that acetylation of histones at specific residues does not elicit a specific transcriptional response, indicating that acetylation might have cumulative effects rather than being encoded.
114. Pokholok, D. K. *et al.* Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**, 517–527 (2005).
115. Shahbazian, M. D. & Grunstein, M. Functions of site-specific histone acetylation and deacetylation. *Annu. Rev. Biochem.* **76**, 75–100 (2007).
116. Santisteban, M. S., Kalashnikova, T. & Smith, M. M. Histone H2A.Z regulates transcription and is partially redundant with nucleosome remodeling complexes. *Cell* **103**, 411–422 (2000).
117. Hogan, G. J., Lee, C. K. & Lieb, J. D. Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. *PLoS Genet.* **2**, e158 (2006).
118. Boeger, H., Griesenbeck, J., Strattan, J. S. & Kornberg, R. D. Removal of promoter nucleosomes by disassembly rather than sliding *in vivo*. *Mol. Cell* **14**, 667–673 (2004).
119. Korber, P., Luckenbach, T., Blaschke, D. & Horz, W. Evidence for histone eviction in *trans* upon induction of the yeast *PHO5* promoter. *Mol. Cell. Biol.* **24**, 10965–10974 (2004).
120. Reinke, H. & Horz, W. Histones are first hyperacetylated and then lose contact with the activated *PHO5* promoter. *Mol. Cell* **11**, 1599–1607 (2003).
121. Adkins, M. W., Howar, S. R. & Tyler, J. K. Chromatin disassembly mediated by the histone chaperone Asf1 is essential for transcriptional activation of the yeast *PHO5* and *PHO8* genes. *Mol. Cell* **14**, 657–666 (2004).
122. Moreira, J. M. & Holmberg, S. Transcriptional repression of the yeast *CHA1* gene requires the chromatin-remodeling complex RSC. *EMBO J.* **18**, 2836–2844 (1999).
123. Zhao, J., Herrera-Diaz, J. & Gross, D. S. Domain-wide displacement of histones by activated heat shock factor occurs independently of Swi/Snf and is not correlated with RNA polymerase II density. *Mol. Cell. Biol.* **25**, 8985–8999 (2005).
124. Zhang, H. & Reese, J. C. Exposing the core promoter is sufficient to activate transcription and alter coactivator requirement at RNR3. *Proc. Natl Acad. Sci. USA* **104**, 8833–8838 (2007).
125. Almer, A., Rudolph, H., Hinnen, A. & Horz, W. Removal of positioned nucleosomes from the yeast *PHO5* promoter upon PHO5 induction releases additional upstream activating DNA elements. *EMBO J.* **5**, 2689–2696 (1986).
126. Cartwright, I. L. & Elgin, S. C. Nucleosomal instability and induction of new upstream protein–DNA associations accompany activation of four small heat shock protein genes in *Drosophila melanogaster*. *Mol. Cell. Biol.* **6**, 779–791 (1986).
127. Armstrong, J. A. & Emerson, B. M. NF-E2 disrupts chromatin structure at human β-globin locus control region hypersensitive site 2 *in vitro*. *Mol. Cell. Biol.* **16**, 5634–5644 (1996).
128. Bu, P., Evrard, Y. A., Lozano, G. & Dent, S. Y. Loss of Gcn5 acetyltransferase activity leads to neural tube closure defects and exencephaly in mouse embryos. *Mol. Cell. Biol.* **27**, 3405–3416 (2007).
129. Shilatifard, A. Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. *Annu. Rev. Biochem.* **75**, 243–269 (2006).
130. Whittle, C. M. *et al.* The genomic distribution and function of histone variant HTZ-1 during *C. elegans* embryogenesis. *PLoS Genet.* **4**, e1000187 (2008).
131. Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.* **319**, 1097–1113 (2002).

Acknowledgements
We thank S. Tan for providing the image for FIG. 1a. Support from National Institutes of Health grant HG004160 is gratefully acknowledged.

DATABASES

Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
PHO5
UniProtKB: http://www.uniprot.org
Bdf1 | H2A.Z | H3.3 | Isw2 | TBP | TFIIB

FURTHER INFORMATION

B.F. Pugh's laboratory homepage:
http://www.bmb.psu.edu/faculty/pugh/pugh.html
Penn State Genome Cartography project:
http://atlas.bx.psu.edu
UCSC Genome Bioinformatics: http://genome.ucsc.edu

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Genetic diseases of connective tissues: cellular and extracellular effects of ECM mutations

John F. Bateman, Raymond P. Boot-Handford‡ and Shireen R. Lamandé**

Abstract | Tissue-specific extracellular matrices (ECMs) are crucial for normal development and tissue function, and mutations in ECM genes result in a wide range of serious inherited connective tissue disorders. Mutations cause ECM dysfunction by combinations of two mechanisms. First, secretion of the mutated ECM components can be reduced by mutations affecting synthesis or by structural mutations causing cellular retention and/or degradation. Second, secretion of mutant protein can disturb crucial ECM interactions, structure and stability. Moreover, recent experiments suggest that endoplasmic reticulum (ER) stress, caused by mutant misfolded ECM proteins, contributes to the molecular pathology. Targeting ER stress might offer a new therapeutic strategy.

Osteogenesis imperfecta
(OI). A genetic bone disorder caused by abnormalities of collagen I structure or synthesis that results in poorly formed and fragile bones. Multiple distinct clinical manifestations range in severity from mild to congenital lethal.

**Murdoch Childrens Research Institute and Department of Paediatrics, University of Melbourne, Royal Childrens Hospital, Parkville, Victoria 3052, Australia.*
‡Wellcome Trust Centre for Cell-Matrix Research, Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester M13 9PT, United Kingdom.
Correspondence to J.B.
e-mail:
john.bateman@mcri.edu.au
doi:10.1038/nrg2520
Published online
10 February 2009

The major connective tissues of the body, such as skin, tendon, ligaments, cartilage and bone, provide the structural and informational framework that is necessary for development. The extensive extracellular matrix (ECM) of connective tissues is a complex interacting network of proteins, glycoproteins and proteoglycans providing the dynamic and essential three-dimensional environment that supports the maintenance, growth and differentiation of cells. In addition to providing a highly organized framework, the ECM mediates signals to and from cells that are involved in important biological processes such as cell differentiation and migration during development, and repair processes.

The diversity and functional importance of the ECM is illustrated by the occurrence of numerous genetic and acquired connective tissue disorders. Mutations in individual ECM genes cause conditions such as osteogenesis imperfecta (OI), numerous chondrodysplasias, Ehlers–Danlos syndrome and Marfan syndrome, and although these conditions are individually rare, collectively they are a considerable health burden. Furthermore, studies on these conditions have been of considerable importance in understanding fundamental aspects of ECM assembly and function.

Here we will review recent advances in our understanding of the molecular genetics and pathophysiology of diseases caused by mutations in ECM genes, including insights into the important pathogenic role of endoplasmic reticulum (ER) stress, which might pave the way to new therapeutic opportunities. Comprehensive information about gene mutations has been covered in earlier reviews and is available in online databases, so rather than repeating information we provide a summary of the mutation types and diseases that result from ECM gene mutations. Our main aim is to bring together a discussion of the molecular pathways of pathophysiology that have been elucidated by studies using cells from patients, genetically manipulated cell lines or mutant mouse disease models. In particular, we highlight the realization that ECM mutations exert important pathogenic effects inside the cell, as well as in the ECM outside the cell.

ECM integrates cells into functional assemblies

The ECM is diverse, with precisely regulated combinations of molecular components providing tissue-specific properties ranging from the rock-hard nature of bone to the elasticity of ligament and skin, and from the longevity of adult articular cartilage to the transient nature of growth plate cartilage. ECM components can be broadly placed into groups: structural components; matricellular proteins that have little structural role but modulate cell–ECM interactions and growth factor and protease activity; cell surface receptors and ancillary proteins that interact with the ECM and perform signalling and structural roles; and proteins involved in ECM homeostasis and remodelling, such as proteinases and their inhibitors. Although the different

Chondrodysplasia
A disturbance in the development of cartilage, primarily affecting the long bones. More than 200 forms are recognized, presenting a clinical range from mild to severe arrested growth and dwarfism, to congenital lethal.

Ehlers–Danlos syndrome
A group of inherited disorders of collagen synthesis and fibril formation that result in a range of pathologies, including joint laxity and hypermobility, and skin and blood vessel fragility.

Marfan syndrome
An inherited disorder presenting with long bone overgrowth, and defects of the heart valves and aorta. It is caused by mutations in the microfibrillar protein fibrillin 1.

Articular cartilage
The permanent cartilage that forms the smooth articulating surface of joints. It is a dense connective tissue with an extracellular matrix rich in collagen II and the proteoglycan aggrecan.

connective tissue matrices vary in composition and detailed architecture, there are basic similarities in the types of molecular components and how these components form interacting structural networks. The major ECM components include collagens, proteoglycans and a large number of non-collagenous glycoproteins and proteins[1–8] (BOX 1).

Many ECM structural components assemble into multimers in the cell. This intracellular assembly is an important prerequisite for the multivalent extracellular interactions that occur between the ECM components which generate the architecturally precise matrix that is crucial for function. It can also be an important determinant of how ECM gene mutations cause disease, as we discuss below. The details of the biosynthetic pathway and multilevel assembly are best understood for the collagen family, and for this reason much of our discussion will focus on collagen as the prototypical ECM protein to illustrate common themes in ECM assembly and disease mechanisms. The central features of intracellular synthesis and assembly of several collagen types and cartilage oligomeric matrix protein (COMP, also known as thrombospondin 5, TSP5) are presented in FIG. 1.

The interactions of ECM with cells integrates them into functional assemblies and provides two-way conduits for signalling and mechanotransduction[9,10]. ECM components interact with cells, and thus connect to the cytoskeleton through a range of cell surface receptors, predominantly integrins[11]. In addition to these direct roles in signalling, many ECM components bind growth factors and thus act as a reservoir controlling their bioavailablity. These important regulatory roles of ECM function are beyond the scope of this article and are covered in several recent reviews[6,12,13].

Box 1 | **Extracellular matrix organization**

A diverse range of extracellular matrix (ECM) structural components have been described, including 28 distinct collagen subtypes, hyalectins, proteoglycans and a large number of non-collagenous proteins. Many of these matrix proteins are modular, and are assembled from a limited set of protein domains, or modules, that are utilized in ECM and non-ECM proteins to provide specific functional or structural characteristics. The major components of the ECM are members of the collagen protein family, which provide the backbone scaffolding that is essential for the interaction of ECM components to provide tissue structure and integrity (reviewed in REFS 5,7,8). The characteristic feature of collagens is that they contain a triple helical collagen domain in which glycine occurs at every third position of the protein sequence. The triple helical domain can be the sole protein module in the mature protein, as is the case for the fibril-forming collagens type I, II and III. These fibrillar collagens have uninterrupted triple helical domains that assemble into the highly organized tensile fibrils of many tissues such as skin (collagen I and III), bone (collagen I) and cartilage (collagen II). However, other members of the collagen family have interruptions in their triple helix, and in many cases the collagen module is only a small component of the mature protein (for example, collagen VI). The distinct domains of collagens drive formation of different molecular structures. Functional collagen fibrils are commonly heterotypic co-assemblies, such as collagens I, III and V in skin, and collagens II, IX and XI in cartilage[3,107]. These composites provide important structural characteristics that are further modified by interactions with small leucine-rich proteoglycans[1,2] and other non-collagenous components to produce the architecturally precise ECM that is crucial for the biomechanical function of the tissue.

Molecular pathology of ECM gene mutations

Many mutations have been characterized in the structural components of the ECM and in the enzymes involved in their post-translational processing and folding. The molecular basis of how these mutations cause the myriad of connective tissue disorders depends on the function of the gene product, its tissue distribution and the nature of the mutation. Despite this phenotypic diversity, unifying features in the molecular mechanisms that lead to tissue pathology are emerging from recent studies. This Review will focus on human disease mutations (TABLE 1) and related mouse models (Supplementary information S1 (table)). A comprehensive description of ECM gene-targeted knockouts is provided by Aszodi *et al.*[14].

Loss-of-function mutations

The most common genetic causes of reduced synthesis of a gene product are mutations that result in the introduction of premature termination codons (PTCs). The presence of a PTC triggers an mRNA surveillance process and nonsense mediated decay (NMD), whereby aberrant mRNAs are distinguished from normal mRNAs and are rapidly degraded[15–18] (FIG. 2). There are many PTC mutations in ECM structural genes, and for several of these it has been directly shown that the PTC mutations result in NMD and haploinsufficiency. These include PTC mutations in collagen I in OI[19,20], collagen II in Stickler syndrome[21,22], collagen VI in Bethlem myopathy[23] and collagen X in metaphyseal chondrodysplasia, Schmid type[24]. Recessive PTC mutations lead to the absence of collagen VI in Ullrich's congenital muscular dystrophy[25,26] and collagen VII in dystrophic epidermolysis bullosa[27,28]. It is likely that the majority of other PTC mutations in ECM genes will also be shown to cause NMD of the mRNA transcribed from the mutant allele.

Although NMD can be 100% efficient[29], more commonly it reduces the abundance of PTC-containing transcripts to approximately 5–25% of the normal allele transcript, so there can be small amounts of mutant truncated protein produced that have the potential to exert a dominant negative or gain-of function effect. The extent of NMD is likely to be an important contributor to the clinical outcome[30], and this can be difficult to predict from the mutation alone. The efficiency of NMD depends on the position of the mutation in the gene relative to sequence elements that designate NMD competency, such as the exon–exon boundaries defined during splicing or, in the case of *COL10A1* (collagen, type X, α1), the position of the PTC relative to the 3′ UTR[31]. NMD competency can also show gene and tissue specificity[29]. In general, dominant heterozygous PTC mutations lead to approximately half the amount of normal protein and have a milder clinical phenotype than structural gain-of-function mutations in the same gene.

Loss-of-function mutations in the genes involved in ECM protein processing, folding and post-translational modification can also result in connective tissue disease. For example, mutations in ADAMTS2, the enzyme that removes the collagen I N-propeptide before fibril formation (FIG. 1), causes recessive Ehlers–Danlos syndrome[32],

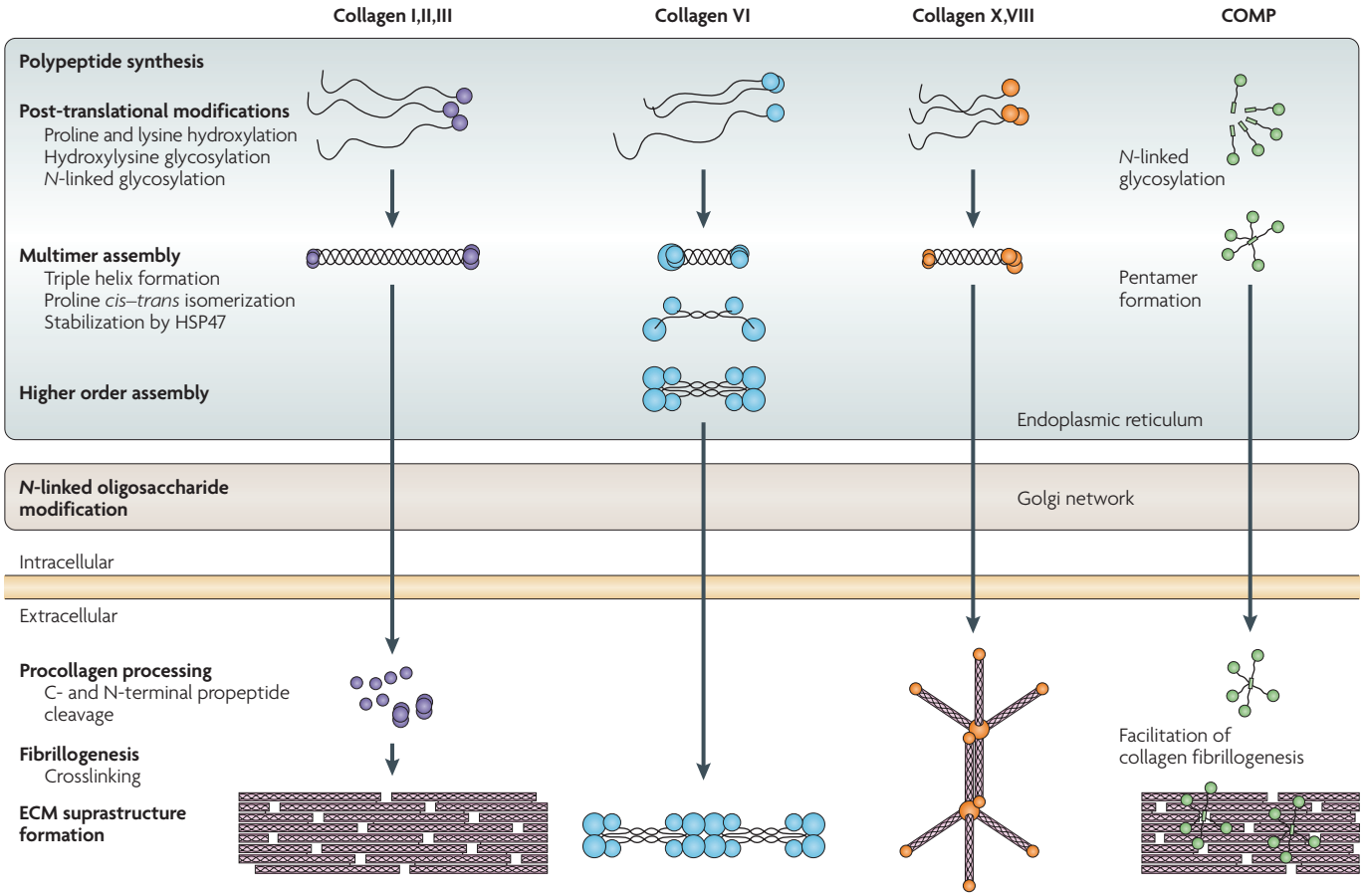


Figure 1 | **Supramolecular assembly pathways.** Major commonalities and differences in the synthesis and assembly pathway of three different functional classes of collagen types — fibril-forming collagens I, II and III, microfibillar collagen VI and network-forming collagens X and VIII — and cartilage oligomeric matrix protein (COMP). Collagen synthesis, chain assembly and formation of the triple helical domain is similar for most collagen types (reviewed in REFS 8,40,41). The collagen precursor chains are co-translationally translocated into the endoplasmic reticulum (ER) lumen, where specific post-translational modifications occur. Three collagen α-chains associate specifically via their C-terminal domains[42] to form heterotrimers or homotrimers. The helical collagens are trafficked via the Golgi network to the plasma membrane, and secreted into the extracellular space[41]. With collagen VI the individual collagen helices are not secreted as monomers but assemble intracellularly into antiparallel overlapping dimers (two triple-helical collagen VI molecules), which then align to form tetramers (four triple-helical collagen VI molecules). The fibril-forming collagens are secreted as precursor forms, called procollagens, with N- and C-terminal non-collagenous domains. These domains are removed by the action of specific proteases, and the collagens are assembled into dense fibrils with a characteristic D-periodicity. The fibril is stabilized by covalent lysine- and hydroxylysine-derived crosslinks[41]. With collagen X and VIII there is no evidence that the N- and C- terminal non-collagenous domains are processed, and they are thought to play a part in the formation of a tetrahedron of four homotrimers. It has been proposed that these tetrahedrons could then form hexagonal lattices by secondary interactions involving terminal and helical sequences[24]. Collagen VI is secreted as tetrameric structures of four collagen VI molecules that aggregate end-to-end to form long thin periodically beaded microfibrils. COMP monomers associate via N-terminal recognition sequences into homopentamers, which, after secretion, can interact with and facilitate collagen I and II fibril formation. COMP pentamers interact with numerous other extracellular matrix (ECM) components, including collagen IX, matrilins and aggrecan.

Growth plate cartilage
A transient cartilage type that drives bone growth and is located at one or both ends of long bones between the epiphysis and the diaphysis. The chondrocytes of the growth plate undergo specific maturation steps leading to hypertrophy and replacement with bone during endochondral bone formation before puberty.

Haploinsufficiency
A condition in a diploid organism in which a single functional copy of a gene results in a phenotype, such as a disease.

Stickler syndrome
A mild inherited chondrodysplasia with early degenerative joint and vertebral changes and often retinal detachment and blindness.

Bethlem myopathy
A genetic disease associated with muscle weakness. This congenital form of muscular dystrophy caused by collagen VI mutations is less severe than the allelic disorder, Ullrich congenital muscular dystrophy.

and mutations in arylsulphatase E[33] and diastrophic dysplasia sulphate transporter (*DTDST*, also known as *SLC26A2*)[34] affect sulphation of glycosaminoglycans and cause chondrodysplasias. Mutations of lysyl hydroxylase 2 (*PLOD2*), a collagen post-translational processing enzyme, cause Bruck syndrome[35], and *PLOD3* mutations have been identified in a patient with complex features that overlap several collagen disorders[36]. Homozygosity or compound heterozygosity for mutations in cartilage associated protein (*CRTAP*) and *LEPRE1* cause abnormalities in collagen I helix formation, resulting in OI[37,38]. *LEPRE1* encodes prolyl-3-hydroxylase 1 (P3H1, also known as leprecan) and forms a molecular complex in the ER with CRTAP and cyclophilin B (CYPB; also known as peptidylprolyl isomerase B, PPIB)[39], and this complex acts as a molecular chaperone for efficient helix formation (FIG. 1). It is likely that this impairment of collagen helix formation has gain-of-function consequences similar to those discussed below for protein structural mutations.

Protein folding and assembly mutations

Extensive studies of dominant structural mutations in ECM components have led to the currently accepted model that the tissue pathology results from the effects exerted by the mutant protein on the ECM. The deleterious effects on the ECM are thought to result from reduced protein levels owing to intracellular degradation of the mutant polypeptide and/or secretion of mutant matrix protein that disrupts the organization of the ECM (FIG. 2). These effects will be discussed first, followed by

Table 1 | **Examples of mutations in ECM structural proteins causing human disease**

ECM component	Gene(s)	Principal tissue(s) affected	Principal disease(s)	Inheritance
Aggrecan	*ACAN*	Cartilage	Spondyloepiphyseal dysplasia, Kimberley type	AD
COMP	*COMP*	Cartilage, ligaments	Multiple epiphyseal dysplasia, pseudoachondroplasia	AD
Collagen I	*COL1A1*, *COL1A2*	Bone	Osteogenesis imperfecta	AD
	COL1A1, *COL1A2*	Skin, joints	Ehlers–Danlos syndrome, type VII	AD
	COL1A2	Skin, joints, heart	Ehlers–Danlos syndrome, cardiac valvular form	AR
Collagen II	*COL2A1*	Cartilage, eyes	Spondyloepiphyseal dysplasia, spondyloepimetaphyseal dysplasia, achondrogenesis, hypochondrogenesis, Kniest dysplasia, Stickler syndrome	AD
Collagen III	*COL3A1*	Blood vessels	Ehlers–Danlos syndrome, type IV	AD
Collagen IV	*COL4A1*	Kidney, skin, basement membranes	Familial porencephaly, hereditary angiopathy	AD
	COL4A3, *COL4A4*	Kidney, skin, basement membranes	Alport syndrome, benign familial haematuria	AR, AD
	COL4A5, *COL4A6*	Kidney, skin, basement membranes	Alport syndrome, leiomyomatosis	X
Collagen V	*COL5A1*, *COL5A2*	Skin, joints	Ehlers–Danlos syndrome, type I, II	AD
Collagen VI	*COL6A1*, *COL6A2*, *COL6A3*	Muscle	Bethlem myopathy, Ullrich congenital muscular dystrophy	AD, AR
Collagen VII	*COL7A1*	Skin, dermal–epidermal junction	Dystrophic epidermolysis bullosa	AD, AR
Collagen VIII	*COL8A2*	Cornea	Fuchs corneal dystrophy	AD
Collagen IX	*COL9A1*, *COL9A2*, *COL9A3*	Cartilage	Multiple epiphyseal dysplasia	AD
	COL9A1	Cartilage	Autosomal recessive Stickler syndrome	AR
Collagen X	*COL10A1*	Cartilage, growth plate	Metaphyseal chondrodysplasia, Schmid type	AD
Collagen XI	*COL11A1*, *COL11A2*	Cartilage, eyes	Stickler syndrome, Marshall syndrome	AD
	COL11A2	Cartilage, ears	Otospondylomegaepiphyseal dysplasia	AD, AR
	COL11A2	Ears	Deafness	AD, AR
Decorin	*DCN*	Cornea	Congenital stromal corneal dystrophy	AD
Elastin	*ELN*	Arteries, skin	Supravalvular aortic stenosis, cutis laxa	AD
Fibrillin 1	*FBN1*	Skeleton, eyes, cardiovascular	Marfan syndrome, ectopia lentis, Shprintzen–Goldberg syndrome, Weill–Marchesani syndrome	AD
Fibrillin 2	*FBN2*	Skeleton	Contractural arachnodactyly	AD
Fibronectin	*FN1*	Kidney	Glomerulopathy	AD
Fibulin 4	*FBLN4*	Skin	Cutis laxa	AR
Fibulin 5	*FBLN5*	Eyes	Age-related macular degeneration	AD
	FBLN5	Skin	Cutis laxa	AD, AR
Laminin	*LAMA2*	Muscle	Congenital muscular dystrophy	AR
	LAMA3, *LAMB3*, *LAMC2*	Skin, dermal–epidermal junction	Epidermolysis bullosa, junctional	AR
	LAMB2	Kidney, eyes	Pierson syndrome	AR
Matrilin 3	*MATN3*	Cartilage	Multiple epiphyseal dysplasia	AD
Perlecan	*HSPG2*	Cartilage, basement membranes	Schwartz–Jampel syndrome, dysegmental dysplasia Silverman–Handmaker type	AR
Tenascin XB	*TNXB*	Skin	Ehlers–Danlos-like syndrome	AR
	TNXB	Skin	Ehlers–Danlos syndrome, type III	AD

AD, autosomal dominant; AR, autosomal recessive; COMP, cartilage oligomeric matrix protein (also known as thrombospondin 5); ECM, extracellular matrix; X, X-linked.

Metaphyseal chondrodysplasia, Schmid type
A form of chondrodysplasia that is caused by mutations in collagen X, a component of growth plate cartilage. Growth plates are structurally altered and the chondrocytes are disorganized, causing mild clinical abnormities of bone growth, such as bowed legs and hip problems.

information that implicates the cellular consequences of disturbances to protein folding as a major gain-of-function component of the molecular pathology of these ECM structural mutations.

Collagen mutations have been well studied and serve to illustrate the effect of structural mutations, such as missense mutations and in-frame deletions, on protein assembly at many different levels during synthesis and secretion. There are several excellent reviews containing detailed information on collagen synthesis and assembly[8,40,41]. Briefly, collagen trimers are assembled in the ER via interactions of trimerization domains that direct selection of the appropriate partner chains to form one of the 28 heterotrimeric or homotrimeric forms (FIG. 1). For most collagens the trimerization domains are at the C terminus of the protein. This initial association provides the correct chain registration required for subsequent formation of the collagen triple helix. The structural basis of this vital self-organizing step has been recently reviewed[42]. Folding of the C-terminal trimerization domains probably involves interactions with ER-resident molecular chaperones such as immunoglobulin-heavy-chain-binding protein (BiP; also known as heat shock 70 kDa protein 5, HSPA5)[43,44].

Not surprisingly, mutations of the C-terminal propeptide of many collagens interfere with folding of the domains, and prevent or severely impede trimer assembly. Heterozygous dominant mutations in the collagen I proα1(I) C-propeptide in patients with OI are the cause of abnormal procollagen trimerization, resulting in delayed triple helix folding and reduced secretion[44]. This compromised assembly leads to intracellular degradation of the mutant misfolded proα1(I) chains via the ER-associated proteasomal pathway[45], and results in a major collagen I deficiency in bone. In the OI mouse model, *Oim*, a C propeptide mutation in the collagen I proα2(I) chain also prevents association of the mutant proα2(I) with the proα1(I) chain. In *Oim*/*Oim* homozygotes this results in production of collagen I that contains only proα1(I) trimers rather than the functional collagen I heterotrimers[46]. A similar mutation in the C-propeptide of collagen II also prevents assembly in the *Dmm*/*Dmm* mouse, which is dwarfed and has major cartilage defects[47].

Mutations in *COL10A1* that cause metaphyseal chondrodysplasia, Schmid type, cluster in the C-terminal trimerization domain[24] — further highlighting the importance of the initial chain association step. The crystal structure of the trimerization domain[48] predicts that mutations that disrupt the hydrophobic core of the domain are likely to prohibit correct folding, resulting in exclusion of affected collagen X chains from trimers. The structure also suggested that some types of trimerization domain mutation could permit trimer formation but perturb subsequent collagen X supramolecular network assembly (FIG. 1) or interactions in the cartilage matrix. However, studies on mutant collagen X in transfected cells and in transgenic mice have shown that both classes of missense mutations cause misfolding and severely compromise trimer assembly[49–52].

The next stage of collagen assembly, formation of the triple helix, is also crucial to collagen function (reviewed in REFS 8,40,41). In most collagen types, helix formation along the repetitive Gly–X–Y collagen domain sequence progresses from the C terminus, where the chains are held in register by trimerized C-terminal domains, towards the N terminus (FIG. 1). Helix formation and stabilization involves *cis–trans* isomerization of prolyl peptide bonds by peptidyl-prolyl *cis–trans* isomerase and collaboration of the ER-resident foldase protein disulphide isomerase (PDI), prolyl-4-hydroxylase (P4H), the CRTAP–CYPB–P3H1 complex[39] and 47 kDa heat shock protein (HSP47, also known as serpin 1)[53]. Prolines in the Y position of collagen Gly–X–Y triplet sequences are hydroxylated by P4H. This step is crucial, as hydroxyproline provides the hydrogen bonding force necessary to stabilize the collagen helix. In addition to prolines, some lysine residues are also hydroxylated by lysyl

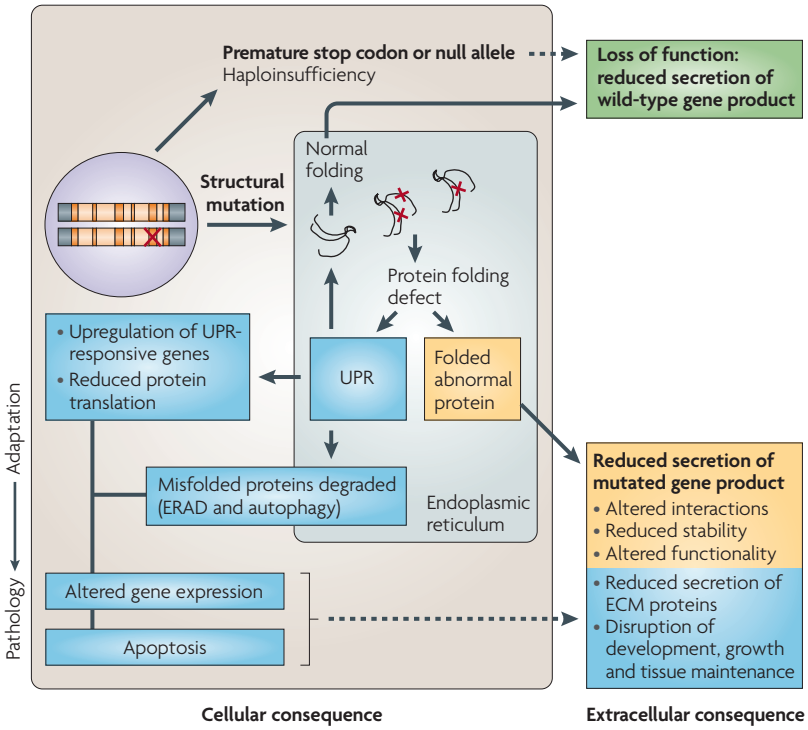


Figure 2 | **The extracellular matrix (ECM) disease paradigm.** The existing model for ECM mutation pathophysiology proposes that the extracellular consequences account for the molecular pathology. Reduced synthesis owing to regulatory mutations or decay of mRNA-containing premature termination mutations results in deficiency of the protein in the ECM, thereby compromising function (green box). However, structural misfolding mutations have a dominant negative effect, leading to partial or complete cellular retention and/or degradation of mutant proteins, and normal proteins being assembled into mutant-containing multimers. This results in a severe protein deficiency and, if the mutant abnormally folded protein is secreted, a further deleterious effect on ECM stability or function (yellow boxes). The new paradigm for understanding ECM mutations also considers cellular consequences that might result from endoplasmic reticulum (ER) stress, such as the unfolded protein response (UPR), which is induced by retention of misfolded proteins in the ER (blue boxes). The UPR is initially an adaptive response but, if unresolved, can lead to changes in gene expression that result in disruption of cellular gene expression patterns, and eventually apoptosis and pathology. The relative contribution of the extracellular and cellular consequences to the molecular pathology is likely to show considerable mutation and gene specificity. ERAD, ER-associated degradation.

hydroxylases and some of these are further modified by addition of galactose or galactosyl-glucose. Formation of the triple helix prevents further post-translational modification. After helix formation, HSP47 stabilizes the helix and prevents aggregation of the collagen in the ER, and is important for efficient secretion, processing and fibril formation[53] (FIG. 1).

Mutations that interfere with the triple helix are by far the most prevalent group of collagen mutations. The most common of these are glycine substitutions, which interrupt the obligatory Gly–X–Y repeat sequence, causing misfolding and a structurally abnormal helix. Glycine mutations in collagen I cause OI; in collagen II they cause a range of chondrodysplasias, including spondyloepiphyseal dysplasia, Kniest dysplasia, achondrogenesis, hypochondrogenesis and Stickler syndrome; and in collagen III they cause Ehlers–Danlos syndrome type IV. Although collagen I glycine mutations will be discussed as the archetype, glycine mutations in the triple helical domain of most collagen types will have similar destabilizing effects, but the clinical consequences will depend on the structural role of the helix in the particular collagen and the role of the collagen type in ECM architecture.

Glycine mutations generally cause major disruptions to helix folding, delaying helix propagation at the site of the mutation. This pause in helix formation exposes the unfolded portions of the chains that lie N-terminal to the mutation to additional post-translational modification, resulting in increased hydroxylation and glycosylation. In OI, 682 of the 832 independent mutations reported (82%) are glycine substitutions in either the collagen I proα1(I) or proα2(I) chains[54]. Substitutions in the most N-terminal 20% of the helix are non-lethal, whereas more C-terminal substitutions are of variable phenotype but are, in general, more severe. This is broadly consistent with predictions from the long-held view that mutations that disturb helix formation closer to the site of propagation (the C terminus) are more disruptive. However, for substitutions in the most C-terminal 80% of the helix there is no apparent correlation between disease severity and the position of the mutation[54], or disease severity and regions of different local helix stability[55], suggesting a more complex relationship between mutations and phenotype.

By contrast, the nature of the substitution is important in disease severity. In *COL1A1,* substitution by amino acids with charged or branched side chains (Asp, Arg or Val) causes more disruption to the helix, and usually results in lethal OI phenotypes[54]. A significant consequence of impaired collagen folding is reduced secretion of trimers containing mutant collagen chains. As just one mutant chain in a trimer will impair helix formation, heterozygous mutations have a dominant negative effect. In heterotrimers, such as collagen I $[\alpha1(I)]_2\alpha2(I)$, three-quarters of the collagen trimers contain one or more abnormal chains if one *COL1A1* allele is mutated. In the case of homotrimers, such as collagen II $[\alpha1(II)]_3$, seven-eights of the trimers will have abnormal helix folding if one *COL2A1* allele is mutated.

For secretion from the cell, the collagen triple helix must be correctly folded. Collagen molecules containing mutant chains are secreted poorly and are largely retained within the ER. Collagen chains containing mutations that affect initial chain association, such as those in the proα1(I) C-propeptide, are removed by retrotranslocation of monomeric unfolded mutant collagen chains into the cytosol followed by proteasomal degradation (ER-associated degradation, ERAD)[44,45]. However, there is no evidence that ERAD degrades molecules containing a triple-helical glycine substitution[56]. Indeed, although collagens containing helix mutations have unstable poorly formed triple helices, they do associate at their C termini and form trimers, and this is likely to preclude them from retrotranslocation and ERAD.

There are no published data on the mechanisms of degradation of collagen trimers that contain helix mutations. However, some clues exist from studies using cells from *Hsp47*-null mice. In HSP47-deficient cells collagen triple helix formation and stability is impaired, and the improperly folded triple helices form insoluble aggregates in the ER[53]. Autophagy, a degradation mechanism commonly deployed to degrade protein aggregates[57], is therefore a likely mechanism for degradation of collagen trimers containing helix mutations. Mutant chain-containing collagen that exits the cell can also have important extracellular affects. If incorporated into collagen fibrils it might have a destabilizing effect and be selectively degraded[58], or it might compromise the interactions of collagen with other ECM ligands[54], disturbing ECM architecture and stability.

Collagen VI provides an example of mutations that can also affect the higher levels of supramolecular assembly. Glycine substitutions and exon-skipping mutations towards the N terminus of the helix can have a severe dominant negative effect by interfering with intracellular formation of the larger multimers, which is necessary for secretion and microfibril formation[59–61] (FIG. 1). Another striking example of how structural mutations can dominantly affect ECM protein assembly is provided by COMP in two skeletal dysplasias, pseudoachondroplasia and multiple epiphyseal dysplasia[62,63]. Structural mutations in COMP affect protein folding and assembly, in many cases causing retention of the misfolded protein within the ER. Because COMP assembles into a pentamer (FIG. 1), almost all the multimers will contain at least one structurally abnormal mutant chain, resulting in intracellular retention of COMP. Accumulation of intracellular COMP can also result in co-retention of interacting partners, including collagen IX and matrilin 3 (REFS 64,65).

Misfolded ECM proteins cause ER stress

In addition to dominant effects of mutations exerted through reduced rates of synthesis and secretion, or disturbed interactions in the ECM as discussed above, misfolded mutant ECM proteins such as COMP, collagens and matrilin 3 have recently been shown to induce significant ER stress and trigger the unfolded protein response (UPR). ER stress and the UPR have been extensively reviewed[66–69], and the main features are summarized in BOX 2. The UPR pathway evolved to allow cells to adjust

Dystrophic epidermolysis bullosa
A severe genetic disorder resulting in extremely fragile skin and recurrent blister formation caused by mutations in collagen VII.

Dominant negative
A form of mutation that interferes with the function of its wild-type allele product.

Bruck syndrome
A recessive form of osteogenesis imperfecta, with joint contractures caused by mutations in the collagen-modifying enzyme lysyl hydroxylase 2.

ER-associated degradation
(ERAD). An intracellular quality control pathway that directs retrotranslocation of normal and misfolded proteins from the endoplasmic reticulum to the cytoplasm for proteasomal degradation.

Autophagy
In autophagy the cell is degraded largely from within, with little or no help from phagocytes. Bulk cytoplasm and organelles are sequestered within double-membrane-bound vesicles. These ultimately fuse with the lysosome and their contents are degraded.

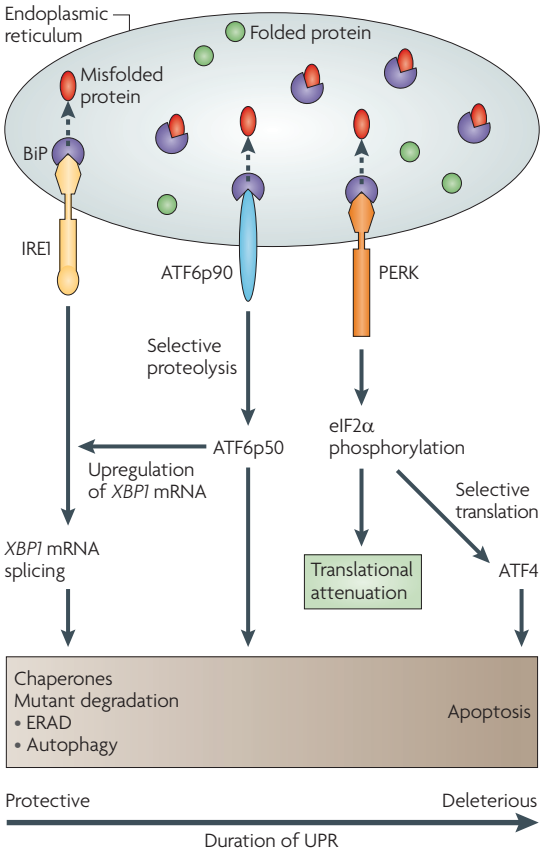
Box 2 | **The unfolded protein response**

One of the major functions of the endoplasmic reticulum (ER) is the correct folding and maturation of proteins and glycoproteins that are destined for secretion. This protein folding factory imposes stringent quality control, such that only correctly modified functional proteins leave the ER[108]. Incorrectly folded proteins are bound by immunoglobulin-heavy-chain-binding protein (BiP; also known as heat shock 70 kDa protein 5, HSPA5). This removes BiP from the ER luminal domains of the three major transmembrane stress sensors, IRE1, PERK and ATF6, which activates the sensors and initiates the unfolded protein response (UPR).

IRE1 activation is transmitted to the cytosolic domain of IRE1, which contains a serine/threonine kinase and site-specific RNase activities. The RNase cleaves X-box binding protein 1 (*XBP1*) mRNA; this splicing leads to the production of an active transcription factor, $XBP1_s$. This transcription factor acts by binding to the UPR-responsive elements in the promoters of a subset of genes to stimulate the synthesis of chaperones as a mechanism to cope with the unfolded protein load.

A second arm of the UPR is mediated by the transcription factor ATF6, which contains a basic leucine zipper domain. The release of BiP from the luminal domain allows uncleaved ATF6 (ATF6p90) to transit to the Golgi network, where it is cleaved to generate an active cytosolic fragment (ATF6p50). ATF6p50 migrates to the nucleus and binds to the promoters of genes containing an ER stress-responsive element.

Release of BiP from the luminal domain of PERK induces dimerization of PERK, which can then phosphorylate the translational initiation factor eIF2α. This prevents the formation of the translational initiation complex, and thus downregulates general translation and reduces the protein folding load. Although phosphorylated eIF2α inhibits translation of most mRNAs, it also promotes translation of a subset of stress response genes, including activating transcription factor 4 (ATF4). ATF4 subsequently upregulates the transcription of numerous genes that are involved with amino acid metabolism and transport, oxidation–reduction reactions, and ER stress-induced apoptosis genes such as *CHOP*[68].



In addition to translational attenuation, misfolded proteins can be removed by ER-associated degradation (ERAD) and/or autophagy[57,109,110], which are complementary processes to reduce the unfolded protein load and promote cell survival. In the ERAD process, the misfolded or unfolded proteins are retrotranslocated from the ER to the cytoplasm, where they are ubiquitylated and degraded by proteasomes. Autophagy is a collection of pathways that result in sections of the cytoplasm, including organelles, becoming sequestered into membrane-bound compartments that then fuse with lysosomes, where their contents are degraded by acid hydrolases.

The initial activation of the UPR is cytoprotective, offering the cells an opportunity to return to protein folding homeostasis. But with increased duration of unfolded protein load the balance in activities of the three pathways changes, and prolonged ER stress results in deleterious effects such as apoptosis.

the folding capacity of the ER to differing protein folding loads. The UPR is deployed as a cytoprotective strategy to restore protein folding homeostasis when misfolded proteins are present in the ER, but it can also contribute to the pathophysiology of many heritable ECM disorders (FIG. 2). Although this is a recent concept for ECM disorders, it is unsurprising in light of the many studies that implicate elevated ER stress and its consequences as significant contributors to the pathology of an increasing number of human disorders[70–73].

In pseudoachondroplasia and multiple epiphyseal dysplasia, *COMP* structural mutations that cause misfolding result in a characteristic distension of the ER[62,63] and retention of mutant COMP. Although this disrupts the normal secretion of crucial cartilage ECM components, it also induces ER stress. Evidence of ER stress comes from analyses of patient cells and cells transfected with mutant *COMP*, which show co-localization of molecular chaperones, such as calnexin, HSP47 (REF. 74), PDI[74,75], calreticulin[75,76] and BiP[75,76], with the mutant COMP in the ER. Furthermore, phosphorylation of the eukaryotic translation initiation factor eIF2α, an ER stress marker, was increased in COS cells expressing mutant COMP[77]. In a knockin mouse model containing a mild pseudoachondroplasia *Comp* mutation, no intracellular accumulation was noted, although a UPR ensued, characterized by upregulation of the chaperones BiP and calreticulin and activation of eIF2α and activating transcription factor 6 (ATF6)[78]. These data suggest that although the extent of the trafficking defect

Proteasome
A large cytoplasmic protein complex that degrades proteins to which ubiquitin has been added by a process that requires ATP.

might be mutation specific, protein misfolding resulting in UPR activation seems to be a common feature with COMP structural mutations. These studies also demonstrate reduced chondrocyte proliferation and increased and spatially dysregulated apoptosis[78]. Apoptosis, the downstream consequence of unresolved ER stress, has also been observed in other *in vivo*[79,80] and *in vitro*[77,81] studies on COMP mutations.

Mutations in matrilin 3 can also cause the pseudoachondroplasia or multiple epiphyseal dysplasia phenotype[63], and recent studies implicate matrilin misfolding and UPR activation in a disease model produced by a knockin mutation of matrilin 3 (REF. 82). In this model, chondrocyte proliferation was reduced, levels of the chaperones BiP and GRP94 were increased, and apoptosis was dysregulated in the growth plates of affected mice. These data strongly support the hypothesis that a component of the pathophysiology of COMP and matrilin 3 misfolding mutations is activation of the UPR. Although a direct link between the UPR and chondrocyte proliferation and apoptosis remains to be proven, it seems likely that the combined effects of the UPR and diminished secretion of functional COMP–collagen IX–matrilin 3 assemblies into the cartilage ECM account for the phenotypes in these disorders.

For fibrillar collagens I and II, evidence is also mounting that misfolding mutations initiate a UPR with deleterious cellular consequences. In OI, *COL1A1* C-terminal trimerization domain mutants bind to BiP and upregulate expression of both BiP and GRP94 (REFS 43,44). These mutant chains are targeted for degradation via the proteasomal ERAD system[45]. A mouse model of OI (*Aga2*) provides further support for the contribution of the UPR to the clinical phenotype[83]. In this mouse model a collagen I C-propeptide mutation causes ER retention of collagen and increases caspase-induced apoptosis and levels of HSP47 and CHOP in osteoblasts, both *in vitro* and *in vivo*. These data provide support for the proposal that changes in cell behaviour as a consequence of UPR activation are an important component of the pathology in conditions caused by problems with collagen trimer association in the ER.

This raises the important question of whether the more common collagen helix mutations also trigger ER stress and whether this contributes to the clinical phenotype of OI (for collagen I mutations), chondrodysplasias (for collagen II mutations) and other collagenopathies. Helix mutations allow initial chain association, but then impair subsequent folding of the triple helix. Early studies suggested that the glycine mutations in the collagen I helix do not bind BiP, but instead bind to another ER-resident foldase, protein disulphide isomerase, which has isomerase and chaperone activities[84]. Studies in an OI mouse model with an engineered *Col1a1* helix glycine mutation provided a preliminary indication that CHOP, a key pro-apoptotic regulator[85], is increased in bone when this mutation is present. As CHOP is upregulated by the UPR, these results suggest that helix glycine mutations can also trigger a form of UPR. However, because BiP was not upregulated in this model, alternative mechanisms for sensing helix misfolding mutations might exist and require further study. Recent studies on cells transfected with a form of collagen II that contains an arginine to cysteine mutation towards the C terminus of the helix[86] demonstrated a UPR characterized by upregulation and binding of BiP to the mutant unfolded protein and the expression of apoptosis markers. More N-terminal arginine to cysteine mutations, or a helical glycine to glutamic acid substitution, did not bind BiP or elicit an apoptotic response[86]. By contrast, BiP and CHOP were upregulated in a mouse expressing a collagen II helical glycine to cysteine mutation[52].

Although these studies support an important role for the UPR in cell dysfunction, the most convincing data so far comes from studies on collagen X mutations in metaphyseal chondrodysplasia, Schmid type. Mutations in the collagen X C-terminal trimerization domain compromise trimer assembly. *In vitro* studies show that the mutant collagen X protein misfolds, forming aberrant disulphide-bonded dimers[50], causing upregulation of ER chaperones including BiP, splicing of X-box binding protein 1 (*XBP1*) mRNA, and ERAD, resulting in little or no collagen X secretion[50]. Activation of the UPR has been confirmed in more detail in transgenic mouse models[51,52], in which mutations in the *Col10a1* trimerization domain are expressed in growth plate cartilage. Expression of the mutant protein led to upregulation of BiP and CHOP, and to XBP1 splicing, indicative of the UPR. Importantly, although CHOP was upregulated, apoptosis did not occur and the ER-stressed chondrocytes survived. However, there were significant changes to hypertrophic chondrocytes, such as cell cycle re-entry and expression of genes from the prehypertrophic stages of cartilage development. This was proposed as an adaptive response to the UPR; by downregulating collagen X expression and reverting to a less mature phenotype the cells survive, although at a significant cost to their normal function.

Because collagen X secretion is reduced as a result of intracellular degradation of the misfolded mutant protein, it was considered possible that the major changes in gene expression could result from the effect of the reduced collagen ECM, rather than from a direct downstream regulatory consequence of the misfolded collagen X causing a UPR. However, this effect can be excluded because heterozygous and homozygous *Col10a1*-null mice do not have any of these changes in cellular differentiation[87]. Recent studies with a knockin *Col10a1* trimerization domain mutation (R.P.B.-H., unpublished data) also show characteristic growth plate expansion, upregulation of BiP, GRP94 and the protein disulphide isomerase ERp72, and other hallmarks of the UPR, along with profound downstream alterations to gene expression patterns.

Implications for therapy

Therapeutic strategies for ECM disorders caused by structural mutations, such as in forms of OI, have focused on approaches to increase expression of the normal ECM product, in this case collagen I, or to suppress mutant protein expression[88]. Cell therapy approaches using mesenchymal stem cells[88,89] with the ability to differentiate

Chondrocyte
Cartilage cells that produce the structural components of cartilage.

Osteoblast
A mesenchymal cell with the capacity to differentiate into bone tissue.

Mesenchymal stem cells
Multipotent mesenchymally derived stem cells that can differentiate into a variety of cell types, including osteoblasts, chondrocytes, myocytes and adipocytes.

into bone cells have been explored *in vitro* and in animal models of OI. Allogeneic mesenchymal stem cells administered in small clinical trials showed poor engraftment, although bone growth improvements were reported in some patients. Gene therapy approaches to inducing increased target gene expression have been explored *in vitro* and in mouse models, but for the treatment of patients these suffer from the current limitations of these techniques[88]. As many of the structural mutations exert a gain-of-function effect either by interference with the interactions, assembly and integrity of the ECM, an important target in the development of therapeutic approaches has involved mutant gene expression knockdown. A number of approaches have been explored, including antisense oligonucleotides, ribozymes and small interfering RNAs[88]. Although these offer potential and are being explored for efficacy, they are hampered by the need to develop mutation-specific knockdown tools to achieve near complete ablation of mutant gene expression, as production of even small amounts of structurally abnormal protein can exert strong dominant negative effects.

The emerging importance of protein folding abnormalities and the concomitant ER stress in the pathology of a range of connective tissue disorders offers the possibility of more 'generic' new treatment strategies. If the misfolded protein load in the ER can be reduced to levels that can be managed by the cell, then the serious deleterious outcomes of an unresolved UPR, such as apoptosis, could be ameliorated. This could be achieved by correcting the protein folding defect, stimulating rapid degradation or blocking translation of the mutant protein[90,91]. One promising approach is the use of pharmacological agents, such as small chemical chaperones, which can stabilize proteins in their native conformation and rescue mutant protein folding and/or trafficking defects[91–94]. Overexpression of the endogenous chaperone BiP can reduce ER stress[95], and recent studies identified a small chemical that induces BiP and protects against ER stress in neurons[96].

Another therapeutic approach that offers possibilities is stimulation of the bulk destruction of ER containing the mutant protein by autophagy using rapamycin, an mTOR inhibitor, or other drugs that enhance autophagy[97–99]. In another approach, a selective inhibitor of dephosphorylation of eIF2α protected cells from ER stress[100]. Manipulation of the downstream consequences of the UPR, such as apoptosis, also offers therapeutic potential in the treatment of heritable ECM disorders. However, it is important to temper our enthusiasm with the cautionary realization that some of these approaches might result in increased secretion of mutant dysfunctional protein with the potential to exert deleterious effects on the ECM. It is therefore vital that we gain a more comprehensive molecular understanding of the contribution of the extracellular and intracellular components to inherited ECM disease pathology.

In this Review we have concentrated on how structural mutations and protein misfolding cause ECM disorders, but it is important to recognize other disease mechanisms, such as those exemplified by fibrillin 1 mutations in Marfan syndrome[101]. These mutations reduce the levels of extracellular fibrillin-rich microfibrils, which normally act as a transforming growth factor-β (TGFβ) reservoir, resulting in disturbances to the normal regulation of TGFβ signalling. In these instances, treatment of mouse models and patients with TGFβ antagonists to attenuate TGFβ signalling is providing important therapeutic benefits[102,103], even though other ECM structural deficiencies and possibly unfolded protein effects are not corrected.

A novel therapeutic strategy for an ECM disease is suggested by studies on the collagen VI knockout mouse muscular dystrophy model. Myofibres from these mice have ultrastructural ER and mitochondrial defects and increased apoptosis, which are thought to be a result of mitochondrial depolarization and Ca^{2+} deregulation[104]. Treatment with cyclosporine A, an inhibitor of the mitochondrial transition pore, rescued the ultrastructural defects and decreased apoptosis[104]. An open pilot trial with cyclosporine A in five patients with collagen VI mutations also showed reduced apoptosis in muscle biopsies[105]. Although these early results are encouraging, improvement in muscle function has not been demonstrated and the molecular basis for the cyclosporine A effect is controversial and requires further study[106].

Perspectives and future directions

The long-standing view has been that mutations cause ECM dysfunction by combinations of two mechanisms, both of which ultimately have an impact extracellularly on the quality and integrity of the matrix that surrounds cells. The first mechanism involves a quantitative reduction in ECM components by mutations affecting synthesis, or by structural mutations causing cellular retention and/or degradation. Second, secretion of mutant protein can disturb the ECM qualitatively, compromising crucial interactions, structure and stability.

However, in this Review we have presented recent evidence suggesting that there is another significant player in the molecular pathology of these disorders: ER stress. This ER stress results from the intracellular effect of misfolded ECM proteins in the ER eliciting the UPR. The relative contribution of each of the intracellular and extracellular components to pathophysiology (FIG. 2) will depend on the mutation and will be context dependant. In most cases it would seem likely that both gain-of-function UPR consequences and alterations to the ECM, either by reduced secretion, altered interactions or composition, will contribute to the disease mechanism. However, new experiments in which ER stress is triggered in hypertrophic chondrocytes *in vivo* by expressing an exogenous misfolding protein under the control of the collagen X promoter are indicating that the initiation of an UPR can, by itself, lead to growth plate cartilage pathology similar to that seen with collagen X misfolding mutations (R.P.B.-H., unpublished data). These findings make it clear that in disorders involving ECM protein misfolding, the relative contribution of the cellular effects of the UPR and its downstream consequences, such as apoptosis and altered gene expression, and the

Allogeneic
In allogeneic transplants, cells, organs or tissues from any human other than self or a monozygotic twin are used for therapeutic purposes.

mTOR
The mammalian target of rapamycin is a serine/threonine protein kinase that regulates cell growth, proliferation, survival, protein synthesis and transcription.

Transforming growth factor-β
(TGFβ). A secreted protein that controls cellular proliferation, differentiation and other functions in most cells. It has a role in immunity, cancer, heart disease and Marfan syndrome.

extracellular dominant negative disturbance of the ECM on pathophysiology must be thoroughly assessed.

There are numerous important questions that will need to be addressed before we can fully understand the molecular pathology. Which UPR-triggering pathways are used and which downstream signalling and gene expression pathways are activated? How are outcomes affected by mutant protein levels and duration of expression? And to what extent are the mutant proteins degraded, and by which pathways? Developing this level of understanding of the role of the UPR in ECM protein misfolding disorders will require additional mouse genetic models in which the UPR and ECM effects can be assessed in the *in vivo* developmental context, along with *in vitro* studies on transfected cells in which protein expression levels and timing can be experimentally manipulated.

1. Iozzo, R. V. The biology of the small leucine-rich proteoglycans. Functional network of interactive proteins. *J. Biol. Chem.* **274**, 18843–18846 (1999).
2. Svensson, L., Oldberg, A. & Heinegard, D. Collagen binding proteins. *Osteoarthr. Cart.* **9** (Suppl. A), S23–S28 (2001).
3. Eyre, D. R. Collagens and cartilage matrix homeostasis. *Clin. Orthop. Relat. Res.* S118–S122 (2004).
4. Alford, A. I. & Hankenson, K. D. Matricellular proteins: extracellular modulators of bone development, remodeling, and regeneration. *Bone* **38**, 749–757 (2006).
5. Heino, J. The collagen family members as cell adhesion proteins. *Bioessays* **29**, 1001–1010 (2007).
6. Lamoureux, F., Baud'huin, M., Duplomb, L., Heymann, D. & Redini, F. Proteoglycans: key partners in bone cell biology. *Bioessays* **29**, 758–771 (2007).
7. Kadler, K. E., Baldock, C., Bella, J. & Boot-Handford, R. P. Collagens at a glance. *J. Cell Sci.* **120**, 1955–1958 (2007).
8. Myllyharju, J. & Kivirikko, K. I. Collagens, modifying enzymes and their mutations in humans, flies and worms. *Trends Genet.* **20**, 33–43 (2004).
Succinct review of the collagen protein family, including supramolecular structure, biosynthesis and mutations in collagens and their modifying enzymes.
9. Nelson, C. M. & Bissell, M. J. Of extracellular matrix, scaffolds, and signaling: tissue architecture regulates development, homeostasis, and cancer. *Annu. Rev. Cell Dev. Biol.* **22**, 287–309 (2006).
10. Gieni, R. S. & Hendzel, M. J. Mechanotransduction from the ECM to the genome: are the pieces now in place? *J. Cell Biochem.* **104**, 1964–1987 (2007).
11. Danen, E. H. & Sonnenberg, A. Integrins in regulation of tissue development and function. *J. Pathol.* **201**, 632–641 (2003).
12. Whitelock, J. M., Melrose, J. & Iozzo, R. V. Diverse cell signaling events modulated by perlecan. *Biochemistry* **47**,11174–11183 (2008).
13. ten Dijke, P. & Arthur, H. M. Extracellular control of TGFβ signalling in vascular development and disease. *Nature Rev. Mol. Cell Biol.* **8**, 857–869 (2007).
14. Aszodi, A., Legate, K. R., Nakchbandi, I. & Fassler, R. What mouse mutants teach us about extracellular matrix function. *Annu. Rev. Cell Dev. Biol.* **22**, 591–621 (2006).
Comprehensive review on the use of mutant mouse models to explore the structure and function of the ECM.
15. Frischmeyer, P. A. & Dietz, H. C. Nonsense-mediated mRNA decay in health and disease. *Hum. Mol. Genet.* **8**, 1893–1900 (1999).
16. Schell, T., Kulozik, A. E. & Hentze, M. W. Integration of splicing, transport and translation to achieve mRNA quality control by the nonsense-mediated decay pathway. *Genome Biol.* **3**, REVIEWS1006 (2002).
17. Weischenfeldt, J., Lykke-Andersen, J. & Porse, B. Messenger RNA surveillance: neutralizing natural nonsense. *Curr. Biol.* **15**, R559–R562 (2005).
18. Isken, O. & Maquat, L. E. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes Dev.* **21**, 1833–1856 (2007).
Reviews the molecular mechanisms of NMD.
19. Willing, M. C. *et al.* Osteogenesis imperfecta type I: molecular heterogeneity for COL1A1 null alleles of type I collagen. *Am. J. Hum. Genet.* **55**, 638–647 (1994).
20. Dalgleish, R. The human type I collagen mutation database. *Nucleic Acids Res.* **25**, 181–187 (1997).
21. Richards, A. J. *et al.* High efficiency of mutation detection in type 1 Stickler syndrome using a two-stage approach: vitreoretinal assessment coupled with exon sequencing for screening COL2A1. *Hum. Mutat.* **27**, 696–704 (2006).
22. Snead, M. P. & Yates, J. R. Clinical and molecular genetics of Stickler syndrome. *J. Med. Genet.* **36**, 353–359 (1999).
23. Lamandé, S. R. *et al.* Reduced collagen VI causes Bethlem myopathy: a heterozygous *COL6A1* nonsense mutation results in mRNA decay and functional haploinsufficiency. *Hum. Mol. Genet.* **7**, 981–989 (1998).
24. Bateman, J. F., Wilson, R., Freddi, S., Lamande, S. R. & Savarirayan, R. Mutations of *COL10A1* in Schmid metaphyseal chondrodysplasia. *Hum. Mutat.* **25**, 525–534 (2005).
25. Lucarini, L. *et al.* A homozygous *COL6A2* intron mutation causes in-frame triple-helical deletion and nonsense-mediated mRNA decay in a patient with Ullrich congenital muscular dystrophy. *Hum. Genet.* **117**, 460–466 (2005).
26. Peat, R. A., Baker, N. L., Jones, K. J., North, K. N. & Lamandé, S. R. Variable penetrance of *COL6A1* null mutations: implications for prenatal diagnosis and genetic counselling in Ullrich congenital muscular dystrophy families. *Neuromuscul. Disord.* **17**, 547–557 (2007).
27. Christiano, A. M., Amano, S., Eichenfield, L. F., Burgeson, R. E. & Uitto, J. Premature termination codon mutations in the type VII collagen gene in recessive dystrophic epidermolysis bullosa result in nonsense-mediated mRNA decay and absence of functional protein. *J. Invest. Dermatol.* **109**, 390–394 (1997).
28. Hovnanian, A. *et al.* Characterization of 18 new mutations in COL7A1 in recessive dystrophic epidermolysis bullosa provides evidence for distinct molecular mechanisms underlying defective anchoring fibril formation. *Am. J. Hum. Genet.* **61**, 599–610 (1997).
29. Bateman, J. F., Freddi, S., Nattrass, G. & Savarirayan, R. Tissue-specific RNA surveillance? Nonsense-mediated mRNA decay causes collagen X haploinsufficiency in Schmid metaphyseal chondrodysplasia cartilage. *Hum. Mol. Genet.* **12**, 217–225 (2003).
30. Khajavi, M., Inoue, K. & Lupski, J. R. Nonsense-mediated mRNA decay modulates clinical outcome of genetic disease. *Eur. J. Hum. Genet.* **14**, 1074–1081 (2006).
31. Tan, J. T. *et al.* Competency for nonsense-mediated reduction in collagen X mRNA is specified by the 3′ UTR and corresponds to the position of mutations in Schmid metaphyseal chondrodysplasia. *Am. J. Hum. Genet.* **82**, 786–793 (2008).
32. Colige, A. *et al.* Novel types of mutation responsible for the dermatosparactic type of Ehlers–Danlos syndrome (type VIIC) and common polymorphisms in the *ADAMTS2* gene. *J. Invest. Dermatol.* **123**, 656–663 (2004).
33. Brunetti-Pierri, N. *et al.* X-linked recessive chondrodysplasia punctata: spectrum of arylsulfatase E gene mutations and expanded clinical variability. *Am. J. Med. Genet. A* **117A**, 164–168 (2003).
34. Dawson, P. A. & Markovich, D. Pathogenetics of the human SLC26 transporters. *Curr. Med. Chem.* **12**, 385–396 (2005).
35. Ha-Vinh, R. *et al.* Phenotypic and molecular characterization of Bruck syndrome (osteogenesis imperfecta with contractures of the large joints) caused by a recessive mutation in *PLOD2*. *Am. J. Med. Genet. A* **131**, 115–120 (2004).
36. Salo, A. M. *et al.* A connective tissue disorder caused by mutations of the lysyl hydroxylase 3 gene. *Am. J. Hum. Genet.* **83**, 495–503 (2008).
37. Morello, R. *et al.* CRTAP is required for prolyl 3- hydroxylation and mutations cause recessive osteogenesis imperfecta. *Cell* **127**, 291–304 (2006).
With reference 38 this study provides new insights into how mutations affecting the collagen folding machinery, in addition to collagen I mutations, can cause collagen defects and OI.
38. Baldridge, D. *et al. CRTAP* and *LEPRE1* mutations in recessive osteogenesis imperfecta. *Hum. Mutat.* **29**, 1435–1442 (2008).
39. Vranka, J. A., Sakai, L. Y. & Bachinger, H. P. Prolyl 3-hydroxylase 1, enzyme characterization and identification of a novel family of enzymes. *J. Biol. Chem.* **279**, 23615–23621 (2004).
40. Canty, E.G. & Kadler, K. E. Procollagen trafficking, processing and fibrillogenesis. *J. Cell Sci.* **118**, 1341–1353 (2005).
41. Kielty, C. M., Hopkinson, I. & Grant, M. E. in *Connective Tissue and its Heritable Disorders. Molecular, Genetic, and Medical Aspects*. (eds Royce, P. M. & Steinmann, B.) 103–147 (Wiley-Liss, Inc., New York, 1993).
42. Khoshnoodi, J., Cartailler, J. P., Alvares, K., Veis, A. & Hudson, B. G. Molecular recognition in the assembly of collagens: terminal noncollagenous domains are key recognition modules in the formation of triple helical protomers. *J. Biol. Chem.* **281**, 38117–38121 (2006).
43. Chessler, S. D. & Byers, P. H. BiP binds type I procollagen pro alpha chains with mutations in the carboxyl-terminal propeptide synthesized by cells from patients with osteogenesis imperfecta. *J. Biol. Chem.* **268**, 18226–18233 (1993).
A seminal paper that identified binding of the ER chaperone BiP to mutant procollagen, providing the first molecular evidence for what subsequently became known as an ER stress response to mutant collagen I in OI.
44. Lamandé, S. R. *et al.* Endoplasmic reticulum-mediated quality control of type I collagen production by cells from osteogenesis imperfecta patients with mutations in the pro alpha 1 (I) chain carboxyl-terminal propeptide which impair subunit assembly. *J. Biol. Chem.* **270**, 8642–8649 (1995).
This paper described the role of ERAD in the degradation of mutant misfolded type I procollagen, a further early indication that ER stress might be involved with these collagen I mutations
45. Fitzgerald, J., Lamandé, S. R. & Bateman, J. F. Proteasomal degradation of unassembled mutant type I collagen pro-alpha1(I) chains. *J. Biol. Chem.* **274**, 27392–27398 (1999).
46. Chipman, S. D. *et al.* Defective pro alpha 2(I) collagen synthesis in a recessive mutation in mice: a model of human osteogenesis imperfecta. *Proc. Natl Acad. Sci. USA* **90**, 1701–1705 (1993).
47. Fernandes, R. J., Seegmiller, R. E., Nelson, W. R. & Eyre, D. R. Protein consequences of the *Col2a1* C-propeptide mutation in the chondrodysplastic Dmm mouse. *Matrix Biol.* **22**, 449–453 (2003).
48. Bogin, O. *et al.* Insight into Schmid metaphyseal chondrodysplasia from the crystal structure of the collagen X NC1 domain trimer. *Structure* **10**, 165–173 (2002).
49. Wilson, R., Freddi, S. & Bateman, J. F. Collagen X chains harboring Schmid metaphyseal chondrodysplasia NC1 domain mutations are selectively retained and degraded in stably transfected cells. *J. Biol. Chem.* **277**, 12516–12524 (2002).
50. Wilson, R., Freddi, S., Chan, D., Cheah, K. S. & Bateman, J. F. Misfolding of collagen X chains harboring Schmid metaphyseal chondrodysplasia mutations results in aberrant disulfide bond formation, intracellular retention, and activation of the unfolded protein response. *J. Biol. Chem.* **280**, 15544–15552 (2005).
51. Ho, M. S. *et al. COL10A1* nonsense and frame-shift mutations have a gain-of-function effect on the growth plate in human and mouse metaphyseal chondrodysplasia type Schmid. *Hum. Mol. Genet.* **16**, 1201–1215 (2007).
52. Tsang, K. Y. *et al.* Surviving endoplasmic reticulum stress is coupled to altered chondrocyte differentiation and function. *PLoS Biol.* **5**, e44 (2007).
An important paper providing the first description of alterations to gene expression and cellular differentiation as a consequence of mutant collagen-induced ER-stress *in vivo*.

53. Ishida, Y. *et al.* Type I collagen in Hsp47-null cells is aggregated in endoplasmic reticulum and deficient in N-propeptide processing and fibrillogenesis. *Mol. Biol. Cell* **17**, 2346–2355 (2006).
54. Marini, J. C. *et al.* Consortium for osteogenesis imperfecta mutations in the helical domain of type I collagen: regions rich in lethal mutations align with collagen binding sites for integrins and proteoglycans. *Hum. Mutat.* **28**, 209–221 (2007).
A comprehensive review of collagen I mutations in OI and discussion of the possible molecular effects of the mutations.
55. Makareeva, E. *et al.* Structural heterogeneity of type I collagen triple helix and its role in osteogenesis imperfecta. *J. Biol. Chem.* **283**, 4787–4798 (2008).
56. Forlino, A., Kuznetsova, N. V., Marini, J. C. & Leikin, S. Selective retention and degradation of molecules with a single mutant alpha1(I) chain in the Brtl IV mouse model of OI. *Matrix Biol.* **26**, 604–614 (2007).
57. Ding, W. X. & Yin, X. M. Sorting, recognition and activation of the misfolded protein degradation pathways through macroautophagy and the proteasome. *Autophagy* **4**, 141–150 (2008).
58. Bateman, J. F. & Golub, S. B. Deposition and selective degradation of structurally-abnormal type I collagen in a collagen matrix produced by osteogenesis imperfecta fibroblasts *in vitro*. *Matrix Biol.* **14**, 251–262 (1994).
59. Baker, N. L. *et al.* Dominant collagen VI mutations are a common cause of Ullrich congenital muscular dystrophy. *Hum. Mol. Genet.* **14**, 279–293 (2005).
60. Lamandé, S. R., Shields, K. A., Kornberg, A. J., Shield, L. K. & Bateman, J. F. Bethlem myopathy and engineered collagen VI triple helical deletions prevent intracellular multimer assembly and protein secretion. *J. Biol. Chem.* **274**, 21817–21822 (1999).
61. Pace, R. A. *et al.* Collagen VI glycine mutations: Perturbed assembly and a spectrum of clinical severity. *Ann. Neurol.* **64**, 294–303 (2008).
62. Posey, K. L., Yang, Y., Veerisetty, A. C., Sharan, S. K. & Hecht, J. T. Model systems for studying skeletal dysplasias caused by TSP-5/COMP mutations. *Cell. Mol. Life Sci.* **65**, 687–699 (2008).
63. Briggs, M. D. & Chapman, K. L. Pseudoachondroplasia and multiple epiphyseal dysplasia: mutation review, molecular interactions, and genotype to phenotype correlations. *Hum. Mutat.* **19**, 465–478 (2002).
64. Holden, P. *et al.* Cartilage oligomeric matrix protein interacts with type IX collagen, and disruptions to these interactions identify a pathogenetic mechanism in a bone dysplasia family. *J. Biol. Chem.* **276**, 6046–6055 (2001).
Describes the dominant negative effect of COMP mutations that cause intracellular retention of its interacting partners.
65. Merritt, T. M., Bick, R., Poindexter, B. J., Alcorn, J. L. & Hecht, J. T. Unique matrix structure in the rough endoplasmic reticulum cisternae of pseudoachondroplasia chondrocytes. *Am. J. Pathol.* **170**, 293–300 (2007).
66. Malhotra, J. D. & Kaufman, R. J. The endoplasmic reticulum and the unfolded protein response. *Semin. Cell Dev. Biol.* **18**, 716–731 (2007).
67. Bernales, S., Papa, F. R. & Walter, P. Intracellular signaling by the unfolded protein response. *Annu. Rev. Cell Dev. Biol.* **22**, 487–508 (2006).
68. Rutkowski, D. T. & Kaufman, R. J. That which does not kill me makes me stronger: adapting to chronic ER stress. *Trends Biochem. Sci.* **32**, 469–476 (2007).
69. Ron, D. & Walter, P. Signal integration in the endoplasmic reticulum unfolded protein response. *Nature Rev. Mol. Cell Biol.* **8**, 519–529 (2007).
Comprehensive review of the molecular signalling induced by protein misfolding in the ER.
70. Zhang, K. & Kaufman, R. J. The unfolded protein response: a stress signaling pathway critical for health and disease. *Neurology* **66**, S102–S109 (2006).
71. Zhao, L. & Ackerman, S. L. Endoplasmic reticulum stress in health and disease. *Curr. Opin. Cell Biol.* **18**, 444–452 (2006).
72. Marciniak, S. J. & Ron, D. Endoplasmic reticulum stress signaling in disease. *Physiol. Rev.* **86**, 1133–1149 (2006).
73. Lin, J. H., Walter, P. & Yen, T. S. Endoplasmic reticulum stress in disease pathogenesis. *Annu. Rev. Pathol.* **3**, 399–425 (2008).
Review of how ER stress and the UPR play an important part in the pathogenesis of many human diseases.
74. Vranka, J. *et al.* Selective intracellular retention of extracellular matrix proteins and chaperones associated with pseudoachondroplasia. *Matrix Biol.* **20**, 439–450 (2001).
75. Hecht, J. T. *et al.* Calreticulin, PDI, Grp94 and BiP chaperone proteins are associated with retained COMP in pseudoachondroplasia chondrocytes. *Matrix Biol.* **20**, 251–262 (2001).
76. Dinser, R. *et al.* Pseudoachondroplasia is caused through both intra- and extracellular pathogenic pathways. *J. Clin. Invest.* **110**, 505–513 (2002).
77. Hashimoto, Y., Tomiyama, T., Yamano, Y. & Mori, H. Mutation (D472Y) in the type 3 repeat domain of cartilage oligomeric matrix protein affects its early vesicle trafficking in endoplasmic reticulum and induces apoptosis. *Am. J. Pathol.* **163**, 101–110 (2003).
78. Pirog-Garcia, K. A. *et al.* Reduced cell proliferation and increased apoptosis are significant pathological mechanisms in a murine model of mild pseudoachondroplasia resulting from a mutation in the C-terminal domain of COMP. *Hum. Mol. Genet.* **16**, 2072–2088 (2007).
Provides direct evidence for triggering of the UPR in a mouse knockin model of a human COMP mutation.
79. Hecht, J. T. *et al.* Chondrocyte cell death and intracellular distribution of COMP and type IX collagen in the pseudoachondroplasia growth plate. *J. Orthop. Res.* **22**, 759–767 (2004).
80. Schmitz, M. *et al.* Transgenic mice expressing D469Δ mutated cartilage oligomeric matrix protein (COMP) show growth plate abnormalities and sternal malformations. *Matrix Biol.* **27**, 67–85 (2008).
81. Weirich, C. *et al.* Expression of PSACH-associated mutant COMP in tendon fibroblasts leads to increased apoptotic cell death irrespective of the secretory characteristics of mutant COMP. *Matrix Biol.* **26**, 314–323 (2007).
82. Leighton, M. P. *et al.* Decreased chondrocyte proliferation and dysregulated apoptosis in the cartilage growth plate are key features of a murine model of epiphyseal dysplasia caused by a *matn3* mutation. *Hum. Mol. Genet.* **16**, 1728–1741 (2007).
83. Lisse, T. S. *et al.* ER stress-mediated apoptosis in a new mouse model of osteogenesis imperfecta. *PLoS Genet.* **4**, e7 (2008).
Describes the UPR and the downstream consequences of apoptosis in an *in vivo* *N*-ethyl-*N*-nitrosourea (ENU) mutant mouse with a collagen I OI mutation.
84. Chessler, S. D. & Byers, P. H. Defective folding and stable association with protein disulfide isomerase/prolyl hydroxylase of type I procollagen with a deletion in the pro alpha 2(I) chain that preserves the Gly-X-Y repeat pattern. *J. Biol. Chem.* **267**, 7751–7757 (1992).
85. Forlino, A. *et al.* Differential expression of both extracellular and intracellular proteins is involved in the lethal or nonlethal phenotypic variation of BrtlIV, a murine model for osteogenesis imperfecta. *Proteomics* **7**, 1877–1891 (2007).
86. Hintze, V. *et al.* Cells expressing partially unfolded R789C/p.R989C type II procollagen mutant associated with spondyloepiphyseal dysplasia undergo apoptosis. *Hum. Mutat.* **29**, 841–851 (2008).
87. Kwan, K. M. *et al.* Abnormal compartmentalization of cartilage matrix components in mice lacking collagen X: implications for function. *J. Cell Biol.* **136**, 459–471 (1997).
88. Millington-Ward, S., McMahon, H. P. & Farrar, G. J. Emerging therapeutic approaches for osteogenesis imperfecta. *Trends Mol. Med.* **11**, 299–305 (2005).
89. Dazzi, F. & Horwood, N. J. Potential of mesenchymal stem cell therapy. *Curr. Opin. Oncol.* **19**, 650–655 (2007).
90. Rochet, J. C. Novel therapeutic strategies for the treatment of protein-misfolding diseases. *Expert Rev. Mol. Med.* **9**, 1–34 (2007).
91. Cohen, F. E. & Kelly, J. W. Therapeutic approaches to protein-misfolding diseases. *Nature* **426**, 905–909 (2003).
92. Ulloa-Aguirre, A., Janovick, J. A., Brothers, S. P. & Conn, P. M. Pharmacologic rescue of conformationally-defective proteins: implications for the treatment of human disease. *Traffic* **5**, 821–837 (2004).
93. Perlmutter, D. H. Chemical chaperones: a pharmacological strategy for disorders of protein folding and trafficking. *Pediatr. Res.* **52**, 832–836 (2002).
94. Arakawa, T., Ejima, D., Kita, Y. & Tsumoto, K. Small molecule pharmacological chaperones: from thermodynamic stabilization to pharmaceutical drugs. *Biochim. Biophys. Acta* **1764**, 1677–1687 (2006).
95. Reddy, R. K. *et al.* Endoplasmic reticulum chaperone protein GRP78 protects cells from apoptosis induced by topoisomerase inhibitors: role of ATP binding site in suppression of caspase-7 activation. *J. Biol. Chem.* **278**, 20915–20924 (2003).
96. Kudo, T. *et al.* A molecular chaperone inducer protects neurons from ER stress. *Cell Death Differ.* **15**, 364–375 (2008).
97. Williams, A. *et al.* Novel targets for Huntington's disease in an mTOR-independent autophagy pathway. *Nature Chem. Biol.* **4**, 295–305 (2008).
Describes a novel mTOR-independent pathway that regulates autophagy, and drugs that can act upon this pathway to stimulate autophagy as a putative therapeutic strategy in some protein misfolding diseases.
98. Sarkar, S. *et al.* Small molecules enhance autophagy and reduce toxicity in Huntington's disease models. *Nature Chem. Biol.* **3**, 331–338 (2007).
99. Rubinsztein, D. C., Gestwicki, J. E., Murphy, L. O. & Klionsky, D. J. Potential therapeutic applications of autophagy. *Nature Rev. Drug Discov.* **6**, 304–312 (2007).
100. Boyce, M. *et al.* A selective inhibitor of eIF2α dephosphorylation protects cells from ER stress. *Science* **307**, 935–939 (2005).
101. Robinson, P. N. *et al.* The molecular genetics of Marfan syndrome and related disorders. *J. Med. Genet.* **43**, 769–787 (2006).
102. Brooke, B. S. *et al.* Angiotensin II blockade and aortic-root dilation in Marfan's syndrome. *N. Engl. J. Med.* **358**, 2787–2795 (2008).
103. Habashi, J. P. *et al.* Losartan, an AT1 antagonist, prevents aortic aneurysm in a mouse model of Marfan syndrome. *Science* **312**, 117–121 (2006).
104. Irwin, W. A. *et al.* Mitochondrial dysfunction and apoptosis in myopathic mice with collagen VI deficiency. *Nature Genet.* **35**, 367–371 (2003).
105. Merlini, L. *et al.* Cyclosporin A corrects mitochondrial dysfunction and muscle apoptosis in patients with collagen VI myopathies. *Proc. Natl Acad. Sci. USA* **105**, 5225–5229 (2008).
106. Hicks, D. *et al.* Cyclosporine A treatment for Ullrich congenital muscular dystrophy: a cellular study of mitochondrial dysfunction and its rescue. *Brain* 16 Nov 2008 (doi:10.1093/brain/awn289).
107. Hansen, U. & Bruckner, P. Macromolecular specificity of collagen fibrillogenesis: fibrils of collagens I and XI contain a heterotypic alloyed core and a collagen I sheath. *J. Biol. Chem.* **278**, 37352–37359 (2003).
108. van Anken, E. & Braakman, I. Versatility of the endoplasmic reticulum protein folding factory. *Crit. Rev. Biochem. Mol. Biol.* **40**, 191–228 (2005).
109. Meusser, B., Hirsch, C., Jarosch, E. & Sommer, T. ERAD: the long road to destruction. *Nature Cell Biol.* **7**, 766–772 (2005).
110. Yorimitsu, T. & Klionsky, D. J. Endoplasmic reticulum stress: a new pathway to induce autophagy. *Autophagy* **3**, 160–162 (2007).

Acknowledgements

This work was supported by grants from National Health and Medical Research Council of Australia. J.F.B. and S.R.L. are National Health and Medical Research Council Research Fellows. We thank C. Little, P. Farlie, A. Fosang and R. Wilson for critical reading of the manuscript.

DATABASES

UniProtKB: http://www.uniprot.org
BiP | collagen X | COMP | GRP94 | matrilin 3

FURTHER INFORMATION

Bateman laboratory homepage: http://www.mcri.edu.au/SkeletalBiology
Online Medelian Inheritance in Man (OMIM): http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
The Human Gene Mutation Database at the Institute of Medical Genetics in Cardiff: http://www.hgmd.cf.ac.uk/ac/index.php
Database of osteogenesis imperfecta and type III collagen mutations: http://www.le.ac.uk/genetics/collagen
Leiden Muscular Dystrophy pages: http://www.dmd.nl

SUPPLEMENTARY INFORMATION

See online article: S1 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Mapping complex disease traits with global gene expression

William Cookson, Liming Liang‡, Gonçalo Abecasis‡, Miriam Moffatt* and Mark Lathrop§*

Abstract | Variation in gene expression is an important mechanism underlying susceptibility to complex disease. The simultaneous genome-wide assay of gene expression and genetic variation allows the mapping of the genetic factors that underpin individual differences in quantitative levels of expression (expression QTLs; eQTLs). The availability of systematically generated eQTL information could provide immediate insight into a biological basis for disease associations identified through genome-wide association (GWA) studies, and can help to identify networks of genes involved in disease pathogenesis. Although there are limitations to current eQTL maps, understanding of disease will be enhanced with novel technologies and international efforts that extend to a wide range of new samples and tissues.

Genome-wide association study
(GWA study). An examination of common genetic variation across the genome designed to identify associations with traits such as common diseases. Typically, several hundred thousand SNPs are interrogated using microarray or bead chip technologies.

**National Heart and Lung Institute, Imperial College London, London SW3 6LY, UK.*
‡Center for Statistical Genetics, Department of Biostatistics, SPH II, Ann Arbor, Michigan 48109-2029, USA.
§CEA / Centre National de Genotypage, 91057 Evry, France.
Correspondence to W.C., L.L., G.A., M.M. or M.L.
e-mails:
w.cookson@imperial.ac.uk;
lianglim@umich.edu;
goncalo@umich.edu;
m.moffatt@imperial.ac.uk;
mark@cng.fr
doi:10.1038/nrg2537

Genome-wide association (GWA) studies of common complex or multifactorial diseases have been spectacularly successful in the last 2 years, with many new loci identified with levels of probability that were once thought unattainable. However, the extraordinary levels of significance of the association signals have yet to be translated into a full understanding of the genes or genetic elements that are mediating disease susceptibility at particular loci.

The functional effects of DNA polymorphism on multifactorial disease can be mediated through several mechanisms. Polymorphisms that alter protein function can have very important effects, such as *NOD2* (nucleotide-binding oligomerization domain-containing 2; also known as *CARD15*) mutations in inflammatory bowel disease[1] and *FLG* (filaggrin) mutations in eczema (atopic dermatitis)[2]. However, systematic study of complex diseases with known non-synonymous SNPs has not yielded many highly significant results[3], and variation in gene expression is probably a more important mechanism underlying susceptibility to complex disease. The abundance of a gene transcript is directly modified by polymorphism in regulatory elements. Consequently, transcript abundance might be considered as a quantitative trait that can be mapped with considerable power. These have been named expression QTLs (eQTLs)[4,5].

There is a substantial gap between SNP associations from a GWA study and understanding how the locus contributes to disease. Further genotyping and statistical analyses are often necessary to identify causal variants, which are then functionally investigated. This Review explores the value of systematic identification of eQTLs as one means of characterizing the function of loci underlying complex disease traits. The combination of whole-genome genetic association studies and the measurement of global gene expression allows the systematic identification of eQTLs. By assaying gene expression and genetic variation simultaneously on a genome-wide basis in a large number of individuals, statistical genetic methods can be used to map the genetic factors that underpin individual differences in quantitative levels of expression of many thousands of transcripts.

The resulting comprehensive eQTL maps provide an important reference source for categorizing both *cis* and *trans* effects of disease-associated SNPs on gene expression. In addition to providing information about the biological control of gene expression, such data aid in interpreting the results of GWA studies. Once the statistical evidence for association of genetic markers to a disease trait has been established, genome-wide eQTL mapping data can be examined to see if the same genetic markers are also associated with quantitative transcript levels of one or more genes — such markers are known as eSNPs. The availability of systematically generated eQTL information provides immediate insight into a probable biological basis for the disease associations, and can help to identify networks of genes involved in disease pathogenesis.

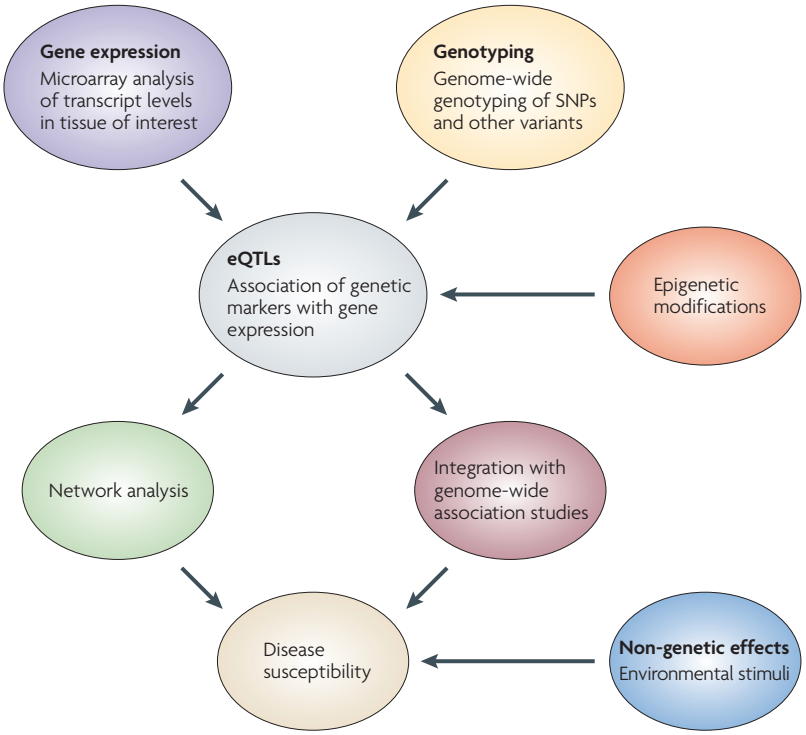


Figure 1 | **eQTL mapping.** Expression QTL (eQTL) mapping begins with the measurement of gene expression in a target cell or tissue from multiple individuals. This information is the substrate for investigating the effects of DNA polymorphism (of any type) on the expression of individual genes. Other factors that can alter transcription, such as epigenetic CpG methylation, may also be mapped. Network analyses build upon strong correlations that are present between transcripts, and allow the identification of modules of genes that mediate complex functions. This information can then be made available and used to interpret genetic associations and mapping information from the study of complex disease.

The potential of genome-wide eQTL identification was shown originally in the yeast *Saccharomyces cerevisiae*[6] and then in humans, animals and plants[4,7]. The history of eQTL mapping has been comprehensively reviewed[7–9], and will not be described in detail here. This Review will show how the combination of genetics and global gene expression can be a powerful tool for systematically unravelling the effects of variation in transcription on disease. First, we briefly introduce the principles and current methods of eQTL mapping and describe the basis of eQTLs. We then explore the relevance of these results to disease gene identification. The limits of current eQTL mapping data are discussed, as are the expected impact of new technologies, international efforts to extend results to new samples and tissues, and how cell lines might be tested with stimuli that are relevant to disease.

eQTL mapping

In practical terms, the starting point for eQTL mapping is the measurement of gene expression in a target cell or tissue from multiple individuals (FIG. 1). This information is the substrate for investigating the effects of DNA polymorphism (of any type) on the expression of individual genes. The use of microarray technology to measure gene expression from many thousand of genes simultaneously has been a principle driving force for systematic mapping of eQTLs[7]. The field is benefiting from progressively more sophisticated platforms for such studies, which are described in the later sections of this Review. Procedures for eQTL mapping are based on the insight that expression levels can be analysed with genetic approaches in the same manner as any other quantitative trait phenotype, such as body weight or blood lipids. In particular, study designs and statistical methods that are used traditionally to map QTLs can be successfully applied to the identification of eQTLs[10–12]. Interpretation of eQTL data can then be developed further by the incorporation of additional biological information, such as epigenetic modifications and analysis of regulatory networks, which are discussed below.

eQTLs are influenced not only by genetic polymorphisms, but also by a range of other biological factors. These can be dissected systematically, starting with the measurement of heritability (H^2).

Heritability. Family studies have shown that many human eQTLs are highly heritable[13,14]. The linkage approach, in which family members are studied, has been valuable in demonstrating that genetic factors have widespread and identifiable influences on eQTLs in humans, and such studies have provided broad localization for some of the underlying genetic factors[15,16]. GWA mapping of common genetic variants that underlie eQTLs has recently become possible owing to the wide availability of high-throughput and low-cost SNP genotyping. These results are particularly relevant to disease mapping that is also focused on common SNPs characterized with similar SNP arrays. Moreover, the interpretation of these eQTL data relies strongly on methodologies that have been developed for disease GWA[13]. For example, a family study of lymphoblastoid cell lines (LCLs) identified nearly 15,000 traits (each corresponding to an individual Affymetrix probe) with an estimated $H^2 > 0.3$, indicating that genetic influences on gene expression seem to be widespread[13]. Other studies have similarly described a high H^2 of many eQTLs in LCLs and other tissues[4,17,18].

Genetic factors (with both *cis*-acting and *trans*-acting effects; see below), are often identified for eQTLs that have high H^2. For example, in the LCL study mentioned above[13], eQTLs for 81% of traits with $H^2 > 0.8$ could be mapped to one or more SNPs at genome-wide significance. However, the SNP map on average accounted for less than 20% of the estimated trait H^2, consistent with results obtained by other studies[16]. This indicates the presence of genetic or other factors affecting familial clustering on transcription that are not detectable in these genetic associations. Factors other than SNPs that might affect H^2 are discussed in more detail below. Further understanding of disease phenotypes can also be gained from analysing whether particular types of genes have more heritable variation in expression level[13,19] (BOX 1).

Cis *and* trans *effects.* Statistical analyses of eQTLs need to take into account that the loci identified can influence gene expression either in *cis* or in *trans*. The definition of

Epigenetic
A mitotically stable change in gene expression that depends not on a change in DNA sequence, but on covalent modifications of DNA or chromatin proteins such as histones.

Heritability
(H^2). The heritability of an individual trait is estimated by the ratio of genetic variance to total trait variance, so that 0 indicates no genetic effects on trait variance and 1 indicates that all variance is under genetic control.

Major histocompatibility complex
(MHC). A complex locus on chromosome 6p that comprises numerous genes, including the human leukocyte antigen genes, which are involved in the immune response.

Gene Ontology
(GO). A widely used classification system of gene functions and other gene attributes that uses a standardized vocabulary. The system uses a hierarchical organization of concepts (an ontology) with three organizing principles: molecular functions (the tasks done by individual gene products), biological processes (for example, mitosis) and cellular components (examples include the nucleus and the telomere).

a *cis* effect is somewhat arbitrary, but *cis*-acting eQTLs are typically considered to include SNPs within 100 kb upstream and downstream of the gene that is affected by that eQTL. This definition becomes more problematic in regions of extended linkage disequilibrium, such as the major histocompatibility complex (MHC) locus.

Detailed analysis of the position of mapped *cis*-acting eQTL effects have shown that these are enriched around transcription start sites and within 250 bp upstream of transcription end sites, and they rarely reside more than 20 kb away from the gene[20]. *Cis*-acting variants also seem to occur more often in exonic SNPs[20]. *Trans* effects are usually weaker than *cis* effects in humans[4,5] and in rats[21], but they are more numerous.

It is not known if *trans* effects are mostly mediated through transcription factor variants or through other mechanisms. 'Master regulators' are *trans*-acting factors with multiple effects on gene expression that have been identified in *S. cerevisiae*[22], in rat tissues[21] and in the human genome[5]. It is of interest that, at least in yeast, master regulators are not enriched for transcription factors, and *trans*-regulatory variation seems to be broadly dispersed across classes of genes with different molecular functions[22].

Other types of variant. The function of DNA can be altered by many mechanisms in addition to SNPs. Transcription can also be modified by copy number variants (CNVs), insertions and deletions, short tandem repeats and single amino acid repeats[23]. A systematic investigation of the effects of CNVs in individuals who are part of the International HapMap project showed that SNPs and CNVs captured 84% and 18%, respectively, of the total detected genetic variation in gene expression but the signals from the two types of variation had little overlap[24]. It has been shown that CNVs in regulatory hot spots in the malaria parasite genome dictate transcriptional variation[25]. It has also been observed that small-scale copy number variation (that is, a single or few copies) can lead to multiple orders of magnitude change in gene expression and, in some cases, switches in deterministic control[26].

Box 1 | **Gene ontology analyses**

Many expression QTLs (eQTLs) are highly heritable. Therefore, Gene Ontology (GO) analyses can be applied to eQTL databases to identify the types of gene that show the most inherited variation in their levels of expression (at least in the cell type studied, usually lymphoblastoid cell lines; LCLs). The most highly heritable GO biological process for eQTLs in LCLs in one study was, unexpectedly, 'response to unfolded proteins', a group containing numerous chaperonins and heat shock proteins. The individual variation in response to unfolded proteins may be an evolutionary response to cellular stress, and these genes could be candidates in the study of neurodegenerative diseases and the ageing processes. Genes that regulate RNA processing, DNA repair and progression through the cell cycle were also exceptionally heritable. The evolutionary advantage of individual variation in these genes is unclear.

As expected, genes with significant heritability are also enriched in GO categories of immune response[13,19]. These highly heritable immune genes may be of particular value for the study of infectious and inflammatory diseases. The most heritable traits can be considered as candidate genes for effects on particular disease traits, but they could also be studied in large population samples, such as those contained in national biobanks, to investigate their actions on unexpected phenotypes.

Epigenetic factors. In addition to DNA sequence variants, gene transcription is also modulated by epigenetic modifications (discussed further in the 'Limitations of mapping studies' section below). For example, non-germ line epigenetic methylation of CpG residues that regulate gene expression is common in the human genome[27]. In a limited study of three chromosomes, 17% of genes can be differentially methylated in their 5′ UTRs and approximately one-third of the differentially methylated 5′ UTRs are inversely correlated with transcription[27]. A further level of complexity comes from post-translational modifications of histones that modulate DNA accessibility and chromatin stability to provide an enormous variety of alternative interaction surfaces for *trans*-acting factors (reviewed in REF. 28).

eQTLs and disease gene mapping

Combining eQTL and GWA studies. One of the most important consequences of eQTL mapping is the link that it provides between genetic markers of disease identified in GWA studies and the expression of a specific gene or genes. In particular, the power of these studies depends upon the identification of specific genetic markers that are simultaneously associated with disease and eQTLs, whereas simply comparing differences in gene expression in cases and controls might not provide sufficient power to detect important differences with the available sample sizes. The value of this is illustrated by several recent investigations in which eQTL analysis was incorporated directly as a component of the GWA study design (included in TABLE 1). The number of GWA studies continues to rise rapidly. In GWA studies to date, 10–15% of the top hits have affected a known eQTL in a public data set (TABLE 1). We will therefore discuss selected instances of these to show the value of the method.

For example, a recent study generated genome-wide transcriptional profiles of lymphocyte samples from participants in the San Antonio Family Heart Study, and showed that high-density lipoprotein cholesterol concentration was influenced by the *cis*-regulated vanin 1 (*VNN1*) gene[15]. Similarly, a study of post-mortem brain tissue identified eQTLs affecting the *MAPT* (microtubule-associated protein tau) and *APOE* (apolipoprotein E) genes, which play an important part in Alzheimer's disease[29].

At the same time as the San Antonio study the results of a GWA study of asthma[13,30] identified a series of SNPs in strong linkage disequilibrium and spanning more than 200 kb of chromosome 17q23. The study showed that these SNPs were strongly associated with the risk of asthma[30]. The region of association contains 19 genes, none of which is an obvious candidate for disease. Examination of eQTL data derived from Affymetrix HU133A arrays[13,30] on the same families showed that the disease-associated SNPs had highly significant ($p < 10^{-22}$) effects in *cis* on the expression of one the genes: *ORMDL3* (ORM1-like 3).

This locus illustrates the utility of combining eQTL and disease mapping studies. Despite the highly significant association with both expression and disease, the

Table 1 | **Disease-linked associations with significant expression QTLs from the literature and public databases**

Study	Trait	Region	Candidate gene(s)	Transcript affected by SNP	Transcript region	Logarithm of odds (LOD) score
Gudbjartsson *et al.*[102]*	Height	7p22	*GNA12*	*GNA12*	7p22	13
		11q13.2	Intergenic	*CCND1*	11q13	7.4
		7q21.3	*LMTK2*	*C17orf37*	17q21	6.0
				HSD17B8	6	6.4
				NDUFS8	11	6.1
		3p14.3	*PXK*	*RPP14*	3	9.2
Göring *et al.*[15]	High-density lipoprotein cholesterol levels	6q21	*VNN1*	*HDL* (serum)	Multiple sites	8.0
Kathiresan *et al.*[40]	Polygenic dyslipidaemia	20q13	*PLTP*	*PLTP*	20q13	16
		15q22	*LIPC*	*LIPC*	15q22	17
		11q12	*FADS1*, *FADS2*, *FADS3*	*FADS1*	11q12	35
				FADS3	11q12	8.0
		9p22	*TTC39B*	*TTC39B*	9p22	7.0
		1p13	*CELSR2*, *PSRC1*, *SORT1*	*SORT1*	1p13	270
				PSRC1	1p13	249
				CELSR2	1p13	80
		12q24	*MMAB*, *MVK*	*MMAB*	12q24	43
		1p31	*ANGPLT3*	*DOCK7*	1p31	27
				ANGPLT3	1p31	11
Libioulle *et al.*[37]	Crohn's disease	5p13	Intergenic	*PTGER4*	5p13	3.0
Barrett *et al.*[36]	Crohn's disease	5q31	*OCTN1*, *SLC22A4*, *SLC22A5*	*SLC22A5*	5q31	Unknown
Hom *et al.*[103]*	Systemic lupus erythematosus	8p23.1	*C8orf13*, *BLK*	*BLK*	8p23.1	20
				C8orf13	8p23.1	28
Harkonason *et al.*[104]*	Type 1 diabetes	12q13	*RAB5B*, *SUOX*, *IKZF4*	*RPS26*	12q13	33
		1p31.3	*ANGPTL3*	*DOCK7*	1p31.3	16
Wellcome Trust Case Control Consortium[105]*	Type 1 diabetes	12q13.2	*ERBB3*	*RPS26*	12q13.2	43.2
Todd *et al.*[106]*	Type 1 diabetes	12q13.2	*ERBB3*	*RPS26*	12q13.2	30.3
Plenge *et al.*[107]*	Rheumatoid arthritis	9q34	*TRAF1-C5*	*LOC253039*	9q34	6.3
Thein *et al.*[108]	Fetal haemoglobin F production	6q23.3	Intergenic	*HBS1L*	6q23.3	6.0
Moffatt *et al.*[30]	Childhood asthma	17q21	Intergenic	*ORMDL3*	17	14
Wellcome Trust Case Control Consortium[105]*	Bipolar disorder	16p12	*PALB2*, *NDUFAB1*, *DCTN5*	*DCTN5*	16p12	9.2
		6p21	*NR*	*HLA-DQB1*	6p21	8.9
				HLA-DRB4	6p21	11
Di Bernardo *et al.*[109]*	Chronic lymphatic leukaemia	2q37	*SP140*	*SP140*	2q37	8.8

*Identified through comparison of the National Human Genome Research Institute's Catalog of Published Genome-Wide Association Studies and the mRNA by SNP Browser v 1.0.1.

predicted expression differences in cases and controls, which was averaged over all genotypes, was not expected to be significant given the sample size: this was in agreement with the observed results[30].

In these data, borderline significant effects were also observed in the expression of the gene neighbouring *ORMDL3*, *GSDML* (gasdermin-like)[30]. Subsequent eQTL studies with the Illumina platform and RT-PCR experiments confirmed that the same SNPs determine eQTLs with both genes. These results focus attention on one or both of these genes as probable candidates for a role in disease pathology. Many additional studies are now underway to investigate the biological functions of these two genes and their relationship to asthma[31,32–35].

Using eQTLs to interpret GWA studies. Such findings have encouraged the use of these eQTL data as a general tool for interpreting results from GWA studies.

Recent analyses of Crohn's disease (CD) illustrate this approach[36,37]. Initially, markers on chromosome 5 were shown to be strongly associated with CD in one GWA scan, but their biological effects could not be readily deduced as they reside in a 1.25 Mb gene desert. Examination of the LCL eQTLs database showed that one or more of these polymorphisms act as a long-range *cis*-acting factor influencing expression of *PTGER4* (prostaglandin E receptor 4), a gene that resides approximately 270 kb proximal to the association region[37]. The homologue of this gene has been implicated in phenotypes similar to CD in the mouse[37,38]. Thus, research is now focused on *PTGER4* as a primary candidate gene for this disease susceptibility locus.

Subsequently, the eQTL approach has been applied systematically in a meta-analysis of GWA studies of CD, and several other interesting results have been obtained[36]. For example, eQTLs were used to address an outstanding question in CD genetics related to the identification of the CD susceptibility gene or genes in the cytokine cluster on chromosome 5q31, where SNPs have an established association with disease[39]. The disease-associated SNPs in the meta-analysis of this region were all shown to be correlated with decreased *SLC22A5* (solute carrier family 22, member 5) mRNA expression levels.

Another CD locus identified in the meta-analysis coincided with the asthma risk locus on chromosome 17, in which the disease markers are also correlated with expression of *ORMDL3* and *GSDML*, as described above. Thus the same genetic variants contribute to susceptibility to both CD and asthma, possibly by perturbing expression of one or both of these genes. Several additional examples of eQTLs within CD susceptibility loci have also been reported[36]. These co-localizations greatly exceed the number that would be expected by chance, suggesting that many of them are indicative of underlying biological processes involved in disease susceptibility[36].

Public GWA study results are available at the National Human Genome Research Institute's Catalog of Published Genome-Wide Association Studies. Examination of these results identifies many other disease associations for which eQTL data provide similar insights (TABLE 1). For example, a recent large study of polygenic dyslipidaemia identified 30 loci with highly significant effects on blood lipid measurements[40]. Examination of gene expression in samples of liver from 957 subjects allowed highly significant eQTLs to be identified for 7 of the 30 loci[40] (TABLE 1). In some cases, the eQTL data give genetic evidence to support a candidate gene for which a role was previously suggested from location and biological hypotheses (such as *GNA12* for height on chromosome 7p22, and *BLK* and *C8orf13* for auto-immune systemic lupus erythematosis on 8p23.1). More often the gene expression data identifies different genes or suggests a particular gene from a number of candidates. Examples of this include the cluster of *trans*-acting genes from the height locus on chromosome 7q21.3, the *RPS26* gene from the type 1 diabetes locus on 12q13.2, and the *DCTN5* gene from the bipolar disorder locus on chromosome 16p12.1.

Not all examples of eQTL findings are straightforward, as exemplified by the association reported between the *SH2B1* (SH2B adaptor protein 1) locus and body mass index (BMI)[41]. In this study, a missense SNP in *SH2B1* was also associated with significant variation in transcript abundances of *EIF3C* (eukaryotic translation initiation factor 3, subunit C) and *TUFM* (Tu translation elongation factor, mitochondrial). When mutated, the homologue of *SH2B1* leads to extreme obesity in mice, apparently because of a failure in proper regulation of appetite. The authors speculate that the *SH2B1* variant has a causal role but is in linkage disequilibrium with a different variant that influences *EIF3C* and *TUFM* mRNA levels; alternatively, regulation of *EIF3C* or *TUFM* mRNA levels could have a causal role instead of, or in addition to, variation in *SH2B1* (REF. 41).

eQTL databases. mRNA by SNP Browser v 1.0.1 is a database of eQTLs from asthma studies[13,30] that allows searches by genes, chromosomal regions and SNPs, and is a good example of how data from this kind of research can be examined. VarySysDB is another public database and contains 190,000 extensively annotated mRNA transcripts from 36,000 loci. VarySysDB offers information encompassing published human genetic polymorphisms for each of these transcripts separately. In addition to SNP effects on transcription, VarySysDB includes deletion–insertion polymorphisms from dbSNP, CNVs from the Database of Genomic Variants, short tandem repeats and single amino acid repeats from H-InvDB and linkage disequilibrium regions from D-HaploDB[23].

MHC locus. Analysis of eQTLs in the MHC locus is of particular interest for studies of diseases in which infection and autoimmunity is a major component[42]. Intense study of the MHC locus over many years has revealed many genes that are duplicated or polymorphic, and DNA variants in the MHC locus have been associated with more diseases than any other region of the human genome[42]. Many disease associations have been attributed to selective binding of processed antigen to the antigen-presenting grooves of human leukocyte antigen (HLA) variants.

The results of eQTL studies on the MHC locus must be interpreted with caution. This is because the high degree of genetic variability and linkage disequilibrium across the MHC locus could introduce some spurious results owing to polymorphism in sequences corresponding to probes used for expression measurements[43] (see below). Nevertheless, global gene expression data has shown very strong effects of particular SNPs on the level of expression of the classical MHC antigens *HLA-A*, *HLA-C*, *HLA-DP*, *HLA-DQ* and *HLA-DR* ($p < 10^{-20}$ to $p = 10^{-30}$)[13]. This confirmed the effect of genetic variation on the level of *HLA-DQ* expression observed previously[44]. The strength of these effects suggests that associations of MHC class I and class II polymorphism might depend on the level of gene transcription as much as restriction of response to antigen[13]. A possible example is type I diabetes, in which the functional effects of the long-recognized association to the class II MHC genes[45] have not been elucidated, despite combined p values of less than 10^{-100}.

Human leukocyte antigen (HLA). A glycoprotein, encoded at the major histocompatibility complex locus, that is found on the surface of antigen-presenting cells and that present antigen for recognition by helper T cells.

Serial analysis of gene expression
(SAGE). A method for quantitative and simultaneous analysis of a large number of transcripts. Short sequence tags are isolated, concentrated and cloned; their sequencing reveals a gene expression pattern that is characteristic of the tissue or cell type from which the tags were isolated.

These results suggest that even in this intensively studied region, the investigation of eQTLs could add to our understanding of the many known genetic associations.

Additional biological interpretation and validation. A genome exerts its functions not through particular genes or proteins, but through highly complex networks that produce a range of responses[46]. As perturbations of such networks underlie the pathogenesis of many diseases[47,48], network analysis incorporating eQTL data has recently provided important novel insights into mechanisms underlying multifactorial diseases[16,17,49] (BOX 2).

Extensive investigations of human populations, animal models and cellular systems are required to provide biological validation of the relationship between specific genes and multifactorial disease traits, even when the relationship is identified through eQTL analysis. Given the substantial effort that is required for validation, careful selection of only the strongest candidates is essential. As shown in the above examples, the combination of GWA studies and eQTL analysis is a powerful way of identifying a small number of candidate genes and pathways. With the deployment of new technologies, such as exon arrays and RNA resequencing, and expansion of the tissue types covered, as described below, we expect future eQTL databases to be even more powerful tools for such identifications.

Box 2 | **Networks and other analytical tools**

Traditional genetics and cellular biology has rested on the assumption that a single stimulus (or DNA variant) when applied to a cell (or gene) will have a single outcome. The reality is that even a simple stimulus will induce changes in transcription in many genes that interact in complex networks, with an outcome that affects many different transcripts and processes.

The networks can be considered to be made up of multiple pathways that act at genetic, genomic, cellular, tissue and whole-organism levels[46]. The technology that is already available to gather global information on gene expression, proteins and metabolites is now allowing the systematic identification of the networks of genes that interact in disease processes[92,93]. Analysis of genetic variants that perturb networks through the effects of expression QTLs (eQTLs) has recently provided important novel insights into mechanisms underlying multifactorial diseases[16,17,49]. This type of analysis may also lead to the systematic identification of transcription modules[94] and the construction of regulatory networks[95]. The potential of genetic mapping approaches to identify networks of genes operating on hematopoietic stem cells[96] and immune responses[97] are amongst the examples that have been discussed in the literature.

The impact of combining eQTL analysis with an investigation of gene networks is shown by the recent detection of genetic variants associated with transcript abundance of a macrophage-enriched network and obesity-related traits in human subjects. Parallel studies in mice and humans identified a network module for obesity-related traits that was enriched for genes involved in the inflammatory and immune response. eQTL mapping was then used to identify *cis*-acting genetic variants associated with this network of genes. The authors characterized these genetic variants in a large cohort of individuals, and showed statistical enrichment for variants that were associated with obesity-related biometric traits[16]. This approach allowed the identification of genetic variants that had minor individual effects on the trait, but that can be identified as a group because of the overall perturbation of the network. Three genes in this network, lipoprotein lipase (*Lpl*), lactamase β (*Lactb*) and protein phosphatase 1-like (*Ppm1l*), were validated by gene knockouts, strengthening the association between this network and metabolic disease traits[49].

A bibliography and a range of statistical routines for network analysis can be found on the Weighted Gene Co-expression Network site.

Potential limitations and future directions

Despite the power of eQTL mapping to help identify the genetic basis of disease, there are many limitations to current methodologies and potential for considerable improvements as technologies develop. The best appreciated technical barriers to optimal eQTL mapping are in the use of microarrays to measure gene expression (BOX 3). Other problems and their potential solutions are discussed below.

Comparisons between microarray platforms. It was assumed that different microarray platforms give broadly comparable results[50]. However, numerous studies are now showing that the overlap in transcript detection between platforms is only ~30–40%, whether considered as presence or absence of detectable transcripts or the absolute level of transcript abundance[51–53]. The same level of discordance appears whether comparisons are made between Affymetrix arrays and serial analysis of gene expression (SAGE)[52], Affymetrix and Illumina arrays[50], Affymetrix and Applied Biosystems arrays[53], or across multiple platforms[51].

Some of this discrepancy may be because individual genes are commonly interrogated by different sequences on different platforms. The situation can be improved when matching of genes is sought using genomic sequence rather than sequences inferred from the UniGene database of transcripts[54]. Concordance between platforms is improved further when probes are compared only when they target overlapping transcript sequence regions on cDNA microarrays or gene chips[55].

These discrepancies may follow from the complex and unpredictable factors that determine hybridization of particular nucleic acids to complementary array-bound sequences[56,57]. In addition, the selection of sequences on microarrays has been strongly biased to the 3′ end of genes, simply because public cDNA databases were first populated with genes identified by 3′ tags.

A consistent conclusion of comparison studies has been that different platforms provide complementary results[51,52], probably because they are all sampling only a selected fraction of the total transcriptome from the cells or tissue under study. The use of multiple platforms to extract all the expression information from a cell or tissue is impractical.

New platforms for measuring gene expression. A more comprehensive measurement of gene expression comes from arrays that interrogate all known human exons. Affymetrix have produced global exon arrays[58], which show a high degree of correspondence in terms of fold changes with their pre-existing 'classical microarrays', suggesting that the additional probe sets on the exon arrays will provide reliable as well as more detailed coverage of the transcriptome[59]. The use of exon arrays allows the identification of tissue-specific alternative splicing events as well as significant expression outside of known exons and well-annotated genes[60]. Exon arrays on other platforms are likely to provide similarly robust results.

Box 3 | **Pitfalls with microarrays**

The use of microarrays to measure gene expression has led directly to the development of expression QTL (eQTL) analyses. However, the microarray approaches that underlie most eQTL studies to date provide only partial gene coverage and have a limited dynamic range for quantitative detection of expression. Specific problems inherent in the use of these microarrays include: the systematic bias that can be introduced during sample preparation, hybridization and measurement of expression; batch to batch variation in array manufacture; and day to day variation in laboratory conditions[98]. These types of effects are probably under-recognized, as exemplified by a report of large-scale differences in gene expression between ethnic groups[98,99]. In this case the highly significant differences in gene expression that the data had suggested between the groups[98] were found to be due to the separate processing of expression measurements in lymphoblastoid cell lines (LCLs) from subjects of European and Asian ancestry.

Cis-eQTL artefacts can also arise from the overlap of SNPs with transcript probes[100]. Alterations in hybridization efficiency owing to the SNP can give an erroneous impression of differences in transcript abundance attributable to the SNP (and to other DNA variants with which it is in linkage disequilibrium)[100]. It has been estimated that 15% of microarray probes for any given gene will overlap with SNPs that are polymorphic in the population under study[100]. However, most coding SNPs in the human genome are uncommon, and it also seems that measurements of abundances are robust against mismatches between the probe and RNA sequences[101]. Although evidence that these artefacts have an effect has been presented[43], it is reassuring that in a large study in humans Emilsson *et al.*[16] found no evidence of systematic or specific hybridization artefacts from SNPs in their eQTL data. Nevertheless, important findings from microarrays need confirmation by specific assays, such as quantitative PCR, that avoid polymorphic sequences. Statistical methodology to account for batch effects, polymorphism and other sources of artefact is discussed by Alberts *et al.*[102]

Most human studies of eQTLs have been performed in LCLs, primarily because LCLs were often created as a source of nucleic acids for genetic studies. However, LCLs can exhibit progressive genomic instability with multiple passages of storage and re-growth, with the resulting potential for artefacts.

Many of the problems that are inherent in the use of microarrays can be solved by massively parallel, ultra-high-throughput DNA sequencing systems (reviewed in REF. 61). These systems allow direct ultra-high-throughput sequencing of RNA, which can then be mapped back to the genome. Sequencing of RNA provides a generic tool that can support a family of assays for measuring the genome-wide profiles of mRNAs, small RNAs, transcription factor binding, chromatin structure, DNase hypersensitivity and DNA methylation status[61]. RNA splices may also be effectively mapped by sequence-based methods.

Despite the formidable promise, ultra-high-throughput sequencing is still not without problems. The machines can produce terabytes of data daily, and make profound demands on bioinformatics for data storage and assembly of reads. Short reads may pose severe problems for the interpretation of transcripts arising from gene families with high homology or repetitive regions of the genome. Nevertheless, it can be anticipated that within 2 years many studies will rely on this technology, and that alternative or complementary approaches, such as large-scale real-time PCR-based expression assays (for example, as described by Watson *et al.*[62] and developed by WaferGen), will continue to evolve.

Limitations of mapping studies. As discussed in the section on heritability, currently mapped loci account for only a portion of the estimated heritability of eQTLs. A similar degree of unattributed or 'dark' heritability has been observed in GWA studies of common complex traits and diseases. A large GWA meta-analysis, for example, recently identified 20 variants that are significantly associated with adult height. The combined effects of the 20 SNPs explained only 3% of height variation, taking into account such factors as age and population[63]. Similarly, a large GWA meta-analysis of Crohn's disease identified 32 loci that significantly affect the disease, which together explained only 10% of the overall variance in disease risk and 20% of the genetic risk[36].

A large portion of the unattributed heritability is expected to result from the effects of multiple loci that are too weak to detect using current sample sizes[18]. This explanation would be consistent with data in yeast, in which only 3% of highly heritable transcript abundances are explained by single-locus (monogenic) inheritance and 50% are consistent with more than five controlling loci of equal effect[64]. Although current SNP arrays provide relatively comprehensive coverage of the genome (more than 80%), some of the unattributed heritability will be due to genetic factors that reside in unmapped regions, or to variation that is not effectively tagged at present, such as CNVs. Dominance and interaction effects may also account for some of the unattributed heritability, as these may be confounded with additive genetic effects in the heritability estimates with some study designs.

A previously described global eQTL study was based on sibling pairs, allowing estimates of heritability for all the transcripts measured[13]. The study suggested that dominance had a minimal effect on gene transcription[13]. Interestingly it seemed that genetic interactions may have important influences on regulation of expression for some genes, but inclusion of interaction effects had a minimal impact on the overall attributable heritability[13].

Epigenetic modifications and other factors that affect transcript abundance might not be accounted for in SNP-based association studies (see 'The basis of eQTLs' section above). Genomic imprinting is a particular case of an epigenetic effect with a parent of origin-dependent pattern. Monoalleleic expression is established at imprinted loci, via epigenetic marks transmitted through the germ line. Several common complex diseases exhibit parent-of-origin effects that might indicate underlying imprinting, including asthma[65], type I diabetes[66,67], rheumatoid arthritis[68], psoriasis[69], inflammatory bowel disease[70] and selective immunoglobin A deficiency[71], but systematic analysis of parent-of-origin effects in eQTL data has not yet been reported.

Finally, transcript abundance is a function of transcript stability as well as transcript production. Many factors mediate transcript stability, particularly in *trans*, either through protein–RNA interaction or through mechanisms mediated by small interfering RNAs (siRNAs)[72]. It seems clear that future studies of disease susceptibility as well as eQTLs will need to take these mechanisms into account.

Additive genetic effects
A mechanism of quantitative inheritance such that the combined effects of genetic alleles at two or more gene loci are equal to the sum of their individual effects.

Gene expression in tissues. Although RNA for eQTL analyses would ideally be obtained from a wide variety of tissues, most human studies of eQTLs have been performed in LCLs, primarily because LCLs were often created as a renewable source of nucleic acids for genetic studies. Gene expression in LCLs, however, represents the particular circumstances of Epstein–Barr virus infection of B-cells and their subsequent uncontrolled growth. LCLs may also exhibit extreme clonality with random patterns of monoallelic expression in single clones[73].

Although only 60% of genes from any particular cell type will also be found in LCLs[4,13], it has been established that LCLs provide information about gene expression for some genes that do not function primarily in these cells[4,74–76]. In addition, a recent comparison of eQTLs derived from the analysis of blood and adipose tissue showed little difference in the number of eQTLs that could be mapped, and there was an approximately 50% overlap of mapped loci from the two RNA sources[16]. Similarly, comparison between four different tissues showed no statistically significant differences in the number of mapped transcripts in experiments involving mapped recombinant inbred strains of mice[18].

Despite the continued utility and convenience of LCL studies of gene expression, it is evident that many of the transcripts expressed in LCLs may be housekeeping genes, and transcripts that determine specialized cell functions and that modify disease may be more parsimoniously distributed. In addition, LCLs are removed from the stimuli that can induce disordered gene transcription in disease. This is exemplified by the differences that are observed in gene expression between LCLs derived from asthmatics and genes known to be expressed in asthmatic airways[30]. These factors all indicate that the direct examination of tissues that are involved in disease can provide much more information than the LCL alone.

Some eQTL studies of human tissue have already been carried out, notably of liver[17], adipose[16,49] and brain[29] tissue. These show that approximately 60% of the transcriptome is expressed in each tissue, and that eQTLs from these tissues may be a valuable source of information for genetic mapping. Data from animal models suggest that tissue samples can allow detection of *trans*-eQTLs that are important in determining the composition of individual tissues[18]. Tissue samples will also allow the use of network analyses to identify the complex interactions that may underlie disease[16,17,49] (BOX 2).

The costs of reagents and the limited availability of appropriate tissues have to date restricted studies in humans to several hundreds of subjects. Although a formal evaluation of optimal study sizes is difficult because of unknown trait heritability, we know empirically that studies with a few hundred subjects have consistently identified numerous eQTLs with vanishingly small *p* values[4,5,75]. It is also clear that subtle effects, particularly in *trans*, would be detected more reliably with larger samples.

It is therefore timely that the promise of eQTLs as a tool for disease genetics has been sufficiently exciting to prompt a National Institutes of Health (NIH) proposal for the ambitious Genotype-Tissue Expression (GTEx) project, a database that might include 1,000 samples from each of 30 different tissues. The GTEx project is currently running as a 2-year pilot study with the primary goal of testing the feasibility of collecting high-quality RNA and DNA from multiple tissues from approximately 160 donors identified through low post-mortem interval autopsy or organ transplant settings. If the pilot phase proves successful, the project will be scaled up to involve approximately 1,000 donors, with the eventual creation of a database to house existing and GTEx-generated eQTL data.

The use of tissues poses a number of problems that need to be resolved. Normal and diseased tissue samples can be difficult to access, and their use requires careful attention to ethical, legal and social issues. Samples taken at post-mortem from many tissues robustly retain their histological architecture and contain RNA that can be of sufficient quality for measurements of gene expression. However, the changes in gene expression that might accompany death or surgical resection have not yet been documented in any detail. Tissues typically consist of different cell types, and their composition can vary inconsistently in the presence of disease. Finally, tissue-specific DNA methylation profiles may affect 20% of genes[27], and are expected to be important in understanding tissue eQTLs.

Although some of these problems may be expected to degrade the information available from the study of any particular tissue, it should be appreciated that they will not systematically lead to false positives in eQTL analyses[17], emphasizing the robustness of the eQTL approach.

Exercising the genome. Tissue biopsies and other samples extend the 'expression space' that can be examined by eQTL studies. They nevertheless still have limitations for functional analyses (particularly in humans as opposed to model organisms) when compared with cells that can be grown freely in culture and manipulated by systematic knock downs.

Although the transcripts in a particular cell under particular conditions reflect only part of the function of a particular genome, the range of transcripts from a given cell type can be widened by stimulating the cell in a variety of ways. The experimental extension of the genome expression space has been called 'exercising the genome'[77], and this strategy can be used to learn much more about gene expression and integrated gene functions. Experimentally, evidence is already emerging that environmental actions on gene expression are profound in humans[78] and model organisms[79,80] (reviewed in REF. 81), and it is reasonable to assume that these components of gene expression can be fruitfully accessed through exposure to relevant stimuli. It is interesting that, in model organisms, environmentally induced changes in gene expression seem to act through prominent *trans* effects[79,80] that may not be present in unstressed cells and tissues.

It is therefore desirable that the genome of human LCLs or primary cells of particular interest be exercised by stimulating their gene expression in different ways. Model stimuli that could be tested in these systems include pro-inflammatory stresses, metabolic stresses (such as high or low glucose, or hypoxia), the response to radiation, the response to signalling molecules (such

Box 4 | **eQTLs and network analyses of cancer**

Mutations that disrupt cell growth control mechanisms are a feature of cancer. In addition, the unchecked cell division that is characteristic of cancer can in time result in many secondary mutations and progressive genomic disorganization[103]. Genetic studies of cancer tissue (so called somatic cell genetics) have been used to identify the most common mutations in various tumours. Global gene expression studies have also been used in many cancer types, typically to identify gene signatures that can predict the clinical outcome[104]. However, most signature-based outcome predictions have not been replicated by independent studies[104], perhaps owing to the innate heterogeneity of cancerous tissue and the problems of deriving statistically stringent results from the measurement of thousands of transcripts in limited numbers of samples. Expression QTL (eQTL) analyses are a powerful tool to identify the functional consequences of the numerous copy number variants (CNVs), deletions and epigenetic modifications that are a feature of neoplastic cells. eQTL mapping allows the identification not only of genes underlying malignant processes[105] but also of genes modifying disease progression[106] and genes modulating individual responses to chemotherapy[107]. Network analyses have not yet been widely applied to the study of cancer, but they have already led to interesting findings, such as the identification of the *ASPM* gene as a molecular target in patients with glioblastoma. The application of network analyses to cancer eQTLs may be expected to greatly alleviate problems with multiple comparisons and to lead to easier biological interpretation of results[108,109]. Direct comparison of the transcript network architecture of cancerous tissue against normal tissues may also allow much deeper understanding of cancer biology.

as neurotransmitters, hormones and peptides) and the response to therapeutic and chemotherapeutic agents.

Conclusions

It is now well established that transcript abundances of genes may be considered as quantitative traits that can be mapped with considerable power, and that assaying gene expression and genetic variation simultaneously on a genome-wide basis in a large number of individuals will provide valuable tools for identifying the function of previously mapped susceptibility alleles underlying common complex diseases.

Although eQTLs are shown to be effective in mapping complex traits, there are many levels of information that are inherent in the measurement of global gene expression that have yet to be accessed, such as the effects of transcript stability, epigenetic effects or environmental stimuli. In addition, larger studies involving thousands of subjects may be necessary to identify weak *trans* effects with the same precision as the more powerful effects that are often observed in *cis*. Although *trans* effects can be relatively weak, the genes they modify (the *trans*-transcriptome) are likely to contain master regulators with wide effects on key processes that might feature more strongly in tissues or in cells subjected to particular environmental stimuli. Many genes are only expressed in particular tissues or at specific times during development. Thus, although systematic studies of eQTLs are already being planned for a wide variety of tissues, other strategies will need to be formed to study particular cell types and tissues at specific stages of differentiation and development.

Understanding the genome of cancer cells and tissues is particularly challenging because the primary lesions that initially drive cellular proliferation are difficult to find when uncontrolled division results in progressive secondary damage to the genome and the transcriptome. eQTL analyses may be of particular value in malignant disease, because they allow a more integrated picture of what is happening in cancer cells (BOX 4).

Good progress is being made in cataloguing the SNPs and other polymorphisms that regulate transcription, and this could be the basis for a systematic listing of regulatory sequences and regulatory proteins. Identifying epigenetic effects is likely to be more difficult, particularly if they are mediated through histone modifications (which are difficult to detect on a large scale) rather than through differential CpG methylation.

The remarkable diversity of human transcriptional regulation raises new questions about the evolutionary value of unexpected variation in genes that mediate basic mechanisms, such as heat shock proteins or genes influencing the cell cycle and DNA repair. 'Inverse genetics' could be used to study the SNPs with the strongest effects on expression of such genes, and to investigate their actions on unexpected phenotypes measured in epidemiological samples.

New analytical techniques, particularly network analyses, promise rapid advances in reducing the complexity of expression data. Modules of co-expressed genes mediating complex functions may also be identified by time-series studies of the response of particular cell types to environmental stimuli[82]. In future, integration of eQTLs with data from large-scale approaches for genome resequencing, from proteomic and metabolomic analyses, from epigenomic studies and from functional screening of genes may provide a powerful set of tools for a systems biology approach to multifactorial disease, as well as a way to identify and biologically validate susceptibility genes[83].

In the future, complex disease geneticists will require integrated public databases. Existing databases include the asthma study database (mRNA by SNP Browser v 1.0.1) and VarySysDB. A more comprehensive database is planned as part of the NIH GTEx project, which will house existing as well as GTEx-generated eQTL data. Future databases should include eQTL maps with SNPs, epigenetic marks, *trans* and *cis* effects, as well as effects that are specific for particular cells, tissues and environmental stimuli. Ultimately, they will also allow browsing for networks, modules and comparisons with model organisms.

1. Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
2. Palmer, C. N. *et al.* Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. *Nature Genet.* **38**, 441–446 (2006).
3. Burton, P. R. *et al.* Association scan of 14,500 non-synonymous SNPs in four diseases identifies autoimmunity variants. *Nature Genet.* **39**, 1329–1337 (2007).
4. Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003). **This paper shows the power of eQTL analysis in humans.**
5. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747 (2004).
6. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
7. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nature Rev. Genet.* **7**, 862–872 (2006).
8. Gilad, Y., Rifkin, S. A. & Pritchard, J. K. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* **24**, 408–415 (2008).
9. Jia, Z. & Xu, S. Mapping quantitative trait loci for expression abundance. *Genetics* **176**, 611–623 (2007).
10. Carlborg, O. *et al.* Methodological aspects of the genetic dissection of gene expression. *Bioinformatics* **21**, 2383–2393 (2005).

11. Kendziorski, C. M., Chen, M., Yuan, M., Lan, H. & Attie, A. D. Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* **62**, 19–27 (2006).
12. Schliekelman, P. Statistical power of expression quantitative trait loci for mapping of complex trait loci in natural populations. *Genetics* **178**, 2201–2216 (2008).
13. Dixon, A. L. *et al.* A genome-wide association study of global gene expression. *Nature Genet.* **39**, 1202–1207 (2007).
14. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nature Rev. Genet.* **9**, 255–266 (2008).
15. Goring, H. H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genet.* **39**, 1208–1216 (2007).
16. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).
This paper illustrates the power of eQTL and network analysis in unravelling complex trait genetics.
17. Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6**, e107 (2008).
18. Petretto, E. *et al.* Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet.* **2**, e172 (2006).
19. Monks, S. A. *et al.* Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* **75**, 1094–1105 (2004).
20. Veyrieras, J. B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4**, e1000214 (2008).
21. Hubner, N. *et al.* Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genet.* **37**, 243–253 (2005).
22. Yvert, G. *et al. Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nature Genet.* **35**, 57–64 (2003).
23. Shimada, M. K. *et al.* VarySysDB: a human genetic polymorphism database based on all H-InvDB transcripts. *Nucleic Acids Res.* **37**, D810–D815 (2008).
24. Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
25. Gonzales, J. M. *et al.* Regulatory hotspots in the malaria parasite genome dictate transcriptional variation. *PLoS Biol.* **6**, e238 (2008).
26. Mileyko, Y., Joh, R. I. & Weitz, J. S. Small-scale copy number variation and large-scale changes in gene expression. *Proc. Natl Acad. Sci. USA* **105**, 16659–16664 (2008).
27. Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genet.* **38**, 1378–1385 (2006).
This paper shows the extent and distribution of methylation in the human genome.
28. Krebs, J. E. Moving marks: dynamic histone modifications in yeast. *Mol. Biosyst.* **3**, 590–597 (2007).
29. Myers, A. J. *et al.* A survey of genetic human cortical gene expression. *Nature Genet.* **39**, 1494–1499 (2007).
30. Moffatt, M. F. *et al.* Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature* **448**, 470–473 (2007).
31. Bouzigon, E. *et al.* Effect of 17q21 variants and smoking exposure in early-onset asthma. *N. Engl. J. Med.* **359**, 1985–1994 (2008).
32. Duan, S. *et al.* Genetic architecture of transcript-level variation in humans. *Am. J. Hum. Genet.* **82**, 1101–1113 (2008).
33. Galanter, J. *et al.* ORMDL3 gene is associated with asthma in three ethnically diverse populations. *Am. J. Respir. Crit. Care Med.* **177**, 1194–1200 (2008).
34. Sleiman, P. M. *et al. ORMDL3* variants associated with asthma susceptibility in North Americans of European ancestry. *J. Allergy Clin. Immunol.* **122**, 1225–1227 (2008).
35. Tavendale, R., Macgregor, D. F., Mukhopadhyay, S. & Palmer, C. N. A polymorphism controlling *ORMDL3* expression is associated with asthma that is poorly controlled by current medications. *J. Allergy Clin. Immunol.* **121**, 860–863 (2008).
36. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genet.* **40**, 955–962 (2008).
A substantial meta-analysis of susceptibility loci underlying Crohn's disease that illustrates the problem of unattributed heritability and the utility of eQTL data in understanding the function of disease-associated SNPs.
37. Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of *PTGER4*. *PLoS Genet.* **3**, e58 (2007).
38. Kabashima, K. *et al.* The prostaglandin receptor EP4 suppresses colitis, mucosal damage and CD4 cell activation in the gut. *J. Clin. Invest.* **109**, 883–893 (2002).
39. Rioux, J. D. *et al.* Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nature Genet.* **29**, 223–228 (2001).
40. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genet.* **41**, 56–65 (2009).
41. 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).
42. Horton, R. *et al.* Gene map of the extended human MHC. *Nature Rev. Genet.* **5**, 889–899 (2004).
43. Alberts, R. *et al.* Sequence polymorphisms cause many false *cis* eQTLs. *PLoS ONE* **2**, e622 (2007).
44. Beaty, J. S., West, K. A. & Nepom, G. T. Functional effects of a natural polymorphism in the transcriptional regulatory sequence of HLA-DQB1. *Mol. Cell. Biol.* **15**, 4771–4782 (1995).
45. Nejentsev, S. *et al.* Localization of type 1 diabetes susceptibility to the MHC class I genes *HLA-B* and *HLA-A*. *Nature* **450**, 887–892 (2007).
46. Sieberts, S. K. & Schadt, E. E. Moving toward a system genetics view of disease. *Mamm. Genome* **18**, 389–401 (2007).
47. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
48. Goh, K. I. *et al.* The human disease network. *Proc. Natl Acad. Sci. USA* **104**, 8685–8690 (2007).
49. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429–435 (2008).
50. Barnes, M., Freudenberg, J., Thompson, S., Aronow, B. & Pavlidis, P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res.* **33**, 5914–5923 (2005).
51. Pedotti, P. *et al.* Can subtle changes in gene expression be consistently detected with different microarray platforms? *BMC Genomics* **9**, 124 (2008).
52. van Ruissen, F. *et al.* Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips. *BMC Genomics* **6**, 91 (2005).
53. Bosotti, R. *et al.* Cross platform microarray analysis for robust identification of differentially expressed genes. *BMC Bioinformatics* **8** (Suppl. 1), S5 (2007).
54. Ji, Y. *et al.* RefSeq refinements of UniGene-based gene matching improve the correlation of expression measurements between two microarray platforms. *Appl. Bioinformatics* **5**, 89–98 (2006).
55. Carter, S. L., Eklund, A. C., Mecham, B. H., Kohane, I. S. & Szallasi, Z. Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC Bioinformatics* **6**, 107 (2005).
56. Sohail, M., Akhtar, S. & Southern, E. M. The folding of large RNAs studied by hybridization to arrays of complementary oligonucleotides. *RNA* **5**, 646–655 (1999).
57. Southern, E., Mir, K. & Shchepinov, M. Molecular interactions on microarrays. *Nature Genet.* **21**, 5–9 (1999).
This review, by the inventor of DNA microarrays, highlights the complexity and unpredictability of the interactions between nucleic acids in solution and target sequences on solid supports.
58. Kapur, K., Xing, Y., Ouyang, Z. & Wong, W. H. Exon arrays provide accurate assessments of gene expression. *Genome Biol.* **8**, R82 (2007).
59. Okoniewski, M. J., Hey, Y., Pepper, S. D. & Miller, C. J. High correspondence between Affymetrix exon and standard expression arrays. *Biotechniques* **42**, 181–185 (2007).
60. Clark, T. A. *et al.* Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol.* **8**, R64 (2007).
61. Wold, B. & Myers, R. M. Sequence census methods for functional genomics. *Nature Methods* **5**, 19–21 (2008).
62. Watson, R. M., Griaznova, O. I., Long, C. M. & Holland, M. J. Increased sample capacity for genotyping and expression profiling by kinetic polymerase chain reaction. *Anal. Biochem.* **329**, 58–67 (2004).
63. Weedon, M. N. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genet.* **40**, 575–583 (2008).
64. Brem, R. B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl Acad. Sci. USA* **102**, 1572–1577 (2005).
65. Moffatt, M. & Cookson, W. The genetics of asthma. Maternal effects in atopic disease. *Clin. Exp. Allergy* **28** (Suppl. 1), 56–61 (1998).
66. Bennett, S. & Todd, J. Human type 1 diabetes and the insulin gene: principles of mapping polygenes. *Annu. Rev. Genet.* **30**, 343–370 (1996).
67. Warram, J. H., Krolewski, A. S., Gottlieb, M. S. & Kahn, C. R. Differences in risk of insulin-dependent diabetes in offspring of diabetic mothers and diabetic fathers. *N. Engl. J. Med.* **311**, 149–152 (1984).
68. Koumantaki, Y. *et al.* Family history as a risk factor for rheumatoid arthritis: a case–control study. *J. Rheumatol.* **24**, 1522–1526 (1997).
69. Burden, A. *et al.* Genetics of psoriasis: paternal inheritance and a locus on chromosome 6p. *J. Invest. Dermatol.* **110**, 958–960 (1998); comment **112**, 514–516 (1999).
70. Akolkar, P. N. *et al.* Differences in risk of Crohn's disease in offspring of mothers and fathers with inflammatory bowel disease. *Am. J. Gastroenterol.* **92**, 2241–2244 (1997).
71. Vorechovsky, I., Webster, A. D., Plebani, A. & Hammarstrom, L. Genetic linkage of IgA deficiency to the major histocompatibility complex: evidence for allele segregation distortion, parent-of-origin penetrance differences, and the role of anti-IgA antibodies in disease predisposition. *Am. J. Hum. Genet.* **64**, 1096–1109 (1999).
72. Grosshans, H. & Filipowicz, W. Molecular biology: the expanding world of small RNAs. *Nature* **451**, 414–416 (2008).
73. Plagnol, V. *et al.* Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. *PLoS ONE* **3**, e2966 (2008).
74. Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. & Kinzler, K. W. Allelic variation in human gene expression. *Science* **297**, 1143 (2002).
75. Cheung, V. G. *et al.* Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genet.* **33**, 422–425 (2003).
76. Gretarsdottir, S. *et al.* The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nature Genet.* **35**, 131–138 (2003).
77. Kohane, I. S., Kho, A. T. & Butte, A. J. *Microarrays for an Integrative Genomics* (MIT Press, Cambridge, Massachusetts, 2002).
78. Idaghdour, Y., Storey, J. D., Jadallah, S. J. & Gibson, G. A genome-wide gene expression signature of environmental geography in leukocytes of Moroccan Amazighs. *PLoS Genet.* **4**, e1000052 (2008).
Although this paper describes a small study, it shows the profound effects of different environments on gene expression in peripheral blood lymphocytes.
79. Li, Y. *et al.* Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS Genet.* **2**, e222 (2006).
80. Smith, E. N. & Kruglyak, L. Gene–environment interaction in yeast gene expression. *PLoS Biol.* **6**, e83 (2008).
81. Gibson, G. The environmental contribution to gene expression profiles. *Nature Rev. Genet.* **9**, 575–581 (2008).
82. Reis, B. Y., Butte, A. S. & Kohane, I. S. Extracting knowledge from dynamics in gene expression. *J. Biomed. Inform.* **34**, 15–27 (2001).
This paper shows the utility of using time-series measurements of gene expression to identify co-regulated modules of genes.

83. Schadt, E. E. & Lum, P. Y. Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Reverse engineering gene networks to identify key drivers of complex disease phenotypes. *J. Lipid Res.* **47**, 2601–2613 (2006).
84. Gudbjartsson, D. F. *et al.* Many sequence variants affecting diversity of adult human height. *Nature Genet.* **40**, 609–615 (2008).
85. Hom, G. *et al.* Association of systemic lupus erythematosus with *C8orf13–BLK* and *ITGAM–ITGAX*. *N. Engl. J. Med.* **358**, 900–909 (2008).
86. Hakonarson, H. *et al.* A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study. *Diabetes* **57**, 1143–1146 (2008).
87. The 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).
88. Todd, J. A. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genet.* **39**, 857–864 (2007).
89. Plenge, R. M. *et al. TRAF1–C5* as a risk locus for rheumatoid arthritis — a genomewide study. *N. Engl. J. Med.* **357**, 1199–1209 (2007).
90. Thein, S. L. *et al.* Intergenic variants of *HBS1L–MYB* are responsible for a major QTL on chromosome 6q23 influencing HbF levels in adults. *Proc. Natl Acad. Sci. USA* (in the press).
91. Di Bernardo, M. C. *et al.* A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nature Genet.* **40**, 1204–1210 (2008).
92. Gardner, T. S., di Bernardo, D., Lorenz, D. & Collins, J. J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**, 102–105 (2003).
93. Sontag, E., Kiyatkin, A. & Kholodenko, B. N. Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data. *Bioinformatics* **20**, 1877–1886 (2004).
94. Li, H. *et al.* Integrative genetic analysis of transcription modules: towards filling the gap between genetic loci and inherited traits. *Hum. Mol. Genet.* **15**, 481–492 (2006).
95. Keurentjes, J. J. *et al.* Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proc. Natl Acad. Sci. USA* **104**, 1708–1713 (2007).
96. Gerrits, A., Dykstra, B., Otten, M., Bystrykh, L. & de Haan, G. Combining transcriptional profiling and genetic linkage analysis to uncover gene networks operating in hematopoietic stem cells and their progeny. *Immunogenetics* **60**, 411–422 (2008).
97. de Koning, D. J., Carlborg, O. & Haley, C. S. The genetic dissection of immune response using gene-expression studies and genome mapping. *Vet. Immunol. Immunopathol.* **105**, 343–352 (2005).
98. Akey, J. M., Biswas, S., Leek, J. T. & Storey, J. D. On the design and analysis of gene expression studies in human populations. *Nature Genet.* **39**, 807–808; author reply 808–809 (2007).
99. Spielman, R. S. *et al.* Common genetic variants account for differences in gene expression among ethnic groups. *Nature Genet.* **39**, 226–231 (2007).
100. Doss, S., Schadt, E. E., Drake, T. A. & Lusis, A. J. Cis-acting expression quantitative trait loci in mice. *Genome Res.* **15**, 681–691 (2005).
101. Hughes, T. R. *et al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnol.* **19**, 342–347 (2001).
102. Alberts, R., Terpstra, P., Bystrykh, L. V., de Haan, G. & Jansen, R. C. A statistical multiprobe model for analyzing *cis* and *trans* genes in genetical genomics experiments with short-oligonucleotide arrays. *Genetics* **171**, 1437–1439 (2005).
103. Halazonetis, T. D., Gorgoulis, V. G. & Bartek, J. An oncogene-induced DNA damage model for cancer development. *Science* **319**, 1352–1355 (2008).
104. Sun, Z., Wigle, D. A. & Yang, P. Non-overlapping and non-cell-type-specific gene expression signatures predict lung cancer survival. *J. Clin. Oncol.* **26**, 877–883 (2008).
105. Walker, B. A. *et al.* Integration of global SNP-based mapping and expression arrays reveals key regions, mechanisms, and genes important in the pathogenesis of multiple myeloma. *Blood* **108**, 1733–1743 (2006).
106. Lastowska, M. *et al.* Identification of candidate genes involved in neuroblastoma progression by combining genomic and expression microarrays with survival data. *Oncogene* **26**, 7432–7444 (2007).
107. Huang, R. S. *et al.* A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc. Natl Acad. Sci. USA* **104**, 9758–9763 (2007).
108. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005). **This paper describes a statistical approach to network analyses and provides a set of software tools for their implementation.**
109. Horvath, S. *et al.* Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc. Natl Acad. Sci. USA* **103**, 17402–17407 (2006).

Acknowledgements

The work was supported by the Wellcome Trust and the EC funded GABRIEL project, the French Ministry of Research and Higher Education and by grants from the National Institutes of Health.

FURTHER INFORMATION

Liming Liang's homepage: http://www.sph.umich.edu/csg/liang
Abecasis laboratory homepage (contains programs for genome-scale data analysis): http://www.sph.umich.edu/csg/abecasis
Catalog of Published Genome-Wide Association Studies: http://www.genome.gov/gwastudies
Database of Genomic Variants: http://projects.tcag.ca/variation
dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP
D-HaploDB: http://orca.gen.kyushu-u.ac.jp
Genotype-Tissue Expression (GTEx): http://nihroadmap.nih.gov/GTEx
H-InvDB: http://www.h-invitational.jp
mRNA by SNP Browser v 1.0.1: http://www.sph.umich.edu/csg/liang/asthma
UniGene: http://www.ncbi.nlm.nih.gov/unigene
VarySysDB: http://www.h-invitational.jp/varygene/home.htm
WaferGen: http://www.wafergen.com

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

FUNDAMENTAL CONCEPTS IN GENETICS

Effective population size and patterns of molecular evolution and variation

Brian Charlesworth

Abstract | The effective size of a population, N_e, determines the rate of change in the composition of a population caused by genetic drift, which is the random sampling of genetic variants in a finite population. N_e is crucial in determining the level of variability in a population, and the effectiveness of selection relative to drift. This article reviews the properties of N_e in a variety of different situations of biological interest, and the factors that influence it. In particular, the action of selection means that N_e varies across the genome, and advances in genomic techniques are giving new insights into how selection shapes N_e.

Genetic drift
The process of evolutionary change involving the random sampling of genes from the parental generation to produce the offspring generation, causing the composition of the offspring and parental generations to differ.

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK.
e-mail: Brian.Charlesworth@ed.ac.uk
doi:10.1038/nrg2526
Published online
10 February 2009

The effective size of a population (N_e) is one of several core concepts introduced into population genetics by Sewall Wright, and was initially sketched in his magnum opus, *Evolution in Mendelian Populations*[1]. Its purpose is to provide a way of calculating the rate of evolutionary change caused by the random sampling of allele frequencies in a finite population (that is, genetic drift). The basic theory of N_e was later extended by Wright[2–5], and a further theoretical advance was made by James Crow[6], who pointed out that there is more than one way of defining N_e, depending on the aspect of drift in question. More recently, the theoretical analysis of the effects of demographic, genetic and spatial structuring of populations has been greatly simplified by the use of approximations that resolve drift into processes operating on different timescales[7].

What biological questions does N_e help to answer? First, the product of mutation rate and N_e determines the equilibrium level of neutral or weakly selected genetic variability in a population[8]. Second, the effectiveness of selection in determining whether a favourable mutation spreads, or a deleterious mutation is eliminated, is controlled by the product of N_e and the intensity of selection. The value of N_e therefore greatly affects DNA sequence variability, and the rates of DNA and protein sequence evolution[8].

The importance of N_e as an evolutionary factor is emphasized by findings that N_e values are often far lower than the census numbers of breeding individuals in a species[9,10]. Species with historically low effective population sizes, such as humans, show evidence for reduced variability and reduced effectiveness of selection in comparison with other species[11]. N_e may also vary across different locations in the genome of a species, either as a result of differences in the modes of transmission of different components of the genome (for example, the X chromosome versus the autosomes[12]), or because of the effects of selection at one site in the genome on the behaviour of variants at nearby sites[13]. An important consequence of the latter process is that selection causes reduced N_e in genomic regions with low levels of genetic recombination, with effects that are discernible at the molecular sequence level[14,15]. BOX 1 summarizes the major factors influencing N_e, which will be described in detail below.

In the era of multi-species comparisons of genome sequences and genome-wide surveys of DNA sequence variability, there is more need than ever before to understand the evolutionary role of genetic drift, and its interactions with the deterministic forces of mutation, migration, recombination and selection. N_e therefore plays a central part in modern studies of molecular evolution and variation, as well as in plant and animal breeding and in conservation biology. In this Review, I first describe some basic theoretical tools for obtaining expressions for N_e, and then show how the results of applying these tools can be used to describe the properties of a single population, and how to include the effects of selection. Finally, I describe the effects of structuring of populations by spatial location or by genotype, and discuss the implications of genotypic structuring for patterns of variation and evolution across the genome.

Describing genetic drift and determining N_e

There are three major ways in which genetic drift can be modelled in the simplest type of population, which are outlined below. These theoretical models lead to a general approach that can be applied to situations of greater biological interest, which brings out the utility of the concept of the effective population size.

Poisson distribution
This is the limiting case of the binomial distribution (see next page), valid when the probability of an event is very small. The mean and variance of the number of events are then equal.

Coalescent theory
A method of reconstructing the history of a sample of alleles from a population by tracing their genealogy back to their most recent common ancestral allele.

Coalescence
The convergence of a pair of alleles in a sample to a common ancestral allele, tracing them back in time.

Fast timescale approximation
Used to simplify calculations of effective population size, by assuming that the rate of coalescence is slower than the rate at which alleles switch between different compartments of a structured population as we trace them back in time.

Panmictic
A panmictic population lacks subdivision according to spatial location or genotype, so that all parental genotypes potentially contribute to the same pool of offspring.

The Wright–Fisher population. To see why N_e is so useful, we need to understand how genetic drift can be modelled in the simple case of a Wright–Fisher population[1,16,17]. This is a randomly mating population, consisting of a number of diploid hermaphroditic individuals (N). The population reproduces with discrete generations, each generation being counted at the time of breeding. New individuals are formed each generation by random sampling, with replacement, of gametes produced by the parents, who die immediately after reproduction. Each parent thus has an equal probability of contributing a gamete to an individual that survives to breed in the next generation. If N is reasonably large, this implies a Poisson distribution of offspring number among individuals in the population. A population of hermaphroditic marine organisms, which shed large numbers of eggs and sperm that fuse randomly to make new zygotes, comes closest to such an idealized situation.

With this model, the rate at which genetic drift causes an increase in divergence in selectively neutral allele frequencies between isolated populations, or loss of variability within a population, is given by $1/(2N)$ (BOX 2). An alternative approach, which has a central role in the contemporary modelling and interpretation of data on DNA sequence variation[7,18], is provided by the theory of the coalescent process (the coalescent theory) (BOX 3). Instead of looking at the properties of the population as a whole, we consider a set of alleles at a genetic locus that have been sampled from a population. If we trace their ancestry back in time, they will eventually be derived from the same ancestral allele, that is, they have undergone coalescence (BOX 3). This is obviously closely related to the inbreeding coefficient approach to drift described in BOX 2, and the rate of the coalescent process in a Wright–Fisher population is also inversely related to the population size.

Box 1 | **Factors affecting the effective size of a population**

- Division into two sexes: a small number of individuals of one sex can greatly reduce effective population size (N_e) below the total number of breeding individuals (N).
- Variation in offspring number: a larger variance in offspring number than expected with purely random variation reduces N_e below N.
- Inbreeding: the correlation between the maternal and paternal alleles of an individual caused by inbreeding reduces N_e.
- Mode of inheritance: the N_e experienced by a locus depends on its mode of transmission; for example, autosomal, X-linked, Y-linked or organelle.
- Age- and stage-structure: in age- and stage-structured populations, N_e is much lower than N.
- Changes in population size: episodes of low population size have a disproportionate effect on the overall value of N_e.
- Spatial structure: the N_e determining the mean level of neutral variability within a local population is often independent of the details of the migration process connecting populations. Limited migration between populations greatly increases N_e for the whole population, whereas high levels of local extinction have the opposite effect.
- Genetic structure: the long-term maintenance of two or more alleles by balancing selection results in an elevation in N_e at sites that are closely linked to the target of selection. In contrast, directional selection causes a reduction in N_e at linked sites (the Hill–Robertson effect).

More realistic models of drift. The assumptions of the Wright–Fisher population model do not, however, apply to most populations of biological interest: many species have two sexes, there may be nonrandom variation in reproductive success, mating may not be at random, generations might overlap rather than being discrete, the population size might vary in time, or the species may be subdivided into local populations or distinct genotypes. In addition, we need to analyse the effects of deterministic evolutionary forces, such as selection and recombination, as well as drift.

The effective population size describes the timescale of genetic drift in these more complex situations: we replace $2N$ by $2N_e$, where N_e is given by a formula that takes into account the relevant biological details. Classically, this has been done by calculations based on the variance or inbreeding coefficient approaches[19–23], but more recently coalescent theory has been employed[7]. In general, the use of N_e only gives an approximation to the rate of genetic drift for a sufficiently large population size (such that the square of $1/N$ can be neglected compared with $1/N$), and is often valid only asymptotically, that is, after enough time has elapsed since the start of the process. Exact calculations of changes in variance of allele frequencies or inbreeding coefficient are, therefore, often needed in applications in which the population size is very small or the timescale is short, as in animal and plant improvement or in conservation breeding programmes[19–21,24].

Determining N_e: a general method. Coalescent theory provides a flexible and powerful method for obtaining formulae for N_e, replacing the term involving N in the rate of coalescence in BOX 3 by N_e, which can then be directly inserted in place of N into the results from coalescent theory (BOX 3). A core approach for estimating N_e under different circumstances is outlined briefly below and is discussed in more detail in the following sections of this Review.

This approach involves the structured coalescent process, in which there are several 'compartments' (such as ages or sexes) in the population from which alleles can be sampled[7,25,26]. Alleles are initially sampled from one or more of these compartments, and the probabilities of allele movements to the other compartments, as we go back in time, are determined by the rules of inheritance. A useful simplification is to assume that alleles flow among the different compartments at a much faster rate than the coalescence of alleles: this is termed the fast timescale approximation. This means that we can treat the sampled alleles as coming from the equilibrium state of the process[7,27–31]. This provides a general formula for the rate of coalescence, which is easy to apply to individual cases[7,28–31].

Determining N_e of a single population

The structured coalescent process can be applied to different biologically important scenarios. In this section, I discuss how it can be applied to panmictic populations (BOX 2), with particular reference to the effects on N_e of variation in offspring number among individuals, the

Box 2 | **Using the Wright–Fisher model to describe genetic drift**

Consider the effects of genetic drift on selectively neutral variants, assuming that the population is closed (there is no migration from elsewhere) and panmictic. We also ignore the possibility of mutation. Assume that there are two alternative variants at an autosomal site, A_1 and A_2, with frequencies p_0 and $q_0 = 1 - p_0$ in an initial generation; these might represent two alternative nucleotide pairs at a given site in a DNA sequence, such as GC and AT.

The state of the population in the next generation can then be described by the probability that the new frequency of A_2 is $i/(2N)$, where i can take any value between 0 and $2N$. $2N$ is used because with diploid inheritance there are $2N$ allele copies in N individuals; if the species were haploid, we would use N. The Wright–Fisher model is identical to the classical problem in probability theory of determining the chance of i successes out of a specified number ($2N$) of trials (a success being the choice of A_2 rather than A_1) when the chance of success on a single trial is q. Tossing an unbiased coin $2N$ times corresponds to the case in which $q = 0.5$.

Probability theory tells us that the chances of obtaining i copies of A_2 in the next generation, corresponding to a frequency of $q = i/(2N)$, is given by the binomial distribution[1,16]. The new mean frequency of A_2 is simply q_0, as drift does not affect the mean. But the frequency in any given population will probably change somewhat, becoming $q_0 + \delta q$, where the change δq has variance $V_{\delta q}$, given by:

$$V_{\delta q} = \frac{p_0 q_0}{2N} \qquad (3)$$

After a further generation, the new frequency will be $q_0 + \delta q + \delta q'$, where $\delta q'$ has a mean of zero and a variance of $(p_0 - \delta q)(q_0 + \delta q)/(2N)$, and so on. If we follow a single population, there will be a succession of random changes in q, until eventually A_2 either becomes fixed in the population ($q = 1$) or is lost ($q = 0$).

From equation 3 above, the rate of increase in variance per generation is proportional to $1/(2N)$. This variance can be thought of as measuring the extent of differentiation in allele frequencies between a large set of completely isolated populations, all of which started with the same initial state. Alternatively, it represents the variation in allele frequencies among a set of independent loci within the genome, all with the same initial state.

An alternative way of looking at drift is to use the concept of identity by descent[84,141,142]. Two different allelic copies of a given nucleotide site drawn from a population are identical by descent (IBD) if they trace their ancestry back to a single ancestral copy. The progress of a population towards genetic uniformity is measured by the probability that a pair of randomly sampled alleles are IBD (a value termed the inbreeding coefficient, f), measured relative to an initial generation in which all the alleles in the population are not IBD. Just as for the variance in allele frequency, the inbreeding coefficient increases at a rate that is governed by $1/(2N)$, and the inbreeding coefficient at a given time is equal to the variance divided by $p_0 q_0$ (REFS 1,5). Approach to uniformity thus occurs at the same rate as increase in variance of allele frequencies.

mode of inheritance and the consequence of changes in population size. By looking at real-life data we see that different methods of estimating N_e can give very different answers if the population size has changed greatly.

Outbreeding populations with constant size. First we consider a population with no inbreeding and a Poisson distribution of offspring number. $1/N_e$ for autosomal (A) inheritance and two sexes (m, male; f, female) is given by:

$$\frac{1}{N_{eA}} \approx \frac{1}{4N_m} + \frac{1}{4N_f} \qquad (1)$$

With a 1:1 sex ratio among breeding individuals, the effective size in this case is approximately equal to the total population size ($N = 2N_f = 2N_m$), so that the population then has the same properties as the Wright–Fisher model. But if the numbers of females and males are not the same, the effective size is much less than N. For example, if there is only a small number of breeding males compared with females, the reciprocal of N_m dominates equation 1, and N_e is close to $4N_m$. This reflects the fact that half of the genes in a new generation must come from males, regardless of their numbers relative to females. This situation is approached in populations of farm animals, where artificial insemination is used in selective breeding, causing serious problems with inbreeding[32].

With nonrandom variation in offspring numbers, but with the same variance in offspring number for the two sexes and a 1:1 sex ratio, we have:

$$\frac{1}{N_{eA}} \approx \frac{(2 + \Delta V)}{2N} \qquad (2)$$

An excess variance in offspring numbers compared with random expectation thus reduces N_e below N (REFS 3,4). Conversely, if there is less than random variation, N_e can be greater than N; it equals $2N$ in the extreme case when all individuals have equal reproductive success. This is important for conservation breeding programmes, as it is desirable to maximize N_e in order to slow down the approach to homozygosity[33]. In animals, a major cause of a nonrandom distribution of reproductive success is sexual selection, when males compete with each other for access to mates[34]. Sexual selection is thus likely to have a major effect on N_e, with the magnitude of the effect being dependent on the details of the mating system[35,36].

The effect of inbreeding. An excess of matings between relatives reduces N_e by a factor of $1/(1 + F_{IS})$ (REF. 30), where F_{IS} is the inbreeding coefficient of an individual, caused by an excess frequency over random mating expectation of matings between relatives[37]. N_e is reduced because inbreeding causes faster coalescence of an individual's maternal and paternal alleles compared with random mating[38]. With partial self-fertilization with frequency S in an hermaphrodite population, the equilibrium inbreeding coefficient is $F_{IS} = S/(2 - S)$ (REF. 19). Selfing causes N_e to be multiplied by a factor of $(2 - S)/2$ if there is random variation in offspring number; this approaches 1/2 for 100% selfing[30,38,39].

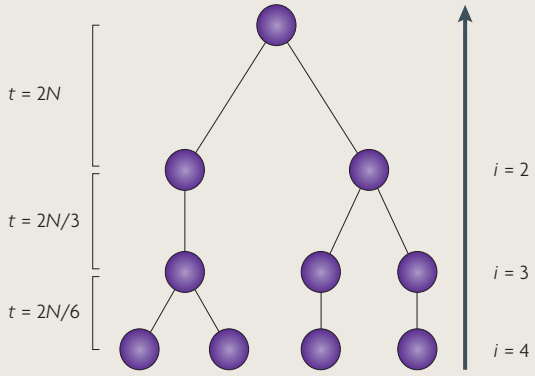
From equation 4 in BOX 3, with N_e replacing N, this result suggests that neutral variability within populations of highly self-fertilizing species, such as *Arabidopsis thaliana* and *Caenorhabditis elegans*, should be reduced to approximately half the value for randomly mating populations of similar size. Indeed, these species do have low levels of genetic variability[40,41] compared with their outcrossing relatives[42,43] (TABLE 1). Additional possible reasons for this low variability are discussed below.

The effects of mode of inheritance. The mode of inheritance can also greatly alter N_e, and hence expected levels of neutral diversity (as shown by the equations in BOX 4). For example, with X-linked inheritance and random mating, a 1:1 sex ratio and Poisson distribution of offspring numbers imply that $N_{eX} = 3N/4$, consistent with the fact that there are only three-quarters as many X chromosomes as autosomes in the population. It is

Binomial distribution
Describes the probability of observing i independent events in a sample of size n, when the probability of an event is p. The mean and variance of the number of events are np and $np(1 - p)$, respectively.

Neutral diversity
Variability arising from mutations that have no effect on fitness.

Box 3 | **The coalescent process**



We consider a sample of alleles at a genetic locus that have been obtained from a population (see the figure; the four bottom circles). For simplicity, assume that no recombination can occur in the locus, as would be true for a mitochondrial genome or Y chromosome, or for a nuclear gene in a region of a chromosome with severely reduced recombination. If we trace the ancestry of the alleles back in time (upward arrow), two of the alleles in the sample will be seen to be derived from the same ancestral allele — they have coalesced at a time in the history of the population when the other two alleles still trace back to two distinct alleles. At this time, there are three distinct alleles from which the sample is descended ($i = 3$). If we continue back in time, the ancestry of the alleles in the sample follows a bifurcating tree, in which the time (t) between successive nodes (points of branching) is dependent on $2N$ and the number of alleles that are present at the later node; with i alleles, the expected time to a coalescent event that generate $i - 1$ alleles is $4N/i(i - 1)$ (REFS 7,18,112,143). This assumes that N is sufficiently large that, at most, one coalescent event can occur in a given generation. The time itself follows an exponential distribution, with a standard deviation equal to the mean. In the figure, t represents the expected times at which the successive coalescent events occur in a Wright–Fisher population, corresponding to the numbers of distinct alleles, i, on the right.

This description of a gene tree is purely theoretical, as gene trees cannot be observed directly. However, the results are relevant to data on population samples, because variation in a sample of allelic sequences reflects mutations that have arisen in different branches of the tree since the most recent common ancestor. To model a sample, we simply allow mutations to occur on the lineages in the gene tree. The simplest model to use is the infinite sites model: the mutation rate probability per generation per site is u, and u is assumed to be low, so that at most one mutation arises per site in the tree[7,18,112,144].

This allows derivations of formulae to predict the values of commonly used measures of variability such as the nucleotide site diversity, that is, the frequency with which a pair of randomly sampled alleles differ at a given nucleotide site. Consider a given pair of alleles taken randomly from the sample. There is a time (t) connecting each of them to their common ancestor. They will be identical at a site if no mutation has arisen over the time separating them from each other, which is $2t$. The probability that a mutation has arisen at that site, and caused them to differ in state, is $2tu$. From the above considerations, t has an expected value of $2N$, so that the net probability of a difference in state at a given nucleotide site is $4Nu$. Averaging over all pairs of alleles in the sample, and over a large number of sites, gives the expected value of the nucleotide site diversity for a sample (π):

$$\pi = 4Nu \qquad (4)$$

In addition to generating simple and useful expressions for the expected level of variability in a sample from a population, coalescent theory allow the computation of the probability distributions of statistics that describe the frequencies of variants in the sample. This permits statistical tests to be applied to data, to test whether the assumptions of the standard model (demographic equilibrium and neutrality) are violated[7,18,112].

therefore common practice to adjust diversity estimates for X-linked loci by multiplying by 4/3 when comparing them with data for autosomal genes; see REF 44 for an example. But the formulae in BOX 4 show that this is an over-simplification. If there is strong sexual selection among males, the effective size for X-linked loci can approach or even exceed that for autosomal loci. N_{eX}/N_{eA} has an upper limit of 1.125; the reason that this ratio can exceed 1 is that autosomes are transmitted through males more often than X chromosomes, and the males' effective population size is small. Surveys of variability in the putatively ancestral African populations of *Drosophila melanogaster* show that the mean silent site nucleotide diversity for X-linked loci is indeed slightly higher than for autosomal loci[45–47], consistent with the operation of very strong sexual selection, although other factors might also be involved[46,48].

For ZW sex determination systems, the predicted difference between males and females is reversed. For Z-linked inheritance, N_{eZ}/N_{eA} with strong sexual selection can be as low as 9/16. Data on DNA sequence variability in introns in domestic chickens gave a ratio of Z-linked to autosomal variability of 0.24, even lower than expected under strong sexual selection[49]. For organelle inheritance, with strictly maternal transmission, N_e is one-quarter of the autosomal value with random variation in offspring number, but is expected to be much larger with sexual selection (BOX 4).

Age- and stage-structure. To calculate N_e for populations in which reproductive individuals have a range of ages or developmental stages, the fast timescale approximation can again be applied. In this case, alleles flow between ages or stages as well as sexes. Expressions can be derived for N_e in an age- or stage-structured population

Table 1 | **Effective population size (N_e) estimates from DNA sequence diversities**

Species	N_e	Genes used	Refs
Species with direct mutation rate estimates			
Humans	10,400	50 nuclear sequences	145
Drosophila melanogaster (African populations)	1,150,000	252 nuclear genes	108
Caenorhabditis elegans (self-fertilizing hermaphrodite)	80,000	6 nuclear genes	41
Escherichia coli	25,000,000	410 genes	146
Species with indirect mutation rate estimates			
Bonobo	12,300	50 nuclear sequences	145
Chimpanzee	21,300	50 nuclear sequences	145
Gorilla	25,200	50 nuclear sequences	145
Gray whale	34,410	9 nuclear gene introns	147
Caenorhabditis remanei (separate sexes)	1,600,000	6 nuclear genes	43
Plasmodium falciparum	210,000–300,000	204 nuclear genes	148

For data from genes, synonymous site diversity for nuclear genes was used as the basis for the calculation, unless otherwise stated.

reproducing at discrete time-intervals, such as annually breeding species of birds or mammals[29–31,50–52]. The results show that N_e is usually considerably less than the number of breeding individuals present at any one time. There is, however, no satisfactory treatment of populations in which individuals reproduce more or less continuously, such as humans and many tropical species[52].

The effect of changes in population size. It is also possible to model changes over time in population size N, while otherwise retaining the Wright–Fisher model[3,4,53]. The expected coalescence time is then similar to that with constant population size, that is, approximately $2N_H$, where N_H is the harmonic mean of N over the set of generations in question (the reciprocal of the mean of the reciprocals of the values of N). This allows the use of N_H instead of N in the equation for expected neutral diversity (BOX 3). For more complex population structures, we can replace the N values for each generation by the corresponding N_e values from BOX 4, provided that the flow between different compartments equilibrates over a short timescale compared with changes in population size.

A population that has recently grown from a much smaller size, such as a population that has recovered from a bottleneck associated with colonization of a new habitat, will thus have a much lower effective size than one that has always remained at its present size, as the harmonic mean is strongly affected by the smallest values in the set[54]. There is increasingly strong evidence for such bottleneck effects in both human[55,56] and *D. melanogaster* populations[46,48,57] that have moved out of Africa.

Estimating N_e for natural and artificial populations. It is obviously of importance to have estimates of N_e, both for practical purposes, such as designing conservation or selective breeding programmes, and for interpreting data on DNA sequence variation and evolution. This can be done simply by using demographic information and substituting into equations of the type shown in BOX 4 (REFS 9,10). More recently, two different approaches that use information on genetic markers have been employed. First, N_e for a large natural population can be estimated from silent nucleotide site diversities, as diversity at equilibrium between drift and mutation depends on the product of mutation rate per nucleotide site, u, and N_e (replacing N by N_e in equation 4 in BOX 3). If the mutation rate is known, either from a direct experimental estimate or from data on DNA sequence divergence between species with known dates of separation, N_e can be estimated as $\pi/(4u)$, where π is nucleotide site diversity. Some examples are shown in TABLE 1. Second, for very small populations, such as those used in animal and plant breeding or in the captive breeding of endangered species, N_e can be estimated from observed changes between generations in the frequencies of putatively neutral variants[9,58–60].

As might be expected from the theoretical results, effective population sizes are often found to be much lower than the observed numbers of breeding individuals in both natural and artificial populations[9,10,61]. The human population, for example, is estimated from DNA sequence variability to have an N_e of 10,000 to 20,000, because of its long past history of small numbers of individuals and relatively recent expansion in size[55,62]. Larger population sizes in the past other than for extant populations have, however, sometimes been inferred from diversity estimates; for example, Atlantic whales, probably reflecting the devastating effects of whaling on their population sizes[63].

The above two genetic methods of estimating N_e can therefore yield very different results if there have been large changes in population size, because the first approach relates to the harmonic mean value of population size over the long period of time required for diversity levels to equilibrate, and the second to the present day population size. A large increase in population size, as in the case of humans, means that the N_e estimated

Box 4 | **Effective population sizes for some common situations**

Using the fast timescale approximation described in the text, formulae for N_e can be derived for various types of discrete generation populations. These provide insights into the effects of different demographic and genetic factors.

Autosomal inheritance:

$$\frac{1}{N_{eA}} \approx \frac{(1+F_{IS})}{4}\left\{\frac{1}{N_f}+\frac{1}{N_m}+\frac{(1-c)^2\,\Delta V_f}{N_f}+\frac{c^2\,\Delta V_m}{N_m}\right\} \quad (5)$$

X-linked inheritance (*Z*-linked inheritance, with female heterogamety, is described by interchanging female and male subscripts, *f* and *m*):

$$\frac{1}{N_{eX}} \approx \frac{1}{9}\left\{\frac{4\,(1+F_{IS})}{N_f}+\frac{2}{N_m}+\frac{4\,(1+F_{IS})\,(1-c)^2\,\Delta V_f}{N_f}+\frac{2c^2\,\Delta V_m}{N_m}\right\} \quad (6)$$

Y-linked inheritance (*W*-linked inheritance, with female heterogamety, is described by replacing the male subscripts, *m*, with the female subscript, *f*):

$$\frac{1}{N_{eY}} \approx \frac{2\,(1+c^2\,\Delta V_m)}{N_m} \quad (7)$$

Maternally transmitted organelles:

$$\frac{1}{N_{eC}} \approx \frac{2\,(1+(1-c)^2\,\Delta V_f)}{N_f} \quad (8)$$

Discrete generations with constant population size are assumed. N_f and N_m are the numbers of breeding females and males, respectively; c is the fraction of males among breeding individuals, that is, $c = N_m/(N_f + N_m)$; ΔV_f and ΔV_m are the excesses of the variances in offspring numbers over the Poisson values for females and males, respectively; F_{IS} is the inbreeding coefficient within the population caused by an excess of matings between relatives over random mating expectation[5,19]. Equations are taken from REF. 30.

from diversity data might be irrelevant to estimates of future changes caused by drift. Care must therefore be taken to apply estimates of N_e only to situations in which they are appropriate.

The simultaneous effects of selection and drift

Although the models outlined above indicate how N_e can be used in models of genetic drift in panmictic populations, in order to understand evolutionary processes more fully we need to include the effects of selection into the models. The effects of selection can be most easily studied by using diffusion equations[16,19,23,64].

Diffusion equations. These provide approximation for the rate of change in the probability of allele frequency q at time t. For diffusion approximations to be valid, the effects of both drift and deterministic forces must both be weak. The evolutionary process is then completely determined by the mean and variance of the change in allele frequency per generation, $M_{\delta q}$ and $V_{\delta q}$, respectively[19,23,64].

The effects of drift in situations can be modelled by $V_{\delta q} = pq/(2N_e)$, where N_e replaces N for non-Wright–Fisher populations in equation 3 in BOX 2. In this context, N_e is known as the variance effective size. Intuitively, it might seem that we can just use the expressions for N_e derived for the neutral coalescent process. However, there are situations in which this is not correct[6,65]. If the population size changes between generations, the rate of the coalescent process depends on the population size in the parental generation, whereas the change in variance depends on the size of the offspring generation. In addition, the binomial expression for $V_{\delta q}$ (equation 3 in BOX 2) is only an approximation when there is selection or when the population does not follow the Wright–Fisher model[22,66,67]. The coalescent N_e that we have used should, however, provide a good approximation to the variance N_e when all evolutionary forces are weak and the population size is constant.

Probability of fixation of a new mutation. A major conclusion from the use of diffusion equations is that the effectiveness of a deterministic force is controlled by the product of N_e and the measure of its intensity[19,23,64]. This principle is exemplified by the probability of fixation of a new mutation, denoted here by Q[8,16,17,64,68] (BOX 5). This is probably the most useful index of the effectiveness of selection versus genetic drift. For a deleterious mutation (with a selection coefficient (s) less than 0), Q is not much below the neutral value when $-N_e s \leq 0.25$; a deleterious mutation has almost no possibility of becoming fixed by drift once $-N_e s > 2$. For a favourable mutation, if $N_e s \leq 0.25$, Q behaves close to neutrally; once $N_e s > 1$, Q is close to that for an infinitely large population, that is, $Q = s(N_e/N)$.

A reduction in N_e below N reduces the efficacy of selection compared with a Wright–Fisher population of size N. This result applies to a wide variety of causes of reduced N_e, as we shall see in the next section. Given the large values of long-term N_e in TABLE 1, weak selection can therefore be very effective in evolution, as was strongly emphasized by Fisher[68]. Indeed, studies of polymorphisms at the sequence level find selection coefficients of a few multiples of $1/N_e$ for many deleterious polymorphic amino-acid variants in human and *Drosophila* populations[56,69–71]; these are sufficient to prevent them becoming fixed in the population with any significant probability. Variants at synonymous or non-coding sites are generally under much weaker selection, with selection coefficients in the order of $1/N_e$ or less[72–75]; this means that drift and mutation as well as selection have a considerable influence on the states of such sites[8,76,77]. There is increasing evidence that the rate of evolution of protein sequences is affected by differences in N_e in the way predicted by theory[11,14,15,78–82].

Determining N_e of a structured population

Having discussed the issue of how to determine the effective size of a population and considered the effects of selection in panmictic populations, the final section of this Review examines how to do this when the population is divided into geographically or genetically defined subpopulations. This is a field that has experienced rapid development in the past few years. New theoretical approaches that use fast timescale approximations have been applied to both spatial and genetic structuring of populations. There is also a growing appreciation of the fact that the genetic structuring of populations with respect to genotypes with different fitnesses implies the existence of differences in N_e values among different parts of the genome of the same species.

Heterogamety
The presence of two different sex-determining alleles or chromosomes in one of the two sexes.

Selection coefficient
(s). The effect of a mutation on fitness, relative to the fitness of wild-type individuals. With diploidy, this is measured on mutant homozygotes.

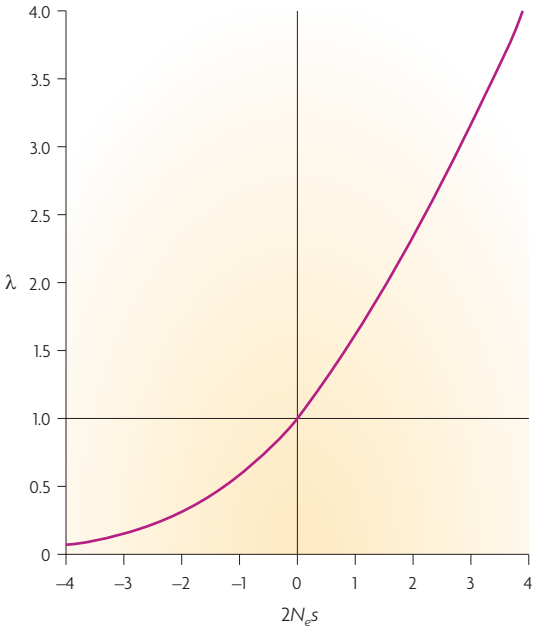
Box 5 | **Fixation probabilities**

The probability of fixation of a mutation is the chance that it will spread through the population and become fixed. In a finite population, even deleterious mutations can become fixed by drift, and favourable ones can be lost. The results of some fairly complex calculations[17,19,64] can be illustrated with the simple case of selection at a biallelic autosomal locus with semi-dominance, such that the relative fitnesses of A_1A_1, A_1A_2 and A_2A_2 are 1, 1 + 0.5s and 1 + s, respectively. s is the selection coefficient, and is negative if A_2 is deleterious and positive if it is advantageous.

If the population size is *N*, and the effective population size is N_e, the probability that a newly arisen mutation to A_2 from A_1 survives in the population and eventually replaces A_1 is given by:

$$Q \approx \frac{N_e s}{N} \frac{1}{\{1 - \exp(-2N_e s)\}} \quad (9)$$

The dependence of *Q* on $N_e s$ is illustrated in the figure. λ is the fixation probability of a semi-dominant mutation, expressed relative to the neutral value (1/2*N*). This is given by *Q* (from the equation above) divided by 1/(2*N*). This also represents the evolutionary rate of substitution of mutations with selection coefficient s, relative to the rate for neutral mutations[8].



The effects of spatial structure on neutral variation. Spatial structure was first studied by classical population genetic methods, extending the methods of BOX 2 to include the effects of geographic subdivision of a metapopulation into partially isolated, local populations[5,83–85]. More recently, the study of neutral variability in a spatially structured population has been simplified by extending the structured coalescent approach described above to a metapopulation consisting of a set (*d*) of discrete local populations (demes) that are interconnected by migration[7] or that are affected by local extinctions of demes and recolonization[7,86].

A useful result applies to the case of 'conservative' migration, that is, when migration among demes leaves their relative sizes unchanged; the mean allele frequency across demes is also unchanged[27,87,88] (the classical island and stepping stone models[83,89,90] are examples of this). Provided that all demes experience some migration events, the mean coalescence time for a pair of alleles sampled from the same deme (T_S) is given by the sum of the effective population sizes over all demes (N_{eT}), so that the mean within-deme nucleotide site diversity is the same as for a panmictic population with this effective population size. This suggests that the mean within-deme nucleotide site diversity for a species is the most appropriate measure to compare the properties of different species.

We might also be interested in describing aspects of variability such as the total amount of variability in a metapopulation, as measured by the mean pairwise nucleotide site diversity among a pair of alleles sampled at random from the metapopulation (π_T) and the corresponding mean coalescence time (T_M) corresponding to what we can call the total effective size of the metapopulation: $N_{eM} = T_M/2$. In contrast to T_S, the value of T_M is highly dependent on the details of the migration process, and can be greatly increased when migration is restricted.

For more general migration models, it is hard to derive an expression for T_M. However, when the number of demes is very large, it is approximately the same as the mean coalescence time for a pair of alleles sampled from two distinct, randomly chosen populations. Wakeley and his collaborators have shown that this large deme number approximation often yields a simple approximate general formula for T_M[7,86,91–93].

Standard tests for departures from neutral equilibrium utilize patterns of variability to detect departures from those predicted by the standard coalescent model; tests of this kind are widely used in studies of DNA sequence variation[7,18]. If such departures are detected, the occurrence of selection or of demographic events, such as changes in population size, is implied. In the case of a metapopulation with a large number of demes, if a sample of *k* alleles is taken by sampling each allele from a separate population, these obey the same coalescent process as alleles sampled from a panmictic population, described in BOX 3. Tests of this kind for a metapopulation are thus best carried out by sampling only one allele from a given population. Similar results also apply to measures of linkage disequilibrium in spatially structured populations. If a single haplotype is sampled from each local population studied, under conservative migration the expected level of linkage disequilibrium between a pair of sites with recombination frequency *r* is controlled by $4N_{eT}r$ in the same way as by $4N_e r$ in the case of a panmictic population[7,94].

The effects of spatial structure on variants under selection. We can also ask how to determine the fixation probability of a mutation under selection in a

Metapopulation
A population consisting of a set of spatially separate local populations.

metapopulation. With semi-dominant or haploid selection (BOX 5), the fixation probability of a new mutation in a structured population consisting of a set of Wright–Fisher populations connected by conservative migration is determined by the product of the selection coefficient and N_{eT} in the same way as by $N_e s$ in a single, panmictic population[87,95,96]. Recent work suggests that an approximate diffusion equation can be derived for more general selection and migration models, using the large deme number approximation just discussed[97–100]. This is useful, as it implies that spatial structure might not have much effect on the fixation probabilities of weakly selected mutations, which are likely to have intermediate dominance coefficients[101], so that the standard models of molecular evolution apply even to highly subdivided populations. Predictions of the effects of differences in effective population sizes on rates of sequence evolution for species[79,81] should therefore use estimates of N_e based on mean within-population diversities.

With dominance, however, population structure can cause important departures from the panmictic results[98,102,103]. Fixation probabilities are reduced for recessive or partly recessive deleterious mutations, and increased for recessive or partly recessive advantageous mutations, relative to the value for a panmictic population with an effective size of N_{eS}. The reverse is true for dominant or partially dominant mutations. The overall effect of population subdivision on the rate of evolution thus depends on both the level of dominance of new mutations, and on the extent to which advantageous or deleterious mutations contribute.

The effects of genetic structure. Investigations of DNA sequence variability have shown that presumptively neutral diversity is not constant across the genome. For example, silent site DNA sequence variability is elevated in the neighbourhood of the highly polymorphic major histocompatibility (MHC) loci of mammals[104], and of the self-incompatibility (SI) loci of plants[105,106]. Conversely, in *D. melanogaster*[14,107,108], humans[109] and some plant species[110], silent site variability correlates positively with the local rate of genetic recombination, and is extremely low in regions where there is little or no recombination. In addition, as already noted, species or populations with high levels of inbreeding often exhibit reduced levels of variation compared with outcrossing relatives[40,41,110], to a much greater extent than the two-fold reduction predicted on a purely neutral model (see above).

The most likely explanation for these patterns, with the possible exception of human populations[109,111], is that N_e is affected by selection occurring at closely linked sites or, in inbreeding populations, sites that rarely recombine with physically distant targets of selection because of the reduced evolutionary effectiveness of recombination in a highly homozygous genome[28]. The concepts and methods used to study the effects of spatial structuring of populations can be used to understand stable genetic structure, whereby different genotypes are maintained in the population, either by long-term balancing selection, or by recurrent mutation to deleterious alleles.

The effects of balancing selection. Long-term balancing selection refers to the situation in which two or more variants at a locus are maintained in the population by forms of selection such as heterozygote advantage or frequency-dependent selection, for much longer than would be expected under neutrality. There is clear evidence for such selection in the cases of the MHC and SI loci mentioned above. What is the effect of balancing selection on neutral variability at linked sites? Consider an autosomal site with two variants, A_1 and A_2, maintained by balancing selection in a randomly mating population with effective population size N_e. A neutral site recombines with the A site at rate r. The flow of neutral variants by recombination between the haplotypes carrying A_1 and A_2 is similar to conservative migration between demes[25,28,112]. High equilibrium levels of differentiation between A_1 and A_2 haplotypes are expected at closely linked neutral sites, for which $N_e r$ is much greater than 1, that is, in the situation equivalent to low migration. This is reflected in a local elevation in the effective population size, equivalent to the elevation of N_{eM} over N_{eT}, producing a local peak of diversity close to the target of balancing selection, as is observed in the cases mentioned above. Coalescence times in this case can be much greater than the time during which the species has existed[113]. Neutral variants that distinguish the selected alleles might then persist across the species boundaries. This is called *trans*-specific polymorphism, and is seen, for example, in the SI polymorphisms of plants[114].

This suggests that polymorphisms maintained by long-term balancing selection could be discovered by scanning the genome for local peaks of silent site diversity and/or polymorphisms that are shared between species. Such scans using the human and chimpanzee genomes have so far been largely negative, suggesting that there are rather few cases of long-term balancing selection[115,116], although some convincing examples have been discovered[117].

Background selection and other Hill–Robertson effects. Another important type of genetic structuring in populations is caused by deleterious alleles maintained by recurrent mutation[118]. These reduce neutral diversity at linked sites because the elimination of a deleterious mutation carried on a particular chromosome also lowers the frequencies of any associated neutral or nearly neutral variants. This process of background selection is one example of the general process known as the Hill–Robertson effect; see REF. 13 for a recent review. This can be understood in terms of N_e as follows. Selection creates heritable variance in fitness among individuals, which reduces N_e (REF. 119). A site that is linked to a selected variant experiences an especially marked reduction in its N_e, because close linkage maintains the effects for many generations[120,121]. In addition to reducing levels of variability, this reduction in N_e impairs the efficacy of selection (see the discussion of fixation probability above). This probably accounts for the observation that the level of adaptation at the sequence level, as well as sequence diversity, often seems to be reduced in low recombination regions of the *Drosophila* genome[14,15,82,122–124].

Semi-dominant or haploid selection
With a diploid species, semi-dominant selection occurs when the fitness of the heterozygote for a pair of alleles is intermediate between that of the two homozygotes; haploid selection applies to haploid species, and is twice as effective as semi-dominant selection with the same selection coefficient.

Dominance coefficient
(h). Measures the extent to which the fitness of a heterozygote carrier of a mutation is affected, relative to the effect of the mutation on homozygous carriers.

Heterozygote advantage
The situation in which the fitness of a heterozygote for a pair of alleles is greater than that of either homozygote. This maintains polymorphism.

Frequency-dependent selection
Situations in which the fitnesses of genotypes are affected by their frequencies in the population. Polymorphism is promoted when fitness declines with frequency.

Background selection
The process by which selection against deleterious mutations also eliminates neutral or weakly selected variants at closely linked sites in the genome.

Hill–Robertson effect
The effect of selection on variation at one location in the genome and on evolution at other, genetically linked sites.

Selective sweep
The process by which a new favourable mutation becomes fixed so quickly that variants that are closely linked to it, and that are present in the chromosome on which the mutation arose, also become fixed.

Another important example of a Hill–Robertson effect is the effect on linked sites of the spread of a selectively favourable mutation. This was called a hitchhiking event by Maynard Smith and Haigh[125], and is now often referred to as a selective sweep[126]. The expected reduction in N_e caused by a single selective sweep is very sensitive to the ratio r/s, where s is the selective advantage to the favourable mutation and r is the frequency of recombination between this mutation and the site whose N_e is being considered, and the reduction in N_e is small unless r/s is much lower than 1 (REFS 125,127). This effect is transient, in the absence of further sweeps in the same region, and resembles the effect of a population bottleneck, as variability will start to recover once the favourable mutation has become fixed[128,129].

The selective sweep model can be extended to allow a steady rate of substitution of favourable variants, at sites scattered randomly over the genome[130–133]. Using empirical estimates of the proportion of amino-acid divergence between species that is due to positive selection, this model provides a good fit to data on sequence variability in *D. melanogaster*[134]. N_e for a typical locus seems to be reduced by a few per cent as a result of ongoing adaptive substitutions of amino-acid mutations. The abundance of weakly selected deleterious amino-acid variants in *Drosophila* populations seems to be sufficiently high for background selection to further reduce N_e for genes with normal levels of recombination by a few per cent[135].

Hill–Robertson effects mean that N_e for a particular location in the genome is highly dependent on its recombinational environment, and that no region is entirely free of the effects of selection at nearby sites, even in genomic regions with normal levels of recombination. Large genomic regions that lack recombination, such as the Y chromosome and asexual or highly self-fertilizing species, are expected to experience the most extreme reductions in N_e (REFS 118,136). This probably accounts for the evolutionary degeneration of Y chromosomes[123,124,137], and the lack of evolutionary success of most asexual and highly inbreeding species[138,139].

Conclusions

From a modest beginning, when Sewall Wright dealt with the process of genetic drift in a population with two sexes, the concept of effective population size has been extended to the status of a unifying principle that encompasses the action of drift in almost any imaginable evolutionary scenario. Over that time, there has been a considerable shift in theoretical methodology, with current formulations using the powerful technology of coalescent theory, and approximations based on separating drift into processes acting on different timescales.

One important advance is that we now have a much clearer appreciation of the role of selection in shaping the effective population size at genetically linked sites than we did 10 years ago. Already, we can be fairly sure that no nucleotide in the compact genome of an organism such as *D. melanogaster* is evolving entirely free of the effects of selection on its effective population size; it will be of great interest to see whether this applies to species with much larger genomes, such as humans, when we make use of the avalanche of data on DNA sequence variation and evolution that will be produced by new sequencing technologies.

However, it is important to note that N_e has some limitations as a tool for understanding patterns of evolution and variation. It is extremely useful for describing expected levels of genetic diversity, and for evaluating the effects of different factors on the efficiency of selection. But certain aspects of genetic variability, such as the distribution of frequencies of individual nucleotide variants across different sites, cannot simply be described in terms of N_e. A given reduction in variability caused by a population bottleneck, a selective sweep or background selection might well be associated with different variant frequency distributions, and so cannot be described by a simple reduction in N_e (REFS 128,129,136,140). Models that describe all aspects of the data are needed in these cases; the challenge is to extend existing models to include increasingly refined estimates of parameters, such as the incidence of selective sweeps and the distribution of selection coefficients against weakly deleterious mutations, into models that can be tested against the data.

1. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931). **A classic founding paper of theoretical population genetics, which introduces the concept of effective population size.**
2. Wright, S. Inbreeding and homozygosis. *Proc. Natl Acad. Sci. USA* **19**, 411–420 (1933).
3. Wright, S. Size of population and breeding structure in relation to evolution. *Science* **87**, 430–431 (1938).
4. Wright, S. *Statistical Genetics in Relation to Evolution (Actualites Scientifiques et Industrielles, 802: Exposés de Biométrie et de la Statistique Biologique. XIII)* 5–64 (Hermann et Cie, Paris, 1939).
5. Wright, S. *Evolution and the Genetics of Populations* Vol. 2 (Univ. Chicago Press, Chicago, Illinois, 1969).
6. Crow, J. F. in *Statistics and Mathematics in Biology* (eds Kempthorne, O., Bancroft, T. A., Gowen, J. W. & Lush, J. L.) 543–556 (Iowa State Univ. Press, Ames, Iowa, 1954).
7. Wakeley, J. *Coalescent Theory. An Introduction* (Ben Roberts, Greenwood Village, Colorado, 2008). **A broad-ranging treatment of coalescent theory and its use in the interpretation of data on DNA sequence variability.**
8. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983). **A somewhat partisan account of how population genetics theory can be used to interpret data on molecular evolution. It provides an excellent summary of the use of the concept of effective population size, and describes results from the use of diffusion equations.**
9. Crow, J. F. & Morton, N. E. Measurement of gene-frequency drift in small populations. *Evolution* **9**, 202–214 (1955).
10. Frankham, R. Effective population size/adult population size ratios in wildlife: a review. *Genet. Res.* **66**, 95–107 (1995). **This reviews evidence for much lower effective population sizes than census sizes in natural populations.**
11. Eyre-Walker, A., Keightley, P. D., Smith, N. G. C. & Gaffney, D. Quantifying the slightly deleterious mutation model of molecular evolution. *Mol. Biol. Evol.* **19**, 2142–2149 (2002).
12. Vicoso, B. & Charlesworth, B. Evolution on the X chromosome: unusual patterns and processes. *Nature Rev. Genet.* **7**, 645–653 (2006).
13. Comeron, J. M., Williford, A. & Kliman, R. M. The Hill–Robertson effect: evolutionary consequences of weak selection in finite populations. *Heredity* **100**, 19–31 (2008). **A review of the theory and data on the effects of selection at one genomic site on variability and evolution at other sites in the genome.**
14. Presgraves, D. Recombination enhances protein adaptation in *Drosophila melanogaster*. *Curr. Biol.* **15**, 1651–1656 (2005). **Reviews data supporting a correlation between recombination rate and neutral or nearly neutral variability in *D. melanogaster*, and presents evidence for reduced efficacy of selection when recombination rates are low.**
15. Larracuente, A. M. *et al.* Evolution of protein-coding genes in *Drosophila*. *Trends Genet.* **24**, 114–123 (2008).
16. Fisher, R. A. On the dominance ratio. *Proc. Roy. Soc. Edinb.* **52**, 312–341 (1922).
17. Fisher, R. A. The distribution of gene ratios for rare mutations. *Proc. Roy. Soc. Edinb.* **50**, 205–220 (1930).
18. Hein, J., Schierup, M. H. & Wiuf, C. *Gene Genealogies, Variation and Evolution* (Oxford Univ. Press, Oxford, 2005).

19. Crow, J. F. & Kimura, M. *An Introduction to Population Genetics Theory* (Harper and Row, New York, 1970).
20. Caballero, A. Developments in the prediction of effective population size. *Heredity* **73**, 657–679 (1994).
21. Wang, J. L. & Caballero, A. Developments in predicting the effective size of subdivided populations. *Heredity* **82**, 212–226 (1999).
22. Nagylaki, T. *Introduction to Theoretical Population Genetics* (Springer, Berlin, 1992).
23. Ewens, W. J. *Mathematical Population Genetics. Theoretical Introduction* Vol. 1 (Springer, New York, 2004).
24. Vitalis, R. Sex-specific genetic differentiation and coalescence times: estimating sex-biased dispersal rates. *Mol. Ecol.* **11**, 125–138 (2002).
25. Hudson, R. R. & Kaplan, N. L. The coalescent process in models with selection and recombination. *Genetics* **120**, 831–840 (1988).
26. Hey, J. A multi-dimensional coalescent process applied to multiallelic selection models and migration models. *Theor. Pop. Biol.* **39**, 30–48 (1991).
27. Nagylaki, T. The strong-migration limit in geographically structured populations. *J. Math. Biol.* **9**, 101–114 (1980).
28. Nordborg, M. Structured coalescent processes on different time scales. *Genetics* **146**, 1501–1514 (1997).
29. Rousset, F. Genetic differentiation in populations with different classes of individuals. *Theor. Pop. Biol.* **55**, 297–308 (1999).
30. Laporte, V. & Charlesworth, B. Effective population size and population subdivision in demographically structured populations. *Genetics* **162**, 501–519 (2002).
This uses the fast timescale approximation to provide a general framework for deriving formulae for effective population size.
31. Nordborg, M. & Krone, S. M. in *Modern Developments in Population Genetics. The Legacy of Gustave Malécot.* (eds Slatkin, M. & Veuille, M.) 194–232 (Oxford Univ. Press, Oxford, 2002).
32. Brotherstone, S. & Goddard. Artificial selection and maintenance of genetic variance in the global dairy cow population. *Phil. Trans. R. Soc. B* **360**, 1479–1148 (2005).
33. Frankham, R., Ballou, J. D. & Briscoe, D. A. *Introduction to Conservation Genetics* (Cambridge Univ. Press, Cambridge 2002).
34. Andersson, M. *Sexual Selection* (Princeton Univ. Press, Princeton, New Jersey, 1994).
35. Nunney, L. The influence of age structure and fecundity on effective population size. *Proc. Roy. Soc. Lond. B* **246**, 71–76 (1991).
36. Nunney, L. The influence of mating system and overlapping generations on effective population size. *Evolution* **47**, 1329–2341 (1993).
37. Wright, S. The genetical structure of populations. *Ann. Eugen.* **15**, 323–354 (1951).
38. Nordborg, M. & Donnelly, P. The coalescent process with selfing. *Genetics* **146**, 1185–1195 (1997).
39. Pollak, E. On the theory of partially inbreeding populations. I. Partial selfing. *Genetics* **117**, 353–360 (1987).
40. Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**, 1289–1299 (2005).
41. Cutter, A. D. Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer *Caenorhabditis elegans*. *Genetics* **172**, 171–184 (2005).
42. Wright, S. J. *et al.* Testing for effects of recombination rate on nucleotide diversity in natural populations of *Arabidopsis lyrata*. *Genetics* **174**, 1421–1430 (2006).
43. Cutter, A., Baird, S. E. & Charlesworth, D. High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*. *Genetics* **174**, 901–913 (2006).
44. Moriyama, E. N. & Powell, J. R. Intraspecific nuclear DNA variation in *Drosophila*. *Mol. Biol. Evol.* **13**, 261–277 (1996).
45. Andolfatto, P. Contrasting patterns of X-linked and autosomal nucleotide variation in African and non-African populations of *Drosophila melanogaster* and *D. simulans*. *Mol. Biol. Evol.* **18**, 279–290 (2001).
46. Hutter, S., Li, H. P., Beisswanger, S., De Lorenzo, D. & Stephan, W. Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosome-wide nucleotide polymorphism data. *Genetics* **177**, 469–480 (2007).
47. Singh, N. D., Macpherson, J. M., Jensen, J. D. & Petrov, D. A. Similar levels of X-linked and autosomal nucleotide variation in African and non-African populations of *Drosophila melanogaster*. *BMC Evol. Biol.* **7**, 202 (2007).
48. Pool, J. E. & Nielsen, R. The impact of founder events on chromosomal variability in multiply mating species. *Mol. Biol. Evol.* **25**, 1728–1736 (2008).
49. Sundström, H., Webster, M. T. & Ellegren, H. Reduced variation on the chicken Z chromosome. *Genetics* **167**, 377–385 (2004).
50. Felsenstein, J. Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics* **68**, 581–597 (1971).
51. Charlesworth, B. *Evolution in Age-structured Populations* 2nd edn (Cambridge Univ. Press, Cambridge 1994).
52. Charlesworth, B. The effect of life-history and mode of inheritance on neutral genetic variability. *Genet. Res.* **77**, 153–166 (2001).
53. Slatkin, M. & Hudson, R. R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **143**, 579–587 (1991).
54. Wright, S. Breeding structure of species in relation to speciation. *Am. Nat.* **74**, 232–248 (1940).
55. Voight, B. F. *et al.* Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl Acad. Sci. USA* **102**, 18508–18513 (2005).
56. Boyko, A. *et al.* Assessing the evolutionary impact of amino-acid mutations in the human genome. *PLoS Genet.* **5**, e1000083 (2008).
57. Haddrill, P. R., Thornton, K. R., Charlesworth, B. & Andolfatto, P. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Res.* **15**, 790–799 (2005).
58. Wang, J. Estimation of effective population sizes from data on genetic markers. *Phil. Trans. R. Soc. B* **360**, 1395–1409 (2005).
A review of methods for using information on genetic variants in populations to estimate effective population size.
59. Waples, R. S. & Yokota, M. Temporal estimates of effective population size in species with overlapping generations. *Genetics* **175**, 219–233 (2007).
60. Jorde, P. E. & Ryman, N. Unbiased estimator for genetic drift and effective population size. *Genetics* **177**, 927–935 (2007).
61. Coyer, J. A., Hoarau, G., Sjotun, K. & Olsen, J. L. Being abundant is not enough; a decrease in effective size over eight generations in a Norwegian population of the seaweed, *Fucus serratus*. *Biol. Lett.* **4**, 755–757 (2008).
62. Wall, J. D. & Przeworski, M. When did the human population size start increasing? *Genetics* **155**, 1865–1874 (2000).
63. Roman, J. & Palumbi, S. R. Whales before whaling in the North Atlantic. *Science* **301**, 508–510 (2003).
64. Kimura, M. Diffusion models in population genetics. *J. App. Prob.* **1**, 177–223 (1964).
65. Kimura, M. & Crow, J. F. The measurement of effective population size. *Evolution* **17**, 279–288 (1963).
66. Ethier, S. & Nagylaki, T. Diffusion approximations of Markov chains with two time scales and applications to population genetics. *Adv. Appl. Prob.* **12**, 14–49 (1980).
67. Nagylaki, T. Models and approximations for random genetic drift. *Theor. Pop. Biol.* **37**, 192–212 (1990).
68. Fisher, R. A. *The Genetical Theory of Natural Selection* (Oxford Univ. Press, Oxford, 1930; Variorum Edn, Oxford Univ. Press, 1999).
Fisher's summing up of his fundamentally important contributions to population genetics theory.
69. Eyre-Walker, A., Woolfit, M. & Phelps, T. The distribution of fitness effects of new deleterious amino-acid mutations in humans. *Genetics* **173**, 891–900 (2006).
70. Keightley, P. D. & Eyre-Walker, A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**, 2251–2261 (2007).
71. Loewe, L. & Charlesworth, B. Inferring the distribution of mutational effects on fitness in *Drosophila*. *Biol. Lett.* **2**, 426–430 (2006).
72. Maside, X., Weishan Lee, A. & Charlesworth, B. Selection on codon usage in *Drosophila americana*. *Curr. Biol.* **14**, 150–154 (2004).
73. Comeron, J. M. & Guthrie, T. B. Intragenic Hill–Robertson interference influences selection on synonymous mutations in *Drosophila*. *Mol. Biol. Evol.* **22**, 2519–2530 (2005).
74. Comeron, J. M. Weak selection and recent mutational changes influence polymorphic synonymous mutations in humans. *Proc. Natl Acad. Sci. USA* **103**, 6940–6945 (2006).
75. Cutter, A. D. & Charlesworth, B. Selection intensity on preferred codons correlates with overall codon usage bias in *Caenorhabditis remanei*. *Curr. Biol.* **16**, 2053–2057 (2006).
76. Li, W.-H. Models of nearly neutral mutations with particular implications for non-random usage of synonymous codons. *J. Mol. Evol.* **24**, 337–345 (1987).
77. Bulmer, M. G. The selection–mutation–drift theory of synonomous codon usage. *Genetics* **129**, 897–907 (1991).
78. Paland, S. & Lynch, M. Transitions to asexuality result in excess amino-acid substitutions. *Science* **311**, 990–992 (2006).
79. Woolfit, M. & Bromham, L. Population size and molecular evolution on islands. *Proc. R. Soc. B* **272**, 2277–2282 (2005).
80. Fry, A. J. & Wernegreen, J. J. The roles of positive and negative selection in the molecular evolution of insect endosymbionts. *Gene* **355**, 1–10 (2005).
81. Charlesworth, J. & Eyre-Walker, A. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. *Proc. Natl Acad. Sci. USA* **104**, 16992–16997 (2007).
82. Haddrill, P. R., Halligan, D. L., Tomaras, D. & Charlesworth, B. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. *Genome Biol.* **8**, R18 (2007).
83. Wright, S. Isolation by distance. *Genetics* **28**, 114–138 (1943).
84. Malécot, G. *The Mathematics of Heredity* (W. H. Freeman, San Francisco, California, 1969).
85. Maruyama, T. *Stochastic Problems in Population Genetics. Lectures in Biomathematics 17* (Springer, Berlin, 1977).
86. Wakeley, J. & Aliacar, N. Gene genealogies in a metapopulation. *Genetics* **159**, 893–905 (2001).
87. Nagylaki, T. Geographical invariance in population genetics. *J. Theor. Biol.* **99**, 159–172 (1982).
88. Nagylaki, T. The expected number of heterozygous sites in a subdivided population. *Genetics* **149**, 1599–1604 (1998).
89. Kimura, M. 'Stepping stone' model of a population. *Ann. Rep. Nat. Inst. Genet.* **3**, 63–65 (1953).
90. Wilkinson-Herbots, H. M. Genealogy and subpopulation differentiation under various models of population structure. *J. Math. Biol.* **37**, 535–585 (1998).
91. Wakeley, J. Nonequilibrium migration in human history. *Genetics* **153**, 1863–1871 (1999).
This explains the large deme number approximation, and applies it to problems in human population genetics.
92. Wakeley, J. The coalescent in an island model of population subdivision with variation among demes. *Theor. Pop. Biol.* **59**, 133–144 (2001).
93. Matsen, F. A. & Wakeley, J. Convergence to the island-model coalescent process in populations with restricted migration. *Genetics* **172**, 701–708 (2006).
94. Wakeley, J. & Lessard, S. Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. *Genetics* **164**, 1043–1053 (2003).
95. Maruyama, T. On the fixation probabilities of mutant genes in a subdivided population. *Genet. Res.* **15**, 221–226 (1970).
96. Maruyama, T. A simple proof that certain quantitities are independent of the geographic structure of population. *Theor. Pop. Biol.* **5**, 148–154 (1974).
97. Cherry, J. L. & Wakeley, J. A diffusion approximation for selection and drift in a subdivided population. *Genetics* **163**, 421–428 (2003).
98. Cherry, J. L. Selection in a subdivided population with dominance or local frequency dependence. *Genetics* **163**, 1511–1518 (2003).
99. Cherry, J. L. Selection in a subdivivided population with local extinction and recolonization. *Genetics* **164**, 789–779 (2003).
100. Whitlock, M. C. Fixation probability and time in subdivided populations. *Genetics* **164**, 767–779 (2003).
101. Garcia-Dorado, A. & Caballero, A. On the average coefficient of dominance of deleterious spontaneous mutations. *Genetics* **155**, 1991–2001 (2000).

102. Whitlock, M. C. Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection. *Evolution* **54**, 1855–1861 (2000).
103. Roze, D. & Rousset, F. Selection and drift in subdivided populations: a straightforward method for deriving diffusion approximations and applications involving dominance, selfing and local extinctions. *Genetics* **165**, 2153–2166 (2003).
104. Shiina, T. *et al.* Rapid evolution of major histocompatibility complex class I genes in primates generates new disease alleles in humans via hitchhiking diversity. *Genetics* **173**, 1555–1570 (2006).
105. Richman, A. D., Uyenoyama, M. K. & Kohn, J. R. Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. *Science* **273**, 1212–1216 (1996).
106. Kamau, E., Charlesworth, B. & Charlesworth, D. Linkage disequilibrium and recombination rate estimates in the self-incompatibility region of *Arabidopsis lyrata*. *Genetics* **176**, 2357–2369 (2007).
107. Begun, D. J. & Aquadro, C. F. Levels of naturally occurring DNA polymorphism correlate with recombination rate in *Drosophila melanogaster*. *Nature* **356**, 519–520 (1992).
108. Shapiro, J. A. *et al.* Adaptive genic evolution in the *Drosophila* genomes. *Proc. Natl Acad. Sci. USA* **104**, 2271–2276 (2007).
109. Hellmann, I., Ebersberger, I., Ptak, S. E., Paabo, S. & Przeworski, M. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72**, 1527–1535 (2003).
110. Roselius, K., Stephan, W. & Städler, T. The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics* **171**, 753–763 (2005).
111. Spencer, C. C. A. *et al.* The influence of recombination on human genetic diversity. *PLoS Genet.* **2**, 1375–1385 (2006).
112. Hudson, R. R. Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7**, 1–45 (1990).
113. Wiuf, C., Zhao, K., Innan, H. & Nordborg, M. The probability and chromosomal extent of *trans*-specific polymorphism. *Genetics* **168**, 2363–2372 (2004).
114. Charlesworth, D., Kamau, E., Hagenblad, J. & Tang, C. *Trans*-specificity at loci near the self-incompatibility loci in Arabidopsis. *Genetics* **172**, 2699–2704 (2006).
115. Asthana, S., Schmidt, S. & Sunyaev, S. A limited role for balancing selection. *Trends Genet.* **21**, 30–32 (2005).
116. Bubb, K. L. *et al.* Scan of human genome reveals no new loci under ancient balancing selection. *Genetics* **173**, 2165–2177 (2006).
117. Baysal, B. E., Lawrence, E. C. & Ferrell, R. E. Sequence variation in human succinate dehydrogenase genes: evidence for long-term balancing selection on *SDHA*. *BMC Biol.* **5**, 12 (2007).
118. Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
119. Robertson, A. Inbreeding in artificial selection programmes. *Genet. Res.* **2**, 189–194 (1961).
120. Santiago, E. & Caballero, A. Effective size of populations under selection. *Genetics* **139**, 1013–1030 (1995).
121. Santiago, E. & Caballero, A. Effective size and polymorphism of linked neutral loci in populations under selection. *Genetics* **149**, 2105–2117 (1998).
122. Marais, G. & Piganeau, G. Hill–Robertson interference is a minor determinant of variations in codon bias across *Drosophila melanogaster* and *Caenorhabditis elegans* genomes. *Mol. Biol. Evol.* **19**, 1399–1406 (2002).
123. Bartolomé, C. & Charlesworth, B. Evolution of amino-acid sequences and codon usage on the *Drosophila miranda* neo-sex chromosomes. *Genetics* **174**, 2033–2044 (2006).
124. Bachtrog, D., Hom, E., Wong, K. M., Maside, X. & De Jong, P. Genomic degradation of a young Y chromosome in *Drosophila miranda*. *Genome Biol.* **9**, R30 (2008).
125. Maynard Smith, J. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).
126. Berry, A. J., Ajioka, J. W. & Kreitman, M. Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. *Genetics* **129**, 1111–1117 (1991).
127. Barton, N. H. Genetic hitchhiking. *Phil. Trans. R. Soc. B* **355**, 1553–1562 (2000).
128. Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. & Stephan, W. The hitchiking effect on the site frequency spectrum of DNA polymorphism. *Genetics* **140**, 783–796 (1995).
129. Simonsen, K. L., Churchill, G. A. & Aquadro, C. F. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429 (1995).
130. Stephan, W., Wiehe, T. H. E. & Lenz, M. W. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theor. Pop. Biol.* **41**, 237–254 (1992).
131. Stephan, W. An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. *Mol. Biol. Evol.* **12**, 959–962 (1995).
132. Kim, Y. Effect of strong directional selection on weakly selected mutations at linked sites: implication for synonymous codon usage. *Mol. Biol. Evol.* **21**, 286–294 (2004).
133. Gillespie, J. H. Genetic drift in an infinite population: the pseudohitchiking model. *Genetics* **155**, 909–919 (2000).
134. Andolfatto, P. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res.* **17**, 1755–1762 (2007).
135. Loewe, L. & Charlesworth, B. Background selection in single genes may explain patterns of codon bias. *Genetics* **175**, 1381–1393 (2007).
136. Kaiser, V. B. & Charlesworth, B. The effects of deleterious mutations on evolution in non-recombining genomes. *Trends Genet.* (in the press).
137. Charlesworth, D., Charlesworth, B. & Marais, G. Steps in the evolution of heteromorphic sex chromosomes. *Heredity* **95**, 118–128 (2005).
138. Normark, B. B., Judson, O. P. & Moran, N. A. Genomic signatures of ancient asexual lineages. *Biol. J. Linn. Soc.* **79**, 69–84 (2003).
139. Barrett, S. C. H. The evolution of plant sexual diversity. *Nature Rev. Genet.* **3**, 274–284 (2002).
140. Gordo, I., Navarro, A. & Charlesworth, B. Muller's ratchet and the pattern of variation at a neutral locus. *Genetics* **161**, 835–848 (2002).
141. Cotterman, C. W. *A calculus for statistico-genetics.* Thesis, Ohio State Univ. (1940).
142. Malécot, G. *Les Mathématiques de l'Hérédité* (Masson, Paris, 1948).
143. Kingman, J. F. C. On the genealogy of large populations. *J. Appl. Prob.* **19A**, 27–43 (1982).
144. Kimura, M. Theoretical foundations of population genetics at the molecular level. *Theor. Pop. Biol.* **2**, 174–208 (1971).
145. Yu, N., Jensen-Seaman, M. I., Chemnick, L., Ryder, O. & Li, W. H. Nucleotide diversity in gorillas. *Genetics* **166**, 1375–1383 (2004).
146. Charlesworth, J. & Eyre-Walker, A. The rate of adaptive evolution in enteric bacteria. *Mol. Biol. Evol.* **23**, 1348–1356 (2006).
147. Alter, S. E., Rynes, E. & Palumbi, S. R. DNA evidence for historic population size and past ecosystem impacts of gray whales. *Proc. Natl Acad. Sci. USA* **104**, 15162–15167 (2007).
148. Mu, J. *et al.* Chromosome-wide SNPs reveal an ancient origin for *Plasmodium falciparum*. *Nature* **418**, 323–326 (2002).

Acknowledgements
B.C. thanks the Royal Society for support from 1997–2007.

FURTHER INFORMATION
Brian Charlesworth's homepage:
http://www.biology.ed.ac.uk/research/institutes/evolution/homepage.php?id=bcharlesworth

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

OPINION

The consequences of asynapsis for mammalian meiosis

Paul S. Burgoyne, Shantha K. Mahadevaiah and James M. A. Turner

Abstract | During mammalian meiosis, synapsis of paternal and maternal chromosomes and the generation of DNA breaks are needed to allow reshuffling of parental genes. In mammals errors in synapsis are associated with a male-biased meiotic impairment, which has been attributed to a response to persisting DNA double-stranded breaks in the asynapsed chromosome segments. Recently it was discovered that the chromatin of asynapsed chromosome segments is transcriptionally silenced, providing new insights into the connection between asynapsis and meiotic impairment.

To produce haploid gametes, germ cells undergo a specialized division cycle, meiosis (FIG. 1), which reduces the number of chromosomes from two sets, one maternally and one paternally derived, to a single set that is a mixture of the parental genomes. The reshuffling (recombination) of the maternal and paternal genomes necessitates extensive self-inflicted DNA damage in the form of DNA double-stranded breaks (DSBs). In most organisms these DSBs are crucial for initiating the intimate pairing of the parental chromosomes, termed synapsis, which in turn facilitates their subsequent repair. During repair some DSBs are processed to form links called chiasmata between the maternal and paternal chromosomes that have a crucial role during subsequent chromosome segregation at the first meiotic division.

It has been known for decades that errors in chromosome synapsis, that is, asynapsis, are associated with impaired fertility and that in mammals males are more severely affected than females[1]. Chromosomal anomalies associated with asynapsis are found in approximately 3% of infertile men[2]. Targeted mutation of meiotic genes has generated many new mouse models with chromosome asynapsis and male-biased sterility[3]. The association between asynapsis and sterility has been attributed to the operation of checkpoints that monitor steps in the meiotic process and that arrest or eliminate cells that have 'got it wrong'; the male–female difference in fertility impairment has been attributed to less efficient checkpoint function during female meiosis[4–6]. Despite the impairment of fertility, these checkpoints are beneficial because they substantially reduce the frequency of unbalanced gametes that generate chromosomally unbalanced conceptions. Indeed, less efficient checkpoint function in women is implicated in the high rate of aneuploidy in human pregnancies, particularly as women get older, that leads to miscarriages or the birth of individuals with chromosomal anomalies, as in Down's syndrome (also known as trisomy 21)[7]. The number of proposed meiotic checkpoints is constantly increasing, but it is checkpoint recognition of unrepaired DSBs in asynapsed chromosome regions that has been most widely invoked to explain the link between asynapsis and meiotic failure in mammals.

Recently it was established that one response to asynapsis in the mouse is the transcriptional silencing of asynapsed chromosomes or chromosome regions[8–10], and this is clearly also the case in humans[11,12]. Here we integrate these new findings into an overview of the consequences of asynapsis in mammals, as not only do they point to transcriptional silencing of crucial meiotic and post-meiotic genes as an important component of asynapsis-related fertility impairment, but they have also identified a substantive male-specific cause.

Asynapsis is linked to meiotic impairment

Because efficient synapsis of homologous chromosomes in mammals is dependent on the formation and processing of DSBs, mutations that interfere with these processes disrupt synapsis; much of the discussion of the consequences of asynapsis in mammals has focused on studies of mice with such mutations[5,13,14]. However, there are numerous chromosomally variant mouse models in which chromosome segments or whole chromosomes fail to achieve homologous synapsis, either because no matching segment is available, or because chromosomal rearrangements impede their coming together[1]. These models are proving invaluable in unravelling the complexities of the consequences of asynapsis.

Irrespective of the underlying cause, asynapsis in males is almost always associated with spermatocyte losses that are due to apoptosis during the pachytene stage of meiotic prophase and/or at the metaphase stage of the first meiotic division, with consequent sub-fertility or sterility[1,15,16]. Although female fertility might often seem to be normal, reproductive lifespan is curtailed as a consequence of oocyte depletion that is already present during the first postnatal week[17–20]. With high levels of asynapsis, sterility is seen in both sexes[21–29].

In the following sections we first discuss two ways by which asynapsis has been proposed to lead to pachytene spermatocyte apoptosis, and then explain how meiotic silencing might cause meiotic or post-meiotic losses. We end by considering the consequences of asynapsis for female meiosis and why female fertility is less severely compromised.

A pachytene response to DNA breaks

The mitotic G2/M checkpoint. DSBs are extremely hazardous lesions and it is imperative that cell division does not occur

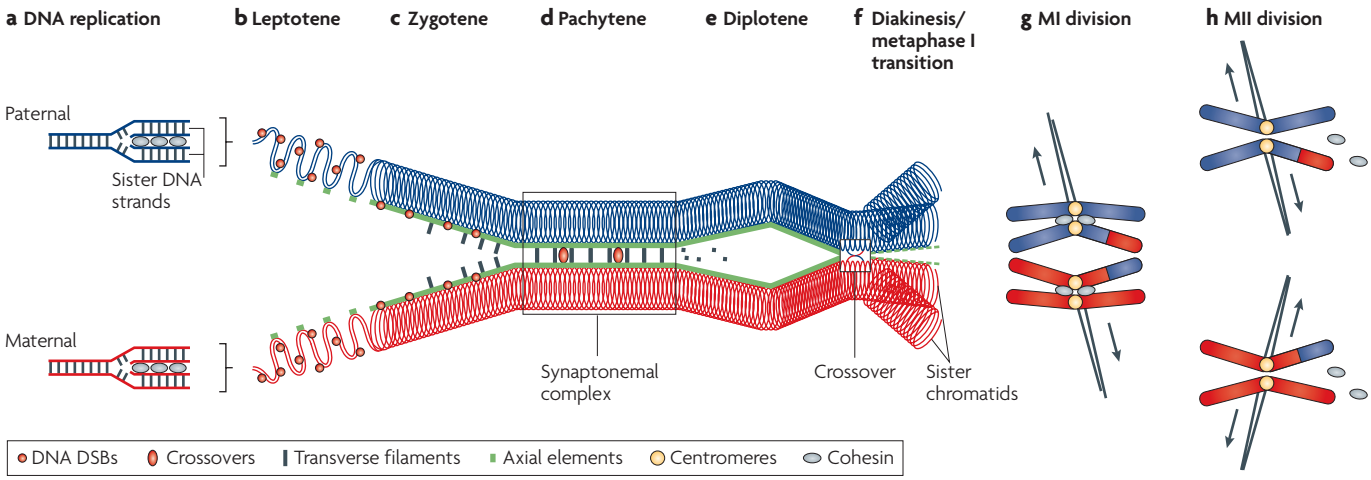


Figure 1 | **An overview of meiosis. a** | During pre-meiotic S phase the DNA of each maternally and paternally derived chromosome is replicated to form two sister DNAs that are held together by cohesins (which remain throughout prophase). **b** | During the leptotene stage hundreds of double-stranded breaks (DSBs; red circles) are introduced into these DNA molecules, and each pair of sister DNA strands begins to assemble a single proteinaceous axis (green). **c** | By the zygotene stage the bulk of the DNA is located in the chromatin loops emanating from the chromosome axes, but the DNA breaks have become axially located. The axes of each pair of homologous maternal and paternal chromosomes begin synapsis via transverse filaments to form a synaptonemal complex; this synapsis is driven by single-stranded DNA tails (not shown), which are generated at the breaks and invade the DNA duplex of the homologue. **d** | The beginning of the pachytene stage is marked by the completion of synapsis. The DNA breaks are repaired, with some of the breaks maturing into crossovers — a minimum of one per chromosome pair. **e** | During the diplotene stage the disassembly of the synaptonemal complex means that the homologous chromosomes are now only held together by the crossovers. **f** | During the transition through diakinesis to the first meiotic metaphase, the axial elements are disassembled and the cohesins that bind the sister chromatids together are removed, except at the centromeres. **g** | The mode of centromere attachment at metaphase of the first meiotic division (MI) ensures that homologues separate with one homologue of each pair passing to each daughter cell. **h** | At the second meiotic division (MII) the remaining cohesion between sister chromatids is lost and the mode of centromere attachment to the spindle ensures that each daughter cell receives one copy of each pair of chromatids.

when DSBs are present. After DNA replication in mitotically dividing cells, DSBs persisting into the ensuing G2 phase undergo homologous recombination repair (HRR) using the intact matching 'sister' DNA molecule as a template. The G2/M checkpoint ensures that once the cell cycle machinery has reached the point when cell division should occur, division is blocked if unrepaired DSBs are still present. This checkpoint has been viewed as very sensitive, but some recent data suggest that the block to cell division might require a threshold number of DSBs, suggested to be around 20 (REF. 30). Key to signalling the presence of unrepaired DSBs are the checkpoint kinases ATM and ATR, and signal amplification involves a number of proteins, including the DNA damage response protein BRCA1 and the phosphorylated form of the variant nucleosomal histone H2AX (γH2AX).

A G2/M-related meiotic pachytene checkpoint? The fact that nearly all the molecular components of the G2/M checkpoint are present during meiosis suggests that a similar checkpoint is operating[31,32]. In male meiosis the first meiotic cell division is scheduled to occur at the end of meiotic prophase, shortly after the long pachytene stage during which meiotic DSB repair should be completed. By analogy to the G2/M checkpoint, it is at this stage that retention of DSB-associated ATM and ATR checkpoint signalling should act to prevent cell division. In females this analogy breaks down because the cells arrest as primordial oocytes at the end of prophase, and progression to the first meiotic division is subsequent to selection from the primordial oocyte pool and the ensuing growth before ovulation; this can be many years after the initial arrest.

Synapsis is linked to accessing a DNA repair template in the homologous chromosome — a feature that distinguishes it from HRR in mitotic G2; asynapsis thus interferes with HRR of the meiotic DSBs. Consequently, chromosome regions that fail to synapse during zygotene retain into the pachytene stage the foci of repair proteins already recruited to the DSBs[27,33]. However, at the zygotene/pachytene transition the key G2/M checkpoint response proteins BRCA1 and ATR, rather than being retained as foci, begin to accumulate along the entire asynapsed chromosomal axis; a phenomenon that was first noted for the asynapsed X and Y chromosome axes in normal males[34–38]. The significance of this axial accumulation is discussed below. In the present context it is reasonable to assume that checkpoint signalling is informing the cell of the persistence of DSBs in the asynapsed chromosomal segments.

Do DNA breaks trigger pachytene spermatocyte apoptosis? In the testis, the association between asynapsis and pachytene spermatocyte apoptosis has been assumed to be a consequence of the checkpoint response to the persisting DSBs. Indirect support comes from observations of the consequences of non-homologous synapsis — a default synapsis that can occur at or after the zygotene/pachytene transition (BOX 1). Non-homologous synapsis leads to loss of the DSB-associated HRR proteins, including ATR and BRCA1, and dephosphorylation of H2AX in the surrounding chromatin, implying that the DSBs have undergone repair and that checkpoint signalling has ceased[9,33,39,40]. There is circumstantial evidence that non-homologous synapsis can circumvent spermatogenic failure[1], so it seems logical to conclude that this is because it leads to DSB repair.

However, the conclusion that checkpoint signalling from the persisting breaks is the trigger for asynapsis-associated pachytene

spermatocyte apoptosis is undermined by other observations. Detailed analysis of meiotic mutants with extensive asynapsis has shown that apoptotic spermatocyte loss occurs at epithelial stage IV of the spermatogenic cycle[16,29,40–43], which equates with the mid-pachytene stage[44]. This is some 2–3 days before pachytene exit and thus it is well before the scheduled time for entry into the first meiotic division. Significantly, *Spo11*-null mouse mutants, which lack meiotic DSBs and consequently have high levels of asynapsis, are also arrested at stage IV[41]. This suggests that stage IV spermatocyte apoptosis might not simply be a checkpoint response to unrepaired meiotic DSBs; evidence that this is indeed the case is provided below.

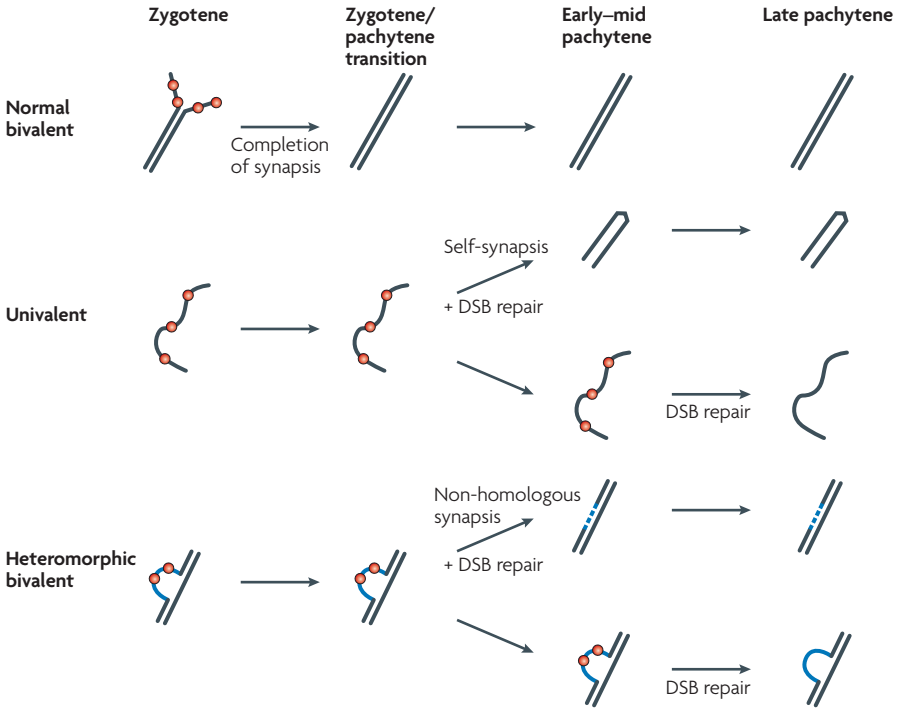
More direct evidence that the persistence of unrepaired DSBs into pachytene is insufficient to trigger stage IV pachytene spermatocyte apoptosis comes from observations of the X chromosome in normal male meiosis, of the asynapsed loops present in association with some chromosome arrangements, and of the added human chromosome 21 derivative of the Down's syndrome mouse model. In each case, DSBs persist into mid-pachytene without triggering apoptosis[33,39,45]. These persisting DSBs are repaired by the end of pachytene, thus shutting down any DSB-dependent checkpoint signalling, and there is progression to the first meiotic metaphase[46] (BOX 1). An important caveat is that in all these examples there are fewer than 20 unrepaired breaks — below the proposed threshold for triggering the mitotic G2/M checkpoint[30].

Failure to inactivate the X and Y

Meiotic sex chromosome inactivation. Global levels of transcription are very low during zygotene[9], presumably owing to the inhibitory effects of chromatin condensation. At the zygotene/pachytene transition in males, when the transcriptional activity of the autosomes is recovering, the X and Y chromosomes are subjected to chromatin modifications that lead to even more complete X and Y transcriptional repression[9,47,48]. This phenomenon, referred to as meiotic sex chromosome inactivation (MSCI), results in absence of X and Y gene transcription throughout pachytene; by contrast, autosomal transcription markedly increases during this period[9,49–53]. A morphological correlate of the transcriptionally inactive XY chromatin domain is the sex body, or XY body, (FIG. 2a) which is a diagnostic feature for pachytene spermatocytes[48,54].

BRCA1 and ATR involvement in MSCI. MSCI is coincident with phosphorylation of H2AX throughout the XY chromatin domain[39] and the finding that *H2afx*-null mice, which lack H2AX, have MSCI failure suggested a causal link[55]. Analysis of male mice carrying a *Brca1* exon 11 deletion, which also show substantial MSCI failure, implicated ATR as the kinase responsible for the phosphorylation of H2AX in the XY chromatin, and the recruitment of ATR was found to be BRCA1-dependent[37,56]. The involvement of ATR has been further supported by the finding that a γH2AX-positive sex body can form in the absence of ATM[57]. Importantly, it has been established that MSCI targets the asynapsed regions of the X and Y chromosomes because they are asynapsed[10]. Our current model for the initiation of MSCI is illustrated in FIG. 2b.

Box 1 | **Non-homologous (heterologous) synapsis and DSB repair in pachytene**



Double-stranded break (DSB)-dependent synapsis during zygotene takes place between homologues and is a prerequisite for timely DSB repair by homologous recombination repair (HRR); consequently, chromosomes or chromosome segments that fail to achieve homologous synapsis during zygotene retain markers of unrepaired DSBs (see figure; red circles). However, after the zygotene/pachytene transition, chromosomes or chromosome segments that have not achieved homologous synapsis can synapse non-homologously[100]. In the case of univalent chromosomes this can occur via self-synapsis to form a fully synapsed hairpin structure, and in the heteromorphic bivalents of heterozygotes for chromosome rearrangements, such as reciprocal translocations, this can occur through synaptic adjustment[101] (blue axes denote the non-homologous segment in the bivalent). This non-homologous synapsis, is (or at least can be) independent of recombinational DSBs, as it is a prominent feature in *Spo11*-null mice[23,102]. One surprising observation is that if non-homologous synapsis of asynapsed regions occurs, this is associated with the loss of the γH2AX and other HRR protein foci, implying that the DSBs have undergone repair[33,39,40]. As far as we are aware the pathway of DSB repair has not been defined, although we would presume the DNA duplex of the sister chromatid is used as the repair template. In agreement with this, Plug *et al.*[33] observed that replication protein A (RPA) reappears in conjunction with the delayed repair; this probably reflects RPA recruitment to the D-loop, which is formed following invasion of the sister DNA duplex[103]. This phenomenon warrants more detailed investigation. Non-homologous synapsis can also circumvent the meiotic silencing of unsynapsed chromatin (MSUC; see main text). Univalents that do not self-synapse, or heteromorphic bivalents that fail to synaptically adjust, retain markers of unrepaired breaks until late in pachytene when the markers disappear, once again indicating that the breaks are repaired[33,40]. This repair is coincident with the repair of DSBs on the asynapsed axis of the X chromosome in normal male meiosis, which must be repaired before proceeding to the first meiotic division.

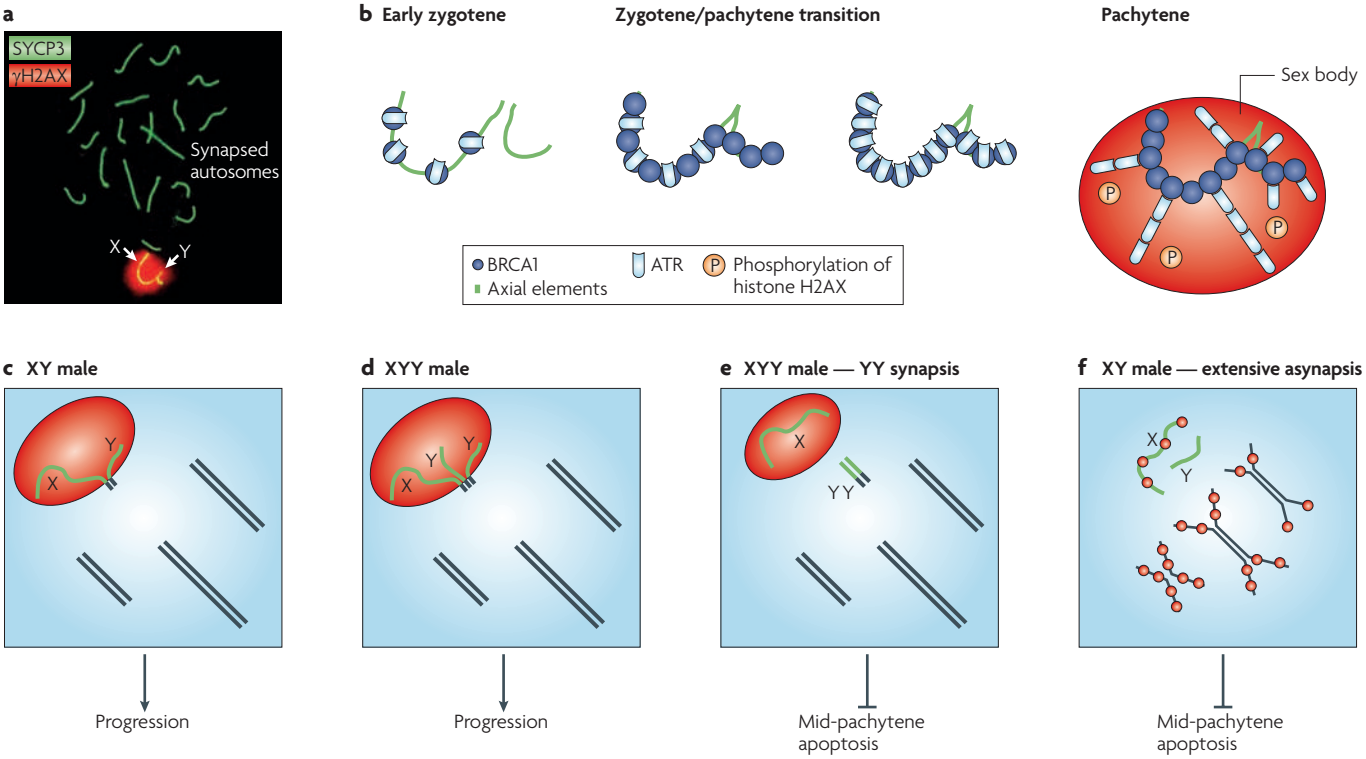


Figure 2 | **Meiotic sex chromosome inactivation (MSCI) and the consequences of its failure. a** | A pachytene spermatocyte stained for the chromosome axial element marker SYCP3 and for the phosphorylated form of the histone variant H2AX (γH2AX), which marks the transcriptionally silenced XY chromatin domain (the sex body). **b** | At the beginning of zygotene the DNA damage response protein BRCA1 and the checkpoint kinase ATR are already present at the sites of all meiotic double-stranded breaks (DSBs), including those on the unpaired X axis. At the zygotene/pachytene transition there is further DSB-independent recruitment of BRCA1 to the asynapsed X and Y axes, which in turn recruits additional ATR; ATR then spreads into the chromatin loops that are associated with the asynapsed X and Y axes and phosphorylates H2AX, triggering the chromatin changes that lead to transcriptional silencing (MSCI). The sex body, which is a diagnostic feature of pachytene spermatocytes, is the morphological manifestation of the silenced X and Y chromatin domain. **c** | Schematic of an XY pachytene spermatocyte with three fully synapsed autosomal bivalents; the asynapsed regions of the XY bivalent are located in the transcriptionally silenced sex body. **d** | An XYY pachytene spermatocyte with an XYY trivalent showing tripartite pseudoautosomal region synapsis; the unsynapsed X and Y axes are located in the sex body. The cells shown in **c** and **d** are able to complete meiotic prophase. **e** | An XYY pachytene spermatocyte with a silenced asynapsed X in the sex body and a fully synapsed transcriptionally active YY bivalent. The inappropriate expression of Y genes is associated with apoptosis. **f** | An XY spermatocyte with extensive autosomal asynapsis. In such cells the X and Y chromosomes are not inactivated, probably because ATR and/or BRCA1 are sequestered at the unrepaired DSBs (red circles) and are thus not available for MSCI. These cells are also eliminated by apoptosis.

MSCI failure leads to pachytene stage IV apoptosis. *H2afx* and *Brca1* mutant males are sterile, owing to overwhelming stage IV pachytene spermatocyte apoptosis. This apoptosis occurs despite only minor disturbances in synapsis with many spermatocytes showing full synapsis[40,55,56], suggesting a link with MSCI failure rather than with asynapsis. However, as the genes that are disrupted are involved in DNA damage responses, it could still be argued that the stage IV apoptosis is due to persisting DSBs that are unrelated to asynapsis. Indeed, the *Brca1* mutant does have γH2AX domains persisting in early pachytene spermatocytes that co-localize with ATR (rarely with the X and Y chromosomes), although the ATR is not axis-associated in contrast to meiotic DSB-associated ATR[37,56]. More direct evidence that MSCI failure leads to stage IV pachytene spermatocyte loss comes from an analysis of MSCI in XYY males and male carriers of the X–autosome translocation T(X;16)16H (REF. 9,10). In some XYY pachytene spermatocytes the two Y chromosomes fully synapse, thus circumventing MSCI; these cells are selectively eliminated at stage IV (FIG. 2d,e). In the T16H carrier males, the X^{16} translocation chromosome sometimes synaptically adjusts to achieve full synapsis with chromosome 16, the X segment escapes MSCI, and these cells are also eliminated in mid-pachytene.

Extensive asynapsis leads to MSCI failure. This link between MSCI failure and stage IV pachytene apoptosis led us to ask whether the stage IV losses in meiotic mutants with extensive asynapsis might also be linked to MSCI failure[16,29,41,43]. *Dnmt3l*-null males proved to be particularly informative because the vast majority of pachytene cells in these mice have some autosomal asynapsis, despite a relatively normal complement of DSBs, but the extent of asynapsis varies over a wide range[40,58]. We found that as the level of asynapsis increased, the recruitment of BRCA1 and ATR to the asynapsed axes of the X and Y chromosomes decreased, probably because ATR and/or BRCA1 was sequestered by the unrepaired DSBs[40]. These males consequently have substantial MSCI failure and pachytene stage IV apoptosis (FIG. 2f). A similar situation pertains in *Msh5*-null males, in which the few *Msh5*$^{-/-}$ cells that survive beyond stage IV have lower levels of asynapsis, but in *Dmc1*-null males there is uniformly severe asynapsis and losses are earlier in stage IV.

These results were foreshadowed in a detailed meiotic study of *Sycp1*$^{-/-}$ males, which have extensive asynapsis and fail to form sex bodies; the authors suggested this might be due to the sequestering of ATR at unrepaired breaks[27]. A more recent microarray-based study looked at transcription in pachytene spermatocytes from a male carrying a reciprocal autosomal translocation involving mouse chromosomes 16 and 17. The authors reported the expected down-regulation of chromosome 17 genes mapping to the asynapsed segment owing to meiotic silencing (see below), but also reported upregulation of X chromosomal genes[59]. Given the limited asynapsis in these mice we suspect that the inferred interference with MSCI is due to occasional non-homologous synapsis of an asynapsed autosomal segment with the X chromosome, rather than to sequestering of ATR and/or BRCA1 at the limited number of unrepaired breaks.

Importantly, there is also extensive MSCI failure in *Spo11*-null males (for reasons that are addressed in the next section), which provides an explanation for the previously puzzling pachytene stage IV apoptosis in the absence of meiotic DSBs[40]. Many other meiotic mutants remain to be analysed, but disruption of sex body formation, which is suggestive of MSCI failure, is prevalent in such mutants[27,60].

Stage IV loss independent of MSCI failure. Recently a mutation of *Trip13* was reported to lead to meiotic failure in male and female mice, and in males the predominant loss was at mid-pachytene[61], specifically at stage IV (J. Schimenti, personal communication). The pachytene spermatocytes exhibited normal synapsis and there were no reported abnormalities in sex body formation; thus MSCI failure is almost certainly not the explanation for the mid-pachytene loss. However, analysis of markers of HRR revealed that there was a problem with DSB repair. This model is distinct from those with MSCI failure as DSB repair is initiated, and it might be that in mammals there is a specific checkpoint response to stalled recombination intermediates similar to that implicated in the Zip1 checkpoint response in yeast[13].

Meiotic and post-meiotic gene silencing

Meiotic silencing of unsynapsed chromatin. In 2005 two groups discovered that in chromosomally variant male and female mice the asynapsed chromosomes or chromosome segments are transcriptionally silenced, and as with MSCI this is associated with the accumulation of BRCA1 and ATR on asynapsed axes, the spreading of ATR into the associated chromatin loops and the phosphorylation of H2AX[8,9]. We now view MSCI as a consequence of this more general silencing mechanism. As suggested by Schimenti[62], we refer to the silenced chromatin domain as unsynapsed chromatin, and the transcriptional silencing as meiotic silencing of unsynapsed chromatin (MSUC); this serves to distinguish it from the mechanistically distinct process of meiotic silencing by unpaired DNA (MSUD) that was identified in the mould *Neurospora crassa*[63]. Because MSUC is not male-limited and probably preceded the evolution of the heteromorphic X and Y chromosomes, we no longer view MSCI as having evolved to protect male meiosis from the consequences of having unrepaired DSBs on the X chromosome axis[47], although it might do so.

MSUC is a meiotic DSB-independent response. *Spo11*-null pachytene spermatocytes have a sex body-like structure that is termed the pseudo sex body because it rarely encompasses X and Y chromatin[39,41,57]. Nevertheless, this domain is a manifestation of a MSUC response, as there is recruitment of BRCA1 and ATR to asynapsed axes in the domain, spreading of ATR to the associated asynapsed chromatin (thus explaining the phosphorylation of H2AX) and there is transcriptional silencing in this domain[40]. It is not yet clear why this MSUC domain does not incorporate all the asynapsed chromatin in this mutant[40], but the MSUC response clearly does not require meiotic (that is, *Spo11*-dependent) DSBs. Recently, in an elegant study confirming the tight link between asynapsis and meiotic silencing, it was suggested that unrepaired DSBs introduced by radiation can enhance the MSUC response[64]. However, this suggestion is hard to reconcile with the data on *Spo11*$^{-/-}$ and *Dnmt3l*$^{-/-}$ males (discussed above), which show that MSUC is meiotic DSB independent and that with increasing numbers of unrepaired DSBs there is progressive impairment of the MSUC response, probably by sequestering ATR and/or BRCA1, with consequent MSCI failure[40].

MSUC as a likely cause of spermatogenic impairment. The pachytene stage lasts approximately 7 days in male mice[65] and is highly transcriptionally active[51]. It is therefore reasonable to assume that MSUC will sometimes silence genes that are crucial for pachytene cell survival; indeed, the progressive silencing of the X and Y chromosomes by MSUC, in conjunction with evolution of the heteromorphic X–Y pair, is thought to have necessitated the generation of autosomally located retrogenes to 'backup' X-encoded genes that are essential for pachytene cell function or survival[66–69]. Furthermore, it is now clear that the transcriptional silencing initiated by the MSUC response is, to a substantial extent, maintained during the otherwise highly transcriptionally active post-meiotic round spermatid stages[10,51,52,70–73]. This applies not only to the X and Y chromosomes during unperturbed spermatogenesis, but also to autosomal segments that are subject to MSUC[10]. Thus, MSUC could also lead to the repression of autosomal genes that are essential for spermatid survival.

Although it seems inevitable that the silencing of autosomal genes owing to MSUC must contribute to spermatogenic failure, direct evidence is as yet lacking. Which mouse models might be used to glean such evidence? Models in which there is extensive asynapsis are precluded because extensive asynapsis interferes with the MSUC response[40], but there are numerous chromosome rearrangements that lead to limited asynapsis and are associated with spermatogenic impairment. Mice that are doubly heterozygous for two semi-identical reciprocal translocations, T(1;13)70H and T(1;13)1Wa (REF. 74), have a 1^{13} bivalent that frequently retains an unsynapsed loop into pachytene (FIG. 3a–c). However, the frequency of loops and the degree of spermatogenic impairment vary markedly between males, with sperm counts significantly correlated with the frequency of fully adjusted 1^{13} bivalents[74]. So why is there an association between synaptic adjustment and fertility? The unadjusted loop has a few (less than 5) unrepaired breaks, as evidenced by the presence of RAD51 foci, and the associated chromatin is also positive for markers of transcriptional silencing; these are lost if there is synaptic adjustment[8,33,39,71,75]. As we observed earlier, a small number of unrepaired breaks in the context of a silenced chromatin domain do not trigger pachytene spermatocyte loss, so we surmise that it is the avoidance of transcriptional silencing that explains the link between synaptic adjustment and improved fertility.

Asynapsis and female meiosis

A pachytene response to DNA breaks. In female mice, meiotic prophase is less accessible for study because it takes place prenatally; consequently there is less information regarding the consequences of asynapsis during the pachytene period in females.

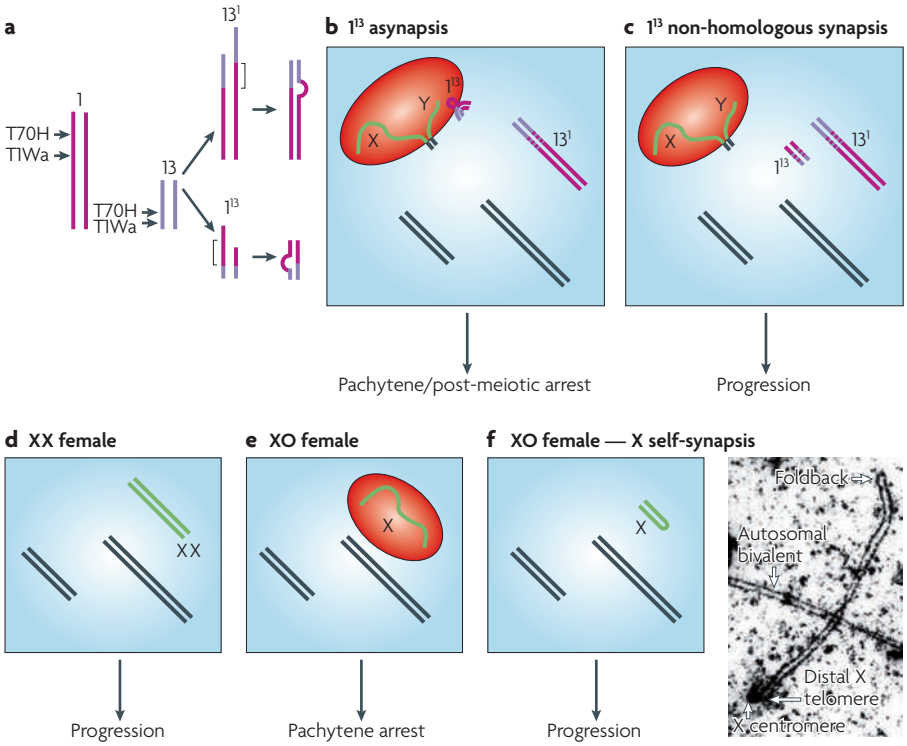


Figure 3 | **Meiotic silencing as a potential cause of meiotic or post-meiotic failure.** **a** | Chromosome 1 and 13 translocation breakpoints are shown, as well as the resulting chromosomal constitution of T70H/T1Wa double translocation heterozygotes. **b** | A pachytene spermatocyte from the double heterozygote in which the 13^1 bivalent is fully synapsed, owing to synaptic adjustment, whereas the much smaller 1^{13} bivalent has the unmatched segment of chromosome 1 present as an asynapsed loop. The chromatin of the asynapsed loop is subject to meiotic silencing (through meiotic silencing of unsynapsed chromatin; MSUC) and this chromatin is usually incorporated in the sex body. The frequency of pachytene spermatocytes with asynapsed loops varies widely between males, and sperm counts are inversely correlated with the frequency of asynapsed loops, suggesting that these cells or their post-meiotic products are eliminated. We suggest that this is due to MSUC-initiated gene silencing in the loop. **c** | A spermatocyte from a double translocation heterozygote in which the 1^{13} bivalent has also adjusted to allow synapsis of the unmatched segment of chromosome 1, which now avoids MSUC-initiated silencing. We presume that these are the cells that go on to form sperm. **d** | A schematic of an XX pachytene oocyte showing the fully synapsed, and thus active, XX bivalent and two autosomal bivalents. **e** | An XO pachytene oocyte in which the asynapsed X univalent is silenced. The X chromosome carries many housekeeping genes that are essential for cell survival, and the death of these cells explains the documented perinatal late pachytene oocyte loss in XO female mice. **f** | An XO pachytene oocyte in which the X chromosome has achieved full (that is, non-homologous) self-synapsis and thus remains active. These oocytes are undoubtedly those that survive the perinatal period of oocyte loss and that contribute to the oocyte pool on which the fertility of XO mice depends. An electron micrograph of a self-synapsed X univalent from a spread pachytene oocyte is shown; a synapsed autosomal bivalent is lying across the univalent X.

However, female meiosis is exempt from the consequences of MSCI failure and thus is better suited for assessing if there is a pachytene response to unrepaired DSBs.

The asynapsed X chromosome of XO females, as in normal males, has DSBs that remain into mid-pachytene; they are then repaired, thus shutting down checkpoint signalling from the DSBs before pachytene exit (P.S.B., S.K.M. and J.M.A.T, unpublished data). Therefore, these mice do not allow us to address the question of whether DSBs that remain beyond the normal time of pachytene exit will trigger pachytene oocyte loss. However, in *Dmc1*-null females, in which there is extensive asynapsis and the DSBs are unrepairable, there is catastrophic oocyte failure by the end of pachytene; this is ameliorated if the mice are also *Spo11* null, implying that the pachytene oocyte loss is at least in part meiotic DSB dependent[76].

MSUC and oocyte loss. In contrast to the silencing of the X and Y chromosomes in pachytene of male meiosis, in females the inactive X of oogonia is reactivated so that both X chromosomes are active throughout prophase[77–79]. In XO female mice there are sufficient oocytes for fertility, but there is substantial excess perinatal pachytene oocyte loss[17], even though any DSBs are by then repaired. So why are some oocytes eliminated while others survive? As the X chromosome lacks a homologue it cannot achieve homologous synapsis, but it does frequently fold back on itself and achieve non-homologous self-synapsis, thus evading MSUC[9,80]. However, if the X fails to achieve synapsis in this way, it is subject to MSUC[8,9] and the autosomal backups for essential X gene functions that are expressed in pachytene spermatocytes are not expressed in XO oocytes. Thus the silencing of essential X genes by MSUC is a probable explanation for the observed excess perinatal oocyte loss in XO females, with the surviving oocytes being those that avoided MSUC by self-synapsis (FIG. 3d–f).

Spo11-null females, like males, are sterile, but oocyte loss occurs both prenatally and postnatally[76]. The lack of meiotic DSBs means that despite the extensive asynapsis, MSUC cannot be impaired by extensive sequestering of ATR and/or BRCA1 at unrepaired breaks; indeed, γH2AX-positive pseudo sex body domains are found in *Spo11*-null pachytene oocytes (P.S.B., S.K.M. and J.M.A.T, unpublished data). Thus the oocyte loss in *Spo11*-null females might be due to the silencing of crucial genes, with the variability in the stage of loss reflecting differences in which genes are silenced in one oocyte relative to another.

Male–female differences

The basis for the more severe effects of asynapsis in males compared with females has been discussed in other reviews[4,5,60]. Here we focus on three factors that we now believe to be of particular importance, and a summary is provided in FIG. 4.

Differing consequences of MSUC failure. In our view the most important factor in the context of meiotic mutants relates to the different consequences of sequestering ATR and/or BRCA1 at the sites of unrepaired breaks in asynapsed chromosomal segments. In males and females with extensive asynapsis there will be an abrogation of the MSUC response; in females this might avoid oocyte losses resulting from transcriptional silencing, whereas in males it leads to MSCI failure and stage IV pachytene apoptosis. These differing consequences of disrupting MSUC are in fact manifest in *H2afx*-null mice or in mice with a homozygous deletion of *Brca1*

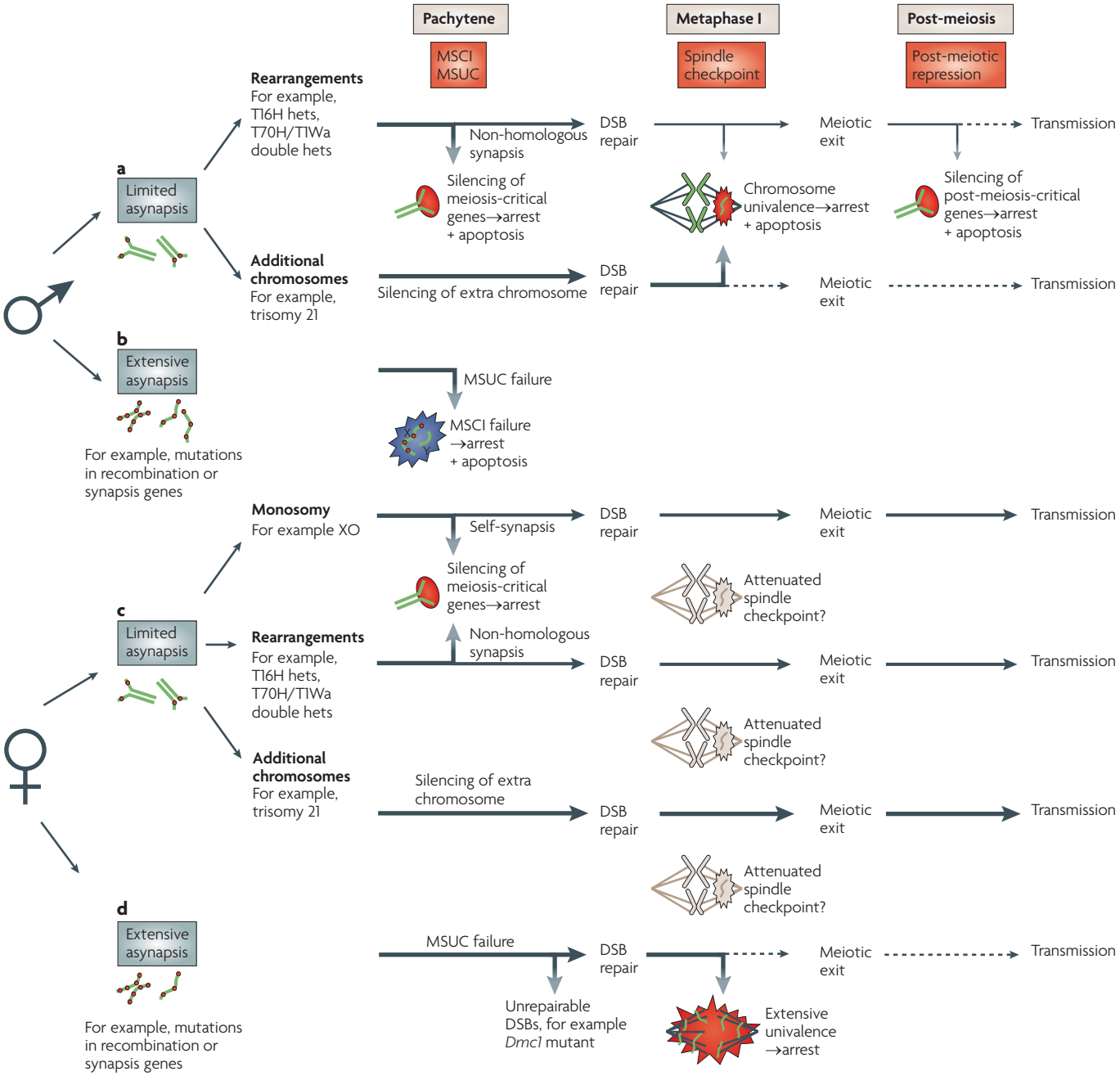


Figure 4 | **The consequences of asynapsis in male and female meiosis.** Vertical grey arrows indicate points at which meiotic loss occurs, the width of the arrows indicating the extent of the loss. **a** | Limited asynapsis in males. In mice with chromosomal rearrangements, silencing of crucial genes by meiotic silencing of unsynapsed chromatin (MSUC) is likely to be a major cause of pachytene or post-meiotic failure; non-homologous synapsis avoids this. Some rearrangements are associated with an increased frequency of univalence, which will cause some loss at metaphase I. With an additional chromosome MSUC is an advantage because it will avoid excess gene activity during pachytene, but univalence at meiosis I (MI) is expected to cause major loss, with only a few cells avoiding elimination. **b** | Extensive asynapsis in males. The sequestering of the checkpoint kinase ATR and/or the DNA damage response protein BRCA1 at the many unrepaired breaks abrogates the MSUC response with consequent meiotic sex chromosome inactivation (MSCI) failure. This leads to apoptosis at mid-pachytene and is a major cause of pachytene loss in males with meiotic mutations that disrupt synapsis. **c** | Limited asynapsis in females. In the case of XO monosomy, the MSUC silencing of the sole X chromosome is undoubtedly the cause of the documented perinatal oocyte loss; the oocytes that survive will be those that achieve self-synapsis and thus avoid MSUC. In the case of rearrangements, MSUC silencing of crucial genes has the potential to lead to losses, although pachytene is a much shorter stage than in males, so this might have less impact. With an additional chromosome the silencing will be an advantage, and the failure to efficiently eliminate oocytes with univalents at MI means that the additional chromosome will be transmitted to progeny. **d** | Extensive asynapsis in females. In females MSUC failure will avoid the potential deleterious effects that result from MSUC silencing of crucial genes. The predicted outcome then depends on whether the DSBs are repaired. In *Dmc1*$^{-/-}$ females the DSBs are not repaired and are implicated in the observed oocyte failure[76], this might be due to a G2/M-related checkpoint response. If the DSBs are repaired at the end of pachytene then extensive univalence at MI is expected to cause spindle assembly defects that prevent further progression, as observed in *Mlh1*$^{-/-}$ females[99]. hets, heterozygotes.

exon 11. These mutations disrupt the MSUC response and in both cases females have no reported meiotic defects and are fertile, whereas males have stage IV pachytene apoptosis and are sterile[40,56,81].

Differing consequences of chromosome univalence. At the first meiotic metaphase the chiasmate chromosome pairs, the bivalents, each need to achieve bipolar attachment to the spindle in order to segregate one chromosome of each pair to each daughter cell (FIG. 1), and there is increasing evidence that in mammals, as in yeasts, there is a spindle assembly checkpoint that monitors this process[82–84].

In males and females with low levels of asynapsis, spermatocytes and oocytes with asynapsed chromosomes can survive the pachytene period if MSUC has not silenced pachytene-critical genes, or if MSUC has been avoided by non-homologous self-synapsis. In both cases the DSBs that are present are repaired, either following non-homologous synapsis or at the end of pachytene. On progression to diplotene these chromosomes lack chiasmate associations and are termed univalents. In males, based predominantly on studies of sex chromosome univalence, it is clear that meiosis I (MI) spermatocytes with univalent chromosomes are efficiently eliminated by apoptosis, although a few do survive[85–89]. It is widely assumed that this apoptosis is a downstream consequence of a MI spindle checkpoint response to the univalents.

It is important to emphasize that the MI spindle checkpoint as defined in yeast causes a delay in progression to anaphase rather than cell death[84]; how the presumed MI checkpoint signalling in male mammals is transduced to trigger an apoptotic response is unknown. Paradoxically, in female mammals, for which there is clear evidence of a MI spindle checkpoint[90], there is no evidence of an apoptotic response and the MI oocytes with univalents are either not eliminated or are eliminated very inefficiently. Thus, in XO female mice the oocytes that avoid elimination during pachytene, by circumventing MSUC through self-synapsis, also survive MI and contribute to the production of XO daughters[91–93].

There is a growing consensus that the high rate of aneuploid conceptions in older women is partly a consequence of loss of bivalent cohesion leading to univalence, which, combined with the absence of an efficient mechanism to eliminate the aberrant oocytes at MI, results in a high rate of aneuploid conceptions[94–96]. There is also evidence that a substantial number of univalents can evade the checkpoint through bipolar attachment to the MI spindle[93,97]; this seems to be a rare occurrence in males and thus might be an important factor in the sex difference in the response to univalents at MI.

Differing consequences of gametic deficiency. The consequences of reduced gamete production are very different between males and females. Thus, substantial reductions in sperm production are reflected in the number of sperm in the ejaculate and have a direct impact on fertility. However, an equivalent reduction in the size of the oocyte pool does not have any immediate impact on the number of eggs ovulated, because the dynamics of oocyte maturation are such that ovulation rate is unaffected by substantial reductions in oocyte pool size; what is reduced is the reproductive lifespan. Unfortunately, oocyte pool size and/or reproductive lifespan are rarely determined when assessing the effects of chromosome anomalies or mutations on female fertility. In cases in which they have been determined, females have been found to be substantially affected[17–20,98].

Conclusions and perspectives

In this article we have made two main propositions: that a failure to silence the X and/or the Y chromosome during pachytene (MSCI failure), rather than a checkpoint response to unrepaired DSBs, is the predominant cause of mid-pachytene spermatocyte loss; and that meiotic silencing of unsynapsed autosomal chromatin is likely to contribute to male and female meiotic losses and, in the male, postmeiotic losses. In this section, we consider a number of unresolved issues that relate to these propositions.

We have suggested that the progressive disruption of synapsis during the evolution of the heteromorphic X–Y pair engendered a MSUC response that resulted in MSCI and necessitated the backing up of essential X-encoded functions as retrogenes on the autosomes. So why should MSCI failure lead to meiotic failure? In the context of the Y chromosome, it is well established that a number of mouse Y genes have diverged considerably from their X progenitors, some becoming testis specific in the process; this may explain why some Y gene products can no longer be tolerated during pachytene. However, it is clear that expression from the X chromosome during pachytene is also cell lethal, which suggests that in conjunction with the backing up of essential X gene functions, spermatocytes have become evolutionarily adapted to the absence of a number of X gene products, so that their expression is now deleterious.

Aside from MSCI failure as a cause of stage IV pachytene loss we suggest that stalled recombination intermediates (as found in *Trip13*$^{-/-}$ spermatocytes) might also trigger stage IV apoptosis. The possibility remains that unrepaired DSBs, when present in sufficient numbers, might also provide a trigger at stage IV, but the only evidence for an effect over and above that due to MSCI failure is found in *Dmc1*$^{-/-}$ males; this mutant is unusual in that there is hyper-resection of the DSBs. In our view, the focus for the future should be on identifying the molecular pathways that link these seemingly disparate triggers with an apoptotic outcome. Barchi *et al.*[41] have hypothesized that apoptotic elimination of spermatocytes at epithelial stage IV might in fact be orchestrated by Sertoli cells that perform some kind of quality surveillance function. From this viewpoint, spermatocytes with MSCI failure or those with incompletely repaired (or perhaps unrepaired) DSBs must have a feature in common that engenders Sertoli cell intervention.

With regard to meiotic silencing, a number of issues need to be resolved. First, our presumption that phosphorylation of H2AX is the precipitating event for MSUC needs to be confirmed, and this would be best achieved by creating a mouse in which the serine that is phosphorylated in H2AX is replaced by an amino acid that cannot be phosphorylated — would this prevent MSUC and thus lead to MSCI failure and stage IV spermatocyte apoptosis? With such a mutant it might also be possible to check whether the oocyte loss in XO female mice is due to the silencing of the X chromosome in those oocytes in which it remains unsynapsed. A greater challenge will be to provide supporting evidence that MSUC leads to pre- and post-meiotic losses in males owing to the silencing of crucial genes. The phenomenon of synaptic adjustment makes this a daunting task because in heteromorphic bivalents only large chromosomal segments fail to adjust completely, and because of partial adjustment the gene content of the unadjusted region is hard to define.

Paul S. Burgoyne, Shantha K. Mahadevaiah and James M. A. Turner are at the Division of Stem Cell Biology and Developmental Genetics, Medical Research Council National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA.

Correspondence to P.S.B.
e-mail: pburgoy@nimr.mrc.ac.uk

doi:10.1038/nrg2505
Published online 3 February 2009

1. de Boer, P. & de Jong, J. H. in *Fertility and Chromosome Pairing: Recent Studies in Plants and Animals* (ed. Gillies, C. B.) 37–76 (CRC, Boca Raton, Florida, 1989).
2. Vincent, M. C. *et al.* Cytogenetic investigations of infertile men with low sperm counts: a 25-year experience. *J. Androl.* **23**, 18–22; discussion 44–45 (2002).
3. Cohen, P. E., Pollack, S. E. & Pollard, J. W. Genetic analysis of chromosome pairing, recombination, and cell cycle control during first meiotic prophase in mammals. *Endocr. Rev.* **27**, 398–426 (2006).
4. Hunt, P. A. & Hassold, T. J. Sex matters in meiosis. *Science* **296**, 2181–2183 (2002).
5. Morelli, M. A. & Cohen, P. E. Not all germ cells are created equal: aspects of sexual dimorphism in mammalian meiosis. *Reproduction* **130**, 761–781 (2005).
6. Wang, H. & Hoog, C. Structural damage to meiotic chromosomes impairs DNA recombination and checkpoint control in mammalian oocytes. *J. Cell Biol.* **173**, 485–495 (2006).
7. Hassold, T. & Hunt, P. To err (meiotically) is human: the genesis of human aneuploidy. *Nature Rev. Genet.* **2**, 280–291 (2001).
8. Baarends, W. M. *et al.* Silencing of unpaired chromatin and histone H2A ubiquitination in mammalian meiosis. *Mol. Cell Biol.* **25**, 1041–1053 (2005).
9. Turner, J. M. *et al.* Silencing of unsynapsed meiotic chromosomes in the mouse. *Nature Genet.* **37**, 41–47 (2005).
10. Turner, J. M., Mahadevaiah, S. K., Ellis, P. J., Mitchell, M. J. & Burgoyne, P. S. Pachytene asynapsis drives meiotic sex chromosome inactivation and leads to substantial postmeiotic repression in spermatids. *Dev. Cell* **10**, 521–529 (2006).
11. Ferguson, K. A., Chow, V. & Ma, S. Silencing of unpaired meiotic chromosomes and altered recombination patterns in an azoospermic carrier of a t(8;13) reciprocal translocation. *Hum. Reprod.* **23**, 988–995 (2008).
12. Sciurano, R., Rahn, M., Rey-Valzacchi, G. & Solari, A. J. The asynaptic chromatin in spermatocytes of translocation carriers contains the histone variant γ-H2AX and associates with the XY body. *Hum. Reprod.* **22**, 142–50 (2007).
13. Hochwagen, A. & Amon, A. Checking your breaks: surveillance mechanisms of meiotic recombination. *Curr. Biol.* **16**, R217–R228 (2006).
14. Roeder, G. S. & Bailis, J. M. The pachytene checkpoint. *Trends Genet.* **16**, 395–403 (2000).
15. Ashley, T. in *Results and Problems in Cell Differentiation* Vol. 28 (ed. McElreavey, K.) 131–173 (Springer, Berlin, 2000).
16. de Rooij, D. G. & de Boer, P. Specific arrests of spermatogenesis in genetically modified and mutant mice. *Cytogenet. Genome Res.* **103**, 267–276 (2003).
17. Burgoyne, P. S. & Baker, T. G. Perinatal oocyte loss in XO mice and its implications for the aetiology of gonadal dysgenesis in XO women. *J. Reprod. Fertil.* **75**, 633–645 (1985).
18. Burgoyne, P. S., Mahadevaiah, S. K. & Mittwoch, U. A reciprocal autosomal translocation which causes male sterility in the mouse also impairs oogenesis. *J. Reprod. Fertil.* **75**, 647–652 (1985).
19. Mittwoch, U., Mahadevaiah, S. K. & Setterfield, L. A. Pachytene pairing and oocyte numbers in mice with two single Robertsonian translocations and the male-sterile compound with monobrachial homology. *Cytogenet. Cell Genet.* **53**, 144–147 (1990).
20. Setterfield, L. A., Mahadevaiah, S. K. & Mittwoch, U. Chromosome pairing and germ cell loss in male and female mice carrying a reciprocal translocation. *J. Reprod. Fertil.* **82**, 369–379 (1988).
21. Pittman, D. L. *et al.* Meiotic prophase arrest with failure of chromosome synapsis in mice deficient for *Dmc1*, a germline-specific RecA homolog. *Mol. Cell* **1**, 697–705 (1998).
22. Yoshida, K. *et al.* The mouse *RecA*-like gene *Dmc1* is required for homologous chromosome synapsis during meiosis. *Mol. Cell* **1**, 707–718 (1998).
23. Romanienko, P. J. & Camerini-Otero, R. D. The mouse *Spo11* gene is required for meiotic chromosome synapsis. *Mol. Cell* **6**, 975–987 (2000).
24. Baudat, F., Manova, K., Yuen, J. P., Jasin, M. & Keeney, S. Chromosome synapsis defects and sexually dimorphic meiotic progression in mice lacking Spo11. *Mol. Cell* **6**, 989–998 (2000).
25. Edelmann, W. *et al.* Mammalian MutS homologue 5 is required for chromosome pairing in meiosis. *Nature Genet.* **21**, 123–127 (1999).
26. Kneitz, B. *et al.* MutS homolog 4 localization to meiotic chromosomes is required for chromosome pairing during meiosis in male and female mice. *Genes Dev.* **14**, 1085–1097 (2000).
27. de Vries, F. A. *et al.* Mouse *Sycp1* functions in synaptonemal complex assembly, meiotic recombination, and XY body formation. *Genes Dev.* **19**, 1376–1389 (2005).
28. Libby, B. J. *et al.* The mouse meiotic mutation *mei1* disrupts chromosome synapsis with sexually dimorphic consequences for meiotic progression. *Dev. Biol.* **242**, 174–187 (2002).
29. de Vries, S. S. *et al.* Mouse MutS-like protein MSH5 is required for proper chromosome synapsis in male and female meiosis. *Genes Dev.* **13**, 523–531 (1999).
30. Deckbar, D. *et al.* Chromosome breakage after G2 checkpoint release. *J. Cell Biol.* **176**, 749–755 (2007).
31. Marcon, E. & Moens, P. B. The evolution of meiosis: recruitment and modification of somatic DNA-repair proteins. *Bioessays* **27**, 795–808 (2005).
32. Burgoyne, P. S., Mahadevaiah, S. K. & Turner, J. M. The management of DNA double-strand breaks in mitotic G_2, and in mammalian meiosis viewed from a mitotic G_2 perspective. *Bioessays* **29**, 974–986 (2007).
33. Plug, A. W. *et al.* Changes in protein composition of meiotic nodules during mammalian meiosis. *J. Cell Sci.* **111**, 413–423 (1998).
34. Keegan, K. S. *et al.* The Atr and Atm protein kinases associate with different sites along meiotically pairing chromosomes. *Genes Dev.* **10**, 2423–2437 (1996).
35. Moens, P. B. *et al.* The association of ATR protein with mouse meiotic chromosome cores. *Chromosoma* **108**, 95–102 (1999).
36. Perera, D. *et al.* TopBP1 and ATR colocalization at meiotic chromosomes: role of TopBP1/Cut5 in the meiotic recombination checkpoint. *Mol. Biol. Cell* **15**, 1568–1579 (2004).
37. Turner, J. M. *et al.* BRCA1, histone H2AX phosphorylation, and male meiotic sex chromosome inactivation. *Curr. Biol.* **14**, 2135–42 (2004).
38. Scully, R. *et al.* Association of BRCA1 with Rad51 in mitotic and meiotic cells. *Cell* **88**, 265–275 (1997).
39. Mahadevaiah, S. K. *et al.* Recombinational DNA double-strand breaks in mice precede synapsis. *Nature Genet.* **27**, 271–276 (2001).
40. Mahadevaiah, S. K. *et al.* Extensive meiotic asynapsis in mice antagonises meiotic silencing of unsynapsed chromatin and consequently disrupts meiotic sex chromosome inactivation. *J. Cell Biol.* **182**, 263–276 (2008).
41. Barchi, M. *et al.* Surveillance of different recombination defects in mouse spermatocytes yields distinct responses despite elimination at an identical developmental stage. *Mol. Cell Biol.* **25**, 7203–7215 (2005).
42. Hamer, G., Kal, H. B., Westphal, C. H., Ashley, T. & de Rooij, D. G. Ataxia telangiectasia mutated expression and activation in the testis. *Biol. Reprod.* **70**, 1206–1212 (2004).
43. Bolcun-Filas, E. *et al.* SYCE2 is required for synaptonemal complex assembly, double strand break repair, and homologous recombination. *J. Cell Biol.* **176**, 741–747 (2007).
44. Ashley, T., Gaeth, A. P., Creemers, L. B., Hack, A. M. & de Rooij, D. G. Correlation of meiotic events in testis sections and microspreads of mouse spermatocytes relative to the mid-pachytene checkpoint. *Chromosoma* **113**, 126–136 (2004).
45. O'Doherty, A. *et al.* An aneuploid mouse strain carrying human chromosome 21 with Down syndrome phenotypes. *Science* **309**, 2033–2037 (2005).
46. Habermann, B. *et al. DAZ* (Deleted in AZoospermia) genes encode proteins located in human late spermatids and sperm tails. *Hum. Reprod.* **13**, 363–369 (1998).
47. Turner, J. M. A. & Burgoyne, P. S. in *The Y Chromosome and Male Germ Cell Biology in Health and Diseases* (eds Lau, Y.-F. C. & Chan, W. Y.) 27–46 (World Scientific, Hackensack, New Jersey, 2006).
48. Handel, M. A. The XY body: a specialized meiotic chromatin domain. *Exp. Cell Res.* **296**, 57–63 (2004).
49. Khil, P. P., Smirnova, N. A., Romanienko, P. J. & Camerini-Otero, R. D. The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nature Genet.* **36**, 642–646 (2004).
50. Monesi, V. Synthetic activities during spermatogenesis in the mouse. *Exp. Cell Res.* **39**, 197–224 (1965).
51. Moore, G. P. DNA-dependent RNA synthesis in fixed cells during spermatogenesis in mouse. *Exp. Cell Res.* **68**, 462–465 (1971).
52. Namekawa, S. H. *et al.* Postmeiotic sex chromatin in the male germline of mice. *Curr. Biol.* **16**, 660–667 (2006).
53. Turner, J. M. Meiotic sex chromosome inactivation. *Development* **134**, 1823–1831 (2007).
54. Solari, A. J. The behaviour of the XY pair in mammals. *Int. Rev. Cytol.* **38**, 273–317 (1974).
55. Fernandez-Capetillo, O. *et al.* H2AX is required for chromatin remodeling and inactivation of sex chromosomes in male mouse meiosis. *Dev. Cell* **4**, 497–508 (2003).
56. Xu, X., Aprelikova, O., Moens, P., Deng, C. X. & Furth, P. A. Impaired meiotic DNA-damage repair and lack of crossing-over during spermatogenesis in BRCA1 full-length isoform deficient mice. *Development* **130**, 2001–2012 (2003).
57. Bellani, M. A., Romanienko, P. J., Cairatti, D. A. & Camerini-Otero, R. D. SPO11 is required for sex-body formation, and Spo11 heterozygosity rescues the prophase arrest of *Atm*$^{-/-}$ spermatocytes. *J. Cell Sci.* **118**, 3233–3245 (2005).
58. Bourc'his, D. & Bestor, T. H. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431**, 96–9 (2004).
59. Homolka, D., Ivanek, R., Capkova, J., Jansa, P. & Forejt, J. Chromosomal rearrangement interferes with meiotic X chromosome inactivation. *Genome Res.* **17**, 1431–1437 (2007).
60. Kolas, N. K. *et al.* Mutant meiotic chromosome core components in mice can cause apparent sexual dimorphic endpoints at prophase or X-Y defective male-specific sterility. *Chromosoma* **114**, 92–102 (2005).
61. Li, X. C. & Schimenti, J. C. Mouse pachytene checkpoint 2 (trip13) is required for completing meiotic recombination but not synapsis. *PLoS Genet.* **3**, e130 (2007).
62. Schimenti, J. Synapsis or silence. *Nature Genet.* **37**, 11–13 (2005).
63. Shiu, P. K., Raju, N. B., Zickler, D. & Metzenberg, R. L. Meiotic silencing by unpaired DNA. *Cell* **107**, 905–916 (2001).
64. Schoenmakers, S. *et al.* Increased frequency of asynapsis and associated meiotic silencing of heterologous chromatin in the presence of irradiation-induced extra DNA double strand breaks. *Dev. Biol.* **317**, 270–281 (2008).
65. Oakberg, E. F. Duration of spermatogenesis in the mouse and timing of stages of the cycle of the seminiferous epithelium. *Am. J. Anat.* **99**, 507–516 (1956).
66. Wang, P. J. X chromosomes, retrogenes and their role in male reproduction. *Trends Endocrinol. Metab.* **15**, 79–83 (2004).
67. Rohozinski, J. & Bishop, C. E. The mouse *juvenile spermatogonial depletion* (*jsd*) phenotype is due to a mutation in the X-derived retrogene, *mUtp14b*. *Proc. Natl Acad. Sci. USA* **101**, 11695–11700 (2004).
68. Bradley, J. *et al.* An X-to-autosome retrogene is required for spermatogenesis in mice. *Nature Genet.* **36**, 872–876 (2004).
69. Zhao, M. *et al. Utp14b*: a unique retrogene within a gene that has acquired multiple promoters and a specific function in spermatogenesis. *Dev. Biol.* **304**, 848–859 (2007).
70. Khalil, A. M., Boyar, F. Z. & Driscoll, D. J. Dynamic histone modifications mark sex chromosome inactivation and reactivation during mammalian spermatogenesis. *Proc. Natl Acad. Sci. USA* **101**, 16583–16587 (2004).
71. van der Heijden, G. W. *et al.* Chromosome-wide nucleosome replacement and H3.3 incorporation during mammalian meiotic sex chromosome inactivation. *Nature Genet.* **39**, 251–258 (2007).
72. Greaves, I. K., Rangasamy, D., Devoy, M., Marshall Graves, J. A. & Tremethick, D. J. The X and Y chromosomes assemble into H2A.Z, containing facultative heterochromatin, following meiosis. *Mol. Cell Biol.* **26**, 5394–405 (2006).
73. Mueller, J. L. *et al.* The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. *Nature Genet.* **40**, 794–799 (2008).
74. Peters, A. H., Plug, A. W. & de Boer, P. Meiosis in carriers of heteromorphic bivalents: sex differences and implications for male fertility. *Chromosome Res.* **5**, 313–324 (1997).

75. van der Laan, R. *et al.* Ubiquitin ligase Rad18^{Sc} localizes to the XY body and to other chromosomal regions that are unpaired and transcriptionally silenced during male meiotic prophase. *J. Cell Sci.* **117**, 5023–33 (2004).
76. Di Giacomo, M. *et al.* Distinct DNA-damage-dependent and -independent responses drive the loss of oocytes in recombination-defective mouse mutants. *Proc. Natl Acad. Sci. USA* **102**, 737–742 (2005).
77. Sugimoto, M. & Abe, K. X chromosome reactivation initiates in nascent primordial germ cells in mice. *PLoS Genet.* **3**, e116 (2007).
78. Chuva de Sousa Lopes, S. M. *et al.* X chromosome activity in mouse XX primordial germ cells. *PLoS Genet.* **4**, e30 (2008).
79. de Napoles, M., Nesterova, T. & Brockdorff, N. Early loss of Xist RNA expression and inactive X chromosome associated chromatin modification in developing primordial germ cells. *PLoS ONE* **2**, e860 (2007).
80. Speed, R. M. Oocyte development in XO foetuses of man and mouse: the possible role of heterologous X-chromosome pairing in germ cell survival. *Chromosoma* **94**, 115–124 (1986).
81. Celeste, A. *et al.* Histone H2AX phosphorylation is dispensable for the initial recognition of DNA breaks. *Nature Cell Biol.* **5**, 675–679 (2003).
82. Wang, W. H. & Sun, Q. Y. Meiotic spindle, spindle checkpoint and embryonic aneuploidy. *Front. Biosci.* **11**, 620–636 (2006).
83. Petronczki, M., Siomos, M. F. & Nasmyth, K. Un menage a quatre: the molecular biology of chromosome segregation in meiosis. *Cell* **112**, 423–440 (2003).
84. Yamamoto, A. *et al.* Spindle checkpoint activation at meiosis I advances anaphase II onset via meiosis-specific APC/C regulation. *J. Cell Biol.* **182**, 277–288 (2008).
85. Ashley, T., Ried, T. & Ward, D. C. Detection of nondisjunction and recombination in meiotic and postmeiotic cells from XY^{Sxr} [XY, Tp(Y)1Ct] mice using multicolor fluorescence *in situ* hybridization. *Proc. Natl Acad. Sci. USA* **91**, 524–528 (1994).
86. Burgoyne, P. S., Mahadevaiah, S. K., Sutcliffe, M. J. & Palmer, S. J. Fertility in mice requires X-Y pairing and a Y-chromosomal 'spermiogenesis' gene mapping to the long arm. *Cell* **71**, 391–398 (1992).
87. Evans, E. P., Burtenshaw, M. D. & Cattanach, B. M. Meiotic crossing over between the X and Y chromosomes of male mice carrying the sex-reversing (*Sxr*) factor. *Nature* **300**, 443–445 (1982).
88. Odorisio, T., Rodriguez, T. A., Evans, E. P., Clarke, A. R. & Burgoyne, P. S. The meiotic checkpoint monitoring synapsis eliminates spermatocytes *via* p53-independent apoptosis. *Nature Genet.* **18**, 257–261 (1998).
89. Eaker, S., Pyle, A., Cobb, J. & Handel, M. A. Evidence for meiotic spindle checkpoint from analysis of spermatocytes from Robertsonian-chromosome heterozygous mice. *J. Cell Sci.* **114**, 2953–2965 (2001).
90. Vogt, E., Kirsch-Volders, M., Parry, J. & Eichenlaub-Ritter, U. Spindle formation, chromosome segregation and the spindle checkpoint in mammalian oocytes and susceptibility to meiotic error. *Mutat. Res.* **651**, 14–29 (2008).
91. Hunt, P., LeMaire, R., Embury, P., Sheean, L. & Mroz, K. Analysis of chromosome behaviour in intact mammalian oocytes: monitoring the segregation of a univalent chromosome during female meiosis. *Hum. Mol. Genet.* **4**, 2007–2012 (1995).
92. LeMaire-Adkins, R., Radke, K. & Hunt, P. A. Lack of checkpoint control at the metaphase/anaphase transition: a mechanism of meiotic nondisjunction in mammalian females. *J. Cell Biol.* **139**, 1611–1619 (1997).
93. Hodges, C. A., LeMaire-Adkins, R. & Hunt, P. A. Coordinating the segregation of sister chromatids during the first meiotic division: evidence for sexual dimorphism. *J. Cell Sci.* **114**, 2417–2426 (2001).
94. Hodges, C. A., Revenkova, E., Jessberger, R., Hassold, T. J. & Hunt, P. A. SMC1β-deficient female mice provide evidence that cohesins are a missing link in age-related nondisjunction. *Nature Genet.* **37**, 1351–1355 (2005).
95. Hunt, P. A. & Hassold, T. J. Human female meiosis: what makes a good egg go bad? *Trends Genet.* **24**, 86–93 (2008).
96. Jones, K. T. Meiosis in oocytes: predisposition to aneuploidy and its increased incidence with age. *Hum. Reprod. Update* **14**, 143–158 (2008).
97. Kouznetsova, A., Lister, L., Nordenskjold, M., Herbert, M. & Hoog, C. Bi-orientation of achiasmatic chromosomes in meiosis I oocytes contributes to aneuploidy in mice. *Nature Genet.* **39**, 966–968 (2007).
98. Lyon, M. F. & Hawker, S. G. Reproductive lifespan in irradiated and unirradiated chromosomally XO mice. *Genet. Res.* **21**, 185–194 (1973).
99. Woods, L. M. *et al.* Chromosomal influence on meiotic spindle assembly: abnormal meiosis I in female *Mlh1* mutant mice. *J. Cell Biol.* **145**, 1395–1406 (1999).
100. Zickler, D. & Kleckner, N. Meiotic chromosomes: integrating structure and function. *Annu. Rev. Genet.* **33**, 603–754 (1999).
101. Moses, M. J. & Poorman, P. A. in *Chromosomes Today* Vol. 8 (eds Bennet, M. D., Gropp, A. & Wolf, U.) 80–103 (Allen and Unwin, London, 1984).
102. Baudat, F., Manova, K., Yuen, J. P., Jasin, M. & Keeney, S. Chromosome synapsis defects and sexually dimorphic meiotic progression in mice lacking Spo11. *Cell* **6**, 989–998 (2000).
103. Wang, X. & Haber, J. E. Role of *Saccharomyces* single-stranded DNA-binding protein RPA in the strand invasion step of double-strand break repair. *PLoS Biol.* **2**, E21 (2004).

DATABASES

Entrez Gene:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
Spo11
OMIM:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
ATM | ATR | BRCA1 | H2AX

FURTHER INFORMATION

Burgoyne group homepage:
http://www.nimr.mrc.ac.uk/devgen/burgoyne
Turner group homepage:
http://www.nimr.mrc.ac.uk/devgen/turner

ALL LINKS ARE ACTIVE IN THE ONLINE PDF