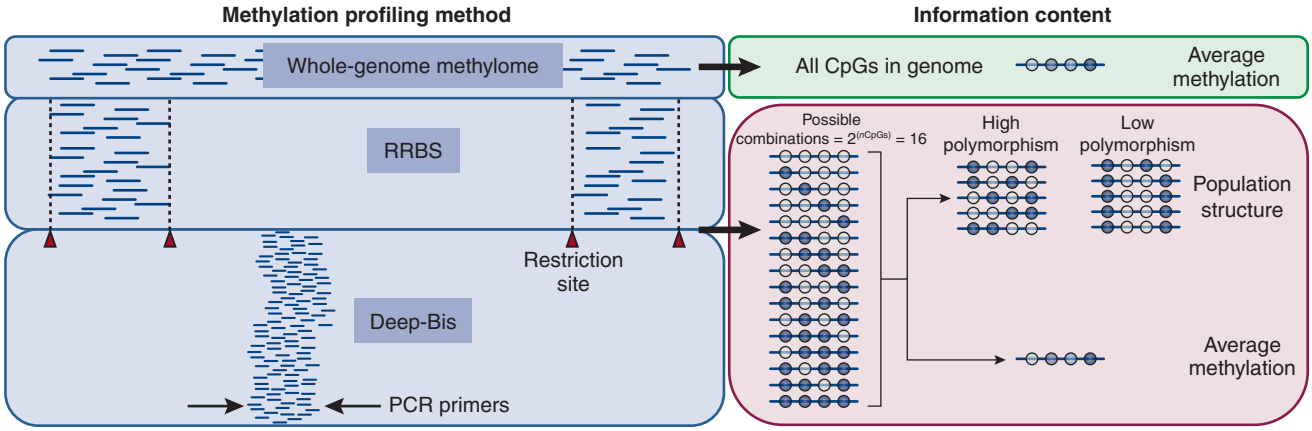

Figure 1 The ability to detect polymorphic methylation patterns depends on the depth of coverage with which a particular region is sequenced. Illustrated on the left are representative sequencing depths. The sensitivity of different profiling methods, which differ with regard to how much of the genome is analyzed, are shown on the right. Note that the sampling distribution of target genomic regions is unbiased for whole-genome bisulfite sequencing (Bis-seq), limited to defined restriction sites for reduced representation bisulfite sequencing (RRBS)[13] and flexible for bisulfite amplicon sequencing (Deep-Bis).

transformation are not the outcome of orchestrated reprogramming of the epigenome but rather of a passive process that progressively invades sensitive regions that are commonly classified as hypermethylated in cancer. A key remaining question is how the results obtained from this *in vitro* system translate into actual carcinogenesis *in vivo*. The study by Tanay and colleagues should provide a framework for experiments addressing this question.

In addition to being relevant for cancer biology, these results further relate to current models that describe how methylation patterns are regulated. One possible explanation for the recurrence of a spatially constrained initiation point of hypermethylation is that loss of binding of a particular factor mediates protection against methylation at this site. Indeed, it has been observed that regions bound by transcription factors are methylation depleted[6–11], and a direct link between lack of protection of certain regions and hypermethylation arising during cellular transformation has been suggested[12]. Although it remains to be determined whether this indeed explains the dynamics observed by Tanay and colleagues, it is evident that tracking at high sequencing depth the dynamics of DNA methylation during normal development and pathogenesis should not only lead to quantitative models of epigenome evolution but may also guide mechanistic studies of the underlying biological process.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Feinberg, A.P. *et al. Nature* **301**, 89–92 (1983).
2. Baylin, S.B. *et al. Nat. Rev. Cancer* **11**, 726–734 (2011).
3. Berman, B.P. *et al. Nat. Genet.* **44**, 40–46 (2012).
4. Adams, D. *et al. Nat. Biotechnol.* **30**, 224–226 (2012).
5. Landan, G. *et al. Nat. Genet.* **44**, 1207–1214 (2012).
6. Stadler, M.B. *et al. Nature* **480**, 490–495 (2011).
7. Wiench, M. *et al. EMBO J.* **30**, 3028–3039 (2011).
8. Xu, J. *et al. Proc. Natl. Acad. Sci. USA* **104**, 12377–12382 (2007).
9. Hodges, E. *et al. Mol. Cell* **44**, 17–28 (2011).
10. Lienert, F. *et al. Nat. Genet.* **43**, 1091–1097 (2011).
11. Brandeis, M. *et al. Nature* **371**, 435–438 (1994).
12. Gebhard, C. *et al. Cancer Res.* **70**, 1398–1407 (2010).
13. Meissner, A. *et al. Nucleic Acids Res.* **33**, 5868–5877 (2005).

Older males beget more mutations

Matthew Hurles

Three papers characterizing human germline mutation rates bolster evidence for a relatively low rate of base substitution in modern humans and highlight a central role for paternal age in determining rates of mutation. These studies represent the advent of a transformation in our understanding of mutation rates and processes, which may ultimately have public health implications.

Under the collective banner of DNA sequence mutation lie a handful of different mutational processes, each of which can be linked to a fundamental cellular process, such as DNA replication, repair or recombination. The mutational process that is dominant in terms of the number of new alleles introduced each generation is base substitution, with an average rate on the order of 1 mutation for every 100 million bases. By contrast, replication slippage at simple tandem repeats occurs at rates that are four to five orders of magnitude higher. Three recent papers[1–3], including one by Campbell *et al.*[1] on page 1277 of this issue, report estimates of the human germline sequence mutation rate, each making use of improved cost-effective genome sequencing technologies but taking different approaches.

Matthew Hurles is at the The Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK.
e-mail: meh@sanger.ac.uk

Measuring mutation rate

A range of experimental approaches have been applied to measure the germline mutation rate, from counting the numbers of new mutations seen in gametes to inferring the number of new mutations that have arisen between two species separated by millions of years of evolution. Strategies based on observing only several generations are limited primarily by the sensitivity and specificity with which new mutations can be identified, whereas those based on

evolutionary comparisons between species are hindered primarily by uncertainty in the estimates of the number of generations over which mutations have arisen.

The three new studies[1–3] span this spectrum of experimental strategies, and each represents a substantial technical achievement in its own right. Kong *et al.*[2] applied high-coverage whole-genome sequencing to 78 parent-offspring trios from Iceland and identified 4,933 potential *de novo* single-nucleotide variants (SNVs). Validation of a small subset (n = 94) of these SNVs indicated an impressive false positive rate of only ~1%. Such specificity might raise concerns of incomplete sensitivity, but the authors estimated a ~2% false negative rate, albeit indirectly and only considering one potential source of false negatives.

Campbell *et al.*[1] identified *de novo* base substitutions based on the whole-genome sequencing of five parent-offspring trios and genotyping data from an extended pedigree. Campbell *et al.* made use of large pedigrees within the well-defined Hutterite founder population, allowing them to identify genomic regions showing inheritance of the same ancestral haplotype from both parents, the common ancestor from whom the haplotype descended and the number of generations separating the two haplotypes. This enabled estimation of a mutation rate from the small numbers of new mutations that distinguish the two closely related haplotypes. This strategy has the technical advantage that most new mutations that have arisen on the ancestral haplotype are observed in both the child and the transmitting parent, which helps to eliminate false positives.

Sun *et al.*[3] genotyped 2,477 autosomal microsatellites, motifs of 1–6 base pairs, in 85,289 individuals from Iceland and identified 2,058 germline mutations. They used this data set to derive an improved statistical model for the evolution of these highly variable microsatellite loci. By incorporating flanking sequence variation into their model and analyzing 23 individuals for whom both microsatellite and whole-genome sequence data were available, they estimated an average base substitution rate over recent human evolutionary history without calibration from the highly contentious fossil record. Sun *et al.* used their model to estimate key evolutionary parameters and attempted to resolve the ancestral relationships among key fossils from around the time of the human-chimpanzee split. However, when they instead used the sequence-based estimates from Kong *et al.*[2] (as described in a note added in proof), their model gave discordant results regarding whether *Sahelanthropus tchadensis* might lie on the human lineage since the split from chimpanzees.

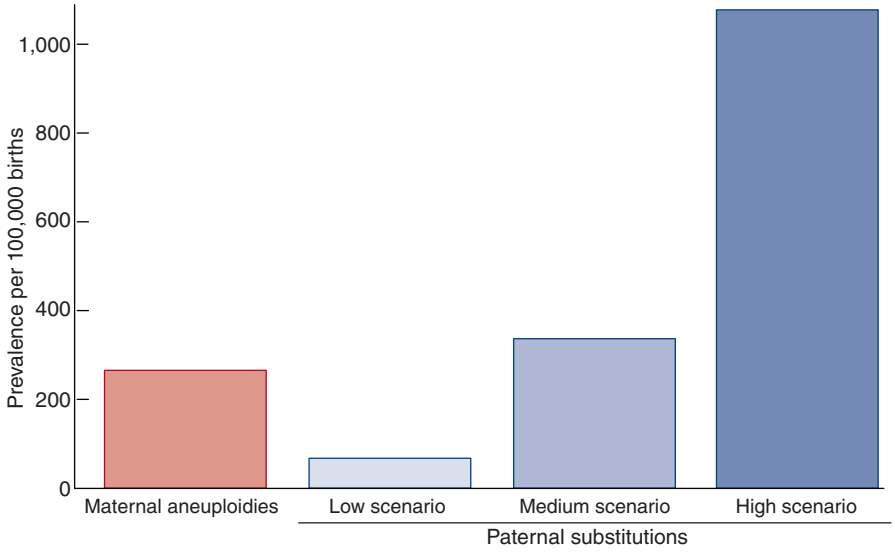

Figure 1 Comparing the birth prevalence of developmental disorders caused by maternal chromosomal aneuploidies and paternal *de novo* base substitutions. The cumulative birth prevalence of chromosomal aneuploidy syndromes of maternal origin (Down, Edwards and Patau syndromes) is obtained by combining maternal age-specific absolute risks[11] with the age distribution of English and Welsh mothers at childbirth[12]. The paternal age-specific numbers of base substitutions are taken from Kong *et al.*[2] and are combined with the age distribution of English and Welsh fathers at childbirth[12]. The disease burden of *de novo* base substitutions is assumed to result from loss-of-function mutations in haploinsufficient genes. The number of currently known haploinsufficient genes is >1% of known genes[13] and is likely considerably underascertained. The proportion of coding and splice-site base substitutions that result in truncating mutations is ~5% (ref. 14), but it has been estimated that ~30% of missense mutations are also likely to be highly evolutionary deleterious[15], the majority of which are likely due to loss of function. Three scenarios were considered, taking into account the proportion of haploinsufficient genes and genes with loss-of-function mutations: low, 2% haploinsufficient and 5% loss of function; medium, 5% haploinsufficient and 10% loss of function; high, 8% haploinsufficient and 20% loss of function.The total length of coding and splicing regions is assumed to be 35 Mb.

The base substitution mutation rates estimated in each of these studies are broadly consistent, although, because of considerable variation in the manner in which uncertainty was accounted for in the estimates, it is difficult to make like-for-like comparisons. All three studies bolster earlier reports[4,5] suggesting that the average mutation rate is about half the much-quoted rate derived from early human-chimpanzee genome comparisons of 2.5×10^{-8} substitutions per base per generation[6]. The two studies based on the identification of mutations in recent generations produced very similar sequence mutation rate estimates of 1.2×10^{-8}. The estimate by Sun *et al.* of the sequence mutation rate is slightly higher at $1.4–2.3 \times 10^{-8}$ but may not be directly comparable to the others, as this estimate is indirect (based on estimates from a microsatellite-based model) and considers an evolutionary longer timeframe (spanning a larger number of generations). Further work is required to resolve this slight discordance.

Paternal age and mutation rate

For over 60 years, the observation that the number of cell replications in the male germline increases with age due to the constant turnover of spermatogonial stem cells, whereas the number of replications in the female germline is fixed at birth, has motivated two hypotheses: (i) that the male germline is more mutagenic than the female and (ii) that mutation rates may increase with paternal but not maternal age[7].

Recent analyses of exome data in families with autism (for example, see ref. 8) have provided support for these two long-standing hypotheses, although they have been limited by the paucity of mutations detected in this small fraction of the genome. Both Kong *et al.*[2] and Campbell *et al.*[1] have now clearly demonstrated that the paternal germline is substantially more mutagenic than the maternal for base substitutions, with Kong *et al.* estimating that the male mutation rate is greater by a factor of 3.9. Sun *et al.* show that the paternal germline is also more mutagenic for replication slippage observed at microsatellites by a factor of 3.3. Kong *et al.* were also able to delineate an approximately linear increase in the numbers of new base substitutions with paternal but not maternal age. This linear increase of ~2 mutations per year is broadly in line with the expectation given the simple (perhaps simplistic) current model for

spermatogonial stem cell turnover that entails 23 replications per year after puberty[9].

These findings are important for comparative population-based studies, as the average mutation rate in the population is strongly influenced by the distribution of parental ages at birth, which are known to vary appreciably, both between contemporaneous populations and in the same population over time. Additional studies are required to investigate the rates of other mutational processes (for example, insertion-deletion events), further examine heterogeneity in mutation rates across the genome, estimate mutation rates in a broader range of populations (and species) and quantify the contribution of environmental and genetic factors to variation in mutation rates between individuals.

Developmental disorders

Advanced maternal age has long been recognized as a major risk factor for developmental disorders resulting from chromosomal aneuploidies (for example, Down syndrome), which has motivated additional prenatal screening for older mothers in some healthcare systems, and we may now consider the relative impact of *de novo* base substitutions of paternal origin in the same light. **Figure 1** shows a preliminary comparison of the prevalence of developmental disorders derived from each parent, which represents the considerable uncertainty associated with the disease burden resulting from paternal base substitutions in the form of credible low, medium and high scenarios. This suggests that the birth prevalence of developmental disorders caused by paternal base substitutions may be at least comparable to that caused by maternal chromosomal trisomies. This analysis can be refined as further understanding is gained of the contribution of *de novo* mutations in developmental disorders and the consequences of increased paternal and maternal age on other mutational processes (for example, see ref. 10). However, it can clearly be appreciated that this revolution in the understanding of mutational processes may well have a broader impact on prenatal screening strategies and on public perception of the consequences of advanced parental age.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Campbell, C.D. *et al. Nat. Genet.* **44**, 1277–1281 (2012).
2. Kong, A. *et al. Nature* **488**, 471–475 (2012).
3. Sun, J.X. *et al. Nat. Genet.* **44**, 1161–1165 (2012).
4. Conrad, D.F. *et al. Nat. Genet.* **43**, 712–714 (2011).
5. Roach, J.C. *et al. Science* **328**, 636–639 (2010).
6. Nachman, M.W. & Crowell, S.L. *Genetics* **156**, 297–304 (2000).
7. Haldane, J.B.S. *Ann. Eugen.* **13**, 262–271 (1947).
8. O'Roak, B.J. *et al. Nature* **485**, 246–250 (2012).
9. Drost, J.B. & Lee, W.R. *Environ. Mol. Mutagen.* **25** Suppl 26, 48–64 (1995).
10. Hehir-Kwa, J.Y. *et al. J. Med. Genet.* **48**, 776–778 (2011).
11. Savva, G.M., Walker, K. & Morris, J.K. *Prenat. Diagn.* **30**, 57–64 (2010).
12. Office for National Statistics. *Review of the National Statistician on Births and Patterns of Family Building in England and Wales, 2008* (Office for National Statistics, Newport, UK, 2009).
13. Huang, N., Lee, I., Marcotte, E.M. & Hurles, M.E. *PLoS Genet.* **6**, e1001154 (2010).
14. Kryukov, G.V., Pennacchio, L.A. & Sunyaev, S.R. *Am. J. Hum. Genet.* **80**, 727–739 (2007).
15. Boyko, A.R. *et al. PLoS Genet.* **4**, e1000083 (2008).

FOXA1 and breast cancer risk

Kerstin B Meyer & Jason S Carroll

Many SNPs associated with human disease are located in non-coding regions of the genome. A new study shows that SNPs associated with breast cancer risk are located in enhancer regions and alter binding affinity for the pioneer factor FOXA1.

Substantial effort has been invested in finding genetic variants that are associated with breast cancer risk. The majority of risk-associated SNPs do not occur in gene-coding sequences, complicating attempts at functional investigation. A new study in this issue by Mathieu Lupien and colleagues shows that the bulk of SNPs associated with breast cancer risk occur at distant enhancer regions and change the binding capacity of FOXA1, a protein required for estrogen receptor-α (ER) function[1].

Disease risk and non-coding regions of the genome

ER is a transcription factor that is expressed in almost three-quarters of all breast cancers. It is the major driving transcription factor in luminal breast cancers and is one of the key targets of endocrine therapies. ER protein dimers regulate gene expression by associating with distant enhancer regions[2] that form chromatin loops with the promoters of target genes[3,4]. Genomic mapping of ER-binding sites in breast cancer showed that an additional protein, FOXA1, also binds at the same enhancer regions, where it functions as a pioneer factor to mediate ER association with compacted DNA[2,5]. Pioneer factors can physically associate with compacted chromatin and facilitate binding of other transcription factors. In the absence of FOXA1, ER cannot interact with DNA, and ER-mediated gene expression is prevented[2].

One of the first genome-wide association studies (GWAS) was conducted for breast cancer and showed that most disease-associated SNPs are located in non-coding regions of the genome[6]. Follow-up studies of individual loci have shown that causative SNPs affect transcription factor binding[7] and the activity of long-range enhancers[8], but a global understanding of the function of non-coding risk-associated SNPs has remained elusive.

Functional consequences of non-coding SNPs

On page 1191 of this issue, Cowper-Sal·lari *et al.*[1] report a systematic approach aimed at identifying potential functions of risk-associated SNPs[1]. They defined a set of risk-associated SNPs including those identified in breast cancer GWAS and the adjacent SNPs that are in linkage disequilibrium. They integrated these risk loci variant sets with genomic data sets of histone modification profiles and transcription factor binding profiles from breast cancer cells. The goal of this data integration was to find transcription factor binding or histone modification that correlates, on a global scale, with the risk loci. They found that histone 3 lysine 4 monomethylation (H3K4me1) marks are correlated with regions encompassing the risk-associated SNPs. Similarly, they found two transcription factors that have binding sites that are also enriched in the risk-associated SNP regions, ER and FOXA1, the two major components of the ER-DNA interaction complex[2,9]. This enrichment seems to

Kerstin B. Meyer and Jason S. Carroll are at Cancer Research UK, Cambridge, UK, and the Department of Oncology, University of Cambridge, Cambridge, UK.
e-mail: jason.carroll@cancer.org.uk