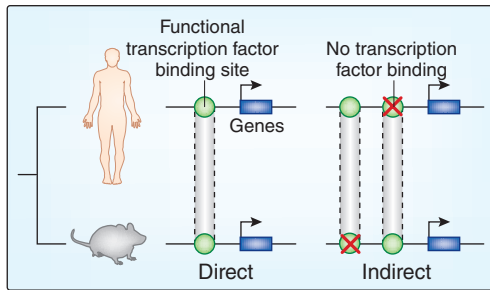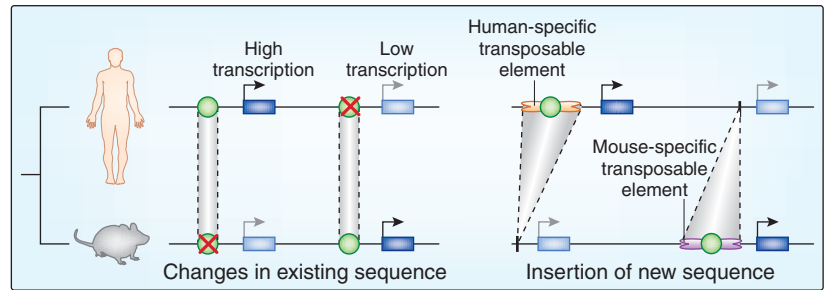**a** Regulatory conservation

b Regulatory rewiring

Figure 1 Conservation and rewiring of regulatory elements between human and mouse genomes. For a given transcription factor, binding site differences between species can have different effects on the transcriptional output of individual genes, resulting in either regulatory conservation or rewiring. (**a**) Transcriptional output is preserved (regulatory conservation) either through direct sequence conservation between species or through indirect conservation by creation of one binding site and destruction of another. (**b**) Transcriptional output is altered (regulatory rewiring) when binding sites are destroyed or created without compensatory nearby changes, or when new binding sites are created by insertion and exaptation of transposable elements.

profiles of major ES cell transcription factors. The study also provides insights into the unexpectedly large role that local binding site turnover, as well as RABS, play in the conservation and rewiring of mammalian regulatory networks. The functional relevance of the new human-specific OCT4 target genes identified by Kunarso *et al.*[1] remains to be determined, but they provide important leads for future studies.

Although sequence conservation has proven useful as a predictor of functional regulatory elements in the genome[2,10], the observations by Kunarso *et al.*[1] are a reminder that it is not justified to assume in turn that all functional regulatory elements show evidence of sequence constraint. It is noteworthy that

whereas OCT4 binding and NANOG binding diverged between human and mouse ES cells, binding of CTCF was highly conserved. Thus, it is expected that other DNA-binding proteins and chromatin marks will fall into a spectrum from strong to weak conservation between these two species. The notion that some regulatory networks have substantially changed in evolution is also supported by recent independent observations of lineage-specific network rewiring in vertebrate preimplantation embryos and adult liver tissue[11,12]. The differences between species identified through these studies highlight the need to complement comparative genomic data with experimental approaches in order to obtain an accurate functional annotation of genomes.

1. Kunarso, G. *et al. Nat. Genet.* **42**, 631–634 (2010).
2. Cooper, G.M. & Brown, C.D. *Genome Res.* **18**, 201–205 (2008).
3. King, D.C. *et al. Genome Res.* **15**, 1051–1060 (2005).
4. Visel, A. *et al. Nature* **457**, 854–858 (2009).
5. ENCODE Project Consortium *et al. Nature* **447**, 799–816 (2007).
6. Bourque, G. *et al. Genome Res.* **18**, 1752–1762 (2008).
7. Wang, T. *et al. Proc. Natl. Acad. Sci. USA* **104**, 18613–18618 (2007).
8. Bejerano, G. *et al. Nature* **441**, 87–90 (2006).
9. Lowe, C.B., Bejerano, G. & Haussler, D. *Proc. Natl. Acad. Sci. USA* **104**, 8005–8010 (2007).
10. Visel, A., Rubin, E.M. & Pennacchio, L.A. *Nature* **461**, 199–205 (2009).
11. Xie, D. *et al. Genome Res.* **20**, 804–815 (2010).
12. Schmidt, D. *et al. Science* **328**, 1036–1040 (2010).

Hints of hidden heritability in GWAS

Greg Gibson

Although susceptibility loci identified through genome-wide association studies (GWAS) typically explain only a small proportion of the heritability, a classical quantitative genetic analysis now argues that considering together all common SNPs can explain a large proportion of the heritability of these complex traits. A related study provides recommendations for the sample sizes needed in future GWAS to identify additional susceptibility loci.

Genome-wide association studies (GWAS) have been highly successful in identifying genetic variants associated with hundreds of complex human traits and diseases, a feat that eluded two decades of linkage mapping. At the same time, the variants identified in GWAS generally only capture a few percent of the estimated heritability for these complex traits, leaving open

Greg Gibson is in the School of Biology, Georgia Institute of Technology, Atlanta, Georgia, USA. email: greg.gibson@biology.gatech.edu

the question of what may explain the remaining heritability. This may include contributions of rare variants, epistasis, epigenetics and genotype–environment interactions[1,2] but may also just imply that complex traits truly are affected by thousands of variants of small effect[3]. On page 565 of this issue, Peter Visscher and colleagues[4] report a joint estimate of the contribution of SNPs across all effect sizes and show that this can explain a large proportion of the heritability for height. On page 570 of this issue, Nilanjan Chatterjee and colleagues[5]

examine recent GWAS for several complex traits and estimate the potential contribution of variants whose effect size is similar to that of discovered SNPs for diverse diseases. These two reports do not attempt to identify further genetic variants that could explain the remaining heritability, but they mount compelling arguments that a large proportion of it can be explained by common variants.

In GWAS, we test for association between the frequency of each of hundreds of thousands of common variants and a given phenotype, call

significant SNPs that exceed a conservative genome-wide threshold for association (usually $P < 5 \times 10^{-8}$), and then test these for evidence of replication in independent cohorts. This necessarily stringent approach misses some, if not most, causal SNPs. It is unlikely that GWAS will ever be sufficiently powered to uncover even the majority of the heritability, but absence of evidence does not imply evidence of absence. For this reason, it is important to evaluate how much of the heritability can, in theory, be captured by common variants: if the answer is less than, say, half, then we have a genuine missing heritability problem; otherwise, this may suggest that most of the genetic variance is simply hidden below the threshold for genome-wide significant associations (**Fig. 1a**).

Inferring combined effects

Yang *et al.*[4] take a classical quantitative genetic approach to this problem by using genome-wide genotypes to infer the genetic relatedness of individuals and then estimate the relationship of this quantity to phenotypic similarity, thus arriving at an estimate of heritability explained[6]. Their model complex trait is height, which has a heritability of ~80%, only approximately 5% of which has been explained by 50 common polymorphisms identified by GWAS[7]. Starting with a Brisbane, Australia cohort of 3,925 unrelated individuals (who have less genetic relatedness than typical second cousins), they in effect linearly regress all of the SNPs from their GWAS onto an appropriately adjusted measure of height.

The raw estimate this produces is that 45% of the total variation for height (and hence 56% of the heritability) in the Brisbane cohort can be explained by common SNPs[4]. There are calibration errors that must be adjusted for, and the likelihood framework Yang *et al.*[4] adopt allows them to do this in a stepwise manner. By accounting for imperfect linkage disequilibrium between tagging and causal variants, Yang *et al.*[4] up their estimate of common variant effects to at least 67% of the genetic contribution to height. A further correction assumes that causal variants are likely to be deleterious and hence at lower frequency than tagging SNPs, and in this case leads to the conclusion that most of the heritability for height can be captured by common variants.

This presents an elegant argument that most of the heritability is hidden rather than missing and, hence, that there is no pressing need to invoke more complex genetic mechanisms to explain height. Whether or not this is also true of susceptibility to common diseases remains to be determined. Furthermore, there is the caveat that this interpretation of the re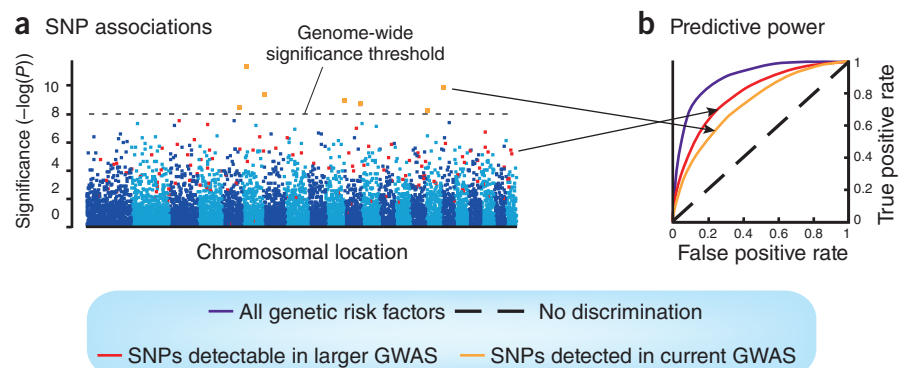sult is a parsimonious one and that a variety of genetic models could also produce similar regressions of phenotypic on genetic relatedness.

Figure 1 Predictive power of genetic variants identified through genome-wide association studies. (**a**) GWAS typically report a visualization of genome-wide association of SNPs using a Manhattan plot of significance against chromosomal location, shown here for a generic common disease. Gold, SNPs detected in current GWAS; red, SNPs that may be detectable in GWAS with larger sample sizes. (**b**) SNPs identified through GWAS as significantly associated with disease susceptibility may be used in a genetic predictive test to classify disease risk in individual. Using SNPs identified from current GWAS increases the area under the curve (AUC), reflecting increased ability to classify risk (gold line) over nondiscrimination (dashed black line). Discovery of additional associated SNPs through larger GWAS (red curve) further increases the AUC, improving the genetic test, but does not approach what would be theoretically possible with full characterization of all genetic variation influencing the disease (purple).

Estimating effect sizes

Park *et al.*[5] ask a more practical question, namely what size GWAS would be required to uncover more risk variants that have similar magnitudes of effect to those uncovered in initial studies. They reexamine recent GWAS for height, Crohn's disease and cancers of the breast, prostate and colorectum to estimate the overall distribution of the effect sizes of associated SNPs, generally with odds ratios in the range of 1.2 to 1.5. They then estimate the number of additional common variants with similar effects that may be expected to be found for these traits, as well as the sample sizes that would be needed to detect these. A clear disparity in the genetic architectures of these diseases emerges: in projections of GWAS including 50,000 cases, they estimate that 140 alleles would explain 20% of the sibling relative risk for Crohn's disease, whereas just 44 alleles would explain 15% of each of the cancers.

The power studies of Park *et al.*[5] suggest that larger GWAS, with on the order of hundreds of thousands of cases, will yield many more validated loci for complex traits. Although this will increase the genetic variance explained, it does not in general seem likely to capture even half of the heritability estimated from sibling relative risks. For the remaining variance, the results of Yang *et al.*[4] suggest that SNPs with smaller effects could collectively add a substantial genetic contribution, although they will remain individually difficult to detect within GWAS and, hence, hidden.

Genetic prediction

Whether it is worthwhile to continue searching for variants with such small effects remains debatable. The International Schizophrenia Consortium[8] found that a collection of thousands of nominally significant SNPs collectively capture over one-third of the heritability for schizophrenia (a disease that has proven particularly refractory to the discovery of large effect alleles), but these only explain a few percent of disease susceptibility in replication cohorts. Sampling variance obscures the estimation of marginal effects in models with numerous genetic variables, so although we can infer that the effects are present, it remains impossible to identify them with any accuracy.

Similarly, Park *et al.*[5] present sobering results, consistent with similar analyses by others[9], that the increases in the proportion of variance explained by GWAS-significant SNPs do not go hand in hand with improved genetic prediction of disease risk. This is measured as sensitivity (true positive rate, reflecting the ability to predict disease) and specificity (false positive rate, reflecting the avoidance of false assignment of risk), often plotted in a receiver operating characteristic (ROC) curve, with the area under the curve taken as a measure of predictive performance (**Fig. 1b**). SNPs discovered in a GWAS can subsequently be used to predict disease risk in an independent population, and this typically provides some level of improvement in classifying risk over random decision making. The identification of many more associated SNPs through larger GWAS will improve prediction, but this does not seem likely to approach what could be obtained given a complete description of all genetic risk factors. Note that although the results

of Yang *et al.*[4] show that joint consideration of all common SNPs increases the proportion of heritability explained, their method does not lead directly to an improved genetic predictor of risk because it does not identify individual loci. In addition, the predictive power of genetic tests will always be constrained by the environmental contributions influencing susceptibility to that particular disease.

Following on these considerations, it seems increasingly unlikely that for most complex diseases genetic prediction alone will ever reach the accuracy traditionally regarded as clinically informative for the population at large. However, these genetic discoveries are clearly bringing genetics into the range of informative stratification that may be useful for individuals in at-risk groups and that will often exceed prediction from family history[10].

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Manolio, T.A. *et al. Nature* **461**, 747–753 (2009).
2. Eichler, E.E. *et al. Nat. Rev. Genet.* **11**, 446–450 (2010).
3. Fisher, R.A. *Trans. Royal Soc. Edinburgh* **52**, 399–433.
4. Yang, J. *et al. Nat. Genet.* **42**, 565–569 (2010).
5. Park, J. *et al. Nat. Genet.* **42**, 570–575 (2010).
6. Haseman, J.K. & Elston, R.C. *Behav. Genet.* **2**, 2–19 (1972).
7. Weedon, M.N. & Freyling, T.M. *Trends Genet.* **24**, 595–603 (2008).
8. Purcell, S.M. *et al. Nature* **460**, 748–752 (2009).
9. Evans, D.M., Visscher, P.M. & Wray, N.R. *Hum. Mol. Genet.* **18**, 3525–3531 (2009).
10. Wray, N.R., Yang, J., Goddard, M.E. & Visscher, P.M. *PLoS Genet.* **6**, e1000864 (2010).

Fungal pathogenicity and morphological switches

P T Magee

The virulence of *Candida albicans*, a major human fungal pathogen, has been considered dependent on the ability to transition between different morphologies. A new study reports a screen of *C. albicans* mutants that demonstrates that pathogenesis can be dissociated from morphological switching and *in vitro* growth rate.

The identification of virulence factors for a fungal pathogen is complicated by the dependence of infection on both the pathogen and host. *C. albicans* adds an extra layer of complexity, as its ability to infect a host has been related to several factors, including a fast growth rate and its ability to switch between a yeast, a hyphal form, a pseudohypha (germ tube) or, under some conditions, a biofilm morphology (**Fig. 1**). Dissecting the genetic factors that influence virulence in *C. albicans* has also been impeded by this pathogen's diploid genome, chromosomal instability and incomplete sexual cycle. On p. 590 of this issue, Suzanne Noble and colleagues report a systematic screen to examine whether the morphological switching capabilities and growth rate of *C. albicans* are essential for pathogenicity[1].

Virulence without switching
Noble *et al.*[1] disrupted 674 candidate genes likely to be involved in pathogenicity, covering ~10% of the *C. albicans* genome and generating ~3,000 homozygous deletion strains[2]. Because *C. albicans* is diploid, each mutant was constructed in two steps, using two markers (*HIS1* and *LEU2*) in sequential transformations to inactivate both alleles of each targeted gene in an auxotrophic strain. A unique DNA sequence tag was incorporated into each mutant to allow signature-tagged mutagenesis[3]. Noble *et al.*[1] assessed virulence

P. T. Magee is in the Department of Genetics, Cell Biology and Development, University of Minnesota, Minneapolis, Minnesota, USA.
e-mail: magee@umn.edu

by means of a commonly used mouse model of systemic infection, tail-vein injection of *C. albicans*. They inoculated mice with pools of 48 different mutants and then measured the fungal load in the kidney at the end of the experiment, classifying a given mutant as attenuated in infectivity if its frequency in the kidney at the end of the experiment was lower than its frequency in the inoculum. They found that, of 103 identified infectivity-attenuated mutants, 48 were still able to carry out morphological transitions. And of the 133 mutants unable to form germ tubes, 80 were scored as pathogenic in the mouse infectivity model. In addition, 42 of 68 mutants with low proliferation rates showed normal infectivity, suggesting that virulence was also not tightly correlated with *in vitro* growth rate.

The finding of Noble *et al.*[1] that pathogenesis can be dissociated from morphological switching and growth rate is surprising because previously the loss of morphological switching had been accompanied by diminished virulence in almost all of the mutants studied. Noble *et al.*[1] suggest that these previous studies may have been confounded by use of the marker *URA3*, a biosynthetic gene essential for virulence and for morphological switching. Differential marker activity depending on chromosomal position would affect both morphological switching and virulence[4]. In addition, many of the mutations previously used to study morphological switching targeted major signaling pathways, such as MAP kinases[5] and protein kinase A[6], which participate in cell functions other than morphological changes, including some functions that are important
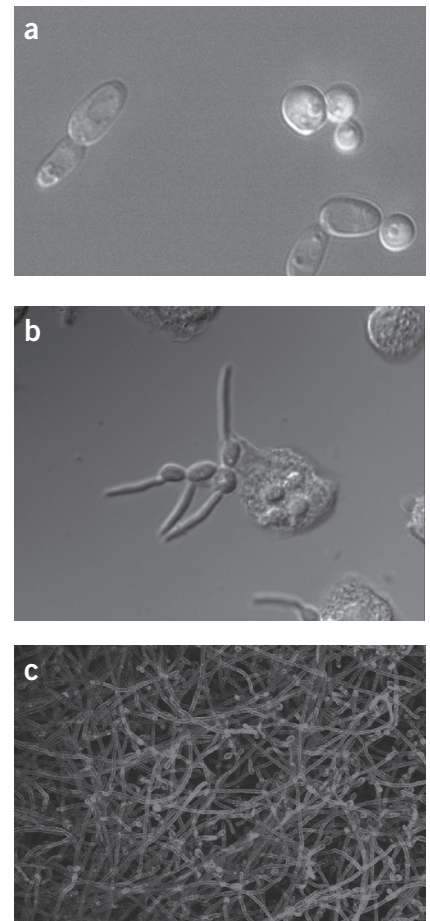
Figure 1 Various morphological forms of *C. albicans*. (**a**) Budding yeast morphology (photo from M. Lohse and R. Zordan). (**b**) Hyphal phase with elongated germ tubes; shown here with an engulfing macrophage (photo from M. Lorenz). (**c**) Biofilm (photo from C.J. Nobile).