upwards of 80% of the messenger RNA for a particular gene and still get a 'normal' plant, because mRNA levels are not always proportional to the amount of protein actually produced (or required). The possibility exists, therefore, that the alterations in gene expression observed in polyploids are compensated for at the translational level. Logic, however, suggests that this is probably not the case, because of the phenotypic success of most polyploids compared to their progenitors, and recent evidence from proteomic studies supports this assumption. In a study of synthetic *Brassica* allopolyploids, Albertin *et al.* [20] discovered a large number of proteins displaying non-additive changes to expression when compared to the parental taxa (305 in stem, 200 in root). They then undertook an *in silico* gene expression analysis to determine whether these proteins were restricted to a particular grouping, such as cellular function and/or localisation, and found that, as with genes identified in transcriptional expression studies, this was not the case. Similarly, Albertin *et al.* [20] found that different proteins within the same complex could have their expression altered in opposing directions — again, as observed in transcriptional studies.

Recent studies therefore suggest that the process of polyploidisation, whether incorporating a hybridisation event or not, has a large-scale impact on gene expression in the new individual, thereby providing raw material upon which selection can act. Small wonder, then, that polyploid species are so numerous — in evolutionary terms, it would seem that two genomes are indeed better than one.

References

1. Grant, V. (1981). Plant speciation, 2nd Edition (New York: Columbia University Press).
2. Liu, Z., and Adams, K.L. (2007). Expression partitioning between genes duplicated by polyploidy under abiotic stress and during organ development. Curr. Biol. *17*, 1669–1674.
3. Adams, K.L., Cronn, R., Percifield, R., and Wendel, J.F. (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc. Natl. Acad. Sci. USA *100*, 4649–4654.
4. Adams, K.L., Percifield, R., and Wendel, J.F. (2004). Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. Genetics *168*, 2217–2226.
5. Rieseberg, L.H., Raymond, O., Rosenthal, D.M., Lai, Z., Livingstone, K., Nakazato, T., Durphy, J.L., Schwarzbach, A.E., Donovan, L.A., and Lexer, C. (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. Science *301*, 1211–1216.
6. Hegarty, M.J., Jones, J.M., Wilson, I.D., Barker, G.L., Coghill, J.A., Sanchez-Baracaldo, P., Liu, G., Buggs, R.J.A., Abbott, R.J., Edwards, K.J., *et al*. (2005). Development of anonymous cDNA microarrays to study changes to the *Senecio* floral transcriptome during hybrid speciation. Mol. Ecol. *14*, 2493–2510.
7. Hegarty, M.J., Barker, G.L., Wilson, I.D., Abbott, R.J., Edwards, K.J., and Hiscock, S.J. (2006). Transcriptome shock after interspecific hybridization in *Senecio* is ameliorated by genome duplication. Curr. Biol. *16*, 1652–1659.
8. Wang, J., Tian, L., Lee, H.-S., Wei, N.E., Jiang, H., Watson, B., Madlung, A., Osborn, T.C., Doerge, R.W., Comai, L., *et al*. (2006). Genomewide nonadditive gene regulation in *Arabidopsis* allotetraploids. Genetics *172*, 507–517.
9. Riddle, N.C., and Birchler, J.A. (2003). Effects of reunited diverged regulatory hierarchies in allopolyploids and species hybrids. Trends Genet. *19*, 597–600.
10. Hammerle, B., and Ferrus, A. (2003). Expression of enhancers is altered in *Drosophila melanogaster* hybrids. Ecol. Dev. *5*, 221–230.
11. Rieseberg, L.H., Sinervo, B., Linder, C.R., Ungerer, M.C., and Arias, D.M. (1996). Role of gene interactions in hybrid speciation: evidence from ancient and experimental hybrids. Science *272*, 741–745.
12. Wang, J., Tian, L., Lee, H.S., and Chen, Z.J. (2006). Nonadditive regulation of FRI and FLC loci mediates flowering-time variation in *Arabidopsis* allopolyploids. Genetics *173*, 965–974.
13. Madlung, A., Masuelli, R.W., Watson, B., Reynolds, S.H., Davison, J., and Comai, L. (2002). Remodeling of DNA methylation and phenotypic and transcriptional changes in synthetic *Arabidopsis* allotetraploids. Plant Physiol. *129*, 733–746.
14. Salmon, A., Ainouche, M.L., and Wendel, J.F. (2005). Genetic and epigenetic consequences of recent hybridization and polyploidy in *Spartina* (Poaceae). Mol. Ecol. *14*, 1163–1175.
15. Lukens, L.N., Pires, J.C., Leon, E., Vogelzang, R., Oslach, L., and Osborn, T. (2006). Patterns of sequence loss and cytosine methylation within a population on newly resynthesized *Brassica napus* allopolyploids. Plant Physiol. *140*, 336–348.
16. Chen, Z.J., and Tian, L. (2007). Roles of dynamic and reversible histone acetylation in plant development and polyploidy. Biochem. Biophys. Acta *1769*, 295–307.
17. Matzke, M., Kanno, T., Huettel, B., Daxinger, L., and Matzke, A.J.M. (2006). RNA-directed DNA methylation and Pol IVb in *Arabidopsis*. Cold Spring Harb. Symp. Quant. Biol. *71*, 449–459.
18. Pikaard, C.S. (2000). The epigenetics of nucleolar dominance. Trends Genet. *16*, 495–500.
19. Liu, B., Brubaker, C.L., Cronn, R.C., and Wendel, J.F. (2001). Polyploid formation in cotton is not accompanied by rapid genomic changes. Genome *44*, 321–330.
20. Albertin, W., Alix, K., Balliau, T., Brabant, P., Davanture, M., Malosse, C., Valot, B., and Theillement, H. (2007). Differential regulation of gene products in newly synthesized *Brassica* allotetraploids is not related to protein function nor subcellular localization. BMC Genomics *8*, 56–70.

School of Biological Sciences, University of Bristol, Woodland Road, Bristol BS8 1UG, UK.
E-mail: simon.hiscock@bristol.ac.uk

DOI: 10.1016/j.cub.2007.08.060

Human Genetics: The Hidden Text of Genome-wide Associations

Genome-wide association studies are finally leading geneticists straight to the genetic susceptibility factors for complex diseases. Several challenges lie ahead, including translation of the findings into practical public health outcomes, and integrating genetic analysis with broader biological understanding.

Greg Gibson[1] and David B. Goldstein[2]

Human genetics is in the midst of a revolution. Testing for association between hundreds of thousands of polymorphisms in several thousand unrelated cases and controls allows the genome to be scanned in an unbiased manner for the major susceptibility variants for complex diseases. Up until eighteen months ago only a handful of gene variants had been securely associated with any common diseases and the majority of published claims of association were at best unsubstantiated, but

more often simply false-positive discoveries. Now that has all changed. Real advances are being made daily and the leading journals weekly report the discovery of multiple clearly validated risk factors.

One recent example of this new reality is the recent paper from the Wellcome Trust Case Control Consortium [1]. The Consortium sought to scan the majority of the human genome for some contribution to seven common diseases: bipolar disorder, hypertension, coronary artery disease, types 1 and 2 diabetes, Crohn's disease, and rheumatoid arthritis. For each disease, half a million single nucleotide polymorphisms (SNPs) were genotyped in 2,000 patients, and the genotype frequencies were compared with a common set of 3,000 controls, namely 'healthy' British subjects. The core result was the identification of 24 highly significant independent association signals. Though not reaching true experiment-wide thresholds in all cases, these sites nevertheless stick out like sore thumbs against a mass of lesser test statistics. Approximately half of the associations have been noted before, providing strong validation for the approach. The other half are novel, pinpointing loci of interest for follow-up, and eight of these have already been independently replicated in new cohorts.

For Crohn's disease, type 2 diabetes, and breast cancer, we now have multiple, multiply validated SNPs that contribute at least several percent each to population attributable risk — let's call them parSNPs. If these variants were not present in the human genome, or if their effects could be therapeutically ameliorated, there would be a measurable reduction in disease incidence. Once tens of thousands of cases have been sampled in a meta-analysis of several studies, parSNPs that provide a relative risk in the vicinity of 1.2 — a twenty percent increase in risk of contracting the disease relative to individuals with the protective genotype — significance can elevate to in excess of 10^{-10}, well beyond any reasonable experiment-wise threshold. If they don't survive such analysis, either they are false positives, perhaps beneficiaries of the 'winner's curse' of sampling bias that can inflate estimates of risk, or they are population-specific.

Genome-wide association results are being widely heralded [2] as a confirmation that many of the precepts (and promises) of 'big science' projects like the human genome and HapMap projects are being realized. After so many years of frustration and disagreement in the field, the sudden appearance of both real discoveries and real standards is most welcome. Henceforth, claims that variation in a particular gene contributes to disease susceptibility must demonstrate significance taking into account all the hypotheses that have been evaluated, and must be replicated using the same SNP and the same phenotype [3]. Understandably, these standards may percolate with different speeds through the different research communities that face specific practical challenges. Naysayers may challenge the efficacy and necessity of the HapMap project as a means to this end [4], but few practitioners will doubt its value and utility. Nevertheless, without wishing to diminish from the scale of these achievements or the transition represented by genome-wide association, it is worth returning to what the real motivation is for genetic studies, and judging progress by that metric.

There are three fundamental reasons to want to map gene variants for common disease. These are: first, to predict risk and therefore allow tailoring of lifestyle choices to risk; second, to improve understanding of disease pathophysiology in a manner that suggests new directions for therapy; and third, to identify subclasses of clinically similar diseases that nevertheless have different genetic etiologies and hence should respond to different therapies.

By the standards of these fundamental goals whole genome studies of common diseases have yet to prove themselves. Quite aside from practical challenges in relation to regulatory oversight and implementation of robust genetic counseling protocols, genome-wide association data are currently explanatory but are not yet usefully predictive. Even in the most successful case, Crohn's Disease, with nine validated associations, most of the genetic risk in the population remains to be explained, and the proportional contribution of parSNPs to family clustering seems to be less than a few percent [1,5]. If there are other factors out there, they are unlikely to individually account for more than one or two percent of susceptibility, and hence some would argue that they will be practically and clinically irrelevant. By contrast, it has been suggested that pharmacogenetics is likely to provide more immediate clinical returns than disease genetics, a good example of which is the recent demonstration that *HLA-B* genotype predicts hypersensitive response to the anti HIV drug abacavir [6].

The term parSNP serves as a reminder that the purpose of genome-wide association is not so much to find individual risk alleles as population attributable risk factors. Like the concept of heritability, which applies to populations not individuals, parSNPs are identified by their influence averaged across individuals within a population. We are not aware of any claims that common disease variants would typically be prognostic, and suspect that geneticists much more commonly imagine that there are likely dozens if not hundreds of variants that can contribute to any given disease. Individuals who happen to inherit an excess of these, and experience damning environmental circumstances, are more susceptible than others. It is the constellation of factors that is potentially predictive, not single variants.

There are some tantalizing suggestions from multivariate modeling that this may be the case. As a class, SNPs in genes involved in steroid hormonal regulation and cell cycle control are more commonly associated with breast cancer than if you query similarly sized sets of SNPs in randomly

chosen loci [7]. Similarly, a recent reanalysis of a low resolution genome-wide association for Parkinson's disease, which initially came up with a single strong candidate parSNP, implicates genes involved in axonal guidance as a class in the etiology of this neurodegenerative disease [8]. The joint probability of association of a signature involving 117 genes with Parkinson's disease is in the range of 10^{-40}.

André Rzhetsky and colleagues [9] arrived at a similar conclusion without even looking at genotypes. They scoured the medical records of 1.5 million patients of the Columbia University Medical System and considered the overlap in disease diagnoses for 124 relatively common diseases. After adjusting for correlations due to age and sex, they find striking tendencies for certain diseases to co-occur in individuals (for example, autism, bipolar depression and schizophrenia). They even estimate that some of these diseases are likely to share as many as 20% or more of their susceptibility alleles.

We are led to conclude that genome-wide association is more likely to have an impact with respect to the second and third objectives presented above. For example, a genome-wide association for asthma supports increased interest in airway remodeling alongside inflammatory hyper-responsiveness [10]; autophagy is implicated in the etiology of inflammatory bowel disease [5]; and the bipolar disorder data [1] place the copious literature pertaining to neurotransmitters and synaptic transmission as mediators of depression in a genome-wide perspective. The impact of genetics on disease sub-classification is yet to be felt, but clearly there is now good reason to expect advances.

A remarkable but yet-to-be remarked upon aspect of the recent deluge of genome-wide association data is the apparent diversity in the yield of associations for different diseases. Why, for example, is it that hypertension and bipolar disorder are devoid of the class of highly significant hits that characterize the immune-related diseases Crohn's disease and type 1 diabetes? A common view is that this as essentially the luck of the draw — some traits happen to have polymorphisms influencing them that are both common and important enough to be detectable in manageable sample sizes, and other traits do not. A more intriguing interpretation is that the data are telling us that some traits are largely resistant to the effects of common genetic variation whereas others are much more susceptible.

While it is perhaps too early to declare a clear pattern, we suspect there may be evolutionary explanations for the differences amongst traits. Contrast, for example, two quantitative traits: blood pressure and HIV viral load following infection. While hypertension produced no clear discoveries in 2000 cases [1], a study of just 500 subjects characterized for HIV-1 viral load identified three variants explaining 14% of the total variation [11]. We doubt these differences are the luck of the evolutionary draw and suspect they may result from a kind of evolutionary canalization. A key difference between HIV-1 set point and blood pressure is that blood pressure is likely to have been close to a long-term optimum, whereas this is not the case for the HIV-1 setpoint.

Canalization refers to the observation that genetic systems have often evolved robustness to genetic and environmental perturbation [12]. Expression of phenotypic variation is suppressed by the network of epistatic interactions among alleles within the normal environmental context that an organism finds itself in. Take the organism outside of this buffering zone, for example, by dramatically changing diet and immune exposure, and cryptic variation can be exposed [13]. Variants that under 'normal' circumstances would not contribute appreciably to disease do so in the modern world because they push individuals closer to the threshold upon which the new environment acts (Figure 1A).

The possibility that (de)canalization helps to explain the heterogeneity of genome-wide association results to date should be considered. A tentative indication that genetic buffering may suppress the effects of mutations that would otherwise lead to disease is seen in the shape of the Q-Q plots from the Wellcome Trust Case Control Consortium study [1] (Figure 1B,C). These plots show the relationship between observed and expected test statistics, and for five of the diseases, the observed curves deviate sharply to higher than expected test statistic values at the upper extreme, due to the parSNPs. For bipolar disorder and hypertension, by contrast, these curves actually plateau off as would be observed if the effects of candidate parSNPs were buffered.

Intriguingly, there is also a difference in the polarity of associations for different classes of disease. Most of the parSNPs are more likely tagging SNPs than the causal sites, so it is not clear what their relationships to the ancestral status of the true disease-promoting variants are. Nevertheless, including moderate evidence for association, almost two thirds of all risk-associated SNPs in the Wellcome Trust Case Control Consortium study [1] are ancestral, implying that the derived allele in the human lineage is offering protection from disease. More strikingly, for hypertension, all of the (moderate) risk alleles are ancestral, and for rheumatoid arthritis 10 of 12 are. The clear exception is Crohn's disease, for which 10 of 16 of the risk-associated parSNPs that we were able to polarize unambiguously are derived.

A prediction of the decanalization hypothesis is that as the prevalence of a disease increases in modern society, the likelihood of genome-wide association uncovering susceptibility alleles also increases. Inflammatory diseases have increased in prevalence recently, arguing that deleterious interactions between the modern environment and derived alleles corrode the evolved buffering, which in turn may expose ancestral

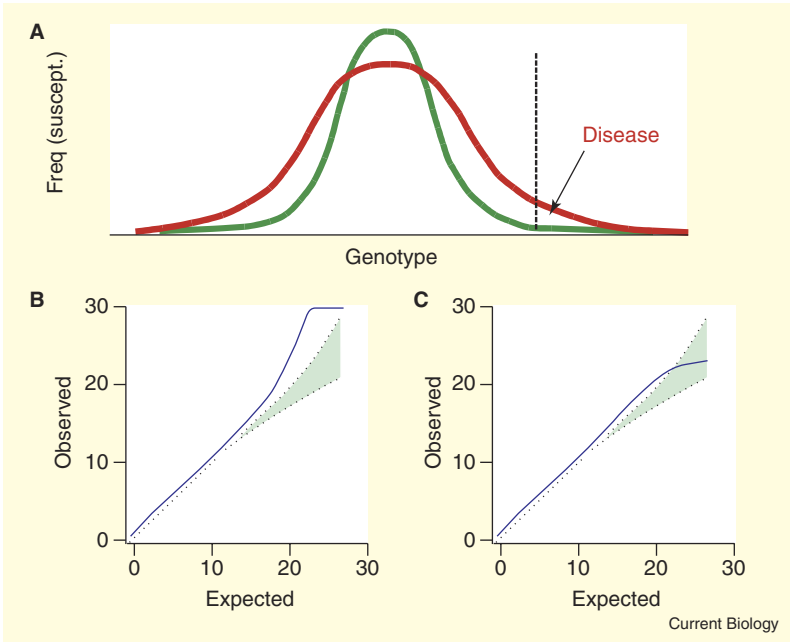

Figure 1. Canalization and genome-wide association.

(A) Model of effect of canalization on disease. Each genotype in a population is associated with a level of disease susceptibility that will typically be normally distributed but with a liability threshold indicated by the vertical broken line. Under normal environmental or genetic circumstances (green), very few individuals are above the threshold, but in an altered environment or non-equilibrium genetic circumstances (red), the variance of the susceptibility increases, and more genotypes give rise to individuals above the threshold, exposing hidden variation. (B) For coronary artery disease, Crohn's disease, types 1 and 2 diabetes, and rheumatoid arthritis, the observed association test statistics (negative log p-values) exceed the expected range of values shown in light green. (C) For bipolar disorder and hypertension, by contrast, high association scores are not seen and the Q-Q plot curve actually flattens off to a plateau, possibly suggesting buffering.

alleles to disease promotion. Our recent demonstration of highly significant associations of derived alleles with HIV setpoint [11] may be another instance of this phenomenon. Not having been exposed to the virus until a few generations ago, there has been no time for the genome to evolve canalization, and additive variation for viral titers segregates in human populations.

This hypothesis puts us slightly at odds with the recommendations of Merikangas and Risch [14] in relation to public health priorities for investment in genomic research. They argued that there is little point in expenditure on genome-wide association for diseases where the genetic contribution pales in comparison to environmental factors, since public health campaigns to modify behavior or otherwise address the environmental risk will be far more effective than treatments that target genetic variation. However, strong environmental perturbations are precisely the conditions under which hidden genetic variants may be revealed and hence under which genome-wide association may be most likely to succeed. The immediate benefit of parSNPs is not as biomarkers for personalized medicine, but rather as entry points into the hidden genetic text of complex disease. A policy consequence is that genetic advances may, counterintuitively, be most obvious where humans are exposed to novel environmental exposures, including nutrition and childhood infection. Findings in at risk populations are likely to carry over into canalized ones, and the key is to seek understanding of all facets of risk.

References

1. The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
2. Altshuler, D., and Daly, M. (2007). Guilt beyond a reasonable doubt. Nat. Genet. *39*, 813–815.
3. NCI-NHGRI Working Group on Replication in Association Studies (2007). Replicating genotype–phenotype associations. Nature *447*, 655–660.
4. Terwilliger, J.D., and Hiekkalinna, T. (2006). An utter refutation of the "Fundamental Theorem of the HapMap". Eur. J. Hum. Genet. *14*, 426–437.
5. Rioux, J.D., Xavier, R.J., Taylor, K.D., Silverberg, M.S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M.M., Datta, L.W., *et al*. (2007). Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat. Genet. *39*, 596–604.
6. Phillips, E., and Mallal, S. (2007). Drug hypersensitivity in HIV. Curr. Opin. Allergy Clin. Immunol. *7*, 324–330.
7. Pharoah, P.D.P., Tyrer, J., Dunning, A.M., Easton, D.F., and Ponder, B.A.J., and the SEARCH Investigators (2007). Association between common variation in 120 candidate genes and breast cancer risk. PLoS Genet. *3*, e42.
8. Lesnick, T.G., Papapetropoulos, S., Mash, D.C., Ffrench-Mullen, J., Shehadeh, L., de Andrade, M., Henley, J., Rocca, W., Ahlskog, J.E., and Maraganore, D.M. (2007). A genomic pathway approach to a complex disease: axon guidance and Parkinson Disease. PLoS Genet. *3*, e98.
9. Rzhetsky, A., Wajngurt, D., Park, N., and Zheng, T. (2007). Probing genetic overlap among complex human phenotypes. Proc. Natl. Acad. Sci. USA *104*, 11694–11699.
10. Moffatt, M.F., Kabesch, M., Liang, L., Dixon, A.L., Strachan, D., Heath, S., Depner, M., von Berg, A., Bufe, A., Rietschel, E., *et al*. (2007). Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. Nature *448*, 470–473.
11. Fellay, J., Shianna, K.V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., *et al*. (2007). A whole-genome association study of major determinants for host control of HIV-1. Science *317*, 944–947.
12. Gibson, G., and Wagner, G. (2000). Canalization in evolutionary genetics: a stabilizing theory? Bioessays *22*, 372–380.
13. Gibson, G., and Dworkin, I. (2004). Uncovering cryptic genetic variation. Nat. Rev. Genet. *5*, 681–690.
14. Merikangas, K.R., and Risch, N. (2003). Genomic priorities and public health. Science *302*, 599–601.

[1]Department of Genetics, North Carolina State University, Gardner Hall, Raleigh, North Carolina 27695-7614, USA.
[2]Center for Population Genomics and Pharmacogenetics, Institute for Genome Science and Policy, Duke University Medical Center, Durham, North Carolina 27708, USA.
E-mail: ggibson@ncsu.edu, d.goldstein@duke.edu

DOI: 10.1016/j.cub.2007.08.044