Copy number variation and human genome maps

Steven A McCarroll

Maps of human genome copy number variation (CNV) are maturing into useful resources for complex disease genetics. Four new studies increase the resolution of CNV maps and seek to locate human phenotypic variation on these maps.

Genetic variation ranges in size from the fine scale of SNPs to the multi-kilobase deletions and duplications that comprise copy number variation (CNV), variation in the number of copies of a genomic segment. Since observations in 2004 that human genomes harbor extensive CNV[1,2], geneticists have sought to determine what segments of the genome are affected by CNV, to develop effective ways to measure the dosage of these segments in individuals' genomes, and to assess how CNV relates to variation in phenotypes. Four new studies[3–6] represent a significant step forward in these efforts.

Humans use map-making to negotiate evolving relationships with the unknown. Like the early cartographic efforts that now populate museums, the series of CNV maps from the 2004 studies[1,2] to the current works[3–6] tell a tale of evolving technology, human foible and grit, and a meandering but increasingly successful effort to apprehend an unseen world.

CNV maps improve

Population-based maps of CNV have to date relied on hybridization of individuals' genomic DNA to microarrays of DNA probes. Many initial studies used arrays of large genomic clones such as bacterial artificial chromosomes (BACs); because BACs are large (150 kb), it was assumed that the CNVs detected were similarly large. This led to exuberant calculations about the scope and gene content of CNV in humans. Studies using higher-resolution arrays showed that most of these CNVs were 5–20 times smaller than initially reported[7,8]. Importantly, most of the CNV in any individual's genome was seen arising not from new mutations but from copy number polymorphisms (CNPs)shared with many other people[7].

Steven A. McCarroll is at the Harvard Medical School Department of Genetics, Boston, Massachusetts, USA.
e-mail: mccarroll@genetics.med.harvard.edu

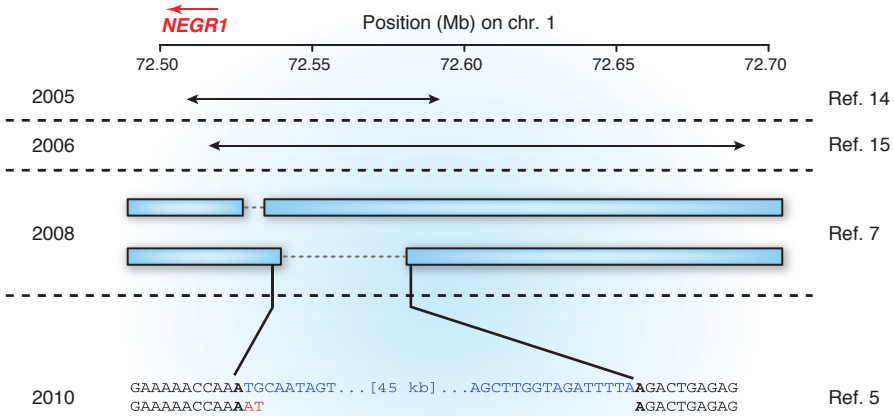

Figure 1 Evolution of maps of structural variation at the *NEGR1* locus. Copy number variation upstream of *NEGR1* was initially identified using fosmid[14] and BAC[15] clones. Analysis using hybrid SNP-CNV arrays subsequently found two common deletion alleles (of 10 kb and 45 kb) segregating at the locus, affecting nonoverlapping sequences and segregating on distinct haplotypes[7]. The 45-kb deletion was found via a tagging SNP to be associated with increased body mass index[12]. In this issue, Conrad *et al.*[5] describe the precise molecular lesion, which involves a single base of microhomology (bold) and replacement of 45 kb of sequence (blue) with 2 bp (red).

The new studies identify several more CNPs and improve the resolution with which CNPs (also called common CNVs) are described—to a resolution of about 100 base pairs (bp) at several thousand loci[3,6] and to single-nucleotide resolution at 320 loci[5]. In a recent study, investigators in the Structural Variation (SV) Consortium used tiling-resolution array comparative genome hybridization (CGH) (42 million probes) in 20 European and 20 Yoruba individuals to identify 11,700 CNV regions larger than 443 bp[1]. On page 385–391 of this issue, Conrad *et al.*[5] use a subset of these data to find the precise breakpoint sequences of 320 CNVs—estimating breakpoint location at a resolution of 50–100 bp using arrays, then sequencing 300-bp genomic fragments spanning these breakpoints. For these 320 CNVs, of which an example locus is illustrated in **Figure 1**, the breakpoint sequence in Conrad *et al.*[5] represents the culmination of years of effort to localize the affected genomic segment.

Conrad *et al.*[5] then used these breakpoint sequences to identify sequence motifs at structural lesions. The ends of 70% of the events showed 1–30 bp of microhomology—far less sequence homology than is required for nonallelic homologous recombination, and therefore implicating a different mechanism. About 30% of the breakpoints contained 1–367 bp of inserted sequence, often sequence from the genomic vicinity. Many such events

were complex combinations of deletion, duplication and inversion of local sequences, consistent with a fork-stalling and template-switching mutational mechanism (known as FoSTES) that produces complex rearrangements[9]. Though FoSTES was described only recently, diverse experimental data already support it[9,10].

On page 400–405 of this issue, Park *et al.*[6] use ultra-high-resolution CGH arrays to map CNV in 30 individuals from Korea, China and Japan. They observe many CNVs for the first time, suggesting that these CNVs have sufficiently low frequency in Europe and West Africa to have gone unsampled in the SV Consortium study. Given the tendency of allele frequency to vary across populations, the authors' map will inform the design of association studies for common CNVs in East Asian populations.

Phenotypic influences

Medieval cartographers annotated unexplored regions on maps with pictures of dragons and serpents, reflecting the human tendency to exoticize the unknown. Similarly, early impressions of CNV were often accompanied by strong predictions about their phenotypic potency. When the variation at many CNV loci was precisely measured and correlated to nearby SNPs, a less exotic picture emerged: most CNVs in the array-accessible regions of the genome were found to be ancient, diallelic polymorphisms, with alternate structural alleles segregating on ancestral SNP haplotypes[7]. The linkage disequilibrium (LD) between SNPs and common CNPs implied that the contribution of most CNPs to human phenotypic variation was already detectable in genome-wide association studies (GWAS) as associations to nearby SNPs[7]. Two GWAS hits were soon shown to involve common CNPs in strong linkage disequilibrium (LD) with the disease-associated SNPs[11,12]. The SV Consortium study[3] substantially expands this list while concurring with the earlier finding that the phenotypic contributions of common, diallelic CNPs—the kinds of CNV that are well measured by array technology—are largely accounted for in GWAS.

In a contemporaneous study, the Wellcome Trust Case Control Consortium (WTCCC) analyzed 3,432 common CNVs for disease association in the 17,000 individuals previously analyzed in WTCCC GWAS studies. The years-long project involved a new array platform and algorithmic innovations in copy number measurement and association testing[4]. But having sailed round a scientific world in search of a new continent, the WTCCC team found that they had instead rediscovered a familiar one—the five CNP-disease associations all related to earlier disease associations of SNPs at the same loci.

Given these findings, why do GWAS discoveries to date explain only 2–20% of the heritability of most common phenotypes[13]? Potential mechanisms could involve rare variants, as well as structural variants that are complex, multiallelic or otherwise hard to measure using array technology. Other explanations may include many common variants of still smaller effect, and interactions among genetic variants and non-genetic factors.

The next maps

For CNV, the next advance in map-making may come with large-scale resequencing studies such as the 1000 Genomes Project. Notably, Conrad *et al.*[5] and Park *et al.*[6] eschewed the false choice between arrays and sequencing, pursuing integrative strategies—Conrad *et al.*[5] by using array data to design a targeted sequencing experiment, and Park *et al.*[6] by using sequence data to calibrate array-based measurements.

The mapping of structural variation in our genomes has in five years traversed an arc that cartography traversed over centuries. What lessons might we draw for the exploration of other new continents, such as rare variants and epigenetics, in the search for the heritable basis of disease? One lesson is that new domains may turn out to be more familiar than they at first appear; many relationships may turn out to have been present in an overlooked form in earlier genome-scale data sets. Also, most discoveries may emerge after initial exuberance gives way to sober exploration. Insobriety about the unknown spurred human ancestors to explore new worlds. In science, a similar insobriety can lead us to mistake the nature of the novel, but it leads to the great efforts that ultimately get the work done. This human proclivity may yet be found to be an adaptive trait. The genetic mechanisms accounting for it—common or rare, structural or single nucleotide—are, at this point, anyone's guess.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Sebat, J. *et al. Science* **305**, 525–528 (2004).
2. Iafrate, A.J. *et al. Nat. Genet.* **36**, 949–951 (2004).
3. Conrad, D.F. *et al. Nature* **464**, 704–712 (2010).
4. WTCCC. *Nature* **464**, 713–720 (2010).
5. Conrad, D.F. *et al. Nat. Genet.* **42**, 385–391 (2010).
6. Park, H. *et al. Nat. Genet.* **42**, 400–405 (2010).
7. McCarroll, S.A. *et al. Nat. Genet.* **40**, 1166–1174 (2008).
8. Perry, G.H. *et al. Am. J. Hum. Genet.* **82**, 685–695 (2008).
9. Lee, J.A., Carvalho, C.M. & Lupski, J.R. *Cell* **131**, 1235–1247 (2007).
10. Zhang, F. *et al. Nat. Genet.* **41**, 849–853 (2009).
11. McCarroll, S.A. *et al. Nat. Genet.* **40**, 1107–1112 (2008).
12. Willer, C.J. *et al. Nat. Genet.* **41**, 25–34 (2009).
13. Manolio, T.A. *et al. Nature* **461**, 747–753 (2009).
14. Tuzun, E. *et al. Nat. Genet.* **37**, 727–732 (2005).
15. Redon, R. *et al. Nature* **444**, 444–454 (2006).

Chipping away at the genetics of smoking behavior

Christopher I Amos, Margaret R Spitz & Paul Cinciripini

Three large consortia present comprehensive analyses that identify genetic factors influencing smoking initiation, intensity and cessation. The genetic architecture of these three phases of smoking behavior appears to be largely distinct.

Nicotine dependence results from an interplay of neurobiological, environmental and genetic factors. Smoking stages are categorized into smoking initiation, current smoking and smoking cessation (**Fig. 1**). Genetic influences at each step in this process have been documented in numerous twin and family studies[1]. Patterns of smoking initiation reflect individual differences in sensitivity to nicotine, the availability of tobacco and social norms. For an individual who has become a habitual smoker, both genetic and psychosocial factors play a role in determining the intensity of smoking, known as smoking dependence, and the ability to quit (cease smoking). On pages 436, 441 and 448 of this issue, three collaborating groups[2–4]—the Oxford-GlaxoSmithKline (Ox-GSK)[3], Tobacco and Genetics Consortium (TAG)[4] and ENGAGE[2] consortia—present results of combined analyses from over 140,000 individuals, bringing new insights into the genetic factors that influence smoking initiation,

Christopher I. Amos, Margaret R. Spitz and Paul Cinciripini are at the Departments of Epidemiology and Behavioral Sciences, University of Texas MD Anderson Cancer Center, Houston, Texas, USA, and the Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas, USA. e-mail: camos@mdanderson.org