Duplications, CNVs and tag SNPs

The enrichment of copy number variants (CNVs) near regions of segmental duplication raises questions regarding their evolutionary origin and the extent to which CNVs reside on common haplotypes in linkage disequilibrium with other forms of variation such as SNPs. To address these questions, Evan Eichler and colleagues *(Am. J. Hum. Genet.*, in the press) used a custom BAC array to analyze CNVs near regions of segmental duplication among the 269 HapMap samples. The most common CNVs were observed across multiple populations and showed a high level of heritability. Fine-scale analysis using high-density oligonucleotide arrays further revealed both diallelic and multiallelic patterns of variation, suggesting that at least some sites have been subject to multiple, independent rearrangement events. Notably, while a few CNVs were strongly correlated with nearby SNPs, most were poorly captured by SNPs from the HapMap panel, which is explained in part by the reduced density of HapMap SNPs near regions of segmental duplication. These findings emphasize the need for denser maps of variation near regions of segmental duplication to adequately assess the role of common structural variants in contributing to phenotypic variation and disease susceptibility. *KV*

Natural history of an infection

Many individuals with cystic fibrosis acquire dangerous *Pseudomonas aeruginosa* infections. Studies have shown that individual *P. aeruginosa* genes are mutated as infections progress, but a genome-wide view of these genetic changes has been lacking. Eric Smith and colleagues have now carried out whole-genome shotgun sequencing of *P. aeruginosa* isolates taken from individuals with cystic fibrosis over a period of years (*Proc. Natl. Acad. Sci. USA*, published online 10 May 2006; doi: 10.1073/pnas0602138103). Smith *et al*. initially compared a single, clonally purified six-month isolate with a comparable 96-month isolate and two publicly available *P. aeruginosa* reference sequences. They identified 68 mutations and estimated that approximately 25% of them affect protein function. An array of virulence factors required for establishment of infection are mutated in the 96-month isolate, possibly owing to immune evasion. Subsequent sequencing of isolates from 29 other individuals with cystic fibrosis revealed strong evidence of positive selection operating on commonly mutated genes, including the multidrug efflux gene *mexZ* and the quorum-sensing regulator *lasR*. Mutation of *lasR* is particularly interesting, as it impairs biofilm formation, which is thought to be essential for ongoing respiratory infection. The authors suggest that these commonly mutated genes may represent new drug targets. *AP*

Mapping malaria resistance in the wild

The malaria mosquito *Anopheles gambiae* is a major vector for transmission of *Plasmodium falciparum* to humans. Kenneth Vernick and colleagues (*Science* **312**, 577–579; 2006) now report a genome-wide linkage scan for loci controlling resistance to *P. falciparum* infection in a natural mosquito population from West Africa. The authors established pedigrees from individual female mosquitoes, allowed the mosquitoes to feed on blood from humans in the same village infected with malaria and scored the number of oocyst-stage parasites in the midgut one week after feeding as a quantitative trait. Among 17 pedigrees genotyped, they found five that segregated a single locus with a major effect on infection intensity, three of which clustered to a common region of chromosome arm 2L. After fine-mapping the so-called *Plasmodium* resistance island to a 15-Mb interval, they applied a series of filters to prioritize candidates and used RNA interference to examine the role of two genes in a laboratory model of *P. berghei* infection. Knockdown of one of these genes, *APL1*, led to a significant increase in parasite load, identifying *APL1* as a strong candidate for contributing to *Plasmodium* resistance in natural mosquito populations. *KV*

Research Highlights written by Emily Niemitz, Alan Packer and Kyle Vogan.

Reprogramming in steps

It has long been known that nuclear transfer or cell fusion can induce somatic cell reprogramming to an alternate cell identity, but the specific activities controlling this process are not clear. Now, Amanda Fisher and colleagues show that acquisition of a new cellular identity and loss an old cellular identity are separable events that are regulated by different enzymatic activities (*J. Cell Science* **119**, 2065–2072; 2006). By inducing lymphocyte nuclei to convert to a muscle identity by fusing human lymphocytes to mouse myotubes, the authors could monitor morphological changes in the lymphocyte nuclei and changes in lineage-specific gene expression. The earliest changes detected in the lymphocyte nuclei were increased nuclear volume and decreased number of chromocenters per nuclei. These changes were followed by activation of expression of myogenic regulatory factors and decreased expression of B cell–specific gene expression. Inhibition of histone deacetylase activity with trichostatin A or valproic acid inhibited the extinction of lymphocyte-specific gene expression, but, interestingly, it did not significantly alter the acquisition of muscle-specific gene expression. This yielded heterokaryons that simultaneously expressed both muscle- and lymphocyte-specific genes. By identifying distinct steps in the reprogramming process, this work shows that loss and gain of cell identities during reprogramming can be separated. *EN*

Phantoms of the transcriptome

The amount of potentially functional RNA encoded in the genome continues to grow. As part of the FANTOM3 project, Martin Frith and colleagues have estimated and analyzed a class of noncoding RNAs in the mouse genome that they term 'pseudo–messenger RNA' *(PLoS Genet.* 2, e23; 2006). The FANTOM collection contains more than 100,000 full-length mouse cDNA sequences, and Frith *et al.* aligned them against all known proteins in the Swiss-Prot database, retaining alignments with frameshifts or internal stop codons. In this manner they identified approximately 10,000 pseudo–messenger RNAs, which they define as those transcripts that appear to encode proteins, but suffer disruptions in the reading frame. While some of these may reflect sequencing errors or cloning artifacts, the authors argue that many of them will be genuine and functional noncoding RNAs. These RNAs include those with large unspliced insertions (possibly transcribed pseudogenes), those with frame disruptions, those harboring transposable elements, and those with internal TGA stop codons that may represent potential selenoproteins. They also identified 159 pseudo–messenger RNAs that have compensatory frameshifts, such that the protein alignment undergoes multiple frameshifts, ending up as fully translatable mRNAs that would nonetheless encode completely unpredicted proteins. *AP*