

deficient in RNA-directed DNA methylation of the target GFP locus but not in the production of siRNAs from the hairpin silencing locus.

To reconcile these data, Kanno *et al.*<sup>7</sup> propose a model in which two distinct Pol IV complexes act at different steps of the TGS pathway (Fig. 1a). The Pol IVa complex (NRPD1a-NRPD2a) transcribes methylated loci to produce single-stranded RNAs (ssRNAs) that are converted to dsRNAs by RDR2 and processed to siRNAs by DCL3 (ref. 11). siRNAs then guide DRD1 and the Pol IVb complex (NRPD1b-NRPD2a) to homologous DNA and facilitate *de novo* DNA methylation and TGS. This model could explain why Kanno *et al.* did not identify *nRPD1a* mutants in their screen<sup>7</sup>. Their two-component TGS system relies on the production of siRNAs from an RDR-independent hairpin-based silencing locus and not from the silenced target locus; therefore, the Pol IVa complex would be dispensable for the production of siRNAs in this system.

#### Uncertainties and paradoxes

Though attractive, this model lacks definitive demonstration. In particular, Pol IVa-trans-

cribed ssRNAs that serve as templates for RDR2 have not been detected<sup>8,9</sup>. In addition, protein fractions containing NRPD2a had no RNA polymerase activity when DNA was used as template<sup>8,9</sup>, indicating that Pol IV may not be a conventional DNA-dependent RNA polymerase. Therefore, alternative models are possible. For example, Pol IVa could use RDR2-generated dsRNA, rather than DNA, as a template. In this model, an amplification loop would perpetuate the production of siRNAs (Fig. 1b). Because of the amplification loop, ssRNA templates would not need to be continuously transcribed from silenced loci by Pol IVa but would instead be produced by Pol II or Pol III if the levels of siRNAs are insufficient to maintain TGS. Conversion of these Pol II- or Pol III-generated transcripts into dsRNA by RDR2 would reinforce the production of siRNAs. Consistent with this hypothesis, AtSN1 transcripts accumulate not only in *dcl3*, *rdr2*, *nRPD1b* and *nRPD2a* mutants<sup>7-9,11</sup> but also in *nRPD1a* mutants<sup>8</sup>, indicating that they are not exclusively Pol IVa transcripts.

Notably, mutants impaired in the heterochromatin siRNA pathway have no visible developmental phenotypes and are

fertile<sup>7-11</sup>. This is unexpected because disrupting association of heterochromatin into chromocenters and methylation at pericentromeric repeats and retroelements should compromise genome stability or markedly perturb the expression of adjacent genes, as observed in *ddm1* mutants<sup>12</sup>. But phenotypic consequences caused by disruptions in this pathway might be detected only after prolonged breeding, as previously observed in *ddm1* mutants, which accumulate phenotypic lesions after three to five generations<sup>13</sup>.

1. Matzke, M.A. & Birchler, J.A. *Nat. Rev. Genet.* **6**, 24–35 (2005).
2. Almeida, R. & Allshire, R.C. *Trends Cell Biol.* **15**, 251–258 (2005).
3. Kawasaki, H. & Taira, K. *Curr. Opin. Mol. Ther.* **7**, 125–131 (2005).
4. Tomari, Y. & Zamore, P.D. *Genes Dev.* **19**, 517–529 (2005).
5. Bartel, D.P. *Cell* **116**, 281–297 (2004).
6. Baulcombe, D. *Nature* **431**, 356–363 (2004).
7. Kanno, T. *et al. Nat. Genet.* **37**, 761–765 (2005).
8. Herr, A.J., Jensen, M.B., Dalmay, T. & Baulcombe, D.C. *Science* **308**, 118–120 (2005).
9. Onodera, Y. *et al. Cell* **120**, 613–622 (2005).
10. Kanno, T. *et al. Curr. Biol.* **14**, 801–805 (2004).
11. Xie, Z. *et al. PLoS Biol.* **2**, 642–652 (2004).
12. Lippman, Z. *et al. Nature* **430**, 471–476 (2004).
13. Kakutani, T., Jeddeloh, J.A., Flowers, S.K., Munakata, K. & Richards, E.J. *Proc. Natl. Acad. Sci. USA* **93**, 12406–12411 (1996).

## Vive la difference!

Charles Lee

**Until very recently, it was widely touted that the complete DNA sequences of any two human beings were 99.9% identical. A new study refutes this notion through a comprehensive comparison of two individual genomes which detects hundreds of new structural genomic variants.**

All humans are genetically similar. These similarities help to define us as a species and are consistent with our recent origin. Until last summer, it was widely accepted that the euchromatin-associated differences between two randomly chosen individuals was limited to ~0.1% of the genome, existing predominantly as SNPs, occurring on average once per 1,000 bases. This view of the human genome was challenged by two studies showing the prevalence of large-scale copy-number variations (LCVs) or copy-number poly-

morphisms (CNPs), which typically involve tens to hundreds of kilobases of DNA, in the human genome<sup>1,2</sup>. These genome-wide studies used array-based comparative genomic hybridization (array CGH) methodologies with limited resolution. On page 727 of this issue, Eray Tuzun and colleagues<sup>3</sup> report an investigation of how different the genomes of two individuals really are and a new *in silico* strategy to compare two genomes at the DNA sequence level.

#### Follow the fosmids

Tuzun *et al.*<sup>3</sup> reasoned that although the reference human genome sequence is a compilation of data from more than eight different DNA libraries (each generated from a different normal individual), ~70% of the reference sequence originated from a single

library (the RPCI-11 BAC library). As such, the reference human genome can be simplistically considered to be predominantly the DNA sequence of a single normal individual—in this case, a male. To compare the genome sequence of this individual with that of another individual, the authors used available clone end-sequence data from a fosmid DNA library (G248 library) generated from an anonymous North American female. The computational analyses were straightforward. Because fosmids are produced using a packaging system that strictly limits the size of their inserts to between 32 and 48 kb, one would normally expect the end sequences of a given fosmid clone to align to the reference sequence with a spacing similar to that in the fosmid clone, and in the same orientation. Any significant deviation from

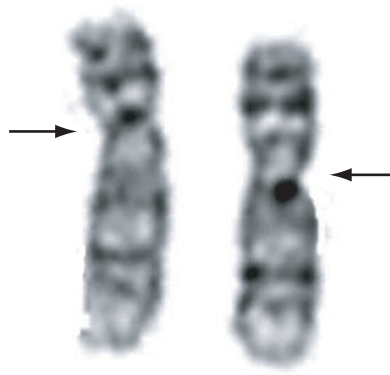
Charles Lee is in the Division of Cytogenetics, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. e-mail: clee@rics.bwh.harvard.edu

the expected spacing, defined by the authors as three standard deviations from the mean size of the fosmid clones, would be suggestive of a locus that has an insertion or deletion with respect to the reference sequence. This *in silico* strategy also allowed the authors to identify the breakpoints of putative inversion variants when the orientation of the end sequences was inconsistent between fosmid end pairs and the reference genome (in this case, National Center for Biotechnology Information (NCBI) build 35), an added advantage over microarray-based methodologies, which are unable to detect balanced chromosomal alterations.

### The dynamic genome

In total, 297 sites of putative structural variation were identified between the two genomes, including 139 insertions, 102 deletions and 56 inversion breakpoints. Assuming that cloning artifacts have been excluded, these numbers are astounding. The size of most of these variants ranged from 8 to 40 kb, but some deletions were estimated to be hundreds of kilobases in size and some inversions up to 2 Mb in length. The nature of the analysis did not allow for detection of insertions into the fosmid-derived genome greater than the clone insert size (~40 kb). Notably, for three-quarters of the structurally variant loci, the authors also found clones in the same fosmid library that had sequence alignments consistent with the reference genome sequence. This suggests that the bulk of the structural variants exist in a heterozygous state in the individual from which the fosmid library was made.

Comparison of the insertions and deletions identified in this study with LCVs and CNPs documented from recent genome-wide array CGH experiments<sup>1,2</sup> and other literature found that 81% (196 of 241) of the loci reported here were new. Part of the reason for this is that the average size of the structural variants identified in this study (~15–30 kb) is below the resolution of the previously used array CGH platforms<sup>1,2</sup>. Consistent with the previous studies, the identified structural variants had an increased bias for known regions of segmental duplication<sup>4,5</sup>. Taken together, the data indicate that structural genome variation in the human genome is ubiquitous, arguing for a much more



**Figure 1** Pericentromeric inversion of chromosome 9. The pericentromeric inversion of chromosome 9, also known as *inv(9)(p11q13)* or *inv(9)*, is one of the first well documented human genomic structural variants<sup>12</sup> and is present in 1–3% of the general population. Normal chromosome 9 is on the left and *inv(9)* is on the right. Arrows indicate the location of the centromere.

dynamic view of the human genome than was previously appreciated<sup>6</sup>.

### Why study structural genomic variation?

Why is it important to study structural genomic variation? For the same reason that substantial efforts and resources have been invested in identifying and cataloging SNPs: these genomic differences may account for some of the differences in individuals' susceptibility to disease or explain why people have different reactions to specific drugs or environmental stimuli. Natural variation in gene expression (contributing to quantitative traits in humans) is well documented<sup>7</sup> and may be due in part to copy-number differences of certain gene loci. Tuzun and colleagues<sup>3</sup> noted that many of the identified sites of structural variation encompass genes that are not essential for viability but could be described as 'environmental sensor genes', associated with drug detoxification, immune response, etc. This is consistent with the growing body of literature that shows how genomic inversions<sup>8,9</sup> and copy-number variations<sup>10</sup> can influence aspects of human adaptability, including disease-associated chromosomal rearrangements, fertility rates and susceptibility to viral infections. Detailed cataloging and characterization of these structural genomic variants will be necessary to facilitate future

studies that investigate the association of each of these specific variants, or a set of them, with a disease or phenotype of interest. Depending on the recurrence frequency of these structural variants, SNP-based linkage disequilibrium studies alone might not find association of these variants with disease<sup>11</sup>. An attempt to consolidate information on structural genomic variants in standardized file formats is currently being made by the Database of Genomic Variants project (<http://projects.tcag.ca/variation/>). It is anticipated that the curated data from this initiative and others (<http://www.humanparalogy.gs.washington.edu/structuralvariation>) will also be available in other public genome browsers such as those from the University of California Santa Cruz, Ensembl and NCBI. Moreover, access to up-to-date information on these genomic variants will be crucial for accurate interpretation of genetic diagnostic tests, especially in an era where testing often precedes biological understanding.

Do we now have a thorough understanding of the extent of structural variation in our genomes? Probably not. Early cytogenetic studies found chromosomal variants (Fig. 1); microarray-based studies identified LCVs and CNPs; and now fosmid analyses have uncovered a finer scale of genomic variation. One must speculate that additional structural genomic variations will be uncovered as more robust methods and complementary studies are done in more individuals. The estimate of 99.9% total DNA sequence identity between any two human beings will need to be continually reassessed as many more structural genomic variants are uncovered over the coming years.

1. Iafrate, A.J. *et al. Nat. Genet.* **36**, 949–951 (2004).
2. Sebat, J. *et al. Science* **305**, 525–528 (2004).
3. Tuzun, E. *et al. Nat. Genet.* **37**, 727–732 (2005).
4. Bailey, J.A. *et al. Science* **297**, 1003–1007 (2002).
5. Cheung, J. *et al. Genome Biol.* **4**, R25 (2003).
6. van Ommen, G.-J. B. *Nat. Genet.* **37**, 333–334 (2005).
7. Morley, M. *et al. Nature* **430**, 743–747 (2004).
8. Stefansson, H. *et al. Nat. Genet.* **37**, 129–137 (2005).
9. Osborne, L.R. *et al. Nat. Genet.* **29**, 321–325 (2001).
10. Gonzalez, E. *et al. Science* **307**, 1434–1440 (2005).
11. Fredman, D. *et al. Nat. Genet.* **36**, 861–866 (2004).
12. de la Chapelle, A. *et al. Am. J. Hum. Genet.* **26**, 746–765 (1974).