Widening the spectrum of human genetic variation

Evan E Eichler

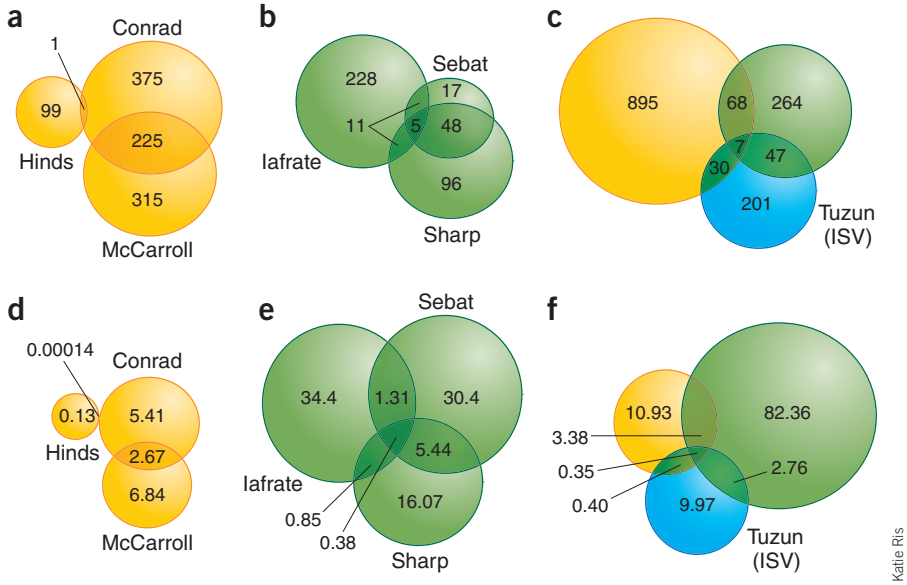
SNP genotyping platforms have been used to discover ~1,000 deletion structural variants within the human genome, with median lengths ranging from 500 bp to 10.5 kb. Analyses of a subset of these provide compelling evidence of linkage disequilibrium with flanking SNPs.

Uncovering the genetic basis of human phenotypic differences requires a comprehensive understanding of all forms of genetic variation. Although there have been tremendous advances in deducing the pattern and nature of single-nucleotide differences[1,2], a similar realization for larger and more complex forms of genetic variation has lagged behind. Three papers reported in this issue of *Nature Genetics* provide critical insight into a broader spectrum of human genetic variation by characterizing deletions in the human population (**Figure 1** and **Table 1**).

From anomalies to discoveries

All three studies[3–5] focus on the identification and characterization of fine-scale deletions by taking advantage of technology designed to catalogue SNPs in the human population. Two of the studies[4,5] use new methodologies to discover deletions by mapping clusters of SNP genotype errors and/or mendelian transmission errors. Conrad and colleagues examine transmission data from 60 parent-offspring trios that seem to violate the rules of mendelian inheritance. McCarroll and colleagues widen the scope to discover clustered sites that do not pass Hardy-Weinberg equilibrium as well as other genotyping errors. Hinds and colleagues, in contrast, use a more direct approach and identify regions of reduced signal intensity by hybridizing long-range PCR products, generated from haploid source material, against an oligonucleotide microarray (**Table 1**). Combined, the three studies have sifted through the flotsam of SNP genotyping data to uncover a treasure trove of ~1,000 deletions. Rigorous experimental validation confirms the majority (80%) of the selected sites reported in these studies. Although there is a considerable range in deletion size, the majority of validated sites are <10 kb in size, which is smaller than most previous studies[6–9].

The analyses of the deletion sites reveal several important trends. Similar to SNPs, deletion variants show greater diversity among individuals of African descent. Also, several lines of evidence suggest that deletion



Katie Ris

Figure 1 Comparison of studies of structural human genome variation. (**a**–**c**) Venn diagram comparing the number of intersecting sites having a minimum of 100 bp overlap. (**a**) Total nonredundant set of 1,015 fine-scale deletions. (**b**) 401 large-scale copy number variations (CNVs). (**c**) 1,503 variants, showing the intersection of fine-scale deletions (**a**) and large-scale CNVs (**b**) with intermediate-size variations found by Tuzun *et al.*[6]. (**d**–**f**) Same classification as **a**–**c**, but comparing the number of intersecting structurally variant base pairs. **d** shows a total of 15.05 Mb of deleted base pairs; **e**, a total of 88.35 Mb; and **f**, a total of 110.15 Mb. Differences in the target size of deletions[3] and specific differences in the detailed methodology[4,5] are likely to explain why only 25% of the deletion sites are shared even though many of the same samples were analyzed[4,5]. A web version of these sites (based on the University of California, Santa Cruz web browser) along with other published studies of structural variation may be found at http://humanparalogy.gs.washington.edu/structuralvariation. Some sites were lost owing to coordinate mapping between different human genome sequence assemblies. Only validated sites were available for Hinds.

Evan E. Eichler is at the Department of Genome Sciences, University of Washington School of Medicine and the Howard Hughes Medical Institute, Seattle, Washington 98195, USA.
e-mail: eee@gs.washington.edu

Table 1 Summary of genome-wide studies of structural variation

Study	Genome coverage	Assay	Variant types	Sample size	Size range	Number of variants	Number validated	Median size (kb)	Comments and limitations
Hinds	100–200 Mb	High-density oligonucleotide hybridization	Deletion	24[a]	70 bp–7 kb	215	100	0.75	Excludes repeat and duplication regions
McCarroll	1.3 million genotyping assays	Clustered genotype errors and mendelian errors	Deletion	269[b]	1 kb–745 kb	541	90	7.0	SNP genotype bias; 'Hapmappable genome'
Conrad	1.3 million genotyping assays	Mendelian errors	Deletion	180[c]	300 bp–1,200 kb	586	100	8.5–10	SNP genotype bias; 'Hapmappable genome'
Tuzun	8× fosmid clone coverage	Paired-end sequence	Deletion, insertion, inversion	1	>6–1,900 kb	297	112[g]	15.2	Reduced power in regions of perfect sequence identity
Sharp	1,986 BACs	BAC-based ArrayCGH	Deletion, insertion[f]	47	>50 kb	160	53[g]	~150	BAC-based; targets duplicated; hotspot regions only
Sebat	85,000 oligonucleotides	ROMA	Deletion, insertion[f]	20	>100 kb[e]	76	11	222	Reduced complexity library hybridzation
Iafrate	5,264 clones	BAC-based ArrayCGH	Deletion, insertion[f]	55[d]	>50 kb	255	18	~150	BAC-based; density is reduced.

[a]Polymorphism Discovery Resource collection ($n = 24$); [b]International HapMap Consortium: CEU, JPT, CHB and YRI samples; [c]60 parent-child trios from CEU and YRI samples; [d]39 normal individuals and 16 individuals with previously characterized karyotype abnormalities. [e]Smaller deletion events may be detected with a higher density of oligonucleotides. [f]Insertion events can be detected only if sequence is represented once in the reference genome or reference BAC clone. These approaches can not detect insertions of *de novo* sequence. [g]Includes validated sites from previous studies. ArrayCGH, array comparative genomic hybridization; ROMA, representational oligonucleotide microarray analysis.

polymorphisms may be under more extreme selection than SNPs. The observed size distribution is biased toward smaller events than expected; both the X chromosome and coding exons are underrepresented for deletions when compared with SNPs[4], and there is an apparent excess of rare deletions when compared to SNPs (although ascertainment bias could not be excluded definitively)[3]. Nevertheless, a growing list of genes seems to intersect with the sites of deletion. Conrad and colleagues, for example, identified 92 genes that were entirely deleted and another 109 genes in which coding sequences were partially eliminated. The set of genes deleted is not randomly distributed in the genome; there is a clear association with segmental duplications among larger deletion events[9,10]. In addition to previously reported associations with immunity, defense, chemosensation and drug detoxification, other functional gene categories emerge in these studies, such as signal transduction, sex hormone metabolism and cancer susceptibility.

A key question about structural variation is whether such events are ancient or recurrent. It is possible that recurrent deletions may appear on different genetic haplotypes due to the known mechanism of repeat-mediated deletions—as highlighted by the frequency of nonallelic homologous recombination among many human genomic disorders[11]. Two of the studies[3,5] use SNP genotype data flanking deletions to address this question specifically. They provide compelling data that many common deletions are in linkage disequilibrium ($r^2 = 0.8$ to 1.0) with flanking SNPs. These findings, along with a recent study that assessed 2,393 smaller insertion and deletion events in 330 individuals[12], are important because they suggest that SNPs may be used as surrogate markers to assess the role of a subset of potential deleterious deletions in genome-wide disease association studies without the need to directly assess the deletion variant itself. In contrast, Conrad provides two intriguing examples in which deletion-bearing haplotypes seem to show heterogeneity in their breakpoints on different genetic backgrounds. This may suggest recurrence, although no obvious intrachromosomal segmental duplications were observed at the boundaries that could explain the instability of these regions.

Tagging larger variants?

Although the linkage disequilibrium data are convincing for fine-scale deletion variants between 500 bp and 10 kb, as well as for smaller insertions and deletions, one should exercise caution in extrapolating these findings to larger, more complex datasets of structural variation. A comparison of these sites of deletion polymorphism documented here with larger intermediate-size structural variation[6] or large-scale copy number variants shows very little overlap[7–9,13]. In addition, the association with segmental duplications is less pronounced than previously reported. This could be due to the fact that regions where one would expect recurrence are underrepresented owing to technical limitations. For example, oligonucleotide-based microarrays specifically designed to detect SNPs frequently exclude repeat-laden regions of the genome at both the long-range PCR and microarray design stages, as hybridization signals cannot be interpreted reliably[14]. Similarly, regions near centromeres, telomeres or segmental duplications are generally not classified as part of the 'Hapmappable genome'[2], and therefore there is a dearth of corresponding genotype data. Based on our current understanding of the molecular basis for genomic disorders, recurrent rearrangement events are most likely to occur between large blocks of virtually identical sequence[11]. The lack of information for these regions in these studies leaves unresolved the question of the linkage disequilibrium and recurrence among the common larger forms of complex genetic variation.

Nevertheless, these studies add another 1,000 sites of deletion to our catalogue of structural variation and provide a methodological framework to begin to test their association with human genetic disease. Surprisingly, the majority of the deletion variants identified in these three studies do not overlap (**Figure 1**). Given the extensive experimental validation presented and the fact that a large number of (often identical) individuals were examined, one may draw two conclusions. First, no single optimized approach has been developed yet to systematically capture all structural variation in the human genome. Second, it is likely that several thousand additional common structural variants await discovery. Their abundance, enrichment in environmental-interaction genes and their apparent attributes in terms of natural selection suggest that these genetic lesions will be important in complex genetic disease. A more systematic, unbiased

approach to discover and genotype such variation is required. A Human Genome Structural Variation Project dedicated to the characterization of not only deletions but also insertions and inversions should become a priority for the human genetics and human genome sequencing communities in an effort to further widen the spectrum of human genetic variation.

1. Hinds, D.A. *et al. Science* **307**, 1072–1079 (2005).
2. The International HapMap Consortium. *Nature* **437**, 1299–1320 (2005).
3. Hinds, D.A., Kloek, A.P. & Frazer, K.A. *Nat. Genet.* **38**, 82–85 (2006).
4. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. *Nat. Genet.* **38**, 75–81 (2006).
5. McCarroll, S.A. *et al. Nat. Genet.* **38**, 86–92 (2006).
6. Tuzun, E. *et al. Nat. Genet.* **37**, 727–732 (2005).
7. Sebat, J. *et al. Science* **305**, 525–528 (2004).
8. Iafrate, A.J. *et al. Nat. Genet.* **36**, 949–951 (2004).
9. Sharp, A.J. *et al. Am. J. Hum. Genet.* **77**, 78–88 (2005).
10. Bailey, J.A. *et al. Science* **297**, 1003–1007 (2002).
11. Stankiewicz, P. *et al. Cold Spring Harb. Symp. Quant. Biol.* **68**, 445–454 (2003).
12. Bhangale, T.R., Rieder, M.J., Livingston, R.J. & Nickerson, D.A. *Hum. Mol. Genet.* **14**, 59–69 (2005).
13. Fredman, D. *et al. Nat. Genet.* **36**, 861–866 (2004).
14. Frazer, K.A. *et al. Genome Res.* **13**, 341–346 (2003).

Amyloid double trouble

John Hardy

A new study shows that some cases of early-onset Alzheimer disease result from duplications of the *APP* locus, which encodes the amyloid β precursor protein. This finding fulfills a 20-year-old prediction that genetic variability in *APP* expression could lead to disease and provides further, perhaps definitive, evidence for the amyloid hypothesis of the disorder.

When Glenner and Wong first reported the isolation and identification of the amyloid β (Aβ) peptide in the meningeal vessels of individuals with Alzheimer disease[1], and later in meningeal vessels of adults with Down syndrome, they wrote[2], "This is the first chemical evidence of a relationship between Down syndrome and Alzheimer's disease…Assuming [Aβ] is a human gene product, it also suggests that the genetic defect in Alzheimer's disease is localized on human chromosome 21." Not only have the direct predictions in these remarkable papers been shown to be essentially accurate, but the implicit prediction that genetic variability in the expression of the normal *APP* gene product could cause Alzheimer disease has now been shown to be correct. On page 24 of this issue, Rovelet-Lecrux *et al.*[3] report several independent duplications of the *APP* locus in French families with a variable, autosomal dominant phenotype intermediate between the pure Alzheimer phenotype seen in most families with *APP* mutations[4] and the cerebral hemorrhage phenotype of Dutch angiopathy associated with the *APP* E693Q (Dutch) mutation[5]. These findings highlight the importance of *APP* gene dosage and provide strong support for the amyloid hypothesis[6], which postulates that accumulation of Aβ in the brain drives Alzheimer disease pathogenesis (**Fig. 1**).

John Hardy is at the Laboratory of Neurogenetics, National Institute on Aging, Porter Neuroscience Building, 35 Convent Drive, Bethesda, Maryland 20892, USA.
e-mail: hardyj@mail.nih.gov

APP overdose

It may seem surprising that it has taken 15 years since the identification of the first *APP* mutations[4,5] for these duplications to come to light, especially since they seem to be relatively common. This is probably because the phenotype is variable, with some cases presenting with a hemorrhagic stroke early in their disease. This phenotype will often lead these individuals to come to the attention of stroke physicians rather than behavioral neurologists and could obscure the familial nature of the underlying dementia syndrome in the absence of pathological examination. The variable phenotype is reminiscent of that associated with the *APP* A692G (Flemish) mutation[7] and has a precedent in other unresolved families with weak evidence for chromosome 21 linkage[8].

The report by Rovelet-Lecrux *et al.* is of interest for several reasons. First, when considered with the report of an individual with Down syndrome who was trisomic distal to the *APP* locus and did not develop Alzheimer pathology[9], it demonstrates that modest (50%) increases in *APP* expression are sufficient to cause disease, with a clinical onset in the fifties. This suggestion is concordant with the notion that genetic variability in the normal disomic expression of *APP* can predispose individuals to late-onset disease[10]. It is difficult to interpret these results in any way other than as strong evidence for the amyloid hypothesis[6]. A clear implication is that individuals homozygous for an *APP* allele with an expression level 25% higher than normal would also be expected to develop Alzheimer disease in their fifties, with smaller increases in *APP* expression presumably leading to disease onset at a later age. Second, in the cases reported by Rovelet-Lecrux *et al.*, as in earlier cases involving copy number increases at the α-synuclein (4q21) locus causing Parkinson disease[11] and the classic cases of *PMP22* duplication (17q11.2) causing Charcot-Marie-Tooth disease Type 1A (ref. 12), several other genes are duplicated without obvious phenotypic consequences. This suggests that the expression of most genes is plastic enough that aberrant dosage does not have severe effects. Third, this study confirms that

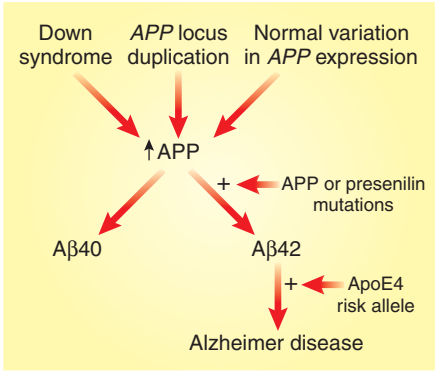


Figure 1 Genetic factors influencing Aβ42 production and Alzheimer disease onset. The amyloid β precursor protein (APP) can be cleaved to yield either of two smaller peptides, Aβ40 and Aβ42, the latter of which contributes to disease pathogenesis. Trisomy 21 (Down syndrome), duplication of the *APP* locus or normal variation in *APP* expression can result in elevated levels of APP, resulting in increased production of Aβ peptides. Mutations in *APP* or in the presenilin genes, which encode the proteases responsible for APP cleavage, can also lead to increased production of Aβ42. Normal variation in ApoE4 influences Aβ42 accumulation, contributing to Alzheimer disease pathogenesis.