Widening the spectrum of human genetic variation

Priit Palta

bioinfo Journal Club

24.04.06

- Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. (2006). **A high-resolution survey of deletion polymorphism in the human genome.** Nat Genet 38, 75-81.

- Hinds, D. A., Kloek, A. P., Jen, M., Chen, X. & Frazer, K. A. (2006). **Common deletions and SNPs are in linkage disequilibrium in the human genome.** Nat Genet 38, 82-85.

- Eichler, E. E. (2006). **Widening the spectrum of human genetic variation.** Nat Genet 38, 9-11.

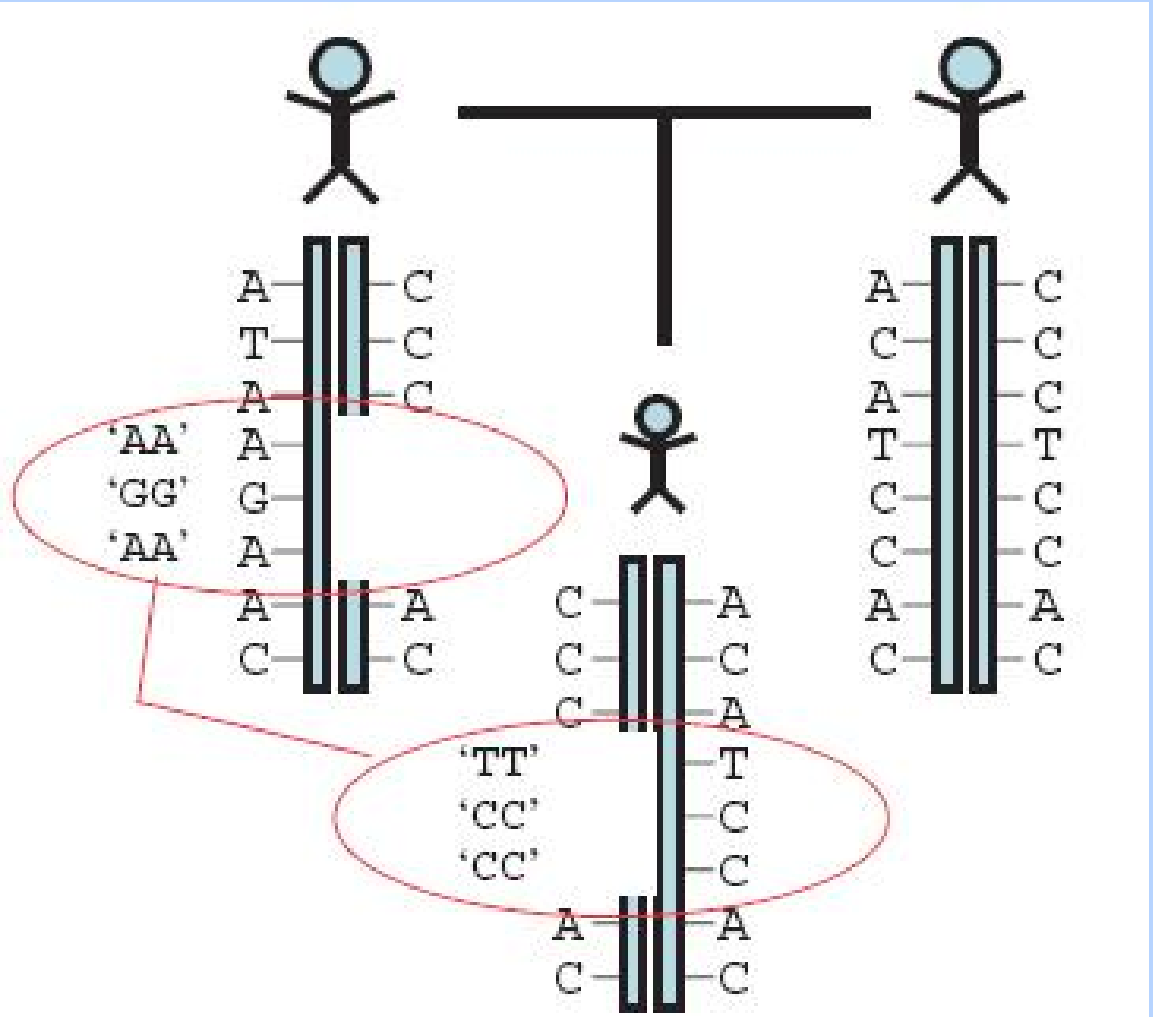
A high-resolution survey of deletion polymorphism in the human genome. Conrad *et al.*

- It has long been known that chromosomal deletions can lead to a variety of serious developmental and malformation disorders, such as DiGeorge and Prader-Willi syndromes, etc.
- While deletions that cause severe diseases are rare in the population, recent work has indicated that less harmful deletions are widespread in the human genome, in many cases deleting genes.
- Deletions may have a significant role
  - in the genetics of complex traits
  - in genome evolution

A high-resolution survey of deletion polymorphism in the human genome. Conrad *et al.*

- The main goal – to obtain a detailed picture of the extent and distribution of deletion variation in the human genome.
- Two sets of thirty parent-offspring trios from HapMap:
  - European derived ‘CEPH’ sample (denoted ‘CEU’) – 30 trios
  - African sample of Yoruba individuals from Ibadan, Nigeria (‘YRI’) – 30 trios

A high-resolution survey of deletion poly-morphism in the human genome. Conrad *et al.*



A high-resolution survey of deletion polymorphism in the human genome. Conrad *et al.*

- Initially total of:
  - 453 regions in the CEU (1,108,950 SNPs)
  - 680 regions in the YRI (1,085,823 SNPs) samples as putative candidate deletion regions
- After removing deletions that appear to be false predictions:
  - 345 candidate deletions in the CEU samples
  - 590 in the YRI

A high-resolution survey of deletion polymorphism in the human genome. Conrad *et al.*

- Two different analytical methods, using different assumptions, suggest that the false positive rate for the predicted deletions is low:
  - 5% of FP rate with 5000 simulated HapMap datasets
  - 14% with empirical method for estimating the FP rate that naturally accommodates correlation in genotyping errors across neighboring SNPs

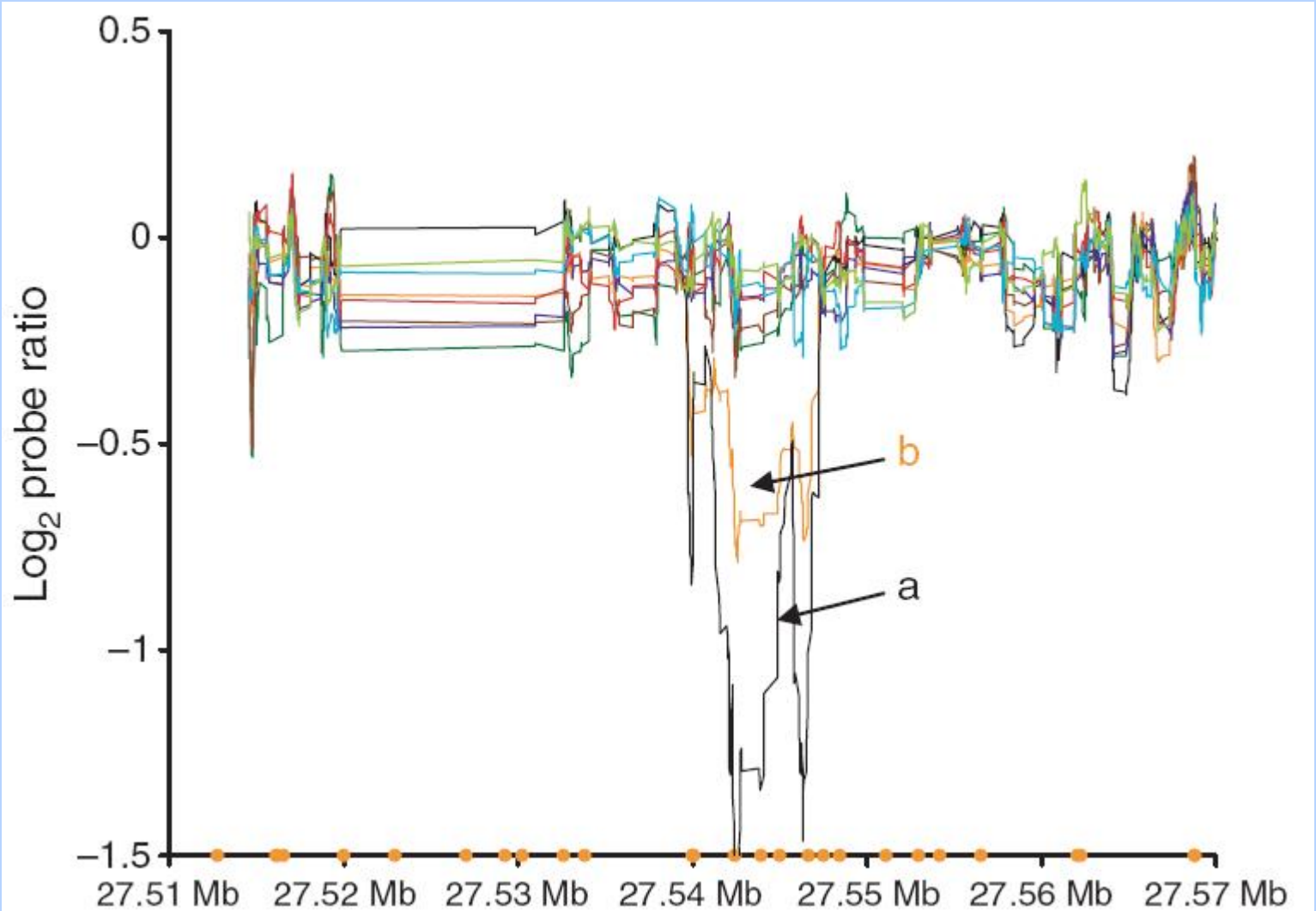
A high-resolution survey of deletion polymorphism in the human genome. Conrad *et al.*

- Validation of deletion detection method:
  - 12 predicted deletion validated using quantitative PCR (concentrations consistent with transmission of a deletion from parent to child)
  - comparative genomic hybridization (CGH) with a custom oligonucleotide microarray (380,000 probes) that tile across all candidate deletions identified in nine HapMap offspring (eight YRI and one CEU)

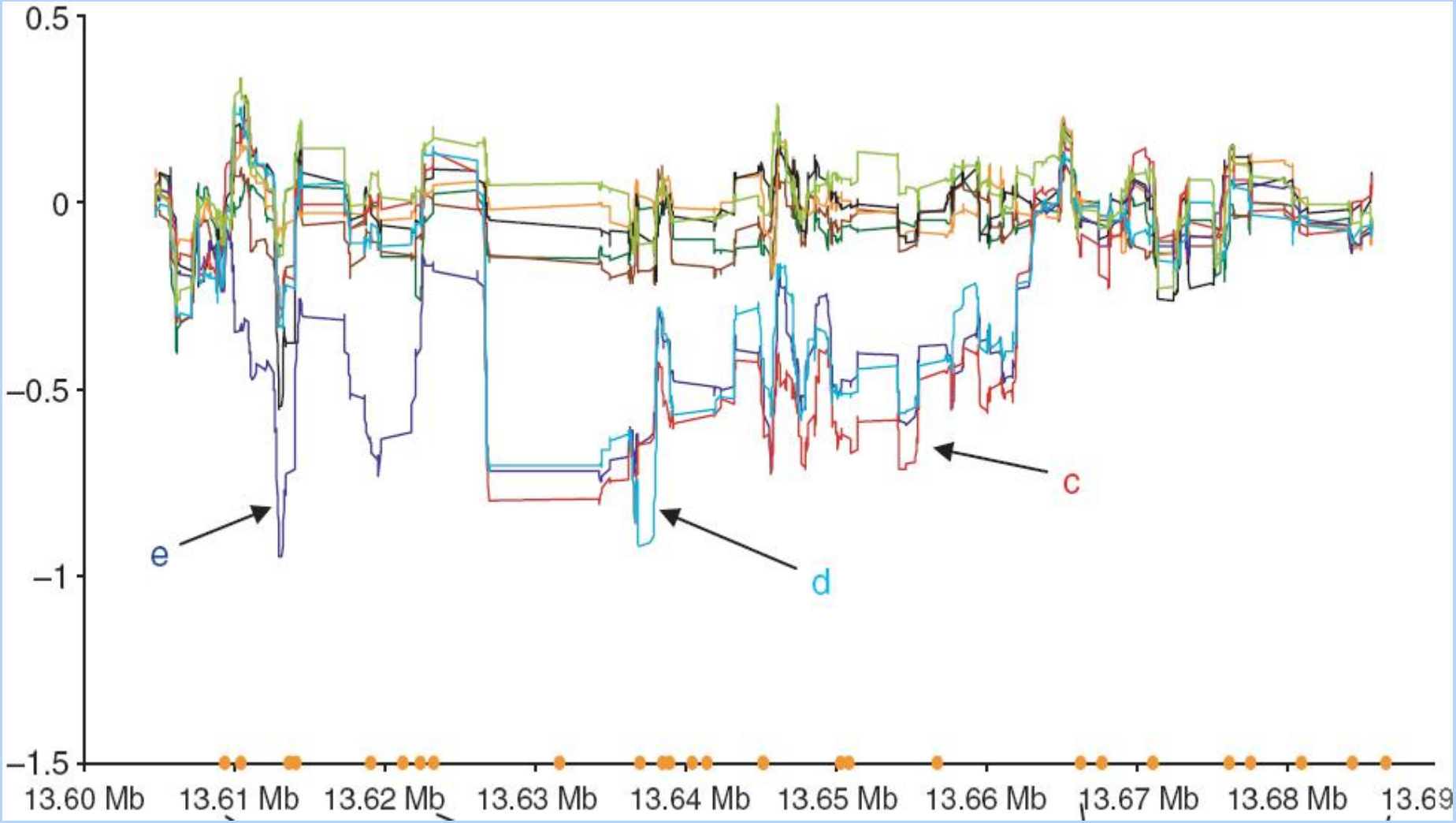
A high-resolution survey of deletion poly-morphism in the human genome. Conrad *et al.*

- Validation of deletion detection method:
  - 12/12 using quantitative PCR
  - 80/93 using CGH (empirical rate of false positives ~14% - consistent with the theoretical prediction)

A high-resolution survey of deletion polymorphism in the human genome. Conrad *et al.*



A high-resolution survey of deletion polymorphism in the human genome. Conrad *et al.*



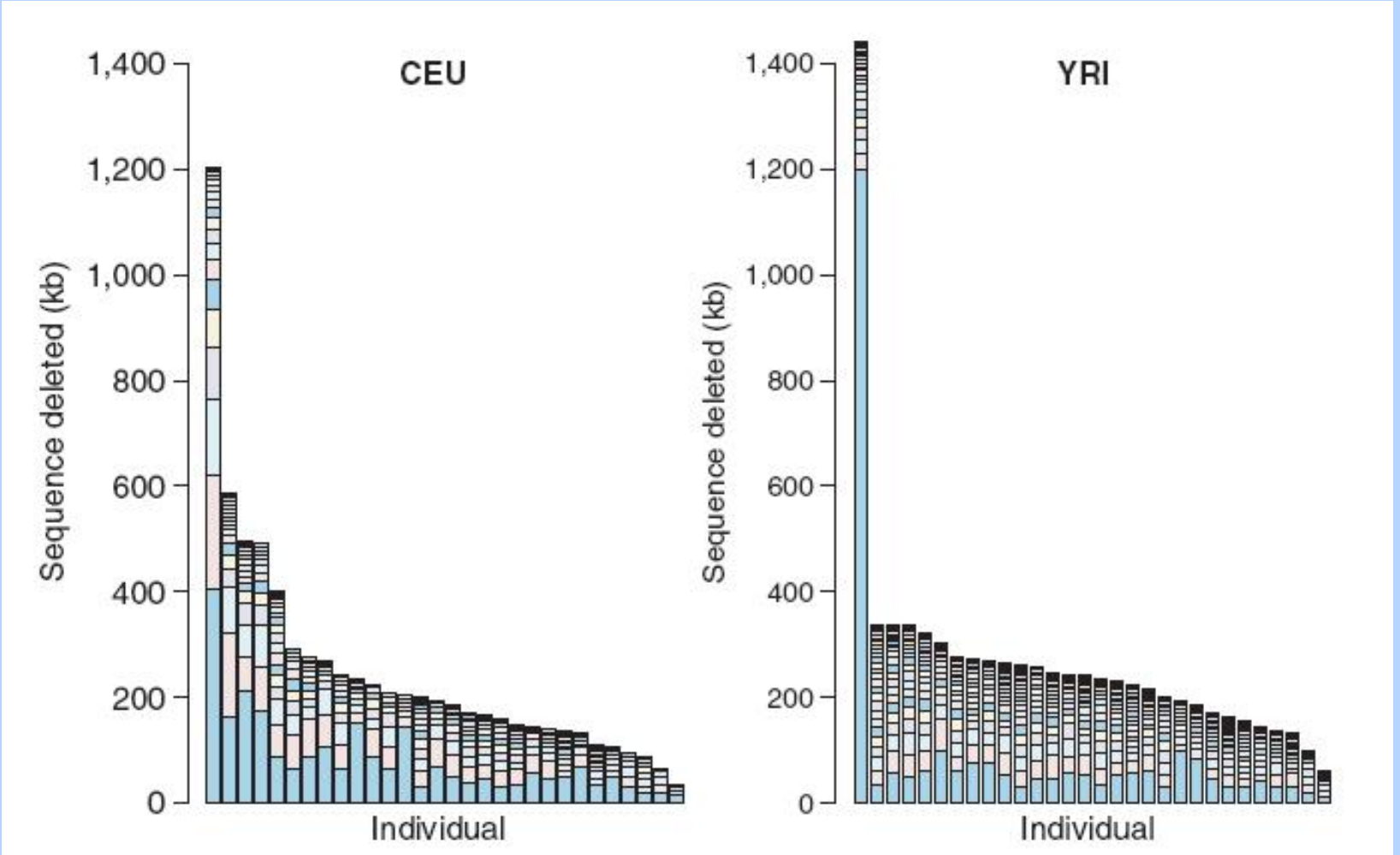
A high-resolution survey of deletion polymorphism in the human genome. Conrad *et al.*

- The lengths of the deletions identified from the SNP data followed an L-shaped distribution, with many small deletions and few large ones:
  - The 345 predicted CEU deletion regions had a median length of 10.6 kb (range: 0.3–404 kb).
  - There were ~70% more deletions detected in the YRI sample, but these were smaller, with a median length of 8.5 kb and with fewer long deletions (range: 0.5–1,200 kb).

A high-resolution survey of deletion polymorphism in the human genome. Conrad *et al.*

- This difference in deletion prevalence cannot be explained by the slightly higher rate of false positives in the YRI sample; it is more likely to reflect the greater diversity commonly found in African-derived populations.

A high-resolution survey of deletion polymorphism in the human genome. Conrad *et al.*



A high-resolution survey of deletion polymorphism in the human genome. Conrad *et al.*

- Found deletions:
  - The 590 YRI deletions occurred at 396 distinct genomic locations and the 345 CEU deletions were at 228 locations
  - 39% were singletons
  - 37 deletions were shared in CEU and YRI
  - 15.7% overlapped/contained segmental duplications
  - 11% matched CNPs from previous studies

A high-resolution survey of deletion polymorphism in the human genome. Conrad *et al.*

- Found deletions:
  - Genic SNPs were strongly underrepresented in deletion regions - 23.6% for YRI and 18.6% for CEU SNPs versus 33.8% in the entire Phase I HapMap data
  - A modest deficit of X-chromosome deletions

A high-resolution survey of deletion poly-morphism in the human genome. Conrad *et al.*

- Found deletions:
  - Affect a large number of genes:
    - The deletions spanned at least part of 267 known and predicted genes
    - Coding sequence was deleted in 201 of these genes
    - 92 genes of these were entirely deleted

Common deletions and SNPs are in linkage disequilibrium in the human genome. Hinds *et al.*

- The main goal – to find out whether common deletion polymorphisms are the result of single mutation events such as SNPs or are due to recurring mutational events such as those resulting in genomic disorders (segmental duplications, etc.)

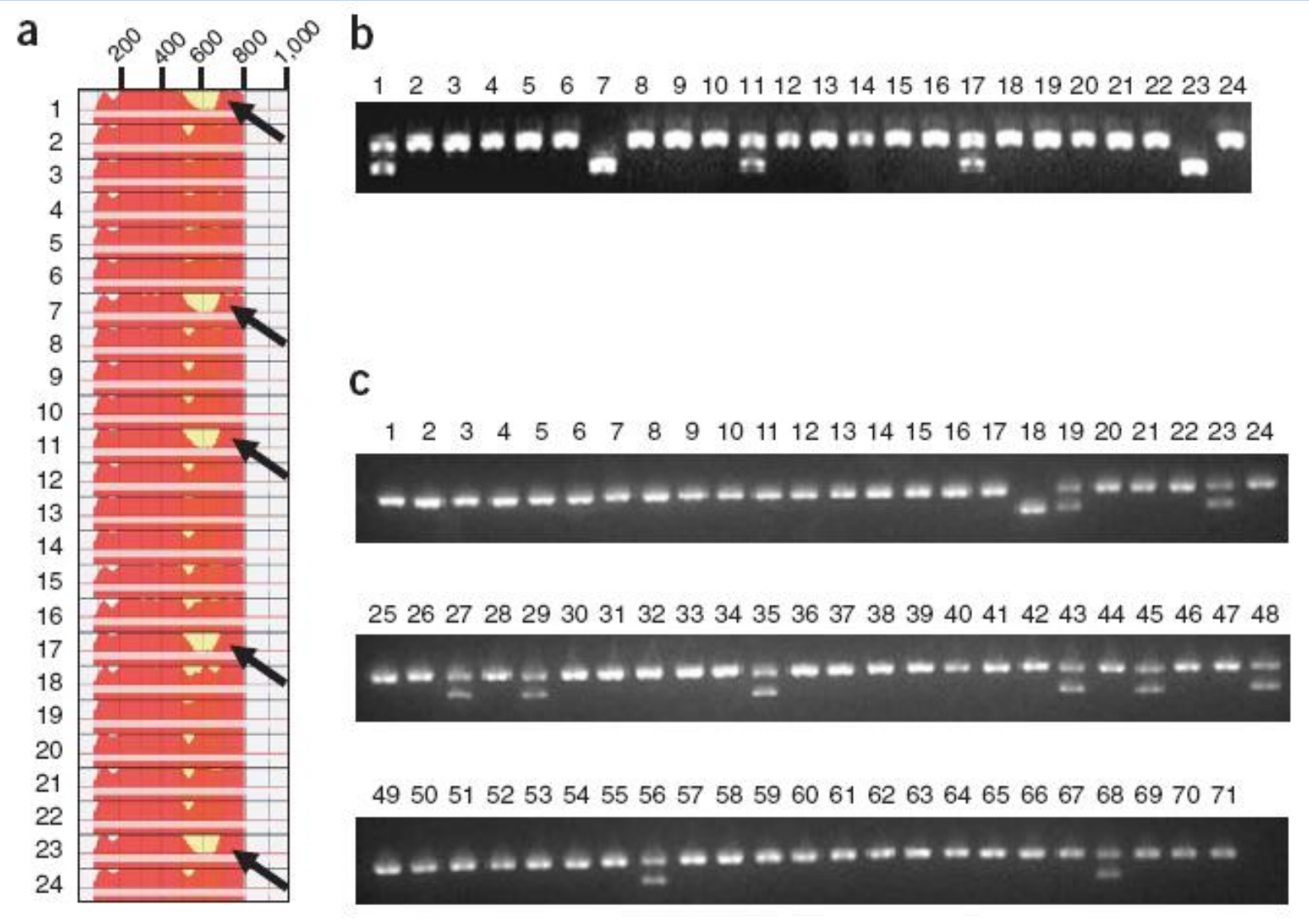
Common deletions and SNPs are in linkage disequilibrium in the human genome. Hinds *et al.*

- To identify human deletions, they used an array-based comparative genomics approach to examine 24 unrelated individuals (termed the ‘Discovery Panel’) obtained from the Polymorphism Discovery Resource

Common deletions and SNPs are in linkage disequilibrium in the human genome. Hinds *et al.*

- Detected 215 candidate deletions ranging from 70 bp to 10 kb
- X of them validated by PCR using diploid genomic DNA from each of the 24 individuals in the Discovery Panel
- 100 deletions selected for linkage analysis

Common deletions and SNPs are in linkage disequilibrium in the human genome. Hinds *et al.*



Common deletions and SNPs are in linkage disequilibrium in the human genome. Hinds *et al.*

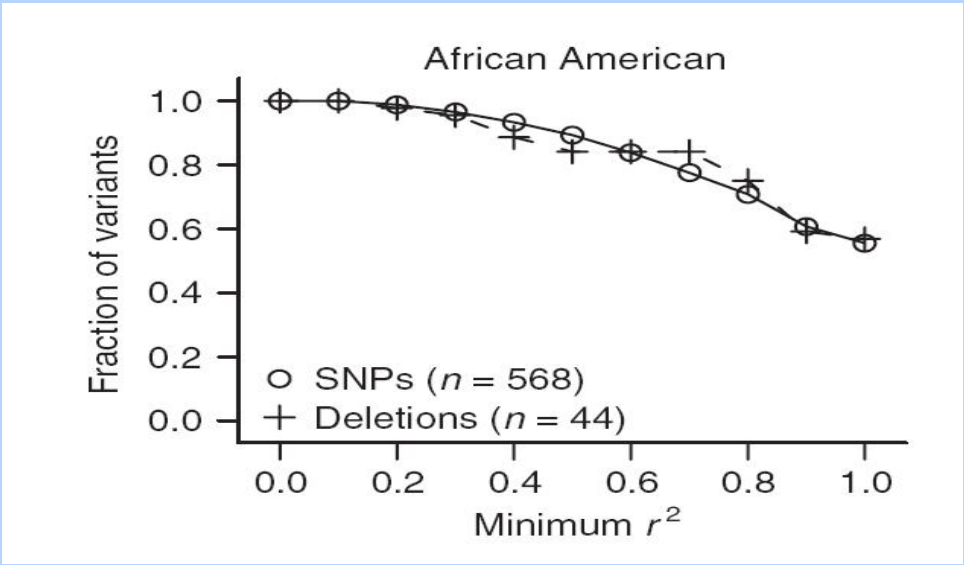
- 71 samples (termed the ‘Diversity Panel’) were obtained from the Human Variation Collection comprise individuals from three populations:
  - 23 African Americans
  - 24 European Americans
  - 24 Han Chinese.

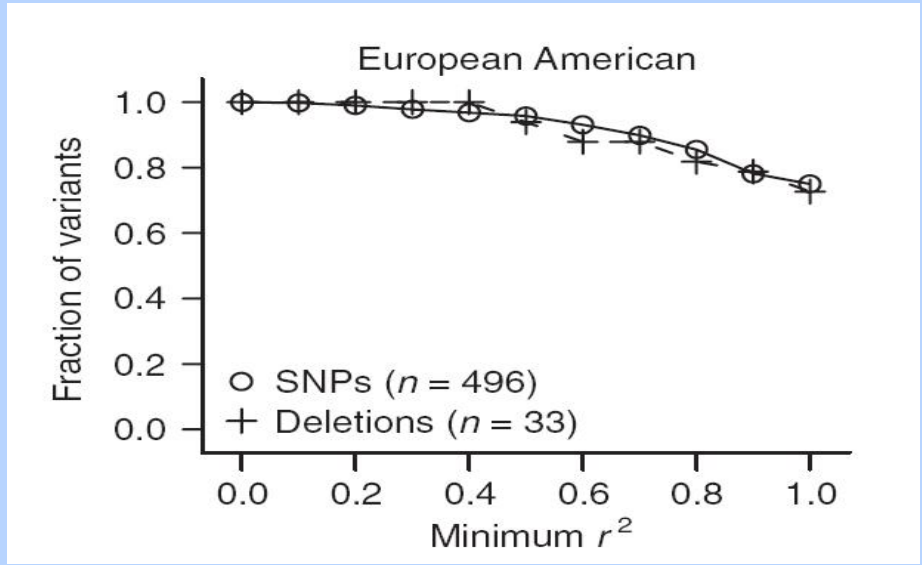
- PCR to determine genotypes for each of the 100 deletions in a larger group of 71 ethnically diverse individuals.

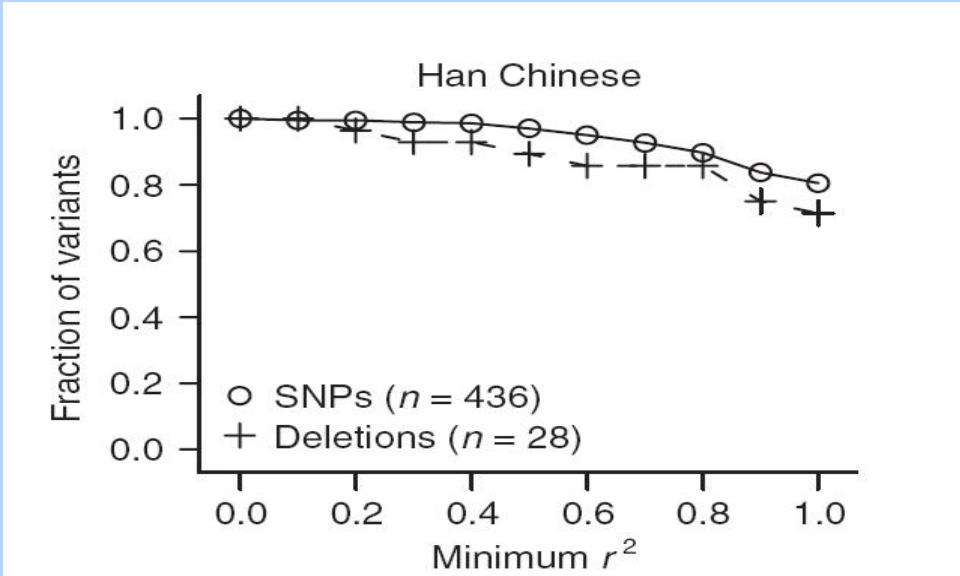
Common deletions and SNPs are in linkage disequilibrium in the human genome. Hinds *et al.*

- For each common deletion, determined the minimum square of the correlation coefficient (r^2) with any genotyped common SNP located within 50 kb on either side of the deletion.
- In each of the three populations, deletion polymorphisms showed comparable linkage disequilibrium to SNPs.

Common deletions and SNPs are in linkage disequilibrium in the human genome. Hinds *et al.*







Common deletions and SNPs are in linkage disequilibrium in the human genome. Hinds *et al.*

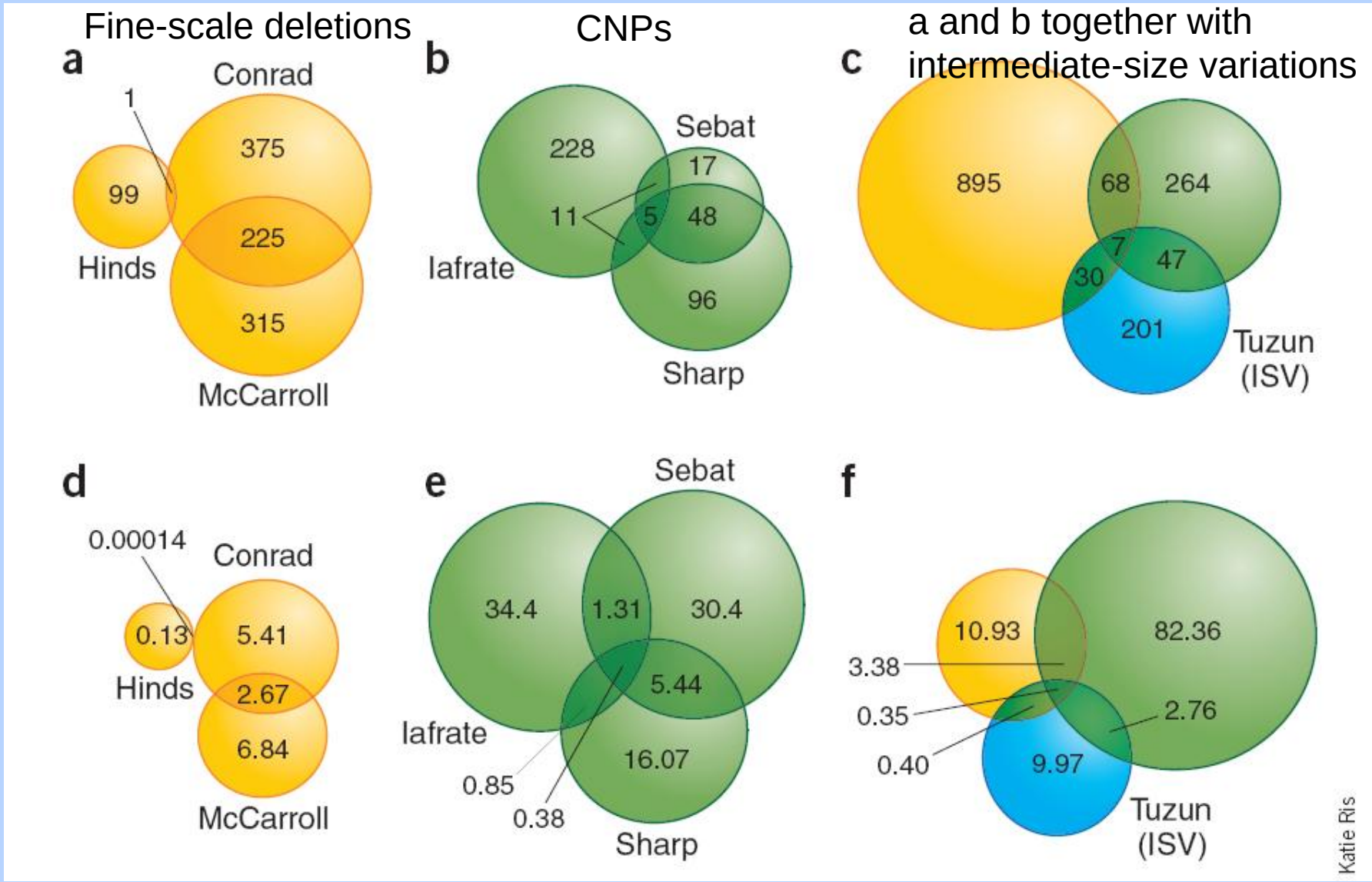
- The set of common intermediate-length deletions identified has linkage disequilibrium patterns similar to SNPs, indicating that these polymorphisms share a similar evolutionary history and suggesting that most intermediate-length deletions, like SNPs, arose once in human history.

Widening the spectrum of human genetic variation. Eichler E.E.

- Uncovering the genetic basis of human phenotypic differences requires a comprehensive understanding of all forms of genetic variation.

- Advances in deducing the pattern and nature of SNPs, but similar realization for larger and more complex forms of genetic variation has lagged behind.

Widening the spectrum of human genetic variation. Eichler E.E.



Widening the spectrum of human genetic variation. Eichler E.E.

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.

Widening the spectrum of human genetic variation. Eichler E.E.

- Oligonucleotide-based microarrays specifically designed to detect SNPs frequently exclude repeat-rich regions of the genome at both the long-range PCR and microarray.

- Regions near centromeres, telomeres or segmental duplications are generally not classified as part of the ‘Hapmappable genome’, and therefore there is a dearth of corresponding genotype data.

Widening the spectrum of human genetic variation. Eichler E.E.

- No single optimized approach has been developed yet to systematically capture all structural variation in the human genome.

- It is likely that several thousand additional common structural variants await discovery.

References:

- Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. (2006). **A high-resolution survey of deletion polymorphism in the human genome.** Nat Genet 38, 75-81.

- Hinds, D. A., Kloek, A. P., Jen, M., Chen, X. & Frazer, K. A. (2006). **Common deletions and SNPs are in linkage disequilibrium in the human genome.** Nat Genet 38, 82-85.

- Eichler, E. E. (2006). **Widening the spectrum of human genetic variation.** Nat Genet 38, 9-11.