LETTERS

Genotype, haplotype and copy-number variation in worldwide human populations

Mattias Jakobsson[1,2]*, Sonja W. Scholz[4,5]*, Paul Scheet[1,3]*, J. Raphael Gibbs[4,5], Jenna M. VanLiere[1], Hon-Chung Fung[4,6], Zachary A. Szpiech[1], James H. Degnan[1,2], Kai Wang[7], Rita Guerreiro[4,8], Jose M. Bras[4,8], Jennifer C. Schymick[4,9], Dena G. Hernandez[4], Bryan J. Traynor[4,10], Javier Simon-Sanchez[4,11], Mar Matarin[4], Angela Britton[4], Joyce van de Leemput[4,5], Ian Rafferty[4], Maja Bucan[7], Howard M. Cann[12], John A. Hardy[5], Noah A. Rosenberg[1,2,3] & Andrew B. Singleton[4,13]

Genome-wide patterns of variation across individuals provide a powerful source of data for uncovering the history of migration, range expansion, and adaptation of the human species. However, high-resolution surveys of variation in genotype, haplotype and copy number have generally focused on a small number of population groups[1–3]. Here we report the analysis of high-quality genotypes at 525,910 single-nucleotide polymorphisms (SNPs) and 396 copy-number-variable loci in a worldwide sample of 29 populations. Analysis of SNP genotypes yields strongly supported fine-scale inferences about population structure. Increasing linkage disequilibrium is observed with increasing geographic distance from Africa, as expected under a serial founder effect for the out-of-Africa spread of human populations. New approaches for haplotype analysis produce inferences about population structure that complement results based on unphased SNPs. Despite a difference from SNPs in the frequency spectrum of the copy-number variants (CNVs) detected—including a comparatively large number of CNVs in previously unexamined populations from Oceania and the Americas—the global distribution of CNVs largely accords with population structure analyses for SNP data sets of similar size. Our results produce new inferences about inter-population variation, support the utility of CNVs in human population-genetic research, and serve as a genomic resource for human-genetic studies in diverse worldwide populations.

The Human Genome Diversity Project (HGDP) was initiated for the purpose of assessing worldwide genetic diversity, providing cell lines maintained at the Centre d'Étude du Polymorphisme Humain (CEPH) for use in population-genetic studies[4]. We genotyped a geographically broad subset of 485 individuals from the HGDP–CEPH panel, with complete inclusion of HGDP–CEPH Africans (Supplementary Fig. 1). After correction for sample size differences across geographic regions[5], 81.17% of SNP alleles were observed in all five of the main regions (Fig. 1a). The next most frequently observed geographic distributions represented alleles found everywhere except Oceania (3.80%), everywhere except the Americas (3.01%), and everywhere except Africa (2.20%). Regionally private alleles were uncommon: 0.91% for Africa, 0.75% for Eurasia (Europe, Central/South Asia and the Middle East, including North Africa), and near zero for other regions.

Genomic analysis of population structure produced higher-resolution inferences than have previously been obtained. In a neighbour-joining population tree based on allele-sharing distance, with one exception, all internal branches were supported by all 1,000 bootstrap replicates across loci (Fig. 1b); nine replicates grouped the Adygei population with Russians and Basques. The tree supports the clustering of each of the main geographic regions and contains a separation of African hunter-gatherers (San, Mbuti and Biaka) from other Africans.

Bayesian cluster analysis[6] was largely concordant with previous analyses of microsatellite and short insertion–deletion polymorphisms[7–9]. Analysis with six clusters revealed groupings corresponding to five geographic subdivisions separated by major barriers, with a cline longitudinally across Asia and with a sixth cluster centred on the Kalash population of Pakistan (Fig. 1c). Within geographic regions, the cluster analysis subdivided groupings that were observed previously with fewer markers[9] (Fig. 1c and Supplementary Fig. 2).

Multidimensional scaling (MDS) separated the populations of different geographic regions (Fig. 1d), including Europe, Central/South Asia and the Middle East, which clustered together in the global bayesian analysis. Within regions, MDS split the individuals of distinct populations into distinct clusters (Supplementary Fig. 3), even in some cases for which bayesian analysis produced little separation between populations. The possibility of placing the MDS graph in approximate geographical orientation, with latitude and longitude representing the vertical and horizontal axes, suggests that geographic distance is a primary determinant of human genetic differentiation[10,11]. This view is supported by a linear increase in genetic distance with geographic distance from East Africa (Fig. 2a).

Linkage disequilibrium (LD), as obtained with the homozygosity-based HR^2 measure[12], declined as a function of physical distance, with the highest values occurring in the Americas, followed by Oceania, East Asia, Eurasia and Africa (Fig. 2b). Only two populations deviated from this pattern—Maya, a potentially admixed group, and Kalash, a population isolate. Although reduced LD has consistently been

[1]Center for Computational Medicine and Biology, [2]Department of Human Genetics, [3]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA. [4]Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, Maryland 20892, USA. [5]Department of Molecular Neuroscience and Reta Lila Weston Institute of Neurological Studies, Institute of Neurology, University College London, Queen Square, London WC1N 3BG, UK. [6]Department of Neurology, Chang Gung Memorial Hospital and College of Medicine, Chang Gung University, Taipei 10591, Taiwan. [7]Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [8]Center for Neurosciences and Cell Biology, Faculty of Medicine, University of Coimbra, 3004-504 Coimbra, Portugal. [9]University of Oxford, Department of Clinical Neurology, John Radcliffe Hospital, Oxford OX3 9DU, UK. [10]Neurogenetics Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland 20892, USA. [11]Unidad de Genética Molecular, Departamento de Genómica y Proteómica, Instituto de Biomedicina de Valencia-CSIC, 46010, Valencia, Spain. [12]Fondation Jean Dausset – Centre d'Étude du Polymorphisme Humain (CEPH), 27 rue Juliette Dodu, 75010 Paris, France. [13]Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia 22908, USA.
*These authors contributed equally to this work.

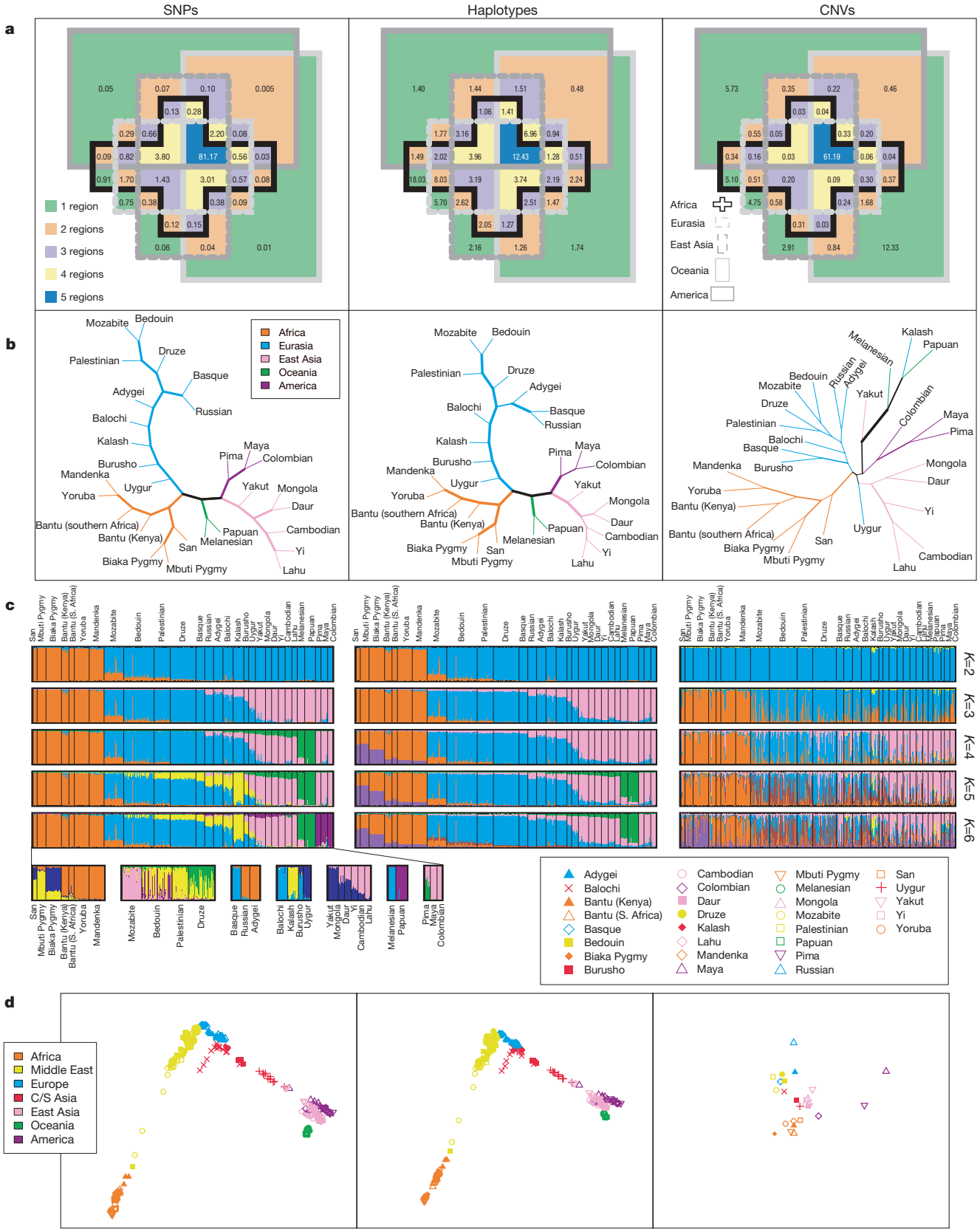

Figure 1 | SNP, haplotype, and copy-number variation across populations. **a**, Venn diagram of the percentages of alleles with particular geographic distributions. **b**, Neighbour-joining trees of population relationships. Internal branch lengths are proportional to bootstrap support. Lines of intermediate thickness represent internal branches with more than 50% bootstrap support, and the thickest lines represent more than 95% support. **c**, Population structure inferred by bayesian clustering. Each individual is shown as a thin vertical line partitioned into K coloured components representing inferred membership in K genetic clusters. The bottom row provides inferred population structure for each geographic region. **d**, MDS representations of genetic distances between individuals (SNPs and haplotypes) and populations (CNVs). C/S Asia, Central/South Asia.

observed in Africa, LD levels in non-African groups have been difficult to rank[13–16]. We observed that, with high precision, LD increased with geographic distance from East Africa (Fig. 2c). This pattern matches the prediction from a model of sequential founder effects during spatial expansion from Africa[11], because such founder effects would be expected to increase LD at each step of the expansion[15,17].

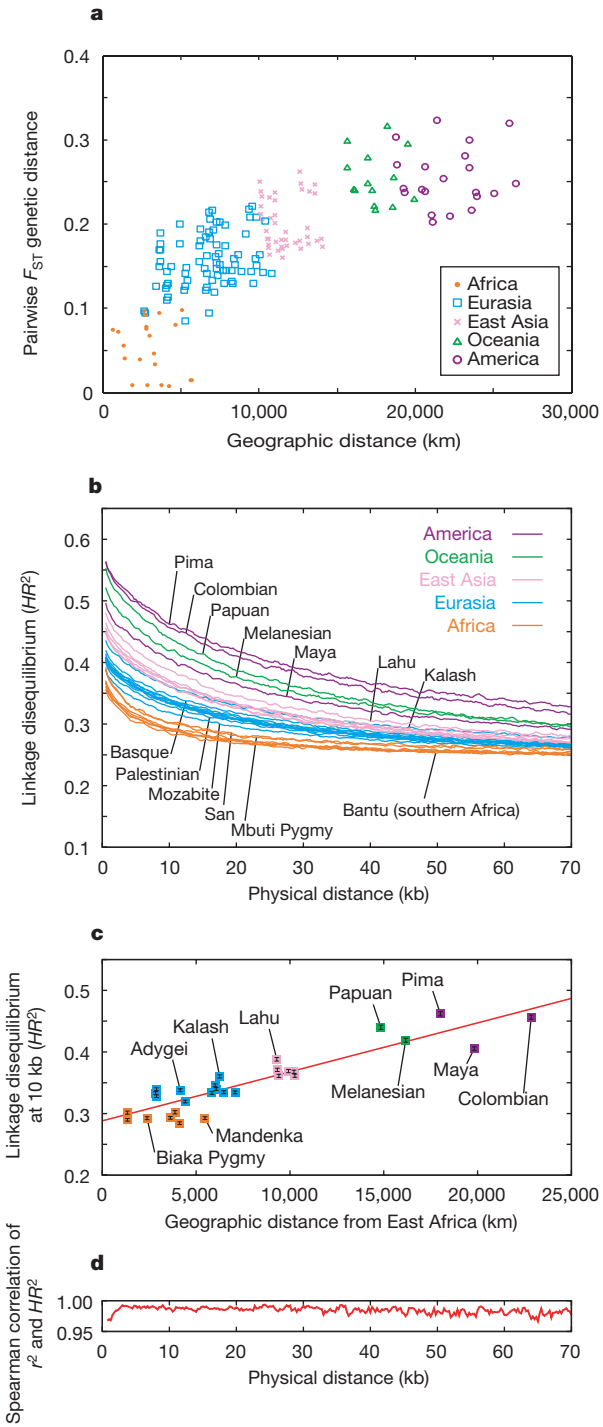

Figure 2 | Genetic distance and linkage disequilibrium. **a**, F_{ST} genetic distance as a function of land-based geographic distance from East Africa. **b**, LD as a function of physical distance. kb, kilobases. **c**, LD as a function of geographic distance from East Africa. Error bars (smaller than symbol size) represent the mean ± 1.96 times the s.e.m. **d**, Correlation of population rank orders by LD, comparing HR^2 applied to unphased data and r^2 applied to phased data. LD calculations are adjusted for sample size differences across populations by sampling five random individuals (HR^2) or ten random haplotypes (r^2) per population at each SNP pair.

To circumvent possible biases in SNP selection procedures[13], we also analysed estimated haplotypes. In comparison with the pattern for HR^2, a nearly identical LD decay was observed with the r^2 measure applied to phased data (Supplementary Fig. 4). The correlation of population ranks by HR^2 and r^2 levels exceeded 0.95 across a wide range of physical distances (Fig. 2d).

For further assessment of haplotype variation, we devised a new approach that avoided the difficulty of choosing window lengths for haplotypic analysis. Variation is summarized locally at each point in the genome by using a collection of 20 'haplotype clusters', each of which represents a group of haplotypes that overlap the point. For every population, frequencies for the various haplotype clusters are estimated at each SNP. Example illustrations of these frequencies are shown in Fig. 3 in the vicinity of the lactase gene (*LCT*). A decrease in haplotype diversity in Europe, particularly in the CEU population (Utah residents with ancestry from northern and western Europe), is apparent from the predominance of a single haplotype cluster well beyond *LCT*. This pattern accords with evidence that *LCT* has recently undergone a selective sweep[1,18,19], because such sweeps are expected to generate high-frequency uninterrupted haplotypes surrounding the selected region. By contrast, the reduced diversity in the Americas and Oceania probably reflects founder events and consequently greater haplotype lengths genome-wide (Supplementary Figs 5–7).

To make use of haplotypes in population structure analysis, we generated ten haplotype cluster data sets, each of which assigned each individual two haplotype clusters at every point along the genome, with both cluster memberships ranging from 1 to 20. The ten data sets were then analysed with the same methods as those used for unphased genotypes, treating distinct clusters in the same manner as distinct alleles.

Only 12.43% of haplotype clusters were observed in all five regions, whereas 18.03% were private to Africa (Fig. 1a). Geographically localized haplotype clusters were considerably more common than localized SNP alleles, with 51.87% of clusters being found in at most two regions, in contrast with 4.66% of SNP alleles. Despite these differences in geographic distributions, the haplotype-based neighbour-joining tree had an identical shape to the SNP-based tree, except for a Basque–Russian–Adygei grouping (Fig. 1b), and haplotype-based and SNP-based MDS plots were extremely similar (Fig. 1d). Bayesian clusters with haplotype data matched those in the unphased analysis, except that the haplotypically diverse Africans quickly split into a cluster partly corresponding to African hunter-gatherers and a cluster for the other African populations, and Native Americans and Kalash did not separate (Fig. 1c). The general agreement of SNP-based and haplotype-based analyses suggests that at the high density considered, unphased SNPs provide considerable population structure information, although haplotype data can contribute an additional informative component for population structure analysis. Haplotype-based subdivision of Africans suggests a preference for splitting the highest-diversity groups over separating relatively isolated populations—Kalash and Native Americans—whose haplotypes largely represent subsets of those seen in neighbouring groups.

In conjunction with SNP typing, we identified CNVs by using PennCNV[20], a CNV-calling program that relies on SNP allele frequencies, SNP spacing, and genotyping signal intensities and allelic intensity ratios normalized by signals for a reference panel. We detected 3,552 CNVs at 1,428 copy-number-variable loci, including 507 loci at which CNVs have not previously been reported. Sufficient reliability of CNV genotypes for population-genetic analysis is supported by the observation that all CNVs detectable by using consecutive heterozygous genotypes on male X chromosomes were also identified from signal intensity (Supplementary Figs 8 and 9), by a combined false-positive and false-negative rate of 9% reported for PennCNV[20], and by a false-positive rate below 0.7% as estimated from duplicate samples[21] (Supplementary Figs 10 and 11). For analyses of population structure (Fig. 1), the CNV data set

was restricted to 396 non-singleton autosomal loci in 405 unrelated individuals.

CNVs tended to have low frequencies worldwide: only one CNV frequency exceeded 10% (Supplementary Fig. 12). Within geographic regions, however, higher-frequency CNVs were more common, especially in Oceania and the Americas (Fig. 4a and Supplementary Fig. 13). Consistent with this trend, three of the four populations with the greatest numbers of CNVs detected per individual occurred in these regions, the fourth being Kalash (Fig. 4b). In contrast with their usual reduced variation[11,13], populations from Oceania and the Americas had more CNV loci and more previously unobserved CNV loci than most other populations. The number of private CNVs was larger for Oceania than for Africa and Eurasia (Fig. 1a), a pattern not observed with SNP and haplotype variation. Private CNVs were more common than private SNP alleles, and for CNVs the percentage observed in all five regions, 61.19%, was smaller than for SNPs. The excess of rare and localized variants is probably due in part to comparison with preselected known SNPs, but it accords with a skew towards rare variants in CNVs observed with other genotyping technologies[22,23]. However, some bias may exist in CNV detection; as a result of difficulties in detecting high-frequency CNVs from comparisons against reference intensities[24], the absence from the reference panel of Kalash and populations from Oceania and the Americas may have increased the potential for identifying CNVs in these groups. In such distinctive populations, unusual intensity signals for deletions or duplications are less likely to have been diluted by inclusion in the reference panel of individuals with an atypical copy number.

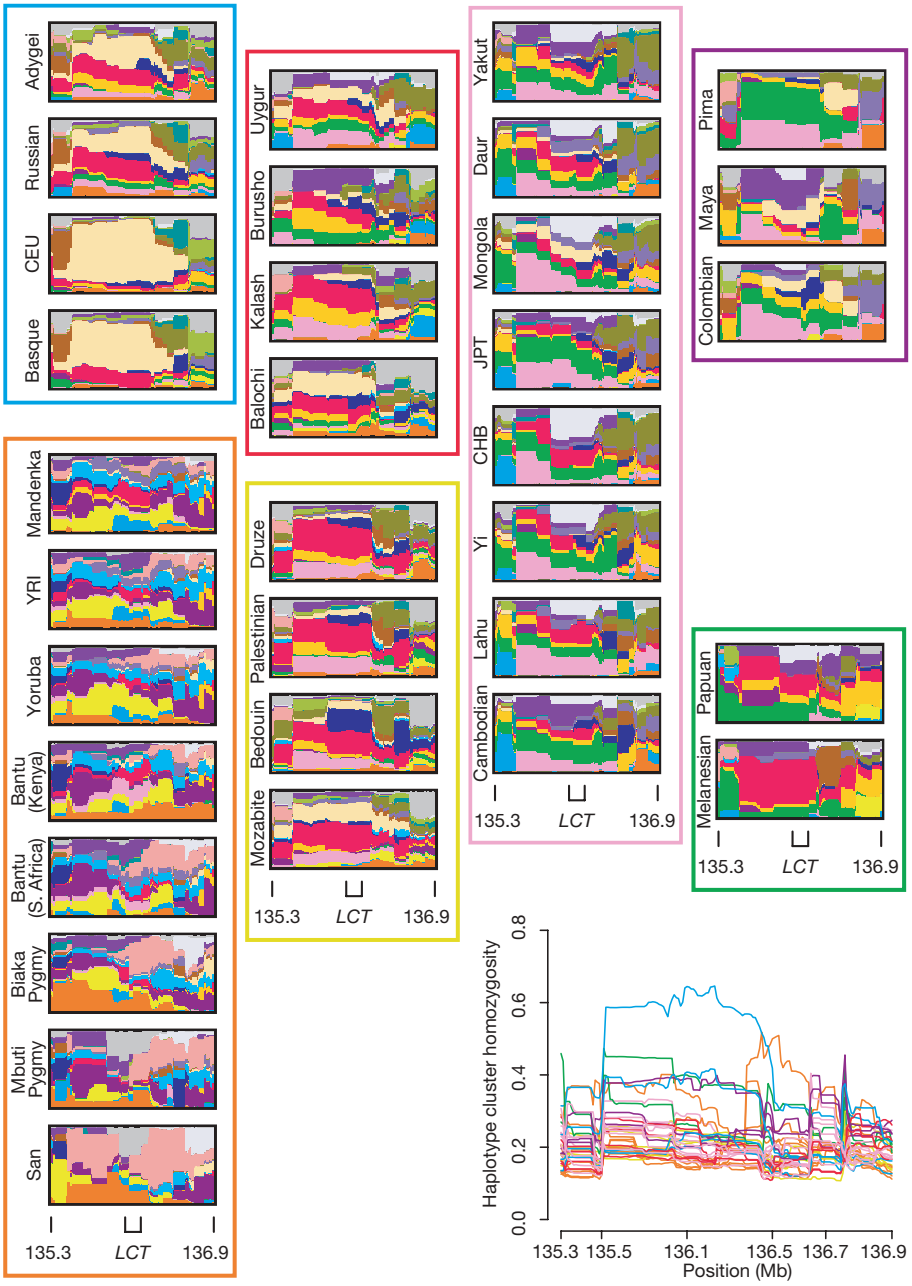

Figure 3 | Haplotype cluster frequencies for 156 consecutive SNPs on chromosome 2 in the region surrounding the *LCT* gene (136.373–136.478 megabases). At each SNP, relative frequencies of haplotype clusters are displayed on a thin vertical line. Each colour depicts a haplotype cluster, and the proportion in a colour gives the frequency of 1 of 20 distinct clusters. Interpretation of colours is made locally, as clustering varies along the chromosome, reflecting a gradual decay of LD. Moving horizontally, changes in colour patterns illustrate the change in haplotypic composition across physical position. CEU, Utah residents with ancestry from northern and western Europe; CHB, Han Chinese from Beijing; JPT, Japanese from Tokyo; YRI, Yoruba from Ibadan, Nigeria.

Partial similarity was observed between population structure inferred for CNVs and that inferred from considerably larger SNP and haplotype data sets. In the population tree, major geographic regions largely formed separate branches, but with different lower-level groupings than in the SNP and haplotype trees, and with less support (Fig. 1b); the unexpected grouping of Kalash, Melanesian and Papuan probably results from long-branch attraction during neighbour-joining analysis of their large numbers of CNVs (Supplementary Tables 1 and 2). Bayesian cluster analysis separated populations from Africa, Eurasia and the combination of East Asia, Oceania and the Americas, but with considerable variation across individuals (Fig. 1c). MDS revealed some degree of geographic clustering, but only after removal of the three outliers that also appear in the population tree (Fig. 1d and Supplementary Fig. 14). The degree of difference between CNV and SNP population structure results is comparable to that obtained with subsets of the SNP data set with the same size as the CNV data set (Supplementary Figs 15 and 16, and Supplementary Tables 3 and 4). Thus, partial correspondence of CNV population structure patterns to those observed for SNPs and haplotypes supports the general reliability of the CNV genotyping and suggests some similarity in the evolutionary history of CNV loci to the histories of other types of marker.

The availability of worldwide high-density SNP data will be important for improving the prospects for disease-gene mapping in a broad set of populations. By employing methods that make use of high-resolution data sets to impute genotypes in study samples[25], it will be possible to increase power to detect associations in diverse populations for which such data have not previously been available. The data also provide the basis for refining informative marker sets in contexts such as multi-population SNP tagging[26], admixture mapping and ancestry inference, and for evaluating SNP tagging of CNVs for disease association tests[3,22]. Because effective tagging may require high r^2 values between markers, and because high r^2 occurs only for markers with similar allele frequencies[27], a difference in SNP and CNV allele frequency spectra suggests that ideal SNP sets for tagging CNVs may require a considerable fraction of rare variants. Finally, our detection of novel copy-number-variable loci in a population panel broader than those used in previous CNV analyses highlights the importance of considering diverse worldwide populations for full characterization of the pattern of human genetic variation.

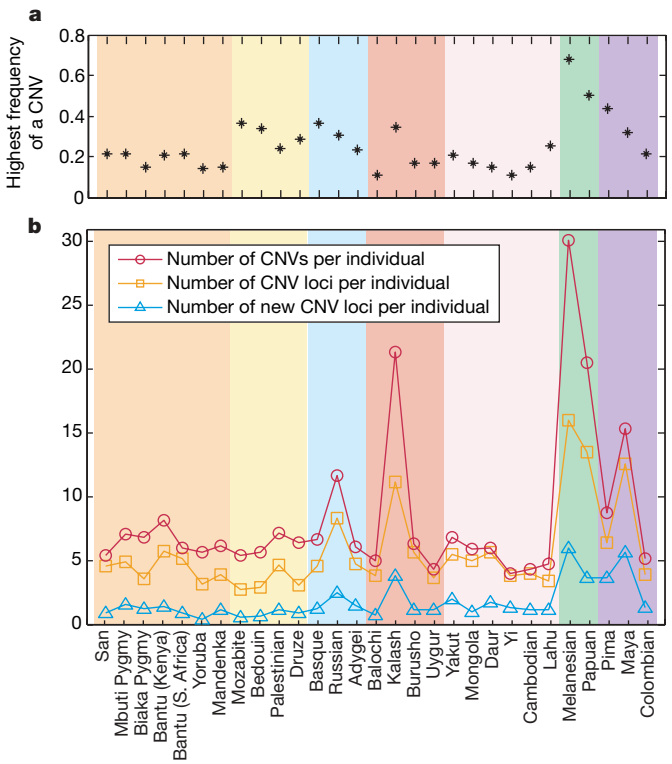

Figure 4 | CNVs across populations, based on 3,552 CNVs at 1,428 copy-number-variable loci. a, Highest frequency of any autosomal CNV in each of 29 populations. **b**, Mean number of CNVs observed per individual. Number of CNVs per individual refers to the number of CNVs considering all individuals in a population, divided by sample size; number of (new) CNV loci refers to the number of (new) CNV loci polymorphic in a population, divided by sample size. To be identified as new, we required that a CNV not overlap with existing CNVs in the Database of Genomic Variants[30] (version hg18.v3). Background colours indicate geographic regions.

METHODS SUMMARY

SNPs. Genotyping used Illumina Infinium HumanHap550 BeadChips. HGDP–CEPH genotypes were augmented with HumanHap550 genotypes of 112 HapMap individuals. Most analyses used 512,762 high-quality autosomal SNPs in 443 unrelated HGDP–CEPH individuals. Data appear at http://neurogenetics.nia.nih.gov/paperdata/public/ and http://www.cephb.fr/hgdp-cephdb/.

Haplotypes. Phasing with fastPHASE[28] used 20 haplotype clusters, combining HGDP–CEPH and HapMap individuals, and employing geographic region labels to enhance accuracy[13]. Relatives were subsequently removed. For each individual, at each SNP, probabilities were obtained for the haplotype cluster memberships of the two unobserved haplotypes of the individual, averaging across individuals to produce cluster 'frequencies' for each population. Haplotype cluster data sets were constructed by taking (for each chromosome) ten independent samples from the conditional distribution of chromosome-wide memberships given the unphased genotypes and the estimated parameters of the model underlying fastPHASE. Cluster data set preparation for population structure analysis ignored geographic labels.

CNVs. CNV detection employed a ten-SNP minimum to increase the reliability of calls[20]. Copy-number-variable loci were identified as regions with CNVs. One-copy changes (one allele duplicated or deleted) were tabulated as one CNV; two-copy changes were tabulated as two CNVs.

Data analysis. Rarefaction computations[5] of mean numbers of variants per locus private to each of 31 combinations of geographic regions used equal samples of 35 chromosomes per region. Percentages shown equal these 31 values, normalized by their sum. Trees were obtained from 1,000 bootstraps across loci; for haplotypes, bootstraps were split evenly across the ten data sets. Bayesian clustering used 40 replicates, using 1% of the SNP and haplotype data to avoid markers in LD. 'Replicates' included different 1% subsets (SNPs, haplotypes), different data sets (haplotypes) and separate runs with identical data (SNPs, haplotypes, CNVs). CLUMPP[29] was used to identify shared modes. For SNPs and CNVs, MDS used allele-sharing distance between individuals; for haplotypes, it used euclidean distance between cluster membership vectors.

Received 2 December 2007; accepted 29 January 2008.

1. The International Haplotype Map Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
2. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
3. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
4. Cann, H. M. *et al.* A human genome diversity cell line panel. *Science* **296**, 261–262 (2002).
5. Kalinowski, S. T. Counting alleles with rarefaction: private alleles and hierarchical sampling designs. *Conserv. Genet.* **5**, 539–543 (2004).
6. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
7. Bastos-Rodrigues, L., Pimenta, J. R. & Pena, S. D. J. The genetic structure of human populations studied through short insertion–deletion polymorphisms. *Ann. Hum. Genet.* **70**, 658–665 (2006).
8. Rosenberg, N. A. *et al.* Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.* **1**, e70 (2005).
9. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
10. Lawson Handley, L. J., Manica, A., Goudet, J. & Balloux, F. Going the distance: human population genetics in a clinal world. *Trends Genet.* **23**, 432–439 (2007).
11. Ramachandran, S. *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl Acad. Sci. USA* **102**, 15942–15947 (2005).

12. Sabatti, C. & Risch, N. Homozygosity and linkage disequilibrium. *Genetics* **160,** 1707–1719 (2002).
13. Conrad, D. F. *et al.* A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genet.* **38,** 1251–1260 (2006).
14. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229 (2002).
15. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411,** 199–204 (2001).
16. Tishkoff, S. A. & Kidd, K. K. Implications of biogeography of human populations for 'race' and medicine. *Nature Genet.* **36,** S21–S27 (2004).
17. McVean, G. A. T. A genealogical interpretation of linkage disequilibrium. *Genetics* **162,** 987–991 (2002).
18. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74,** 1111–1120 (2004).
19. Tishkoff, S. A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genet.* **39,** 31–40 (2007).
20. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17,** 1665–1674 (2007).
21. Wong, K. K. *et al.* A comprehensive analysis of common copy-number variations in the human genome. *Am. J. Hum. Genet.* **80,** 91–104 (2007).
22. Locke, D. P. *et al.* Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79,** 275–290 (2006).
23. Sharp, A. J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77,** 78–88 (2005).
24. Scherer, S. W. *et al.* Challenges and standards in integrating surveys of structural variation. *Nature Genet.* **39,** S7–S15 (2007).
25. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* **3,** e114 (2007).
26. Need, A. C. & Goldstein, D. B. Genome-wide tagging for everyone. *Nature Genet.* **38,** 1227–1228 (2006).
27. Eberle, M. A., Rieder, M. J., Kruglyak, L. & Nickerson, D. A. Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. *PLoS Genet.* **2,** e142 (2006).
28. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78,** 629–644 (2006).
29. Jakobsson, M. & Rosenberg, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23,** 1801–1806 (2007).
30. Zhang, J., Feuk, L., Duggan, G. E., Khaja, R. & Scherer, S. W. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet. Genome Res.* **115,** 205–214 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the Biological Resource Center at the Fondation Jean Dausset – CEPH for preparing HGDP–CEPH diversity panel DNA samples, and S. Chanock and A. Hutchinson for assistance with the DNAs. This work was supported in part by NIH grants, by a postdoctoral fellowship from the University of Michigan Center for Genetics in Health and Medicine, by grants from the Alfred P. Sloan Foundation and the Burroughs Wellcome Fund, by the National Center for Minority Health and Health Disparities, and by the Intramural Program of the National Institute on Aging. The study used the Biowulf Linux cluster at the National Institutes of Health (http://biowulf.nih.gov).

Author Contributions N.A.R. and A.B.S. wish to be regarded as joint last authors.

Author Information The array data described in this paper are deposited in the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) under accession number GSE10331. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to N.A.R. (rnoah@umich.edu) or A.B.S. (singleta@mail.nih.gov).