nature genetics

Structural haplotypes and recent evolution of the human 17q21.31 region

Linda M Boettger[1–4], Robert E Handsaker[1,2,4], Michael C Zody[1,2] & Steven A McCarroll[1,2]

Structurally complex genomic regions are not yet well understood. One such locus, human chromosome 17q21.31, contains a megabase-long inversion polymorphism[1], many uncharacterized copy-number variations (CNVs) and markers that associate with female fertility[1], female meiotic recombination[1–3] and neurological disease[4,5]. Additionally, the inverted H2 form of 17q21.31 seems to be positively selected in Europeans[1]. We developed a population genetics approach to analyze complex genome structures and identified nine segregating structural forms of 17q21.31. Both the H1 and H2 forms of the 17q21.31 inversion polymorphism contain independently derived, partial duplications of the *KANSL1* gene; these duplications, which produce novel *KANSL1* transcripts, have both recently risen to high allele frequencies (26% and 19%) in Europeans. An older H2 form lacking such a duplication is present at low frequency in European and central African hunter-gatherer populations. We further show that complex genome structures can be analyzed by imputation from SNPs.

Simple, common deletion and duplication polymorphisms have been typed in large cohorts[6,7], have been found to segregate on SNP haplotypes[6,7] and have been associated with many human phenotypes via proxy SNPs[7–10]. Complex structural mutations have also been described in specific individuals and cancers[11]. By contrast, little is known about genomic loci that show population-level complexity—that is, loci at which germline structural mutations in many different ancestors have given rise to complex patterns of variation. Such loci have not been analyzed in HapMap[12,13] or the 1000 Genomes Project[14,15], as they are assumed to require reconstruction of each structural form from genomic clones or FISH.

We hypothesized that extensive structural information is present in the statistical relationships among genome structural features in populations. A population genetics approach to reconstructing structural forms of a complex locus would require two capabilities. First, individual structural features would need to be accurately typed in population cohorts. This is increasingly possible due to widely available genome sequence data[14] and approaches for typing structural polymorphism in such data[10]. Second, it would be necessary to infer the haplotypes formed by multiple structural features; such relationships might be made visible in haplotype sharing by relatives or statistical phasing in populations. We sought to understand the extent to which such an approach could reveal genome structures at 17q21.31.

We first examined what structural features of 17q21.31 vary in populations. In addition to the known inversion at 17q21.31, we used array and whole-genome sequencing (WGS) data to identify several segments of 17q21.31 that show distinct patterns of population-level variation in copy number (**Fig. 1a** and Online Methods). We located most of the boundaries of these segments at high resolution by finding read-depth transitions and then (where possible) breakpoint-spanning reads in data from the 1000 Genomes Project pilot[14] (Online Methods, **Supplementary Fig. 1** and **Supplementary Table 1**). This analysis defined three common, overlapping duplication polymorphisms: (i) duplication α, a 150-kb duplication (covering the 5′ end of the *KANSL1* gene, also known as *KIAA1267*), previously partially characterized on the H2 haplotype of 17q21.31 (ref. 16); (ii) duplication β, a novel, longer (300-kb) duplication overlapping duplication α but segregating separately from it; and (iii) duplication γ, a highly multi-allelic 218-kb duplication at the distal end of the 17q21.31 inversion, covering much of the *NSF* gene.

We next sought to type each of these structural features in populations. To address longstanding challenges in measuring the integer copy number of multi-allelic duplication CNVs[17], we deployed two new methods: (i) analysis of read depth applying the Genome STRiP algorithm[10] to WGS data from 946 unrelated individuals sampled in the 1000 Genomes Project[14] (**Fig. 1b–d** and Online Methods) and (ii) a droplet-based approach to digital PCR (ddPCR)[18] to analyze 120 parent-offspring trios from HapMap (**Fig. 1e–g** and Online Methods). These measurements of integer copy number, which varied from 2 to 8, were 99.1% concordant across 234 genotypes in overlapping samples, validating both methods (**Fig. 1h–j**). The integer copy numbers of these segments were then deconvolved into the contributions of the three overlapping duplication polymorphisms (**Fig. 1a**, **Supplementary Figs. 2** and **3** and **Supplementary Tables 2–9**). The state of the inversion polymorphism was inferred from more than 100 SNPs that seem to be in perfect linkage disequilibrium (LD) with the inversion.

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Program in Genetics and Genomics, Graduate Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to S.A.M. (mccarroll@genetics.med.harvard.edu).

Received 8 November 2011; accepted 1 June 2012; published online 1 July 2012; doi:10.1038/ng.2334

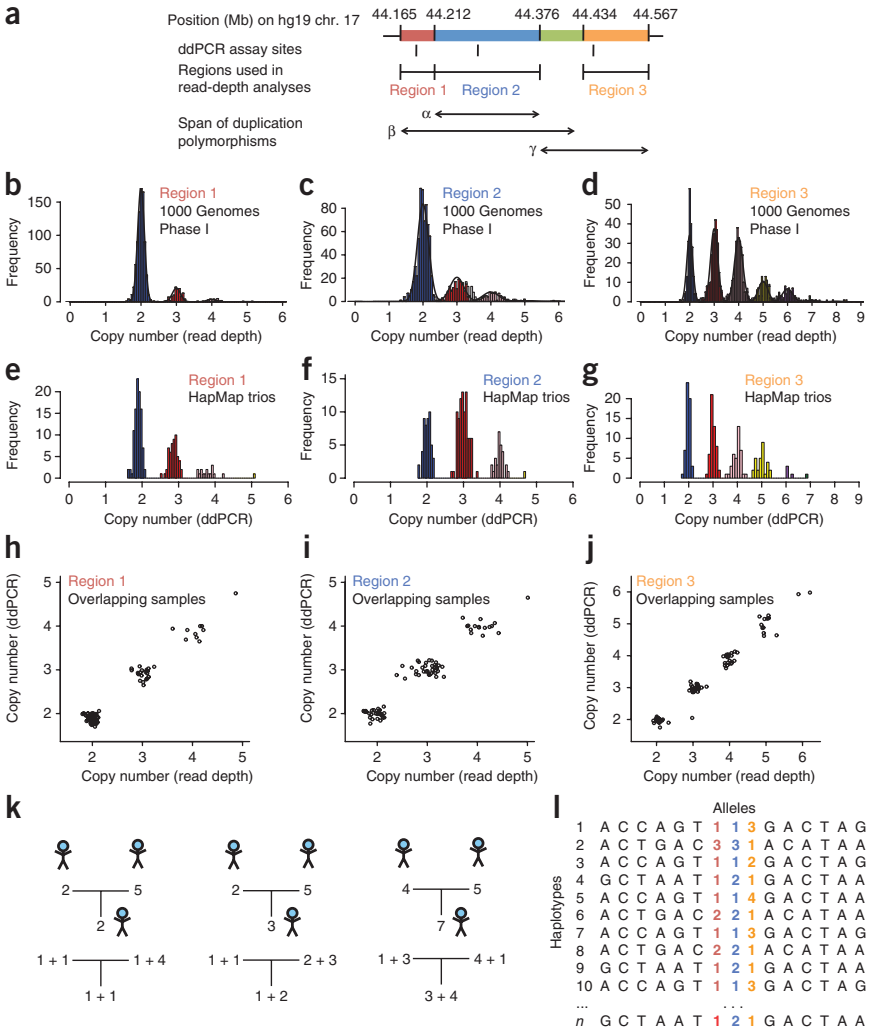

Figure 1 Inference of complex CNV and SNP haplotypes at the 17q21.31 locus. (**a**–**g**) Copy number of three copy number–variable segments of 17q21.31 (**a**) was measured in populations using two approaches: analysis of read depth in WGS libraries available for 942 individuals from the 1000 Genomes Project Phase 1, which we applied to measure copy number of region 1 (**b**), region 2 (**c**) and region 3 (**d**), and a ddPCR approach, which we applied to analyze parent-offspring trios from HapMap at specific sites within region 1 (**e**), region 2 (**f**) and region 3 (**g**). (Note that the frequencies of these copy-number classes are not identical in **b**–**d** and **e**–**g**, as their frequencies stratify by population, and the samples surveyed only partially overlap.) (**h**–**j**) Determinations of copy number were concordant for genomes analyzed by both methods in region 1 (**h**), region 2 (**i**) and region 3 (**j**). (**k**) Analysis of the segregation of copy-number levels in trios, which allowed the contribution of transmitted and untransmitted chromosomes to diploid copy number to be determined in most trios. (**l**) Example reference haplotypes created by phasing of CNV alleles with one another and with SNPs.

We then sought to understand these complex patterns of variation in terms of structural haplotypes—that is, which structural features segregate together. By inferring the number of copies of each CNV segment that segregated on transmitted and untransmitted haplotypes in each trio, we determined the chromosomal phase of each of these segmental copy numbers with respect to one another, with respect to the inversion polymorphism and with respect to SNPs across the locus (**Figs. 1k**,**l** and **2**, Online Methods and **Supplementary Table 10**). All four structural features (the three duplications and the inversion) were highly polymorphic, but they segregated as only nine common haplotypes (**Fig. 2**). Applying a maximum-likelihood model to unphased copy-number measurements that we derived from 1000 Genomes Project sequence data (**Fig. 1b**–**d**), we inferred frequencies of these haplotypes in 12 populations (**Fig. 2** and **Supplementary Table 11**).

The above analyses established the copy-number content of segregating haplotypes but did not establish the genomic locations of structural features. We inferred their locations from both sequence data (breakpoint-spanning reads) and LD to SNPs. Both forms of evidence indicated that duplications β and γ were tandem duplications (**Supplementary Table 1**). In contrast, duplication α seemed to be dispersed to a site 600 kb away from the original copy (**Fig. 2**, haplotype H2.α2), an observation that was reported earlier[16] and that we confirmed by reconstructing a clone spanning an earlier gap in the H2 sequence (**Supplementary Note**).

Knowledge of the structural haplotypes led to a structural phylogeny (**Fig. 2**) and candidate structural history (**Supplementary Note**), yielding several insights about the 17q21.31 locus. Although the H2 inversion form of 17q21.31 was reported to harbor little diversity[1], we found that human populations also possess an older, structurally distinct H2 haplotype at low frequency. Multiple lines of evidence indicate that this rarer H2 structural form (H2.α1) is the ancestral H2 structure. First, H2.α1 resembles the H1 structure more closely than does the previously described H2 structure[1,16] (H2.α2) (**Fig. 2**, **Supplementary Fig. 1** and **Supplementary Note**). Second, we identified H2.α1 haplotypes in central African hunter-gatherer populations (including two Mbuti Pygmies and one Biaka Pygmy, in 13 and 21 individuals sampled from those populations by the Human Genome Diversity Panel); such populations could have harbored H2.α1 over the long period (estimated at 2–3 million years[1,16]) during which H2 diverged from H1. The H2.α1 structure provides a potential missing link that would explain how the inversion could have occurred by a simple non-allelic homologous recombination event (**Supplementary Fig. 4** and **Supplementary Note**).

The H2 inversion state is common in west Eurasians and rare in most other populations, which has been attributed to recent positive selection[1]. We found that other structural variations at 17q21.31 show even greater population differentiation. Two distinct duplications (α and β in **Figs. 1a** and **2**), each affecting the 5′ coding exons of the *KANSL1* gene, have arisen independently on the H1 and H2 backgrounds. Both duplications have reached high allele frequency (19% and 26%, respectively) in west Eurasian populations, together comprising almost half of all forms of chromosome 17 in Europeans; however, we observed the α duplication only once and did not observe the β duplication in 502 east Asian chromosomes and 316 African chromosomes analyzed in data from Phase 1 of the 1000 Genomes Project (**Fig. 2** and **Supplementary Tables 4**–**9** and **11**), placing both duplications among the human genome's most population-differentiated

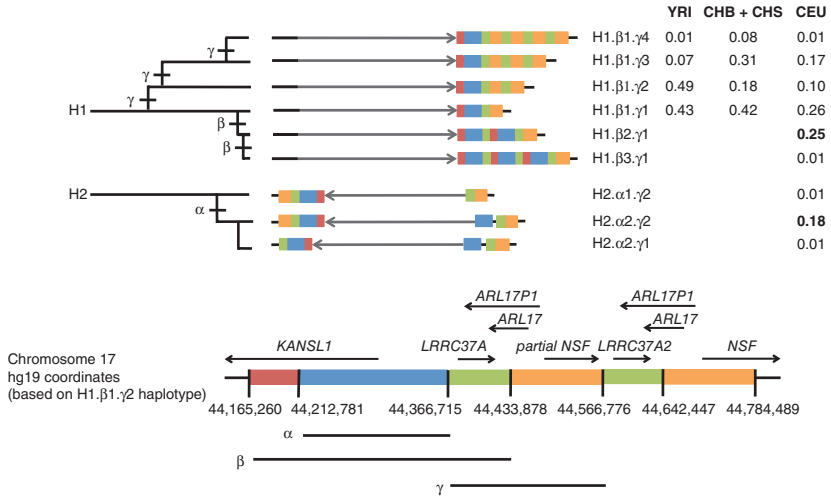

Figure 2 Structural forms of the human 17q21.31 locus and their population frequencies. Each haplotype is represented in a simplified form to highlight major structural differences. The schematic (bottom) indicates which genomic segment is represented by each color (detailed schematics in **Supplementary Figs. 1** and **7** and **Supplementary Table 1**). The gray arrows indicate the orientation of the unique inverted region within 17q21.31. The α, β and γ structural polymorphisms segregate as the nine common haplotypes shown. The table (right) lists allele frequencies for the nine structural haplotypes in different populations. YRI, Yoruba in Ibadan, Nigeria; CHB, Han Chinese in Beijing; CHS, Han Chinese South; CEU, Utah residents of Northern and Western European ancestry. Genotype and allele frequencies in 12 populations are available as **Supplementary Tables 2–9**. Most of these haplotypes correspond one to one to haplotypes identified in a parallel study by Steinberg *et al.*[30]: H1.β1.γ1 corresponds to H1.1; H1.β1.γ2 to H1.2; H1.β1.γ3 to H1.3; H1.β2.γ1 to H1D; H1.β3.γ1 to H1D.3; H2.α1.γ1 to H2.1; H2.α1.γ2 to H2.2; and H2.α2.γ2 to H2D.

polymorphisms (Online Methods and **Supplementary Fig. 5**). The α and β duplications have reached these highly differentiated allele frequencies in parallel at the same locus and in the same populations, in a pattern similar to that observed at other loci (such as the *LCT* and *APOL1* loci in African populations)[19,20] that have undergone recent selection.

We estimated two dates for each duplication: the time to coalescence of contemporary haplotypes and the age of the duplication events. The first can be estimated from the divergence of sequences flanking the duplications, the second by comparing the sequences of the duplication copies. To generate these data, we selectively captured and sequenced the 17q21.31 region in H1.β2 and H2.α2 homozygotes. We estimate the coalescence of the sampled β-duplicated H1 chromosomes at 12,000 years ago. Divergence of otherwise unique sequences within the β duplication suggests that the duplication itself occurred 20,000–27,000 years ago. For the α-duplicated H2 chromosomes, we estimate an average coalescence of 17,000 years ago, but the duplication itself seems to have occurred much earlier (>1 million years ago) than its rise to high frequency in west Eurasia (see **Supplementary Table 12** and **Supplementary Note** for details of dating and discussion of uncertainty surrounding the dates).

The parallel increases in frequency of duplication α (on H2) and duplication β (on H1) in the same populations invite the hypothesis that they could influence a common phenotype. Both duplications involve the 5′ exons of *KANSL1* (also called *KIAA1267*, *MSL1v1* and *CENP-36*). We found that both α and β duplications give rise to novel *KANSL1* transcripts (which we confirmed by RT-PCR and sequencing) in which the 5′ exons of *KANSL1* fuse to cryptic exons that terminate the coding sequence. Notably, a similar truncation in the *Drosophila melanogaster* ortholog of *KANSL1*, *GC4699/E(nos)*, was identified in a mutagenesis screen for modifiers of *Nanos* and was found to enhance the effect of a *Nanos* hypomorph on age-dependent female fertility and germline stem cell differentiation[21]. The precise role of *KANSL1* in these processes is unknown, although the encoded protein is found within the MOF-MSL1v1 chromatin-modifying complex[22,23]. Proteins arising from these novel *KANSL1* transcripts would contain the coiled-coil domain

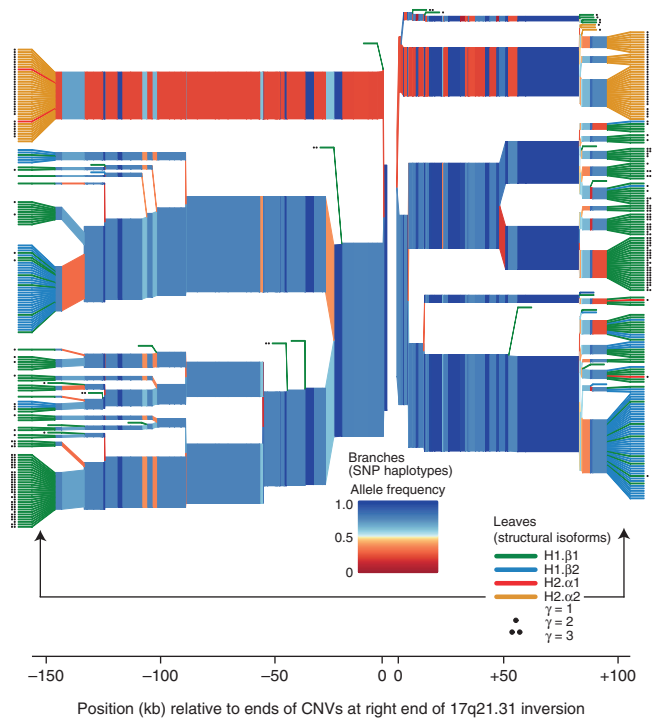

Figure 3 Structural forms of 17q21.31 segregate on specific SNP haplotype backgrounds. The plot shows homozygosity and divergence (due to mutation and recombination) of the SNP haplotypes on which each structural form segregates in the European (CEU) trios analyzed in HapMap 3. The polymorphic CNV copies at the right end of the 17q21.31 inversion (**Fig. 2**) reside between the two origins of this plot (center). SNPs on the left half of the plot therefore reside within the unique inverted region of 17q21.31, whereas SNPs on the right half of the plot are distal to the 17q21.31 inversion. Branch points represent markers at which the depicted haplotypes diverge due to mutation and/or recombination with other haplotypes. In the plot, the structures are represented on the leaves in order to clarify their relationships to SNP haplotypes, but the variable parts of these CNVs actually reside (in genomic space) within the gap at center between the two origins on the plot. The structural forms segregate on characteristic SNP haplotypes, both inside and outside the inversion region. Statistical imputation of structural alleles uses SNPs on both sides of the CNVs together with more distant markers not shown here.

Table 1 Imputation of 17q21.31 structural states from SNP data

	SNP panel used for imputation							
	1000 Genomes SNP genotypes from low-coverage sequencing		HapMap 3 array-based SNP genotypes		Illumina 1M array-based SNP genotypes		SNP 6.0 array-based SNP genotypes	
Structural feature imputed	Imputation	Tag SNP	Imputation	Tag SNP	Imputation	Tag SNP	Imputation	Tag SNP
Copy number of α duplication	0.99	0.96	1.00	0.96	1.00	0.96	0.99	0.96
Copy number of β duplication	0.93	0.49	0.79	0.30	0.80	0.30	0.77	0.30
Copy number of γ duplication	0.84	0.27	0.80	0.30	0.80	0.30	0.68	0.16
Inversion state (H1 versus H2)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Shown is the correlation (r^2) of experimental determinations of the state of each structural feature in each genome with either (i) imputed, probabilistic 'dosages' of each structural feature or (ii) the state of the single most-correlated proxy SNP (tag SNP) from each reference panel. Imputation-based predictions were calculated across 330 leave-one-out simulations. Imputation requires a panel of reference haplotypes with accurate SNP genotypes and structural alleles; we provide such a resource for 17q21.31 (**Supplementary Note**).

of KANSL1 but would lack its PEHE domain (**Supplementary Fig. 6**), which is needed to interact with MOF[22,23].

It is important to understand how genome structures relate to variation in phenotypes. Complex genome structures are not currently typed in sequencing- or array-based genome-wide association studies. Structurally complex regions are assumed to be poorly captured by LD to SNPs[24,25]. However, our analysis suggested that the structural diversity at 17q21.31 arose from a definable series of structural mutations (**Fig. 2**); each mutation likely arose on a specific haplotype and may continue to segregate on that haplotype. Such haplotypes might be identified by combinations of many SNPs.

We analyzed the SNP haplotypes on which each 17q21.31 structural form segregates in European populations (**Fig. 3**). The structural forms of 17q21.31 were strongly associated with SNP haplotypes on both sides of the distal end of the 17q21.31 inversion, where the polymorphic CNV copies reside (**Fig. 3**). These results suggested that it might be possible to capture 17q21.31 structural diversity through statistical imputation from SNPs[26–28]. We therefore constructed and evaluated the first imputation resource for a structurally complex locus. We created reference haplotypes from the 94 trio founders of the Utah residents of Northern and Western European ancestry (CEU) sample and a composite cohort of 373 unrelated individuals from the 1000 Genomes Project Phase 1 data, phasing structural variation along with 934 reference SNP haplotypes and removing any SNPs that fell within CNVs (Online Methods).

We evaluated imputation efficacy for each structural feature using leave-one-out tests. In each test, we selected a different test individual from the reference cohort and removed this individual's structural variation data from the reference genotypes. The test individual was always a CEU trio founder for whom we had separately determined CNV states by ddPCR (**Fig. 1e–g**) and trio-based phasing (**Fig. 1k**,**l**). In each simulation, we used the rest of the reference genotypes as an unphased panel together with the backbone SNP data from the test individual to phase all the data and then impute (using Beagle[29]) the states of the structural alleles into the test individual. We then compared this prediction to the independently derived experimental data.

As a metric of imputation efficacy, we evaluated the statistical correlation (r^2) of the experimentally determined structural state with the imputation-based, probabilistic dosage of each structural feature (**Table 1** and **Supplementary Tables 13–15**). This metric estimates the efficacy of imputation; $1/r^2$ gives the proportional increase in sample size that would be required (in additive tests of association) to recover the statistical power obtained by explicitly typing each variant. For the four large structural features analyzed (the α, β and γ duplications and the inversion), imputation from low-coverage genome sequence data yielded structural determinations that correlated strongly (r^2 = 0.99, 0.93, 0.84 and 1.00, respectively) with the true diploid copy number (**Table 1**). This efficacy was only modestly lower using earlier panels of SNPs typed in GWAS (**Table 1**). Imputation was able to capture the multi-allelic CNVs substantially better than individual SNPs were (**Table 1**). These results suggest that imputing reference haplotypes into available SNP data will allow structural forms of 17q21.31, and perhaps many other such loci, to be evaluated for relationships to human phenotypes.

We have described a population genetics approach for characterizing structurally complex and diverse genome variations. Our approach is complementary to existing methods based on FISH and clone reconstruction. Drawing upon population-level sequence data sets, this approach will yield models of how structurally multiallelic loci vary in populations. Our results provide motivation for the creation of integrated SNP-structure haplotype maps that will allow complex genome structures to be imputed into many other genomes using available SNP data. Our results and methods offer new ways of analyzing complex genome structures and relating them to human disease.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequence data are available at the Sequence Read Archive (SRA) under accession SRA052055.

Note: Supplementary information is available on the online version of the paper.

ACKNOWLEDGMENTS

J. Korn provided an early version of software for visualizing haplotype diversity. N. Rohland and T. Mullen contributed expertise on laboratory experiments. We thank N. Patterson, D. Reich, D. Altshuler, E. Lander, B. Browning, J. Korn, J. Gray, C. Patil, G. Genovese, A. Sekar and S. Grossman for helpful conversations and/or comments on the manuscript. This work was supported by a Smith Family Award for Excellence in Biomedical Research to S.A.M., by the National Human Genome Research Institute (U01HG005208) and by startup resources from the Harvard Medical School Department of Genetics.

AUTHOR CONTRIBUTIONS

S.A.M., L.M.B. and R.E.H. conceived the strategy for population genetics dissection of structurally complex loci. L.M.B. performed all laboratory experiments and multiple computational analyses, including the estimation of haplotype frequencies, delineation of CNV regions and alignment of next-generation sequence data. R.E.H. performed computational analyses of the 1000 Genomes Project data, including finding breakpoint-spanning reads for CNVs and integrated analyses of SNP-CNV haplotypes. M.C.Z. performed analyses of sequence data to determine large-scale structures, estimate coalescence and mutation dates and reconstruct the evolutionary history of the locus. R.E.H. and L.M.B. developed the imputation strategy. S.A.M., L.M.B., R.E.H. and M.C.Z. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2334.

1. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
2. Chowdhury, R., Bois, P.R., Feingold, E., Sherman, S.L. & Cheung, V.G. Genetic analysis of variation in human meiotic recombination. *PLoS Genet.* **5**, e1000648 (2009).
3. Fledel-Alon, A. *et al.* Variation in human recombination rates and its genetic determinants. *PLoS ONE* **6**, e20321 (2011).
4. Skipper, L. *et al.* Linkage disequilibrium and association of *MAPT* H1 in Parkinson disease. *Am. J. Hum. Genet.* **75**, 669–677 (2004).
5. Simón-Sánchez, J. *et al.* Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.* **41**, 1308–1312 (2009).
6. McCarroll, S.A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).
7. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
8. McCarroll, S.A. *et al.* Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. *Nat. Genet.* **40**, 1107–1112 (2008).
9. Willer, C.J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41**, 25–34 (2009).
10. Handsaker, R.E., Korn, J.M., Nemesh, J. & McCarroll, S.A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).
11. Quinlan, A.R. & Hall, I.M. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet.* **28**, 43–53 (2012).
12. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
13. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
14. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
15. Mills, R.E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
16. Zody, M.C. *et al.* Evolutionary toggling of the *MAPT* 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008).
17. McCarroll, S.A. Copy-number analysis goes more than skin deep. *Nat. Genet.* **40**, 5–6 (2008).
18. Hindson, B.J. *et al.* High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal. Chem.* **83**, 8604–8610 (2011).
19. Tishkoff, S.A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39**, 31–40 (2007).
20. Genovese, G. *et al.* Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329**, 841–845 (2010).
21. Yu, L., Song, Y. & Wharton, R.P. E(nos)/CG4699 required for *nanos* function in the female germ line of *Drosophila*. *Genesis* **48**, 161–170 (2010).
22. Smith, E.R. *et al.* A human protein complex homologous to the *Drosophila* MSL complex is responsible for the majority of histone H4 acetylation at lysine 16. *Mol. Cell. Biol.* **25**, 9175–9188 (2005).
23. Li, X., Wu, L., Corsa, C.A., Kunkel, S. & Dou, Y. Two mammalian MOF complexes regulate transcription activation by distinct mechanisms. *Mol. Cell* **36**, 290–301 (2009).
24. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
25. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
26. Browning, S.R. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet.* **124**, 439–450 (2008).
27. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
28. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
29. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
30. Steinberg, K.M. *et al.* Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat. Genet.* published online: doi:10.1038/ng.2335 (1 July 2012).

ONLINE METHODS

Identification of CNV segments. We used a combination of array and sequence data to find the breakpoints of each CNV in the 17q21.31 region. Using array-based data, we identified the approximate span of CNV segments at kilobase resolution. We then refined the boundaries of these segments to 100-bp resolution by comparing read-depth profiles. Ultimately, the precise breakpoints of these rearrangements were identified by searching 1000 Genomes Project data. Details of this analysis are provided in the **Supplementary Note**, and schematics of the structure are available in **Supplementary Figures 1** and **7**.

Analysis of CNVs using droplet-based digital PCR. To determine integer copy number of CNV segments (regions 1–3), we used a droplet-based digital PCR method[17]. We designed a pair of PCR primers and a dual-labeled fluorescence-FRET oligonucleotide probe to both the CNV locus and a two-copy control locus (primer sequences are given in **Supplementary Table 16**). Genomic DNA and primers and probes for both assays were compartmentalized into droplets in an oil-aqueous emulsion (QuantaLife). We performed PCR amplification on the emulsion and then counted the number of droplets that were positive and negative for each fluorophore with a droplet reader (QuantaLife). Absolute copy number for the CNV locus was determined by comparing droplet counts of the CNV locus to the two-copy control locus. This method is described in greater detail in the **Supplementary Note**.

Analysis of CNVs using WGS read depth (Genome STRiP). WGS read depth was also used to determine copy number in regions 1, 2 and 3. We adapted the Genome STRiP genotyping method[10] to analyze duplications in low-coverage sequencing data from 1000 Genomes Project Phase 1. Details of this analysis are available in the **Supplementary Note**.

Inference of inversion state. The ancient, megabase-long inversion polymorphism at 17q21.31 has resulted in a large number of fixed differences between the two inversion states because opposite alleles of the inversion cannot viably recombine with each other within the inverted region. These two inversion states therefore define two long haplotypes, H1 and H2, with hundreds of fixed differences between them. We refined these haplotypes using low-coverage data from Phase I of the 1000 Genomes Project, finding 1,886 sites that are in perfect LD ($r^2 = 1$), even in the large ascertainment (1,000 individuals) afforded by these data. Although the megabase-long inversion polymorphism has been reported to 'toggle' on the longer timescales of primate evolution[18], no study to date has reported any discordance in humans between the cytogenetic orientation of this megabase-long segment and the state of these inversion proxy SNPs. We therefore used these long SNP haplotypes to diagnose inversion state (**Supplementary Tables 2** and **3**). We observed two individuals in 1000 Genomes Project data for whom SNP genotypes in specific segments within the inverted region suggested an H1/H2 type that was discordant from that suggested by the rest of their SNPs at the locus. The clustering of these discordant SNPs suggested that these individuals' genomes reflect gene conversion or double recombination events that occurred within the inverted regions; these individuals were not included in subsequent analyses.

Determination of haplotypic contributions to diploid copy number and heuristic phasing in trios. Inferring haplotypic contributions to diploid copy number (**Fig. 1j**) was addressed with a joint maximum-likelihood analysis of genotypes, allele frequencies and inheritance patterns in trios. Each population was analyzed separately. We considered all possible combinations of integer copy number (on each of the four haplotypes in a trio: paternal transmitted, paternal untransmitted, maternal transmitted, maternal untransmitted) that were consistent with the diploid copy-number measurements from all three trio members from ddPCR. See the **Supplementary Note** for details of this analysis and a description of the expectation-maximization algorithm.

Inference of frequency of copy-number alleles in populations. For regions 1, 2 and 3, diploid copy number was first measured using read depth from WGS by the Genome STRiP algorithm[10] in low-coverage WGS data from populations of unrelated individuals from the 1000 Genomes Project. Allele frequencies were determined from genotype frequencies with an expectation-maximization algorithm similar to that described in the **Supplementary Note**, but without the additional constraints provided by inheritance in trios (**Supplementary Tables 2–4**).

Statistical phasing of structural and fine-scale variation in populations. We determined phased structural genotypes for 47 CEU trios on the basis of our model of the nine structural haplotypes, the ddPCR diploid copy-number estimates for regions 1, 2 and 3, and the assayed H1/H2 inversion state in each sample and by assuming Mendelian inheritance in the trios. Phased haplotypes for the founders in these trios are listed in **Supplementary Table 10**. In a similar manner, we determined phased structural genotypes for 373 additional unrelated samples (where genotype could be determined without the benefit of trio inheritance constraints) from 1000 Genomes Project data, using diploid copy-number estimates from read-depth genotyping in the 1000 Genomes Project Phase 1 low-coverage sequence data and determining H1/H2 inversion state on the basis of a tag SNP, rs17660065. We used this set of resolved structural haplotypes for 467 individuals (the 373 individuals from the 1000 Genomes Project plus the 94 trio founders) to evaluate imputation of the structural haplotypes from the genotypes of nearby SNPs using Beagle[29]. We evaluated imputation accuracy through a series of leave-one-out trials in which we withheld information on one individual from the reference panel and then imputed the structural haplotypes for that individual on the basis of their genotypes at surrounding SNPs. Details of allelic encoding, evaluation methodology and estimation of CNV dosages are available in the **Supplementary Note**.

Identification of *KANSL1* fusion gene RNA transcripts. We designed primers (**Supplementary Table 16**) to amplify the *KANSL1* fusion gene transcript created by the α duplication using information from genomic breakpoints and publicly available RNA sequence data[31]. Primer design to amplify the *KANSL1* fusion gene transcript created by the β duplication was informed by genomic breakpoints and mRNA clone BC006271 (GenBank), which is likely a complete fusion transcript resulting from the β duplication. These analyses are discussed in detail in the **Supplementary Note**.

Dating the coalescence of duplication-containing chromosomes. We performed targeted capture and sequencing of the 17q21.31 region for three individuals homozygous for duplication α and four individuals homozygous for duplication β. Excluding haplotypes that showed evidence of recombination with other structural forms, we computed the average pairwise diversity among chromosomes with the α duplication and separately among chromosomes with the β duplication. To estimate a coalescence date, diversity between sampled humans was compared to human-chimpanzee divergence over the same region. A date was calculated by calibrating the divergence with a human-chimpanzee speciation time of 6 million years ago. Details of the coalescence analysis and direct duplication dating are available in the **Supplementary Note**. A phylogenetic tree constructed from the unique portion of the α duplicon on H2 is shown (**Supplementary Fig. 8**).

Analysis of allele frequency differentiation between European and non-European populations. To evaluate the stratification of each duplication, we calculated the fraction of SNPs (from 1000 Genomes Project Phase 1 data) that had similarly high derived allele frequency in the European populations sampled (379 samples from the CEU, Finnish from Finland (FIN), British from England and Scotland (GBR), Iberian populations in Spain (IBS) and Toscani in Italia (TSI) population groups; **Supplementary Table 17**) and that had similarly low derived allele frequency across the non-European populations sampled (471 samples from the CHB, CHS, Japanese in Tokyo, Japan (JPT), Luhya in Webuye, Kenya (LWK) and YRI population groups). Details of this analysis as well as analysis of allele frequency differentiation within European populations are provided in the **Supplementary Note**.

Analysis of tag SNPs. To compare the efficacy of imputation to that achieved using a single, best tag SNP, we computed the correlation (r^2) between the diploid copy number of each CNV (with the α, β and γ duplications and the H1-H2 state encoded as 0, 1 or 2) and the dosage of each SNP in each reference panel (encoded as 0, 1 or 2). For the Illumina 1M and Affymetrix 6.0 comparisons, we considered only the subset of SNPs from the HapMap 3 panel that was present on each array. We report the highest r^2 across all SNPs in the panel.

31. Montgomery, S.B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).

doi:10.1038/ng.2334