nature genetics

Using population admixture to help complete maps of the human genome

Giulio Genovese[1–4], Robert E Handsaker[1,2,4], Heng Li[1,2], Nicolas Altemose[2], Amelia M Lindgren[5], Kimberly Chambert[1,4], Bogdan Pasaniuc[6], Alkes L Price[1,6], David Reich[2], Cynthia C Morton[1,5,7], Martin R Pollak[1,3], James G Wilson[8] & Steven A McCarroll[1,2,4]

Tens of millions of base pairs of euchromatic human genome sequence, including many protein-coding genes, have no known location in the human genome. We describe an approach for localizing the human genome's missing pieces using the patterns of genome sequence variation created by population admixture. We mapped the locations of 70 scaffolds spanning 4 million base pairs of the human genome's unplaced euchromatic sequence, including more than a dozen protein-coding genes, and identified 8 new large interchromosomal segmental duplications. We find that most of these sequences are hidden in the genome's heterochromatin, particularly its pericentromeric regions. Many cryptic, pericentromeric genes are expressed at the RNA level and have been maintained intact for millions of years while their expression patterns diverged from those of paralogous genes elsewhere in the genome. We describe how knowledge of the locations of these sequences can inform disease association and genome biology studies.

Physical maps of the human genome, including the sequence of most of its euchromatic portions[1,2], are basic resources in human genetics and genomics research: they provide the framework for the analysis of sequence data, and they enable genome-scale analysis of SNPs, copy number variants (CNVs), epigenetic phenomena and gene expression.

Yet, physical maps of the human genome remain incomplete. Almost 30 Mb of euchromatic genome sequence that are apparently human—observed in human whole-genome sequence data[3,4], containing human ESTs[5,6] and homologous to other mammalian genome sequences—are either absent from or have no assigned locations in current assemblies of the human genome[7,8].

These 'missing pieces' of the reference human genome are a likely source of mistaken inference in today's analyses of genome sequence data[9]. Sequence reads arising from the missing pieces may be discarded as non-alignable or incorrectly assumed to arise from paralogous sequences in the known, assembled part of the human genome. Sequences missing from the reference human genome might also help answer questions in human genetics research, such as what is the source of the genetic signals that have been ascertained (but not yet fine mapped to causal variation or causal genes) by linkage, association and CNVs.

Here, we describe an approach for applying admixture mapping to localize the human genome's missing pieces at megabase-pair scales using the patterns of sequence variation that have been created by the isolation and subsequent remixture of human populations. We report the successful mapping of ~5 Mb of unplaced human euchromatic sequences, including many protein-coding genes. We find that most of these sequences are euchromatic islands within the genome's heterochromatic oceans, including centromeres and the short arms of the acrocentric chromosomes, and that they almost always consist of segmental duplications (sometimes recent, sometimes millions of years old) of sequence present elsewhere in the reference genome.

The construction of large-scale genome models (or assemblies) uses physical sequence overlaps between genomic clones[10]. Clones are assembled into larger scaffolds on the basis of overlapping sequences at their ends.

By contrast, mapping based on statistical relationships among variants can provide information that is complementary to physical mapping, as it does not require a continuous path of sequences to be cloned and uniquely assembled. Before physical mapping was feasible, linkage among alleles was used to construct the first genetic maps of the human genome based on RFLPs[11,12] and later to build and improve genetic maps based on microsatellite markers[13,14].

A unique kind of long-range information—finer in resolution than linkage in families, yet longer in reach than linkage disequilibrium (LD) in populations—is present in many of the world's admixed

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [3]Division of Nephrology, Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts, USA. [4]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [5]Department of Obstetrics, Gynecology and Reproductive Biology, Brigham and Women's Hospital, Boston, Massachusetts, USA. [6]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [7]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [8]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi, USA. Correspondence should be addressed to G.G. (giulio.genovese@gmail.com) or S.A.M. (mccarroll@genetics.med.harvard.edu).

Received 24 July 2012; accepted 31 January 2013; published online 24 February 2013; doi:10.1038/ng.2565

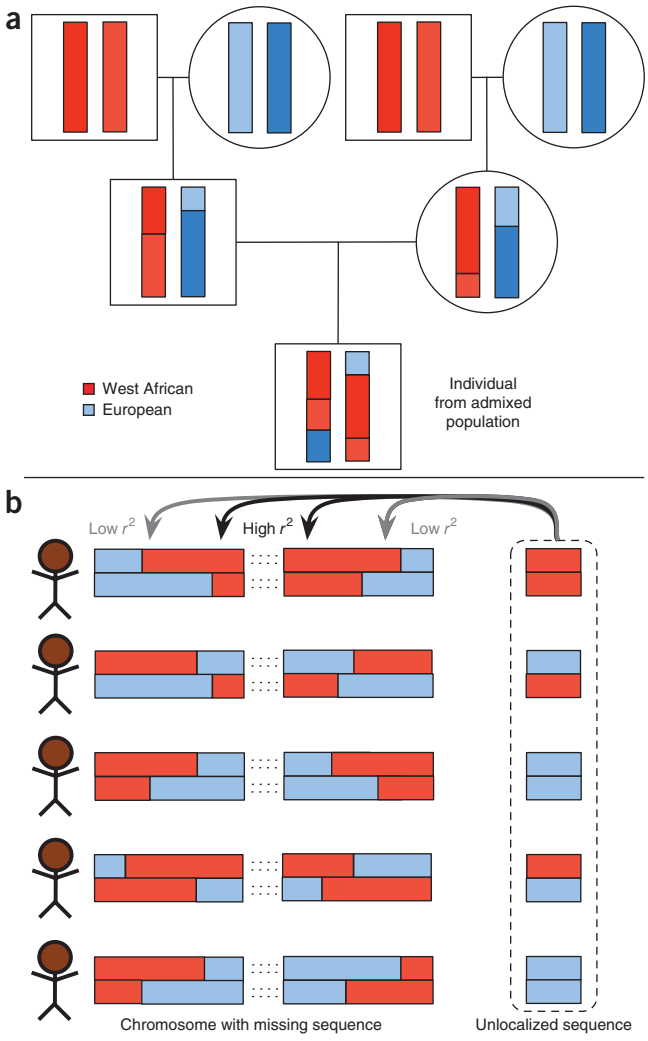


Figure 1 Admixture mapping of the human genome's missing pieces. (**a**) Chromosomes of West African descent have recombined with chromosomes of European descent through admixture to form mosaic genomes in African Americans. (**b**) Localization of genomic missing pieces, including unlocalized scaffolds and cryptic segmental duplications, by admixture mapping. Wherever allele frequencies have been influenced by genetic drift in the ancestral populations, statistically significant correlation between genotype and local ancestry allows the unplaced genomic sequence to be mapped to its correct location.

populations. Whenever human populations have been reproductively isolated for long periods of time (such as Africans and Europeans) and then remixed (such as African Americans), the genomes of the descendants are mosaics of segments that derive from ancestors from the two ancestral populations (**Fig. 1a**). The divergence of the sequences in the ancestral populations gives rise to sequence variation that is informative about the ancestry of each segment. Long-range 'admixture LD' has been used to map genetic factors that segregate at different frequencies in different populations[15,16] and to identify genomic sites of recombination in African Americans[17,18].

We reasoned that population admixture could also be used to map the locations of unmapped human genome sequences. Provided that the sequence in a genomic missing piece is variable, that this variation was subject to genetic drift and that the extent of this drift is known in the two ancestral populations, we could infer the ancestral origin of a missing piece—whether it has been inherited from each individual's European or African ancestors—with varying levels of statistical certainty, in a large panel of admixed individuals. By comparing such ancestry profiles for the genome's missing pieces to similar determinations across the known mapped and assembled sequences that make up the majority of these individuals' genomes, each missing piece could in principle be connected to the genomic location at which it resides, even if we lack a continuous path of cloned, assembled sequence with which to make such a connection (**Fig. 1b**).

Specifically, we can test ancestry-informative SNPs for correlation between their genotypes and inferred local ancestry across the genome, estimated using available genome-wide genotypes[19]. This is different from and potentially much more powerful than detecting LD between genotypes at two SNPs, as the correlation between genotypes and local ancestry is expected to be much stronger (than that between SNPs) at genetic distances up to a few cM, and the distance between unmapped missing pieces and the nearest parts of the reference genome may be substantial. Furthermore, we estimated statistical mapping power from allele frequencies in the ancestral populations and found that it was substantial, even for admixed population samples of even a few hundred individuals (**Supplementary Figs. 1–3** and **Supplementary Note**). Thus, admixture mapping could in principle connect sequences that are physically farther apart than the size of most genomic clones (20–180 kb) and LD blocks (15–50 kb).

RESULTS

Sources of the missing pieces

We used 3 sources of unplaced genome sequence: (i) the current reference genome (hg19), which contains 59 unplaced contigs (~5 Mb of euchromatic sequence) for which the correct location is either only known at the chromosomal level or not known at all; (ii) the HuRef genome[20], assembled by random shotgun sequencing of a single individual, containing an even larger number of unplaced scaffolds (~3.5 Mb of euchromatic sequence in 28 scaffolds >100 kb in length and ~7 Mb of euchromatic sequence in 698 scaffolds >10 kb in length); and (iii) sequence from BAC and fosmid clones available from GenBank[21] (Online Methods).

Mapping the human genome's missing pieces

If an ancestry-informative SNP resides on an unmapped contig, we can map the location of the contig by admixture mapping of the SNP. We (i) aligned all unmapped sequence reads from the 1000 Genomes Project[22,23] to unplaced scaffolds from HuRef, (ii) identified polymorphic sites across these unplaced sequences and (iii) computed genotypes at each locus in all European (CEU) and West African (YRI) samples (Online Methods). We selected 314 ancestry-informative SNPs whose genotypes had Pearson's correlation $r^2 > 15\%$ with local ancestry. We then genotyped these SNPs in a cohort of 380 African-American participants from the Jackson Heart Study[24] (JHS), selecting this sample size on the basis of initial analyses of the predicted power to map each SNP as a function of the number of available genotypes (Online Methods and **Supplementary Fig. 3**).

We successfully performed admixture mapping of 139 SNPs (**Supplementary Fig. 4** and **Supplementary Table 1**), assigning locations for 70 previously unlocalized scaffolds (**Fig. 2** and **Supplementary Table 2**). We never observed SNPs from the same scaffold mapping to different locations, as could be the case if the scaffold were itself misassembled. Sequences mapped by this approach comprised a total of ~4 Mb of euchromatic sequence that had not been included or mapped in hg19.

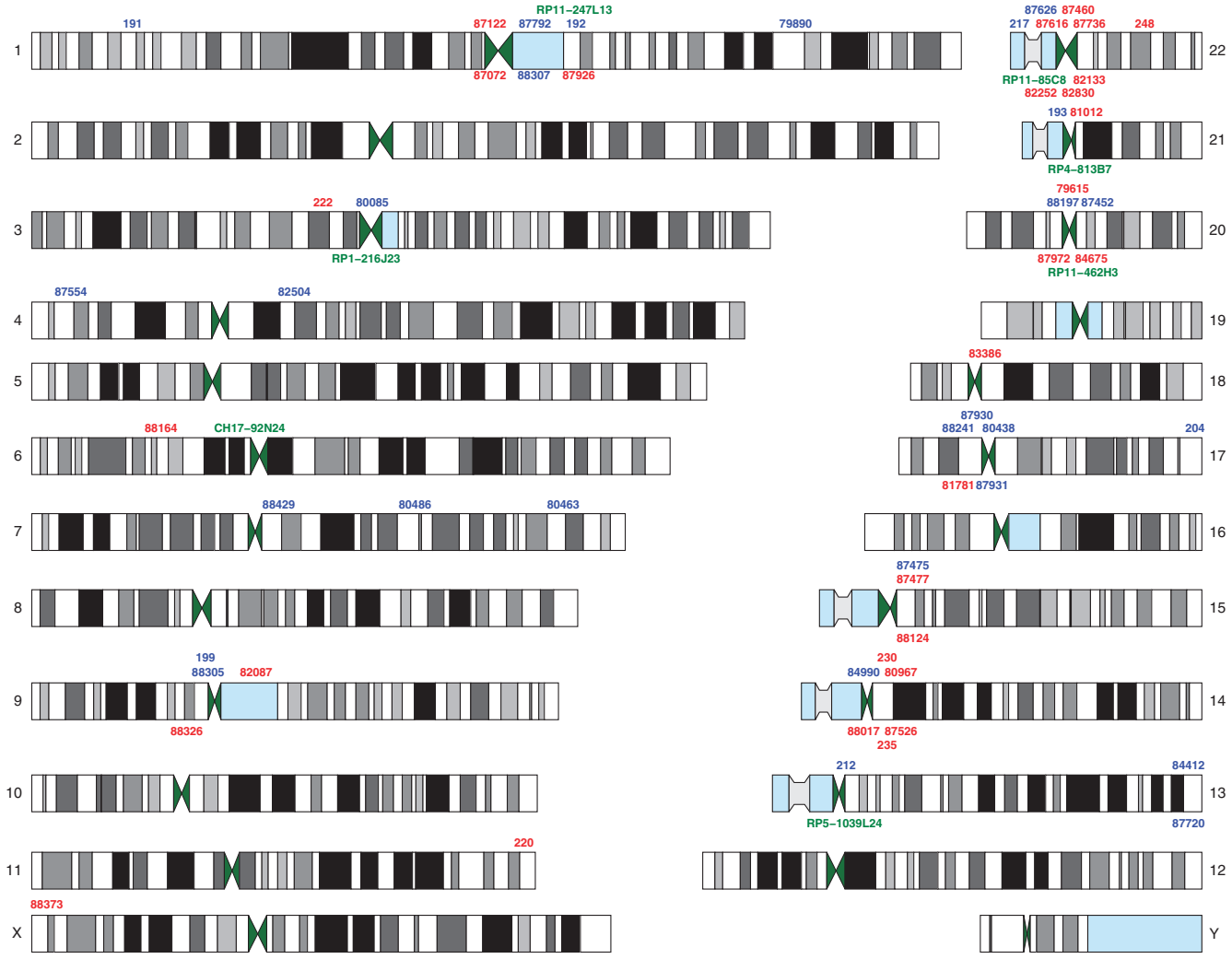


Figure 2 Approximate locations of previously unplaced genome sequence scaffolds that were mapped by our approach. Contigs from hg19 are labeled with three digits and stand for GL000###, and scaffolds from HuRef are labeled with five digits and stand for SCAF_11032791#####. Scaffolds with available chromosomal assignment or FISH data are denoted in blue; other scaffolds are denoted in red. Green indicates BAC clones that we mapped through SNPs from the CARe, ICDB or HapMap data sets. No scaffold was mapped to a location incompatible with FISH data. Mappings in the pericentromeric regions of acrocentric chromosomes indicate any location either in the pericentromeric regions or the short arms.

Identifying additional cryptic missing pieces

An additional set of cryptic missing pieces might be entirely missing from human genome reference sequences (might not even be described as unlocalized sequences nor present in HuRef) but exist instead as cryptic segmental duplications (or paralogs) of known genomic sequences and have been incorrectly assumed to represent the same genomic sequence as their known paralogs.

We reasoned that admixture mapping could also be used to identify cryptic segmental duplications. A SNP that is annotated in the assembled part of the human genome might in fact exist on a cryptic paralogous sequence elsewhere. Therefore, the identification of SNPs that admixture map to a different genomic location than their annotated location might indicate the presence of these SNPs at another genomic location on a cryptic segmental duplication.

To identify mismapped SNPs, we analyzed genome-wide SNP data from two large African-American cohorts. Among the 906,703 SNPs from the Affymetrix 6.0 array genotyped in ~7,800 individuals from the Candidate gene Association Resource (CARe) cohort[25] and the 566,714 SNPs from the Illumina HumanHap550 array genotyped in ~1,800 individuals from the Illumina iControlDB (ICDB) cohort, we identified, respectively, 121 and 15 SNPs that admixture mapped to genomic locations far from their HapMap[26] annotations of physical location (**Supplementary Table 3** and **Supplementary Note**).

Approximately half of these mismapped SNPs belonged to a single region, an approximately 360-kb segmental duplication from 16q22.2 to 1q21.1 involving the *HYDIN* gene[27–29], confirmed by FISH and previously found to give rise to false genome-wide association signals at 16q22.2 that in fact arose from true association at the Duffy locus at 1q23.2 (ref. 30) (**Supplementary Tables 4** and **5**).

Excluding the *HYDIN* paralog, incorrect mapping for ~30 SNPs can be explained by known segmental duplications[31–37], whereas, for the remaining ~40 mismapped SNPs, the most likely explanation is that they reside on sequence missing from the reference genome. (Of the ~30 SNPs that we simply remapped from one known segmental duplication copy to another, 10 corresponded to sites previously used as single unique nucleotides[38] (SUNs) to distinguish known segmental duplications. By definition, none of the remapped SNPs with which we identified novel segmental duplications corresponded to SUNs.)

Table 1 Segmental duplications localized by admixture mapping

Chr.	Position	Band	Gene	Size (kb)	Chr.′	Position′	Band′	Scaffold	Divergence	CARe[a]	ICDB[b]	HapMap[c]	FISH[d]
1	83598160–83955427	1p31.1	*POMZP3*	~400	7	76182346–76575579	7q11.23	NA	~1.4%	6	1	+	–
1	206072708–206558788	1q32.1	*FAM72/SRGAP2*	~240	1	14388000–1440957834	1q21.1	NA	~0.6%	3	0	–	–
2	37958019–38003219	2p22.2	NA	~45	22	NA	22q11.1	SCAF_1103279187616	~4.0%	3	0	+	–
2	91737476–91880745	2p11.1	*OTOP1*	~140	1	NA	1q21.1	RP11-247L13	~1.2%	2	0	+	–
2	133120083	2q21.2	NA	~115	20	NA	20q11.21	RP11-462H3	>2.0%	1	1	+	+
3	612223–663367	3p26.3	NA	~50	22	NA	22q11.1	GL000217	~2.0%	1	0	–	+
3	75761051–75871577	3p12.3	*ZNF717*	>110	21	NA	21q11.2	RP4-813B7	>5.0%	1	0	–	–
4	25709–68702	4p16.3	*ZNF595*	~40	22	NA	22q11.1	RP11-85C8	~0.5%	1	0	–	–
4	3536207–3636136	4p16.3	*FLJ35424*	~100	9	NA	9p11.2	SCAF_1103279188214	~3.0%	1	0	+	+
4	190470115–190684480	4q35.2	NA	~215	21	NA	21q11.2	GL000193	>2.0%	2	0	–	–
5	21506326–21573437	5p14.3	NA	~65	6	58137660–58139549	6p11.2	CH17-92N24	~1.5%	0	0	+	+
6	256518–382461	6p25.3	*DUSP22*	~125	16	NA	16p11.2	NA	~0.1%	0	1	–	–
6	57204729–57435462	6p11.2	*PRIM2*	~230	6	NA	6p11.2	SCAF_1103279188350	~2.0%	0	0	–	+
6	57204729–57608453	6p11.2	*PRIM2*	~400	6	NA	6q11.1	SCAF_1103279188263	~2.0%	0	0	–	+
6	57369236–57608453	6p11.2	*PRIM2*	~240	3	NA	3p11.1	SCAF_1103279180085	~2.0%	3	0	+	+
6	57401565–57570618	6p11.2	*PRIM2*	>170	3	NA	3p11.1	RP1-216J23	~2.0%	3	0	+	+
6	57447574–57575919	6p11.2	*PRIM2*	~130	6	NA	6p11.2	SCAF_1103279188406	~2.0%	0	0	–	+
12	147380–188194	12p13.33	*FAM138*	>40	20	62947067–62965512	20q13.33	SCAF_1103279187960	~1.2%	1	0	–	–
13	19020001–19167977	13q11	*ANKRD30BP2*	~200	21	14447204–14594419	21q11.2	NA	~0.8%	3	0	+	–
14	19817857–20194548	14q11.2	*POTEH/POTEM*	~400	2	16085071–16459525	22q11.1	NA	~0.6%	8	0	+	–
16	70845287–71202573	16q22.2	*HYDIN*	~360	1	146341167–146400000	1q21.1	GL000192	~0.6%	58	8	+	–
21	10971951–11032242	21p11.1	*TPTE*	>60	13	NA	13q11	RP5-1039L24	~0.2%	1	1	–	–
21	11083847–11156072	21p11.1	*BAGE*	>80	13	NA	13q11	NA	NA	2	0	–	–

Chr., chromosome; NA, not available. Genomic positions and bands are based on hg19 coordinates and localization of the ancestral copy of the duplication, respectively. Protein-coding gene(s) overlapping the duplication are shown. The estimated size of the duplication is given. Column titles marked with prime symbols include information on the derived copy of the duplication, with the genomic scaffold containing the sequence in the derived copy of the duplication is indicated. The estimated sequence divergence between the ancestral and derived copies of the duplication is given.

[a]Number of Affymetrix 6.0 SNPs remapped in the CARe data set. [b]Number of Illumina SNPs remapped in the ICDB data set. [c]Whether independent evidence of the cryptic duplication was confirmed by interchromosomal LD from HapMap genotypes. [d]Whether a FISH experiment was performed to validate the duplication.

To understand the relationships between these cryptic paralogs and unplaced scaffolds from large sequencing efforts, we cross-referenced the locations of these SNPs with alignments of unlocalized sequence from HuRef and GenBank. We identified 18 sequences >40 kb in length each containing 1 or more of the mismapped SNPs. Twelve of these 18 regions (spanning ~1.3 Mb of euchromatic sequence) could not be explained by segmental duplications already annotated in the reference genome; these indicate the presence of cryptic segmental duplications.

To critically evaluate these findings by an independent method, we used the principle that cryptic segmental duplications should give rise, for SNPs called from sequencing data, to excess heterozygosity that does not follow simple models of Hardy-Weinberg equilibrium between pairs of alleles. We searched for such a signal—annotated SNPs that behave more like paralogous sequence variants (PSVs)—in data from the 1000 Genomes Project pilot and confirmed all of these regions (Online Methods and **Supplementary Table 6**). For 8 of the 12 cryptic segmental duplications, we could find no mention in the literature. We further confirmed six of them by interchromosomal LD analysis using HapMap genotypes (**Table 1**).

We determined for each region whether the alternate allele of any of the mismapped SNPs was present in any of the BAC clones aligning to that region, by aligning sequences from BAC clones retrieved from GenBank to the hg19 reference genome. For SNPs in six of these regions, we could identify BAC clones carrying the alternate allele, suggesting that these clones harbor the sequence where these SNPs

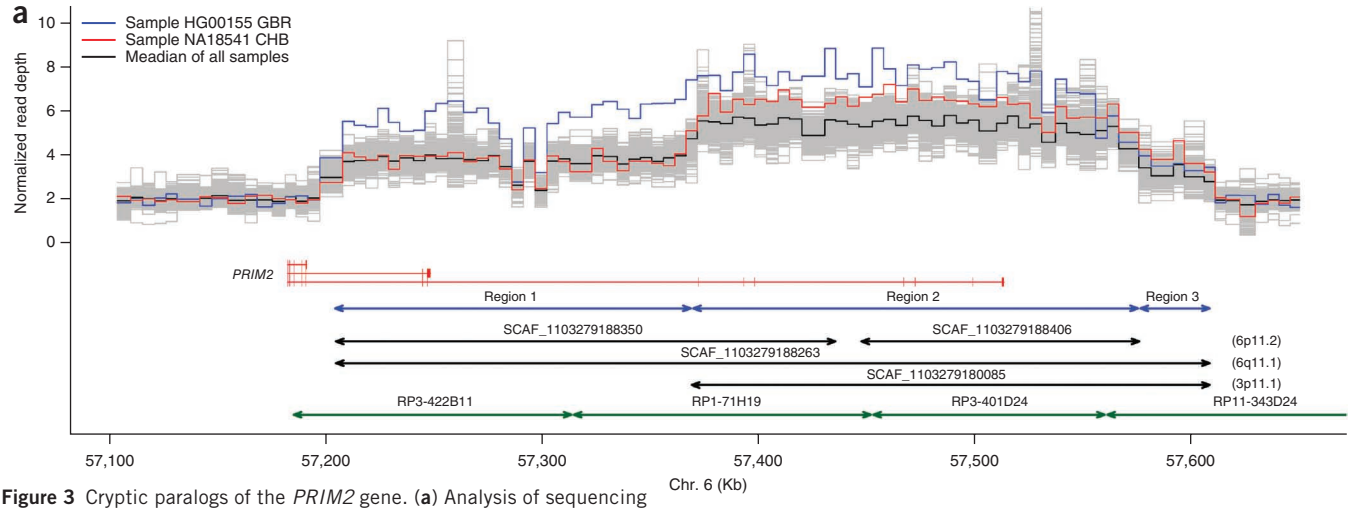
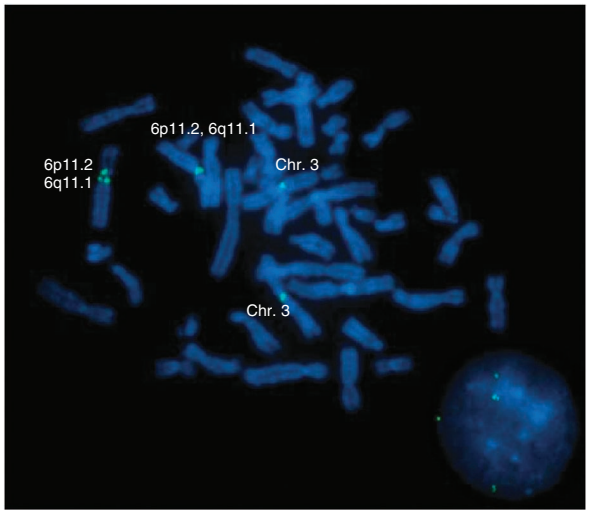


Figure 3 Cryptic paralogs of the *PRIM2* gene. (**a**) Analysis of sequencing coverage depth in data from the 1000 Genomes Project suggests the presence of three segments (blue arrows) with higher copy number. Although the copy number of each segment seems to be fixed in most genomes, at least two genomes show extra copy number gain at two of the three segments (HG00155 GBR at regions 1 and 2 and NA18541 CHB at regions 2 and 3), suggesting a model in which there are two additional copies of this locus in most human genomes, one copy containing regions 1 and 2 and another copy containing regions 2 and 3. Blue arrows indicate the regions, black arrows indicate alignment of HuRef scaffolds within these regions, and green arrows indicate the BAC clones overlapping these regions and used in the reference assembly. (**b**) FISH analysis of *PRIM2* and its cryptic paralogs. Fosmid clone WI2-0569M11 overlapping *PRIM2* (G248P8956G6 aligned to chr. 6: 57,417,677–57,467,167) hybridized to two distinct locations in the pericentromeric region of chromosome 6, 6p11.2 and 6q11.1, and to a third location in the pericentromeric region of chromosome 3, confirming the two additional partial copies of the *PRIM2* gene missing from the reference genome.



actually reside (**Table 1**). For one of these regions containing the gene *PRIM2*, further analysis indicated an intrachromosomal duplication in the pericentromeric region of chromosome 6 and an additional interchromosomal duplication in the pericentromeric region of chromosome 3 (**Supplementary Note**). We confirmed the existence of this triplication by the presence of excess sequence read depth across this region in low-coverage data from the 1000 Genomes Project (**Fig. 3a** and **Supplementary Fig. 5**) and FISH analysis (**Fig. 3b**). We also observed that the copy in the reference genome is a hybrid of the two copies on chromosome 6 owing to a misassembly (**Supplementary Fig. 6** and **Supplementary Note**).

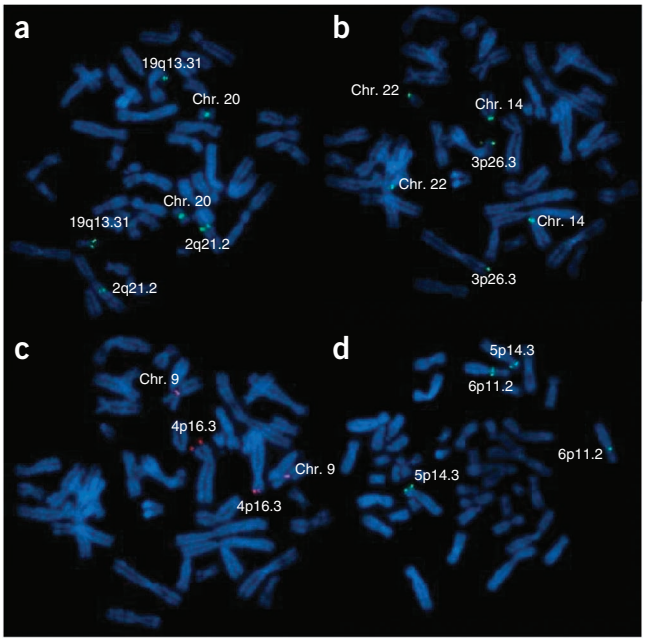


Figure 4 FISH analysis confirmed the presence of cryptic segmental duplications. (**a**) Fosmid clone WI2-1750D05 (G248P87673B3 aligned to chr. 2: 133,062,362–133,104,847) hybridized to 19q13.31 and to the centromeric region of chromosome 20, as predicted by admixture mapping. (**b**) WI2-1656E10 (G248P83226C5 aligned to chr. 3: 613,680–650,737) hybridized to the centromeric/acrocentric regions of chromosomes 14 and 22, as predicted by admixture mapping. (**c**) WI2-0903H06 (G248P8635D3 aligned to chr. 4: 3,573,606–3,614,890) hybridized to the centromere of chromosome 9, as predicted by admixture mapping. (**d**) WI2-1022I06 (G248P82546E3 aligned to chr. 5: 21,531,026–21,568,722) hybridized to 6p11.2.

Pericentromeric locations of the missing pieces

Despite the fact that most of the 300 or so gaps[8] in the reference human genome exist in interstitial regions, most of the sequence we were able to localize mapped not to interstitial gaps but to cytogenetically defined heterochromatic regions of the human genome. Among the mapped scaffolds, 57 of 70 mapped to pericentromeric regions (**Fig. 2** and **Supplementary Table 2**). Among the remapped SNPs

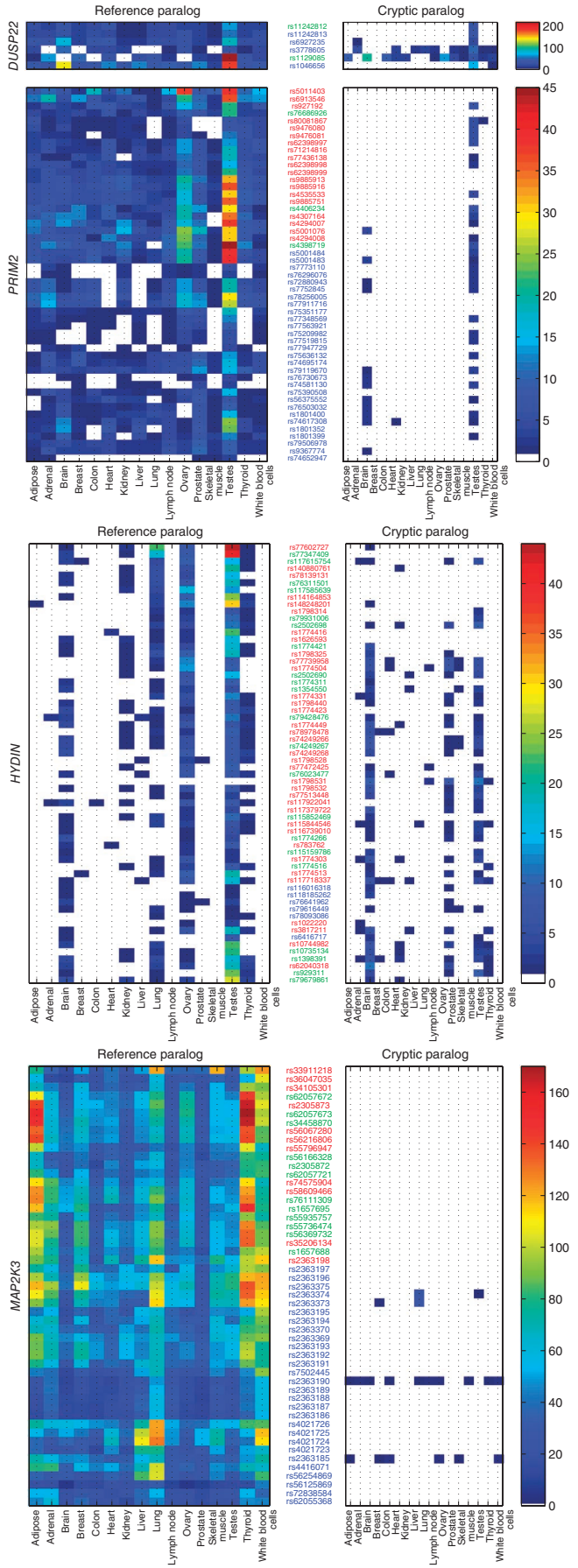


Figure 5 Expression of cryptic gene paralogs from pericentromeric regions of the human genome. PSVs were used to distinguish the expression of the *DUSP22*, *PRIM2*, *HYDIN* and *MAP2K3* genes from the expression of their cryptic paralogs in RNA-seq data from diverse human tissues. PSVs in the UTRs are represented by blue text, PSVs predicted to change the protein product of the paralog are shown in red, and synonymous PSVs are shown in green. The color of each box indicates the number of RNA-seq reads that can be assigned to one paralog or the other using the PSV.

identifying cryptic segmental duplications, 40 of 70 mapped to pericentromeric regions. (In all these cases, the resolution of the mapping was limited to the pericentromeric region identified.)

We sought to confirm these pericentromeric mappings using both published and new cytogenetic data. Of the 70 scaffolds we mapped successfully, 17 were among 29 scaffolds that were previously analyzed by FISH (**Supplementary Information** of ref. 39 and **Supplementary Table 8** of ref. 20). All 17 of these admixture mappings were consistent with 1 of the often multiple locations suggested by FISH (**Fig. 2** and **Supplementary Table 2**). Although confirmatory, this result also emphasizes the discerning power of admixture mapping over techniques based on hybridization, as the latter can yield ambiguous results when clones contain segmental duplications or other kinds of repeats.

We also performed additional FISH experiments to critically evaluate the mappings of five novel cryptic paralogous sequences for which no previous FISH data existed. In all (5/5) cases, FISH confirmed the presence of the additional copy in the predicted pericentromeric region (**Fig. 4** and Online Methods).

A further prediction of these mappings to pericentromeric regions involves the sequence content of the respective scaffolds. If these genomic missing pieces are indeed euchromatic islands in heterochromatic oceans, then they might frequently contain heterochromatic beaches consisting of the satellite sequences associated with human centromeres. To evaluate this prediction, we measured the amount of sequence classified as heterochromatic satellite on each scaffold. The great majority of the scaffolds that admixture mapped to pericentromeric regions (50/57) contained more than 5% satellite sequence (Online Methods, **Supplementary Fig. 4** and **Supplementary Table 2**), compared with almost none (1/13) of the scaffolds that admixture mapped to interstitial regions ($P = 0.003$).

Another prediction of these pericentromeric mappings is that, given earlier data indicating that recombination within centromeres is likely to be heavily repressed[40], scaffolds mapping to the same pericentromeric regions might show LD with one another. We identified pairs of SNPs (from distinct scaffolds) with LD not due to admixture and ~500 SNP pairs from distinct scaffolds for which both SNPs mapped to the same genomic regions (**Supplementary Table 7**). In no instance did these LD-based relationships among scaffolds disagree with our mappings from admixture.

To understand how the pericentromeric missing pieces relate to the known human genome, we aligned their sequences to hg19; virtually all scaffolds mapping to pericentromeric regions were found to consist of one or more segmental duplications of mapped euchromatic sequence, with 2–5% sequence divergence (**Supplementary Table 2**). This suggests that a large fraction of these sequences arrived at their current locations by a process of segmental duplication in primate ancestors[41].

Our mapping of these cryptic segmental duplications to centromeric regions is consistent with an earlier finding that most chromosome arms (35/43) have greater density of known interchromosomal duplications in the proximity of centromeres than is observed farther away from centromeres[42]; both results seem to reflect a tendency of interchromosomal duplications to deposit sequence at and around centromeres.

Are the missing pieces copy number variable?

Although the cryptic, pericentromeric euchromatic regions described here have not been purposefully interrogated in earlier CNV studies, they may have been indirectly interrogated via assays that targeted paralogous sequences in the known, assembled parts of the human genome. This seems the likely scenario, as almost all of the mismapped SNPs we identified from genotyping arrays (63/70, not including the *HYDIN* locus) fell within CNVs reported in the Database of Genomic Variants (DGV)[43] (**Supplementary Table 3**), despite the fact that DGV CNVs together cover less than a third of the human genome.

Given the sequence divergence over the identified cryptic paralogs (often greater than 2%), these additional copies are likely to have fixed in the ancestors of all humans. Identifying CNVs over these sequences at a greater rate than for the rest of the genome might therefore indicate the instability of sequences in pericentromeric regions rather than a persistent state of polymorphism of these additional copies in the human population after the duplication event. To evaluate the copy number variability of four selected paralogous region pairs, we analyzed the read depth of coverage and paralogous sequence variation using data from the 1000 Genomes Project (Online Methods). We identified common CNVs affecting the segmental duplications from the 2p22.2, 4q35.2 and *DUSP22* loci (**Supplementary Figs. 7–9**), and we found evidence for CNVs affecting either of the *PRIM2* cryptic paralogs (**Fig. 3a** and **Supplementary Fig. 5**). In each case, we could confirm, using PSVs, that the cryptic paralogs rather than the paralogs present in the reference genome accounted for the observed copy number variability (**Supplementary Note**), consistent with CNVs having arisen in the pericentromeric paralogs.

Expression of protein-coding genes from pericentromeric regions

Cryptic, pericentromeric paralogs of known protein-coding genes could in principle be either pseudogenes or expressed, intact genes. To test whether cryptic paralogs of coding genes are expressed at the RNA level, we analyzed RNA sequencing (RNA-seq) data from the Human BodyMap 2.0 project. We focused on reads aligning to the *DUSP22*, *PRIM2*, *HYDIN*, *MAP2K3* and *KCNJ12* genes, all of which appear to have cryptic paralogs in pericentromeric regions (**Fig. 3** and **Supplementary Figs. 5**, **9** and **10**). To distinguish RNA arising from reference gene copies from RNA arising from the cryptic paralogs, we focused on reads covering PSVs identifiable from genomic DNA sequence (many of which were previously misannotated as SNPs); this makes it extremely likely that sequence differences observed in RNA have a genomic origin (**Fig. 5** and Online Methods). We identified expressed RNA for all of the paralogs except *MAP2K3* (**Fig. 5**).

The expression of cryptic, pericentromeric gene copies showed several kinds of relationship to the expression of their paralogs. Both *DUSP22* and its recently duplicated paralog were expressed and showed similar distribution across tissues. In contrast, the cryptic paralogs of *PRIM2*, which contain only exons 6–14 of the original transcript (**Fig. 3a**), gave rise to shorter transcripts that were expressed exclusively in the brain and testes (**Fig. 5**). For *HYDIN*, which is expressed in brain and several other tissues, this analysis indicated that the cryptic paralog at 1q21.1 was expressed in the brain, consistent with its earlier observation in a brain cDNA library[28]. For *KCNJ12*, we detected expression of the pericentromeric paralog *KCNJ18* in testes (**Supplementary Fig. 11**), *KCNJ18* is also expressed in skeletal muscle and is essential to muscle function[44]. The tissue specificity observed for paralogous copies is also evidence that these observations are not the result of sequencing errors at putative PSV sites.

These results suggest that many of these cryptic, pericentromeric gene paralogs are expressed genes and that their expression patterns can differ from those of their known paralogs.

DISCUSSION

We have described a population-based approach for helping to assemble the rest of the euchromatic human genome, even when missing pieces are separated from known euchromatic sequence by extensive heterochromatic sequence. Because our approach uses data that are widely available or are quickly becoming so, its power will increase quickly in the coming years. We anticipate that this approach will help complete physical maps of the human genome.

Analysis of ancestry-informative markers in unlocalized scaffolds can be used to map the genomic locations of these scaffolds with a physical resolution comparable to that of FISH but with unambiguous mapping to individual loci, in a highly scalable way that will become inherently more powerful as sequence data sets grow. (Many aspects of the genome assembly will continue to require other methods—for example, our approach does not determine the physical orientation of novel sequence with respect to the chromosome.) Using this approach, we mapped ~4 Mb of unplaced euchromatic sequence, most of which we found to be embedded in the heterochromatic regions of the genome. These regions are not included in the current human reference genome, and, with two exceptions, they do not overlap with any of the current patches included in the latest revision (**Supplementary Table 8**).

One limitation of our approach is that it relies on novel sequence having been correctly assembled and distinguished from paralogous sequence. Most sequences from HuRef unplaced scaffolds have a divergence greater than 2% from their closest paralogs; owing to limitations of shotgun sequencing assembly, paralogous segments with <2% sequence divergence are likely to be under-represented in human genome assemblies[45]. Unfortunately, owing to their short read lengths, current whole-genome next-generation sequencing approaches do not provide better assemblies for such regions than those obtained with capillary-based sequencing approaches[46]. Nonetheless, we showed that admixture mapping of the SNPs ascertained in such regions can still allow the discovery and mapping of these cryptic paralogous sequences.

Our results have several potential implications for the mapping of disease-relevant genes in humans, particularly wherever genetic signals map near pericentromeric regions, assembly gaps and segmental duplications. (i) CNVs frequently straddle or are flanked by ambiguous regions of the genome assembly. For example, deletions and duplications at 1q21.1 reported to affect ~1.5 Mb of genomic sequence associate with cardiac developmental defects[47], schizophrenia[48,49], mental retardation, autism, congenital anomalies[50] and abnormal head size[51]. Fully defining the gene content of these CNVs will require interrogating the missing sequence hidden in the assembly gaps at 1q21.1. (ii) Some regions implicated in genome-wide association studies may require reanalysis in light of the results here. For example, human height associates with rs17511102 and other markers in a long noncoding RNA (lincRNA)-containing segment of 2p22.2 (ref. 52) for which we found a cryptic segmental duplication (and paralogous lincRNA) in the pericentromeric region of chromosome 22. Following up this association will require that markers throughout the region be reassigned to the correct paralogous gene copies. (iii) The *SERPINB6* gene was associated with a clinical phenotype through homozygosity mapping by the identification of an homozygous region terminated by the heterozygous genotype of the rs7762811 SNP[53], which our results suggest is incorrectly assigned to 6p25.3, although it in fact resides at 16p11.2, leading to a slight underestimation of the correct homozygous region. (iv) The genes affected by cryptic segmental duplications may be functionally important and critical to include and explicitly model in exome sequencing studies. For example, mutations in *KCNJ18*,

a gene missing from the reference genome, have been shown to cause thyrotoxic hypokalemic periodic paralysis[44]. (v) An admixture mapping study found that African Americans with multiple sclerosis are more likely than healthy African Americans to have European ancestry around the centromere of chromosome 1 (ref. 15), a region to which our work has assigned more than a megabase of novel sequence.

We showed that CNVs are more common over cryptic paralogs missing from the reference genome, most likely owing to the physical instability of pericentromeric regions. We also showed that paralogous genes in these cryptic, pericentromeric duplications are transcribed, sometimes with patterns of expression that diverge from those of their paralogs, and therefore potentially serve unique biological functions.

The presence of duplicated regions complicates genome assemblies and SNP and CNV discovery (**Supplementary Figs. 12–24**). Notably, *HYDIN* and *PRIM2* are among the most difficult genes to reconstruct using *de novo* assembly from short sequence reads[54]. *PRIM2* and *KCNJ12* are among the genes with the largest number of misidentified nonsynonymous SNPs[55], most likely owing to the identification of PSVs as SNPs.

Approximately 6% of the human genome reference is currently considered unreliable for variant discovery by the 1000 Genomes Project[23], owing to dearth or excess read coverage or poor alignment of sequence reads. Most of the regions we identified as harboring a cryptic segmental duplication (**Table 1** and **Supplementary Table 6**) fall in this inaccessible part of the human genome. While waiting for a more complete version of the human genome reference, the 1000 Genomes Project now aligns sequence data to an expanded genome reference that includes additional unlocalized sequences (termed 'decoy sequences') to reduce false alignments in regions with cryptic segmental duplications. These additional sequences consist mainly of sequenced clones discarded by the Human Genome Project and sequence from the HuRef assembly (~30% of decoy sequences consist of HuRef unlocalized scaffolds). Of course, the eventual goal of such projects will be the alignment of all human sequence reads to their actual physical locations.

In completing maps of the human genome, the important remaining challenges include mapping the human genome's structure at all scales, fully cataloging the genome's sequence content and appreciating how sequences are ordered and arranged along chromosomes. As the scientific community works toward a complete reference assembly of the human genome[56], the analysis of genome-wide data from admixed populations will add unique value and help complete understanding of the human genome's structure and evolution.

URLs. HuRef unplaced scaffolds, ftp://ftp.tigr.org/pub/data/huref/; GenBank database, ftp://ftp.ncbi.nih.gov/genbank/; database of Genotypes and Phenotypes (dbGaP), http://www.ncbi.nlm.nih.gov/gap; Illumina iControlDB, http://www.illumina.com/science/icontroldb.ilmn; HapMap interchromosomal LD, ftp://ftp.ncbi.nlm.nih.gov/hapmap/inter_chr_ld/; Illumina Human BodyMap 2.0 data, http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE30611; decoy sequences, ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/; UCSC Genome Browser, http://genome.ucsc.edu/; RepeatMasker, http://www.repeatmasker.org/.

METHODS

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

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

ACKNOWLEDGMENTS

This study was supported by grants RC1 GM091332-01 (S.A.M. and J.G.W.), R01 HG006855 (S.A.M.) and R01DK54931 (G.G. and M.R.P.) from the US National Institutes of Health and by a Smith Family Foundation Award for Excellence in Biomedical Research (S.A.M.).

The Jackson Heart Study is supported and conducted in collaboration with Jackson State University (N01-HC-95170), University of Mississippi Medical Center (N01-HC-95171) and Touglaoo College (N01-HC-95172) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute for Minority Health and Health Disparities (NIMHD), with additional support from the National Institute on Biomedical Imaging and Bioengineering (NIBIB).

The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C and HHSN268201100012C).

The Coronary Artery Risk Development in Young Adults Study (CARDIA) is conducted and supported by the NHLBI in collaboration with the University of Alabama at Birmingham (N01-HC95095 and N01-HC48047), the University of Minnesota (N01-HC48048), Northwestern University (N01-HC48049) and the Kaiser Foundation Research Institute (N01-HC48050).

MESA, MESA Family and the MESA SHARe project are conducted and supported by the NHLBI in collaboration with the MESA investigators. Support for MESA is provided by contracts N01-HC-95159, through N01-HC-95169, and RR-024156. Funding for MESA Family is provided by grants R01-HL-071051, R01-HL-071205, R01-HL-071250, R01-HL-071251, R01-HL-071252, R01-HL-071258 and R01-HL-071259. MESA Air is funded by the US Environmental Protection Agency (EPA)–Science to Achieve Results (STAR) Program Grant RD831697. Funding for genotyping was provided by NHLBI contracts N02-HL-6-4278 and N01-HC-65226.

This manuscript does not necessarily reflect the opinions or views of ARIC, CARDIA, JHS, MESA or the NHLBI.

AUTHOR CONTRIBUTIONS

G.G. and S.A.M. conceived the project, designed the analyses and wrote the manuscript. G.G. performed the analysis of the CARe, ICDB, JHS and BodyMap 2.0 data sets. R.E.H. performed the sequence read depth analysis of selected regions. H.L. performed the alignments of HuRef scaffolds and GenBank clones. N.A. contributed the analysis of the HuRef unplaced scaffolds. A.M.L. performed the FISH experiments. K.C. organized and contributed to the design of the Sequenom experiment. B.P., A.L.P. and D.R. provided advice for the local ancestry inference. C.C.M. participated in the interpretation of the FISH experiments. M.R.P. participated in planning discussions for the linkage analysis. J.G.W. participated in planning discussions, coordinated interactions with JHS and edited the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).
4. Kidd, J.M. *et al.* Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nat. Methods* **7**, 365–371 (2010).
5. Kirsch, S. *et al.* Interchromosomal segmental duplications of the pericentromeric region on the human Y chromosome. *Genome Res.* **15**, 195–204 (2005).
6. Lyle, R. *et al.* Islands of euchromatin-like sequence and expressed polymorphic sequences within the short arm of human chromosome 21. *Genome Res.* **17**, 1690–1696 (2007).
7. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
8. Lander, E.S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187–197 (2011).
9. Pickrell, J.K., Gaffney, D.J., Gilad, Y. & Pritchard, J.K. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics* **27**, 2144–2146 (2011).
10. Eichler, E.E., Clark, R.A. & She, X. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat. Rev. Genet.* **5**, 345–354 (2004).
11. Botstein, D., White, R.L., Skolnick, M. & Davis, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).

12. Donis-Keller, H. *et al.* A genetic linkage map of the human genome. *Cell* **51**, 319–337 (1987).
13. Weissenbach, J. *et al.* A second-generation linkage map of the human genome. *Nature* **359**, 794–801 (1992).
14. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).
15. Reich, D. *et al.* A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nat. Genet.* **37**, 1113–1118 (2005).
16. Winkler, C.A., Nelson, G.W. & Smith, M.W. Admixture mapping comes of age. *Annu. Rev. Genomics Hum. Genet.* **11**, 65–89 (2010).
17. Hinch, A.G. *et al.* The landscape of recombination in African Americans. *Nature* **476**, 170–175 (2011).
18. Wegmann, D. *et al.* Recombination rates in admixed individuals identified by ancestry-based inference. *Nat. Genet.* **43**, 847–853 (2011).
19. Seldin, M.F., Pasaniuc, B. & Price, A.L. New approaches to disease mapping in admixed populations. *Nat. Rev. Genet.* **12**, 523–528 (2011).
20. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
21. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Sayers, E.W. GenBank. *Nucleic Acids Res.* **39**, D32–D37 (2011).
22. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
23. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
24. Taylor, H.A. Jr. *et al.* Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. *Ethn. Dis.* **15**, S6-4-17 (2005).
25. Musunuru, K. *et al.* Candidate gene association resource (CARe): design, methods, and proof of concept. *Circ. Cardiovasc. Genet.* **3**, 267–275 (2010).
26. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
27. Martin, J. *et al.* The sequence and analysis of duplication-rich human chromosome 16. *Nature* **432**, 988–994 (2004).
28. Doggett, N.A. *et al.* A 360-kb interchromosomal duplication of the human *HYDIN* locus. *Genomics* **88**, 762–771 (2006).
29. Kim, J.I., Ju, Y.S., Kim, S., Hong, D. & Seo, J.S. Detection of *HYDIN* gene duplication in personal genome sequence data. *Genomics Inform.* **7**, 159–162 (2009).
30. Reiner, A.P. *et al.* Genome-wide association study of white blood cell count in 16,388 African Americans: the Continental Origins and Genetic Epidemiology Network (COGENT). *PLoS Genet.* **7**, e1002108 (2011).
31. Guipponi, M. *et al.* Genomic structure of a copy of the human *TPTE* gene which encompasses 87 kb on the short arm of chromosome 21. *Hum. Genet.* **107**, 127–131 (2000).
32. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. & Eichler, E.E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
33. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
34. Bailey, J.A. *et al.* Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.* **70**, 83–100 (2002).
35. Golfier, G. *et al.* The 200-kb segmental duplication on human chromosome 21 originates from a pericentromeric dissemination involving human chromosomes 2, 18 and 13. *Gene* **312**, 51–59 (2003).
36. Ruault, M., Ventura, M., Galtier, N., Brun, M.E. & Archidiacono, N. *BAGE* genes generated by juxtacentromeric reshuffling in the Hominidae lineage are under selective pressure. *Genomics* **81**, 391–399 (2003).
37. Dennis, M.Y. *et al.* Evolution of human-specific neural *SRGAP2* genes by incomplete segmental duplication. *Cell* **149**, 912–922 (2012).
38. Sudmant, P.H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
39. BAC Resource Consortium. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409**, 953–958 (2001).
40. Mahtani, M.M. & Willard, H.F. Physical and genetic mapping of the human X chromosome centromere: repression of recombination. *Genome Res.* **8**, 100–110 (1998).
41. Samonte, R.V. & Eichler, E.E. Segmental duplications and the evolution of the primate genome. *Nat. Rev. Genet.* **3**, 65–72 (2002).
42. She, X. *et al.* The structure and evolution of centromeric transition regions within the human genome. *Nature* **430**, 857–864 (2004).
43. 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).
44. Ryan, D.P. *et al.* Mutations in potassium channel Kir2.6 cause susceptibility to thyrotoxic hypokalemic periodic paralysis. *Cell* **140**, 88–98 (2010).
45. Eichler, E.E. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17**, 661–669 (2001).
46. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* **108**, 1513–1518 (2011).
47. Christiansen, J. *et al.* Chromosome 1q21.1 contiguous gene deletion is associated with congenital heart disease. *Circ. Res.* **94**, 1429–1435 (2004).
48. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
49. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
50. Mefford, H.C. *et al.* Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N. Engl. J. Med.* **359**, 1685–1699 (2008).
51. Brunetti-Pierri, N. *et al.* Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat. Genet.* **40**, 1466–1471 (2008).
52. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
53. Sırmacı, A. *et al.* A truncating mutation in *SERPINB6* is associated with autosomal-recessive nonsyndromic sensorineural hearing loss. *Am. J. Hum. Genet.* **86**, 797–804 (2010).
54. Alkan, C., Sajjadian, S. & Eichler, E.E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2011).
55. Ju, Y.S. *et al.* Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nat. Genet.* **43**, 745–752 (2011).
56. Church, D.M. *et al.* Modernizing reference genome assemblies. *PLoS Biol.* **9**, e1001091 (2011).

ONLINE METHODS

Alignment of HuRef genome and GenBank BAC and fosmid clones. To align the HuRef genome and sequenced BAC and fosmid clones to the hg19 reference genome, we first downloaded all available sequence from The Institute for Genomic Research and GenBank websites (downloading scaffold-not-in-chromosome.fasta files and all gbpri* files, respectively; see URLs), and we then used Burrows-Wheeler Aligner (BWA)[57] (with bwa bwasw) for alignments against hg19. We identified repeats classified as satellite sequences on HuRef unplaced scaffolds using RepeatMasker (see URLs). Satellite sequence consists of large arrays of tandemly repeated units of noncoding DNA. The amount of satellite and missing sequence is reported for each unplaced scaffold (**Supplementary Fig. 4** and **Supplementary Table 2**). To identify within these resources the presence of cryptic segmental duplications—that is, sequence missing from the current reference genome but present in a diverged, duplicated form—we aligned all available contigs from HuRef and GenBank clones against hg19 (**Supplementary Table 2**).

Alignment and variant calls for 1000 Genomes Project data. For genotyping from sequence reads, we selected all the CEU and YRI samples available in the 1000 Genomes Project[22,23]. Unmapped reads were aligned against the HuRef unplaced scaffolds using BWA[58] (with bwa aln/sampe). Genotype calling in the unplaced contigs was performed using the Genome Analysis Toolkit[59] (GATK) with default settings for the UnifiedGenotyper walker.

Strategy for admixture mapping. To map the location of a SNP, genotypes were first adjusted by regressing for the amount of global West African ancestry for each sample. The adjusted genotypes were then tested for correlation with local ancestry across the genome using a one-tailed Pearson's correlation test. If the correlation of the genotypes with global West African ancestry was positive, a right-tailed test was chosen; otherwise, a left-tailed test was chosen. The location corresponding to the smallest P value was then recorded for each SNP, together with the location corresponding to the smallest P value in a different chromosome. All these steps were performed using custom scripts from MATLAB (2011b, The MathWorks).

It is intuitive to expect that the genotyping of SNPs over paralogous sequences, only one of which will be expected to be polymorphic, will often be incorrect, as it will not be possible to correctly infer the homozygous state for the alternate allele, leading to failure of the called genotypes to satisfy Hardy-Weinberg equilibrium, among other things. This is not always so for genotyping arrays, however, as the genotyping of SNPs is often based on a two-dimensional Gaussian mixture model over summarized probe intensities for each of the two alleles[60], enabling the correct distinction of the three possible genotypes, even without modeling the presence of a cryptic paralog.

SNP selection, sample selection and Sequenom genotyping. From all detected SNPs in hg19 unplaced contigs and HuRef unplaced scaffolds, we filtered out SNPs at loci for which the number of reads with mapping quality of 0 was at least four and at least 10% of all reads covering the site. We also filtered out clusters of four SNPs within a window size of 10 bp. The rationale is that, in loci with ambiguous alignment, it is possible to call SNPs that actually belong to a paralogous region of the genome. Variants called in loci where many SNPs cluster together have a higher chance of being an artifact of misaligned reads originating from paralogous regions that are not present in the reference genome used for alignment. This methodology maximizes the chances that a SNP belongs to the unplaced scaffold where it is called. From the filtered list, up to seven ancestry-informative SNPs were chosen for each contig for which genotype was estimated to have Pearson's correlation coefficient with the amount of local European ancestry satisfying $r^2 > 15\%$. SNPs were further filtered to fit within ten Sequenom plexes, prioritizing the degree of correlation with ancestry. We selected 380 samples from JHS[24], which had been genotyped with the Affymetrix 6.0 array and analyzed with HAPMIX[61]. To achieve the maximum possible mapping resolution, we exclusively selected samples with at least 62 detected crossovers between ancestry groups (maximum of 115).

Most likely owing to the repetitiveness of the flanking sequences for which primers were designed, 86 assays failed completely; of the remainder, 53 failed the Hardy-Weinberg equilibrium test ($P < 1 \times 10^{-6}$), and 175 passed. Nevertheless, we could still reliably identify the locations of 139 SNPs (Pearson's correlation test $P < 1 \times 10^{-6}$), 106 of which had passed and 33 of which had failed the Hardy-Weinberg equilibrium test, showing that SNPs with unreliable genotypes can still be informative for mapping purposes (**Supplementary Fig. 4** and **Supplementary Table 1**). By analyzing for each successfully mapped SNP the best correlation between the adjusted genotype and local ancestry on chromosomes other than the one where the SNP mapped, we estimated that the selected conservative P-value threshold of 1×10^{-6} gives a false discovery rate lower than 1%.

Analysis of cryptic paralogs from 1000 Genomes Project pilot data. To identify regions with an excess of PSVs suggesting the presence of large cryptic segmental duplications, we searched for SNPs across the reference genome whose probabilistic genotype from 1000 Genomes Project pilot low-pass sequencing data failed the Hardy-Weinberg equilibrium test[62] (using bcftools view -c). We identified variants that failed the equilibrium test ($P < 1 \times 10^{-6}$) in CEU and YRI samples, grouped them together if they were <5 kb apart (using custom MATLAB scripts) and listed all resulting regions of >40 kb in size (**Supplementary Table 6**).

FISH. Peripheral blood mononuclear cells were stimulated with phytohemagglutinin and harvested. Metaphase spreads were prepared by standard protocols. Fosmid clones spanning the regions of interest were selected for FISH mapping using the UCSC Genome Browser (see URLs). Fosmids were labeled with either SpectrumOrange- or SpectrumGreen-conjugated dUTP using a nick translation kit (Abbott Molecular). Labeled pairs were hybridized overnight to metaphase chromosome preparations. After washes with 4× SSC/0.1% Tween, 2× SSC/0.3% Tween and phosphate-buffered detergent, chromosomes were counterstained with DAPI and analyzed by epifluorescence with a Zeiss Axioplan2 microscope and Applied Imaging CytoVision software.

Analysis of sequence read depth from 1000 Genomes Project data. To assess the copy number variability of the missing reference segments, we used an updated version of Genome STRiP[63] to analyze read depth. Normalized read depth was measured by comparing the number of DNA fragments with sequencing reads aligned to the reference genome in a given region to the expected read depth per haploid copy on the basis of (i) the total sequencing depth for each sample, (ii) the alignability of each position, based on whether it would be uniquely mapped by a perfect 36-bp read and (iii) sequencing bias due to GC content.

We performed normalization for GC bias empirically, similar to the method described in ref. 38. We first identified a 588-Mb subset of the autosomal reference sequence with no known evidence of copy number variability to use as a baseline. We removed all positions within 200 bp of the annotated CNV regions listed in DGV and segmental duplications listed in the UCSC browser, repeats annotated by RepeatMasker and assembly gaps, yielding a subset that is highly likely to be copy number invariant in the majority of people. This reference subset was divided into 400-bp windows and stratified by the GC fraction within each window, and the observed read depth at each GC fraction was compared to the total read depth across all windows to yield a GC normalization curve for each sequencing library.

Given a genomic locus, the estimation of diploid copy number for each sample was performed by fitting a Gaussian mixture model with sample-specific variance to the observed and expected read depth for each sample[63], allowing the model to fit as many copy number classes as needed at each locus.

To analyze genome regions with known paralogs in sequences not in the hg19 reference (notably, 2p22.2), we used BWA[58] (with bwa aln/sampe) to realign the 1000 Genomes Project reads from the genomic region to a synthetic reference containing the original reference sequence plus the sequence for the extra paralog. Estimation of copy number was then carried out as described above.

Analysis of RNA sequence expression data. To compare the expression of different paralogs of the *DUSP22*, *PRIM2*, *HYDIN*, *MAP2K3* and *KCNJ12* genes, we first identified PSVs over the predicted mRNA for these genes, looking at all heterozygous loci called for 1000 Genomes Project pilot high-coverage samples NA12878 CEU, NA12891 CEU, NA12892 CEU, NA19238 YRI, NA19239 YRI and NA19240 YRI, and then determined, when possible, which allele belonged

to each paralog (**Supplementary Tables 9–13**). Once we obtained a list of all PSVs, we counted reads from the Illumina Human BodyMap 2.0 project for each of the alleles observed at the locus using GATK[59] (with default settings for the UnifiedGenotyper walker and custom scripts). To validate the findings and filter out possible artifacts, sequence reads were further manually analyzed using the Integrative Genomics Viewer[64] (IGV).

57. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
58. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
59. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
60. Korn, J.M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260 (2008).
61. Price, A.L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* **5**, e1000519 (2009).
62. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
63. 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).
64. Robinson, J.T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

doi:10.1038/ng.2565