Title

An integrated map of structural variation in 2,504 human genomes

Journal

Nature, 526(7571)

ISSN

0028-0836

Authors

Sudmant, Peter H

Rausch, Tobias

Gardner, Eugene J

et al.

Publication Date

2015-10-01

DOI

10.1038/nature15394

Peer reviewed

Nature. Author manuscript; available in PMC 2016 April 01.

Published in final edited form as:
Nature. 2015 October 1; 526(7571): 75–81. doi:10.1038/nature15394.

An integrated map of structural variation in 2,504 human genomes

A full list of authors and affiliations appears at the end of the article.

[#] These authors contributed equally to this work.

Summary

Structural variants (SVs) are implicated in numerous diseases and make up the majority of varying nucleotides among human genomes. Here we describe an integrated set of eight SV classes comprising both balanced and unbalanced variants, which we constructed using short-read DNA sequencing data and statistically phased onto haplotype-blocks in 26 human populations. Analyzing this set, we identify numerous gene-intersecting SVs exhibiting population stratification and describe naturally occurring homozygous gene knockouts suggesting the dispensability of a variety of human genes. We demonstrate that SVs are enriched on haplotypes identified by genome-wide association studies and exhibit enrichment for expression quantitative trait loci. Additionally, we uncover appreciable levels of SV complexity at different scales, including genic loci subject to clusters of repeated rearrangement and complex SVs with multiple breakpoints likely formed through individual mutational events. Our catalog will enhance future studies into SV demography, functional impact and disease association.

Introduction

SVs, including deletions, insertions, duplications and inversions, account for most varying base pairs (bp) among individual human genomes[1]. Numerous studies have implicated SVs

[@]Correspondence and requests for materials should be addressed to eee@gs.washington.edu (EEE) and korbel@embl.de (JOK) .
[*]**Joint senior authors**
[†]A full list of participants and institutions is in the supplementary material.

Author contributions

SV discovery & genotyping: R.E.H., P.H.S., T.R., E.J.G., A.Ab., K.Y., F.H., K.C., G.D., K.W., M.H.-Y.F., S.K., C.A., S.A.M., R.E.M., M.B.G., S.E.D., E.E.E., J.O.K.; SV merging & haplotype-integration: T.R., R.E.H., M.H.-Y.F., E.G., A.Me., S.McC.; SV validation: R.E.H., A.Ab., G.J., M.H.-Y.F., A.M.S., M.K.K., A.Ma., S.K., M.M., M.J.P.C., S.M., P.C., S.E., J.M.K., B.R., J.A.W., F.Y., T.Z., M.A.B., R.E.M., A.B., C.L., E.E.E., J.O.K.; additional analyses: A.Au., C.M., E.C., E.D., E.-W.L., F.K., J.H., Y.Z., X.S., F.P.C., M.M., M.J.P.C., G.M., S.M., D.A., T.B., J.C., Z.C., L.D., X.F., M.G., J.M.K., H.Y.K.L., Y.K., X.J.M., B.J.N., A.N., R.A.G., M.P., M.R., R.S., D.M.M., M.W., N.F.P., A.Q., E.S., A.S., A.A.S., A.U., C.Z., J.Z., W.Z., J.S., O.S.; data management & archiving: L.C, X.Z-B, P.F.; display items: P.H.S., T.R., E.J.G., A.A., Y.Z., J.H., M.H.-Y.F., K.Y., M.B.G., A.B., O.S., R.E.M., S.E.D., E.E.E., J.O.K.; organization of Supplementary Material: G.D., J.O.K., P.H.S., R.E.M.; SV Analysis group co-chairs: C.L., E.E.E., J.O.K.; manuscript writing: P.H.S., T.R., E.J.G., J.H., R.E.M., M.B.G., O.S., S.E.D., E.E.E., J.O.K.

Data deposition statement: Sequencing data, archive accessions and supporting datasets including GRCh37 variant call files comprising the extended SV Analysis Group release set, a “readme” describing differences to the phase 3 marker paper variant release[16], and a GRCh38 version of our callset, are available at www.1000genomes.org/phase-3-structural-variant-dataset. DGV archive accession: estd219.

Competing financial interests statement: E.E.E. is on the scientific advisory board (SAB) of DNAnexus, Inc. and is a consultant for Kunming University of Science and Technology (KUST) as part of the 1000 China Talent Program. P.F. is on the SAB of Omicia, Inc.

Europe PMC Funders Author Manuscripts

in human health with associated phenotypes ranging from cognitive disabilities to predispositions to obesity, cancer and other maladies[1,2]. Discovery and genotyping of these variants remains challenging, however, since SVs are prone to arise in repetitive regions and internal SV structures can be complex[3]. This has created challenges for genome-wide association studies (GWAS)[4,5]. Despite recent methodological and technological advances[6-9], efforts to perform discovery, genotyping, and statistical haplotype-block integration of all major SV classes have so far been lacking. Earlier SV surveys depended on microarrays[10] as well as genomic and clone-based approaches limited to a small number of samples[11-15]. More recently, short-read DNA sequencing data from the initial phases of the 1000 Genomes Project[8,9] enabled us to construct sets of SVs, genotyped across populations, with enhanced size and breakpoint resolution[6,7]. Previous 1000 Genomes Project SV set releases, however, encompassed fewer individuals and were largely[6] or entirely[8] limited to deletions, in spite of the relevance of other SV classes to human genetics[1,2,4].

The objective of the Structural Variation Analysis Group has been to discover and genotype major classes of SVs (defined as DNA variants ≥50 bp) in diverse populations and to generate a statistically phased reference panel with these SVs. Here we report an integrated map of 68,818 SVs in unrelated individuals with ancestry from 26 populations (Supplementary Table 1). We constructed this resource by analyzing 1000 Genomes Project phase 3 whole-genome sequencing (WGS) data[16] along with data from orthogonal techniques, including long-read single-molecule sequencing (Supplementary Table 2), to characterize hitherto unresolved SV classes. Our study emphasizes the population diversity of SVs, quantifies their functional impact, and highlights previously understudied SV classes including inversions exhibiting marked sequence complexity.

Results

Construction of our phase 3 SV release

We mapped Illumina WGS data (~100 bp reads, mean 7.4-fold coverage) from 2,504 individuals onto an amended version[8] of the GRCh37 reference assembly using two independent mapping algorithms—BWA[17] and mrsFAST[18]—and performed SV discovery and genotyping using an ensemble of nine different algorithms (ED Figure 1 and Supplementary Notes). We applied several orthogonal experimental platforms for SV set assessment, refinement and characterization (Supplementary Table 2) and to calculate the False Discovery Rate (FDR) for each SV class (Table 1). Callset refinements facilitated through long-read sequencing enabled us to incorporate a number of additional SVs into our callset, including an additional 698 inversions and 9,132 small (<1 kbp) deletions, compared to the SV set released with the 1000 Genomes Project marker paper[16]. As a result, our callset differs slightly relative to the marker paper's SV set[16] (see Supplementary Table 2). We merged individual callsets to construct our unified release (Table 1), comprising 42,279 biallelic deletions, 6,025 biallelic duplications, 2,929 mCNVs (multi allelic copy-number variants), 786 inversions, 168 nuclear mitochondrial insertions (NUMTs), and 16,631 mobile element insertions (MEIs— including 12,748, 3,048 and 835 insertions of *Alu*, L1 and SVA (SINE-R, VNTR and *Alu* composite) elements, respectively).

SV non-reference genotype concordance estimates ranged from ~98% for biallelic deletions and MEI classes to ~94% for biallelic duplications. 60% of SVs were novel with respect to the Database of Genomic Variants (DGV)[19] (50% reciprocal overlap criterion, Figure 1a), whereby 71% of SVs (50% reciprocal overlap) and 60% of collapsed copy-number variable regions (CNVRs, 1 bp overlap) were novel compared to previous 1000 Genomes Project releases[6,8], reflecting methodological improvements and inclusion of additional populations. Novel SVs showed enrichment for rare sites, which we detected down to an autosomal allele count of '1'. And while variations in FDR estimates were evident with SV size and VAF (variant allele frequency), we consistently estimated the FDR at ≤5.4% when stratifying deletions and duplications by size and frequency, including for rare SVs with VAF<0.1% (ED Figures 1, 2). A comparison with deep-coverage Complete Genomics (CG) sequencing data indicated an overall sensitivity of 88% for deletions and 65% for duplications, with the false negatives driven largely by the relatively lowered sensitivity for ascertaining small SVs in Illumina sequencing data (Figure 1b, ED Figure 3). The average per-individual sensitivity was similar for deletions (89%) and slightly lower for duplications (50%). For MEI classes, estimated sensitivities ranged from 83%–96% (Table 1) compared to the 1000 Genomes Project pilot phase where a different MEI detection tool was used[20]. For inversions, we estimated an overall sensitivity of 32% based on variants with a positive validation status recorded in the InvFEST database[21], with an increased sensitivity of 67% for inversions <5 kbp in size.

We performed breakpoint assembly using pooled Illumina WGS and Pacific Biosciences (PacBio) sequencing data[22], and additionally performed split-read analysis[23] of short reads, to resolve the fine-resolution breakpoint structure of 37,250 SVs (29,954 deletions, 357 tandem duplications, 6,919 MEIs, and 20 inversions; Supplementary Table 3). Breakpoint assemblies showed a mean boundary precision of 0–15 bp for all SV types with the exception of inversions and duplications for which we achieved mean precision estimates of 32 bp and 683 bp, respectively (Table 1, Figure 1c).

Population genetic properties of SVs

We explored the population genetic properties of SVs among five continental groups—Africa (AFR), the Americas (AMR), East Asia (EAS), Europe (EUR) and South Asia (SAS). The bulk of SVs occur at low frequency (65% exhibit VAF<0.2%) consistent amongst individual SV classes (ED Figures 2, 3). While rare SVs are typically specific to individual continental groups, at VAF≥2% nearly all SVs are shared across continents (Figure 1d, ED Figure 3). Notably, we identified 1,075 SVs with VAF>50% (889 biallelic deletions, 2 biallelic duplications, 90 mCNVs, 88 MEIs and 6 inversions) encompassing 5 Mbp, sites of interest for future updates to the human reference genome. We estimated the mutation rate for each SV class using Waterson's estimator of θ, for example, ascertaining a mutation rate of 0.113 deletions per haploid genome generation, a threefold higher estimate compared with previous reports[10,24], likely due to our increased power for detecting variants <5 kbp (Supplementary Note).

We found that 73% of SVs with >1% VAF and 68% of rarer SVs (VAF>0.1%) are in linkage disequilibrium (LD) with nearby single nucleotide polymorphisms (SNPs) (r^2>0.6);

however, the proportion of variants in LD highly depends on the SV class (Figure 1e, ED Figure 4). For example, only 44% of all biallelic duplications with VAF>0.1% were in LD with a nearby SNP (r^2>0.6), in agreement with previous findings[10,25,26]. Notably, we observed a striking depletion of biallelic duplications amongst common SVs ($P<2\times10^{-16}$, KS-test; ED Figure 5) with most common duplications classified as multi-allelic SVs (*i.e.*, mCNVs). This behavior suggests extensive recurrence of SVs at duplication sites consistent with what was recently observed in a smaller cohort of 849 individuals[27]. These LD characteristics suggest duplications are currently under-ascertained for disease associations using tag-SNP-based approaches.

Based on our haplotype-resolved SV catalog, we observed that individuals of African ancestry exhibit, on average, 27% more heterozygous deletions than individuals from other populations (mean of 1,705 vs. 1,342) consistent with SNPs[28] (ED Figure 5). The relative proportion of deletion- versus SNP-affected sequence, however, showed a 13% excess in non-African compared to African populations (ratio 1.64 vs. 1.45). Principal component analyses with different SV classes generally recapitulated continental population structure and admixture (ED Figure 6 and Supplementary Note). Our analysis further allowed us to identify a catalog of 6,495 ancestry-informative MEI markers of potential value to population genetics history and forensics research (ED Figure 5, Supplementary Table 4).

Since population stratification can be used as a signature to detect adaptive selection, we additionally identified SVs varying in VAF amongst different populations. For each SV site we calculated a Vst statistic, a measure highly correlated with Fst (the fixation index)[29] that can be applied to assess population stratification of biallelic and multi-allelic SVs[29]. We observed 1,434 highly stratified SVs (>0.2 Vst, corresponding to 2.9 standard deviations (s.d.) from the mean; Supplementary Table 5) among which 578 intersected gene coding sequences (CDSs). Among these were several SVs associated with regions previously reported to be under positive selection, such as *KANSL1* mCNVs (ED Figure 6) that tag a European-enriched inversion polymorphism associated with increased fecundity[30]. Most of the population-stratified sites, however, have not been previously described and are, thus, potential targets for future investigation of SVs undergoing adaptive selection or genetic drift. These include, for example, a 14.5 kbp intronic duplication of *HERC2* enriched in East Asians (Vst=0.62 EAS-EUR).

Functional impact of SVs

We analyzed the intersection of deletions binned by VAF with various classes of genic and intergenic functional elements (Figure 2a, ED Figure 7). The CDSs, untranslated regions (UTRs) and introns of genes, in addition to ENCODE[31] transcription factor binding sites and ultrasensitive noncoding regions, showed a significant depletion ($P<0.001$; permutation testing in each VAF bin) compared to a random background model. In general, these elements are more depleted (in terms of fold change) in common VAF bins compared to rarer deletion alleles in keeping with purifying[10] (or in some cases background[32]) selection. Genes more intolerant to mutation (as measured from SNP diversity, residual variation intolerance score (RVIS)[33] <20) exhibited the most pronounced depletion ($P<0.001$; permutation testing between pairs of RVIS-score categories). All other SV classes exhibited

similar signatures of selection; when compared to deletions these depletions were, however, more attenuated (Figure 2b, ED Figure 7). Additional assessment of the site frequency spectrum showed that as deletion sizes increase these SVs become more rare ($p<2.2e-16$; linear model, F-test), evidence of purifying selection against events more likely intersecting functional elements. Duplications, by comparison, did not exhibit such trend, consistent with reduced selective constraints (Supplementary Note).

We additionally analyzed 5,819 homozygous deletions to search for gene knockouts naturally occurring in human populations. Among these we identified 240 genes (corresponding to 204 individual deletion sites), which based on the observation of homozygous losses in normal individuals appear “dispensable” (Supplementary Table 6). Most of the underlying deletions were found in more than one human population, and for only one (0.5%) we observed evidence for the putative involvement of uniparental disomy in the homozygosity (Supplementary Note). The majority (>80%) of these homozygous gene losses were novel compared to a previous analysis based on DGV variants[19], or recent clinical genomics studies (Supplementary Note). As expected, genes affected by homozygous loss were not highly conserved and were relatively tolerant to other forms of genetic variation (RVIS=0.74 compared to OMIM disease genes showing RVIS=0.43; p=9.4e-25; Mann-Whitney test). Moreover, the set was functionally enriched for glycoproteins (Benjamini Hochberg corrected p-value=1.6×10^{-3}, EASE [Expression Analysis Systematic Explorer] score) and genes harboring immunoglobulin domains (Benjamini Hochberg corrected p-value=1.0×10^{-5}, EASE score).

We next quantified the functional impact of SVs using expression quantitative trait loci (eQTL) associations as a surrogate[34,35]. Based on transcriptome data from lymphoblastoid cell lines derived from 462 individuals[36] (the gEUVADIS consortium), we tested 18,969 expressed protein-coding genes for *cis*-eQTL associations, considering 1 Mbp candidate regions upstream and downstream of CDSs. A joint eQTL analysis using SNPs, InDels and SVs with VAF>1% identified 54 eQTLs with a lead SV association (denoted SV-eQTL) and 9,537 eQTLs with a lead SNP/InDel association (10% FDR). For an additional 166 eQTLs with lead associations to SNPs or InDels, we observed SVs in LD ($r^2>0.5$) seven times more than when using random variants matched for LD structure, distance to the transcription start site, and VAF, suggesting that a larger number of eQTLs are likely impacted by SVs (ED Figure 8, Supplementary Table 7). In proportion to the number of variants tested, SV classes were up to ~50-fold enriched for SV-eQTLs (p=2.84×10^{-39}, one-sided Fisher’s exact test; Supplementary Table 8). Large SVs were associated with increased effect size, for example, a twofold increase in effect size for genic SVs >10 kbp versus variants <1 kbp (P=0.0004; t-test; ED Figure 8). Taken together, although SNPs contribute more eQTLs overall, our results suggest SVs have a disproportionate impact on gene expression relative to their number.

Among those 220 eQTLs having either an SV-eQTL or an SV in LD with the lead SNP/InDel, most were due to deletions (55% of associations) followed by mCNVs (19%) (Supplementary Table 8). Although SV-eQTLs with the largest effect sizes tended to overlap with CDSs, such as for the dual specificity phosphatase 22 (*DUSP22*) gene (Figure 2c), we also observed several expression-associated SVs strictly intersecting upstream noncoding

sequences, including an mCNV upstream of *ZNF43* (Figure 2d) possibly mediated through variation of a *cis*-regulatory element. We additionally considered the impact of accounting for SVs when constructing personalized reference genomes for transcriptome analysis. To illustrate this, we considered RNA read alignments for the sample NA12878, comparing the standard reference genome with GRCh37-derived personalized references constructed using NA12878 SNPs, or using NA12878 SNPs and SVs. Using such an approach, we observed marked changes in expression for 525 exons (±10 reads, ≥1-fold change relative to the standard reference), 24 of which could be attributed to the inclusion of SVs into the personalized reference (Supplementary Table 9).

The relevance of SVs to eQTLs suggests that a number of disease associations previously detected by GWAS may be attributable to SVs, which are difficult to assess directly in GWAS. To test this hypothesis we compared 12,892 previously reported SNP-based GWAS hits to SVs identified in our dataset, identifying 136 candidate SVs in strong LD (r^2>0.8) with GWAS variants, which represents a 1.5-fold enrichment when compared to a VAF and haplotype size-matched background set and a 3-fold enrichment for deletions >20 kbp (*P*=0.004) (Figure 2e and Supplementary Note). Approximately a third of these candidate GWAS associations (39) were novel, impacting phenotypes such as colorectal cancer and bone mineral density (Supplementary Table 10). Interestingly, 64% of these novel associations were mediated by deletions <1 kbp, a size-range for which our study has improved power over previous surveys, which more than doubles (from 18 to 40) the number of SVs <1 kbp in strong LD with a GWAS lead SNP. Thus, our SV resource could facilitate discovery of numerous additional disease-linked SVs.

SV clustering and complexity

Advances in Illumina sequencing towards longer read lengths (~100 bp vs. 36 bp)[6] in conjunction with the population-level data allowed us to perform an in-depth investigation of SV complexity and clustering. We identified 3,163 regions where SVs appeared to cluster (>2 SVs mapping within 500 bp; Supplementary Table 11). To reduce redundancy caused by multiple overlapping calls per sample, we calculated distinct CNVRs per cluster by merging calls per sample and haplotype and then counting the distinct CNVRs produced across samples (average 6.4 ± 7.2 CNVRs per cluster). We identified 30 genomic regions with an excess of CNVRs (>4 s.d. or >36 CNVRs per cluster). This clustering effect was not correlated with segmental duplications (r=0.02) and only partially explained by SNP diversity (r=0.15; ED Figure 9). CNVR clusters showed enrichment near late-replicating origins (p=0.013, permutation test) and at cytogenically defined ‘fragile’ sites (p=0.0017; permutation test). Although the proportion of gene content in regions exhibiting excessive SV clustering was significantly reduced when compared to a null distribution (p<0.000001, permutation test), 1,881 of 3,163 such regions (59%) intersected one or more genes (Supplementary Table 11). This includes a region comprised of 47 SVs (ranking 2nd out of the 30 genomic regions with >4 s.d.) encompassing the pregnancy-specific glycoprotein gene family (Figure 3a), a set of genes thought to be critically important for maintenance of pregnancy[37]. Other SV clusters associated with genes (*e.g.*, *IMMP2L*, *CHL1* and *GRID2*) have been implicated as potential risk factors for disease, including neurodevelopmental disorders[38].

We additionally specifically assessed the complexity of the 29,954 deletions with resolved breakpoints and found that 6% (1,822) intersected another deletion with distinct breakpoints. A larger fraction (16% or 4,813 of assembled deletion sites) showed the presence of additional inserted sequence at deletion breakpoints. We grouped 1,651 deletions with mean size of 3.1 kbp and at least 10 bp of additional DNA sequence between the original SV site boundaries into five broad classes (Figure 3b, Supplementary Table 12). The most common class (*N*=501, 30.3%), termed *Ins with Dup and Del*, comprised deletions exhibiting a recognizable duplicated sequence interval within the respective inserted sequence. Notably, in many cases (*N*=191) the inserted sequences comprised two or more apparent sequence duplications at the deletion boundaries (a class denoted *Ins with MultiDup and Del*). Additional classes commonly observed include *Inv and Del* (inversion with adjacent deletion*; N*=9) and *MultiDel*—a class where two or more adjacent deletions are separated by at least one sequence "spacer" of up to ~204 bp in length (*N*=370). However, not all complex SVs fit into these classes, with 214 sites forming distinct patterns corresponding to multiple classes or exhibiting increased complexity. Template-switching mechanisms could explain the notable complexity of these SVs[3]. Indeed, microhomology patterns were typically present between the breakpoints of deletions and the respective boundaries of insertion templates at these sites (ED Figure 9) consistent with formation through single mutational events (Supplementary Note). Across the complex sites assessed, 871 (53%) showed evidence for a local template (≥10 bp match, within 10 kbp), whereas for 41 the insertion was presumably templated from a distal region (≥22 bp match, >10 kbp away), including 17 sites where the DNA stretch was likely derived from RNA templates (Supplementary Table 13).

To further characterize SV breakpoint complexity, we employed two alternative approaches that do not rely on low-coverage Illumina read assembly. We first examined 7,804 small deletions for breakpoint complexity using split-read analysis[23] (Figure 3c) and identified 664 (median size: 67 bp) exhibiting complexity, 64 of which contained insertions ≥3 bp that may be derived from a nearby template (Supplementary Table 14, ED Figure 9). We additionally realigned long DNA reads from a single individual (NA12878)[22] sequenced by high-coverage PacBio (median read length=3.0 kbp) and Moleculo (median=3.2 kbp) single-molecule WGS around deletions from our release set (Figure 3d). Out of 766 deletions in NA12878 investigated with this approach, 62 exhibited complexity showing three to six breakpoints (Supplementary Table 12). A deletion of exon 3 of the serine protease inhibitor *SPINK14*, for example, was accompanied by an inversion of an internal segment of the SV sequence (Figure 3d panel *i*). In contrast to the smaller proportion of deletions showing breakpoint complexity, the majority of inversions assessed in NA12878 (19/28) exhibited multiple breakpoints.

To further explore inversion sequence complexity, we performed a battery of targeted analyses, leveraging PacBio resequencing of fosmids (targeting 34 loci), sequencing by Oxford Nanopore Minion (60 loci) and PacBio (206 loci) of long-range PCR amplicons, and data for 13 loci from another sample (CHM1) sequenced by high-coverage PacBio WGS[14]. Altogether we verified and further characterized 229 inversion sites, 208 using long-read data and 21 by PCR (Supplementary Table 15) increasing the number of known validated inversions[21] by >2.5-fold. Remarkably, only 20% of all sequenced inversions characterized

in this manner were “simple” (termed *Simple Inv*), exhibiting two breakpoints (Figure 3e), including a 2 kbp inversion on chromosome 4 intersecting a regulatory exon of the Ras homolog family member *RHOH* (Figure 3d panel *iv*). The majority of inversions (54%) corresponded to inverted duplications (*Inverted Dup*; Figure 3d panel *iii*). In nearly all cases, these involved duplicated stretches <1 kbp inserted within 5 kbp of the alternate copy, suggesting a common mechanism of SV formation (ED Figure 10). The remaining inversions comprised *Inv and Del* events (14%), *MultiDel* events exhibiting inverted spacers (7%), and more highly complex sites (5%; Figure 3d panel *ii*). The appreciable inversion complexity uncovered here is most likely due to a mutational process forming complex SVs, potentially involving DNA replication errors[3], rather than due to recurrent rearrangement, as our analyses failed to detect corresponding intermediate events in 1000 Genomes Project samples.

Discussion

We present the most comprehensive set of human SVs to date as an integrated resource for future disease and population genetics studies. We estimate individuals harbor a median of 18.4 Mbp of SVs per diploid genome, an excess contributed to a large extent by mCNVs (11.3 Mbp) and biallelic deletions (5.6 Mbp; Table 1). When collapsing mCNV sites carrying multiple copies as well as homozygous SVs onto the haploid reference assembly, a median of 8.9 Mbp of sequence are affected by SVs, compared to 3.6 Mbp for SNPs. Furthermore, 37,250 SVs have mapped breakpoints amounting to >113 Mbp of SV sequence resolved at the nucleotide-level. By mining homozygous deletions we identified over two hundred nonessential human genes, a set enriched for immunoglobulin domains that hence may reflect variation in the immune repertoire underlying inter-individual differences in disease susceptibility.

We demonstrate that SV classes are disproportionally enriched (by up to ~50-fold) for SV-eQTLs, although only 220 SVs were either found as lead eQTL association or in high LD with the respective lead SNP. While this corresponds to proportionally fewer associations relative to SNPs compared to a prior estimate based on array technology[34], this may be explained by the reliance of this prior estimate on bacterial artificial chromosome arrays, which ascertain large SVs (>50 kbp) that associate with strong effect size, as well as by the relative scarcity of SNPs tested in an earlier study[34] (HapMap Phase I)[39]. We further expand the number of candidate SVs in strong LD with GWAS hits by ~30% (39/136 novel associations implicating SVs as candidates) and find that GWAS haplotypes are enriched up to threefold for common SVs, which emphasizes the relevance of ascertaining SVs in disease studies. The large number of novel SVs smaller than 1 kbp in length associated with previously reported GWAS hits highlights the importance of increasing sensitivity for SV detection and genotyping at this size range. Additionally, the large number of rare SVs captured by our resource may be of value for disease association studies investigating rare variants.

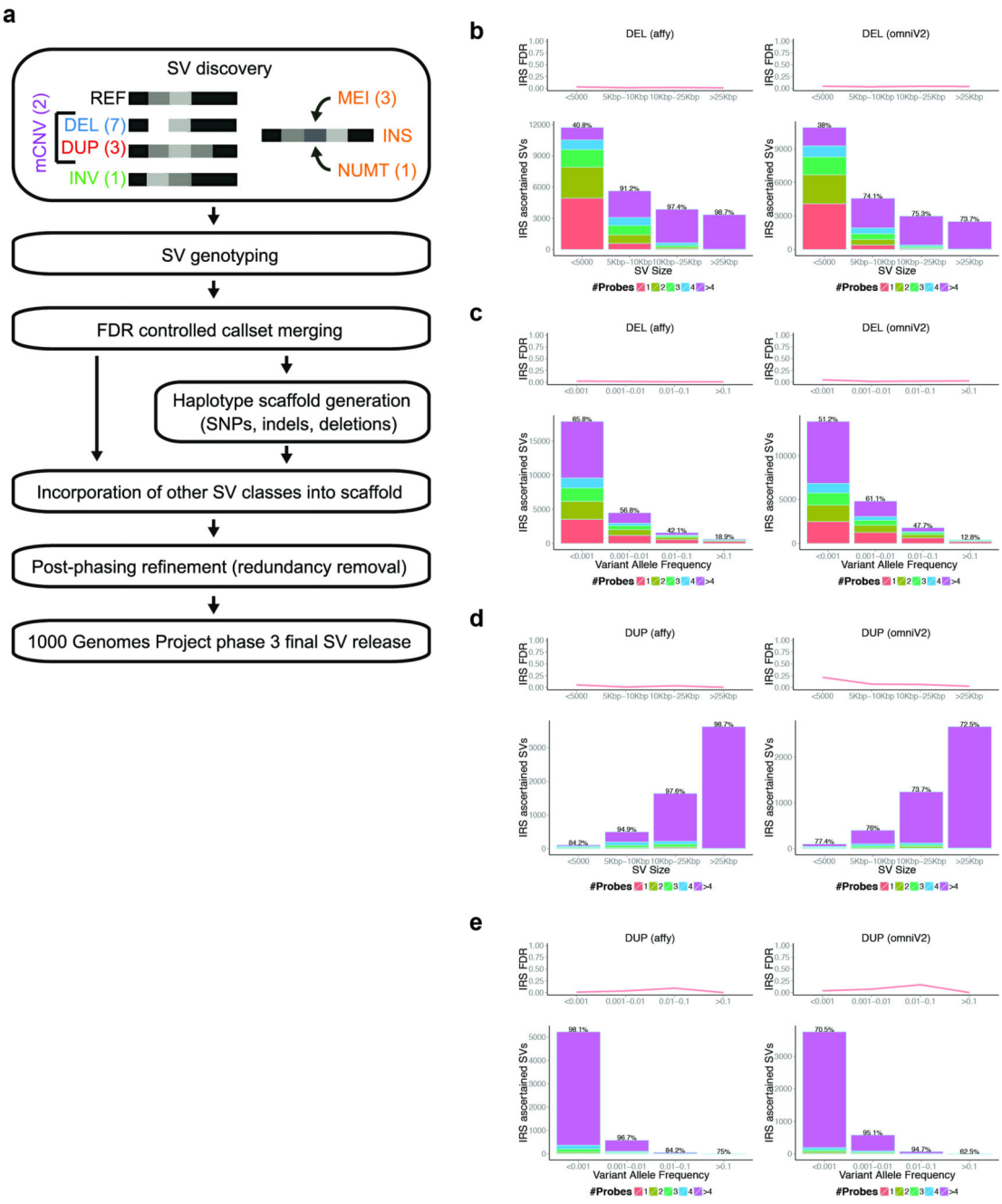
Our deep population survey has identified hotspots of SV mutation that cannot be accounted for by deep coalescence or segmental duplication content. We describe hitherto undescribed patterns of SV complexity, particularly for inversions. These patterns indicate that other

Europe PMC Funders Author Manuscripts

more complex mutational processes outside of non-allelic homologous recombination, retrotransposition, and non-homologous end-joining played an important role in shaping our genome. In spite of this, it remains difficult to fully disentangle the contributions of SV mutation rates and selective forces to the observed variant clustering. The findings presented here leveraged substantial recent technological advances, including increases in Illumina read length and developments in long-read DNA technologies. SV discovery remains a challenge nonetheless, and the full complexity and spectrum of SV is not yet understood. Our analyses, for example, are largely based on 7.4-fold Illumina WGS and, thus, are underpowered to capture much of the complexity of variation, including SVs in repetitive regions, non-reference insertions, and short SVs at the boundaries of the detection limits of read-depth and paired-end-based SV discovery[4]. Furthermore, while many SVs in our callset are statistically phased, the diploid nature of the genome is non-optimally captured by current analysis approaches, which mostly rely on mapping to a haploid reference. We envision that in the future, the use of technology allowing substantial increases in read lengths over the current state-of-the-art will enable genomic analyses of truly diploid sequences to facilitate targeting these additional layers of genomic complexity. Until this is realized, our SV set represents an invaluable resource for the construction and analysis of personalized genomes.

Europe PMC Funders Author Manuscripts

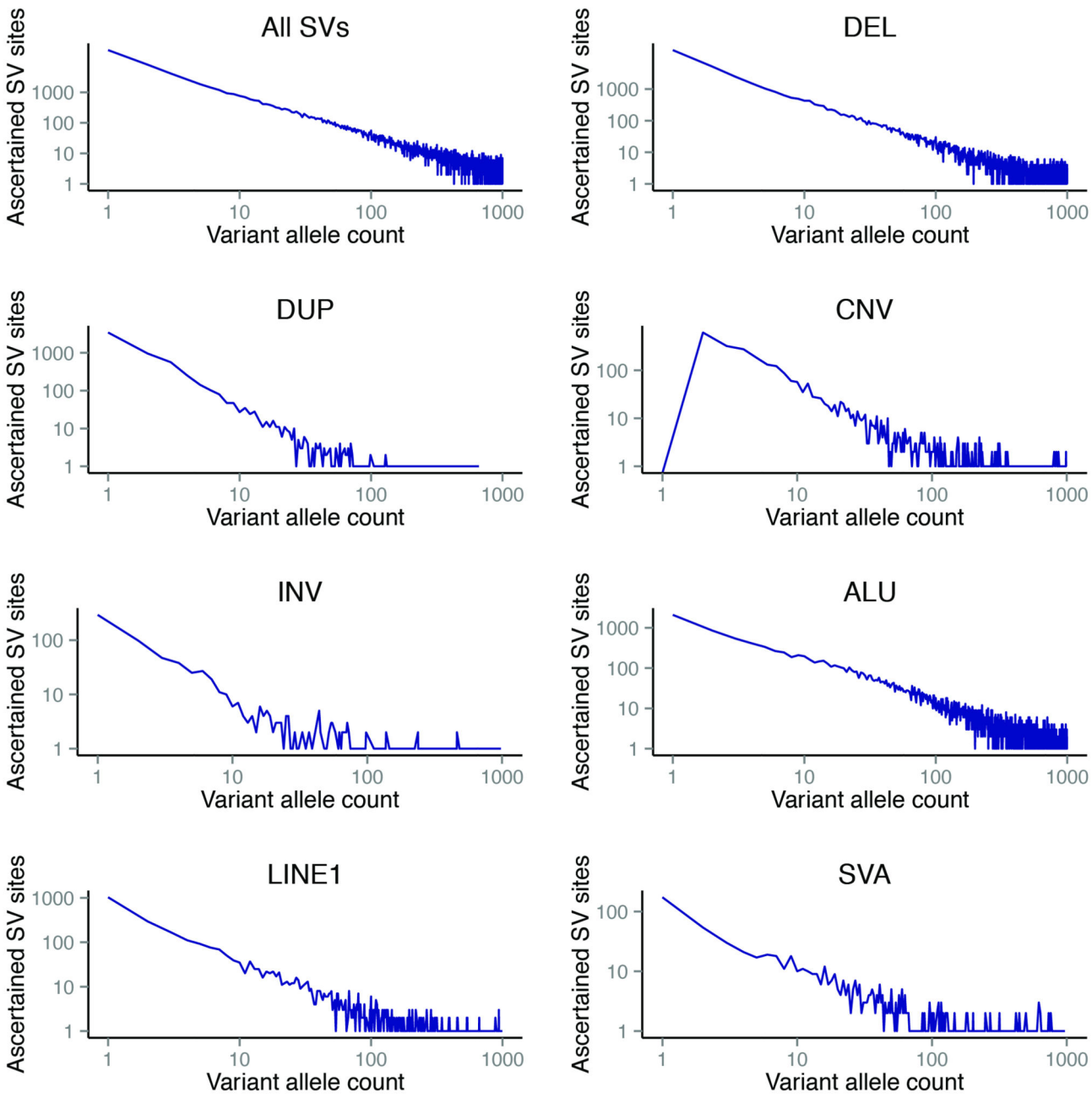
Extended Data



ED Figure 1.

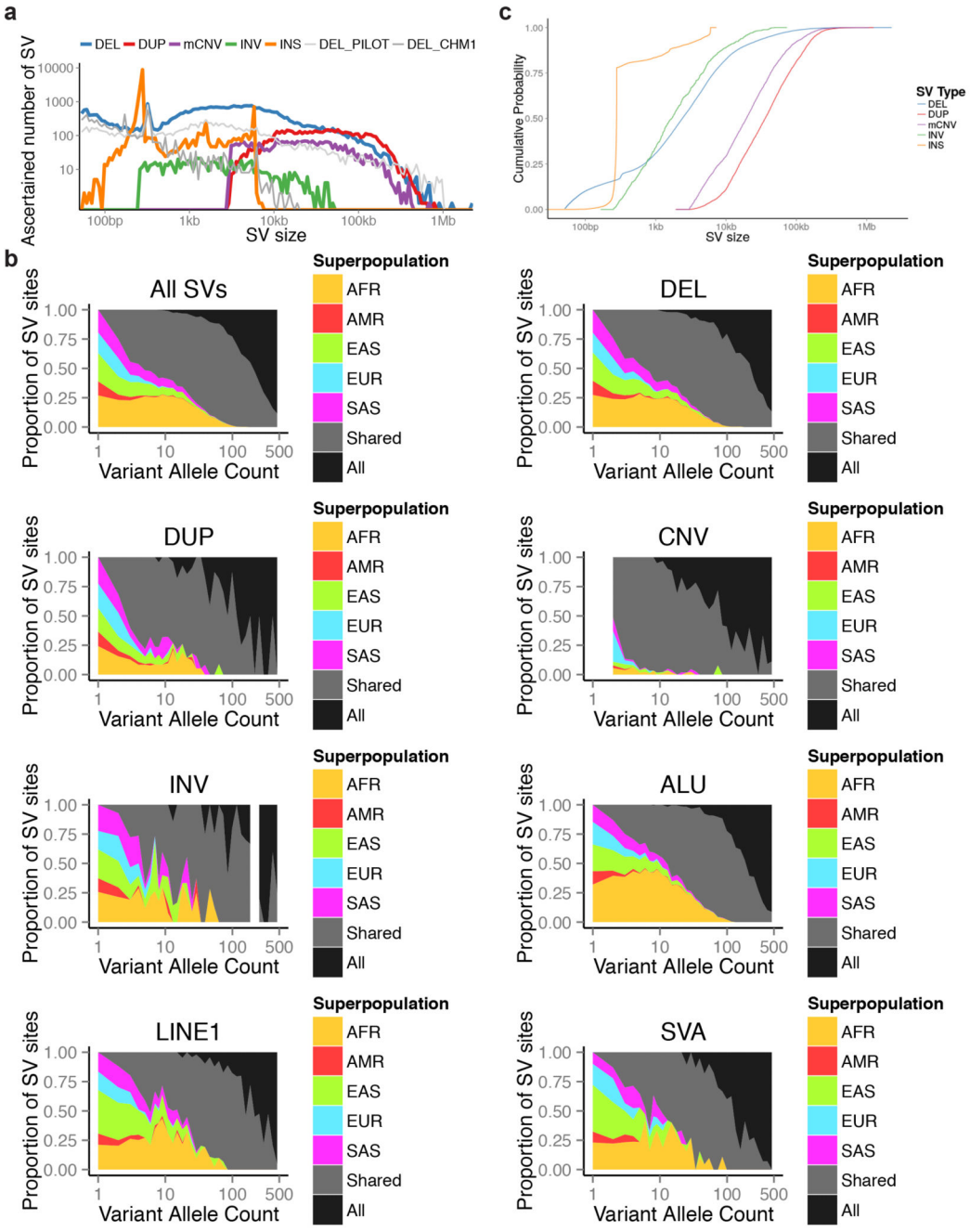
(**A**) Approach used for constructing our SV release set. (**B**) Intensity rank sum (IRS) validation results for deletions in different size bins. (**C**) IRS validation results for deletions in variant allele frequency (VAF) bins. (**D**) IRS results for duplications in different size bins. (**E**) IRS validation results for duplications in VAF bins. Based on Affymetrix SNP6 array

probes, the IRS FDR for all SV length and VAF bins was ≤5.4%, requiring at least 100 SVs per bin with an IRS assigned p-value.



ED Figure 2.

This figure shows the number of SV sites in our phase 3 release relative to allele frequency expressed in terms of allele count. SVs down to an allele count of 1 (corresponding to VAF=0.0002) are represented in our phase 3 SV set (with the exception of mCNVs, denoted 'CNV' in this figure, which are defined as sites of multi-allelic variation thus requiring allele count ≥2 – hence no mCNV sites are ascertained for allele count = 1).

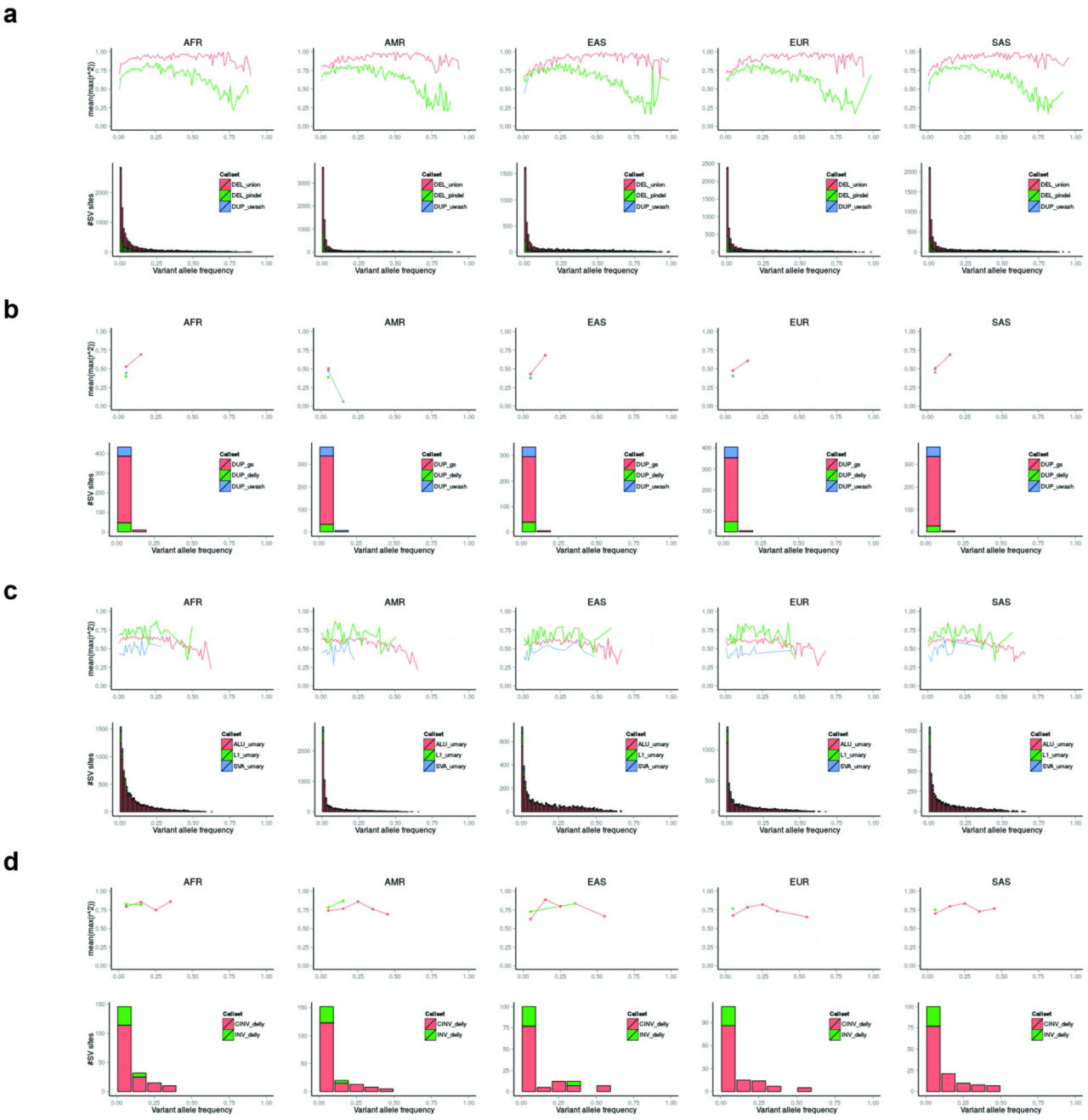


ED Figure 3.

(A) Variants ascertained in the 1000GP pilot phase (Mills et al., light gray) as well as the recent publication of SVs ascertained by PacBio sequencing in the CHM1 genome (Chaisson et al., gray) are displayed for comparison in this SV size distribution figure (INS, used as abbreviation for MEIs and NUMTs in this display item). (**B)** Population distribution of SV allele sharing across continental groups for different SV classes. (**C)** Cumulative distributions of the number of events as a function of size by SV class.

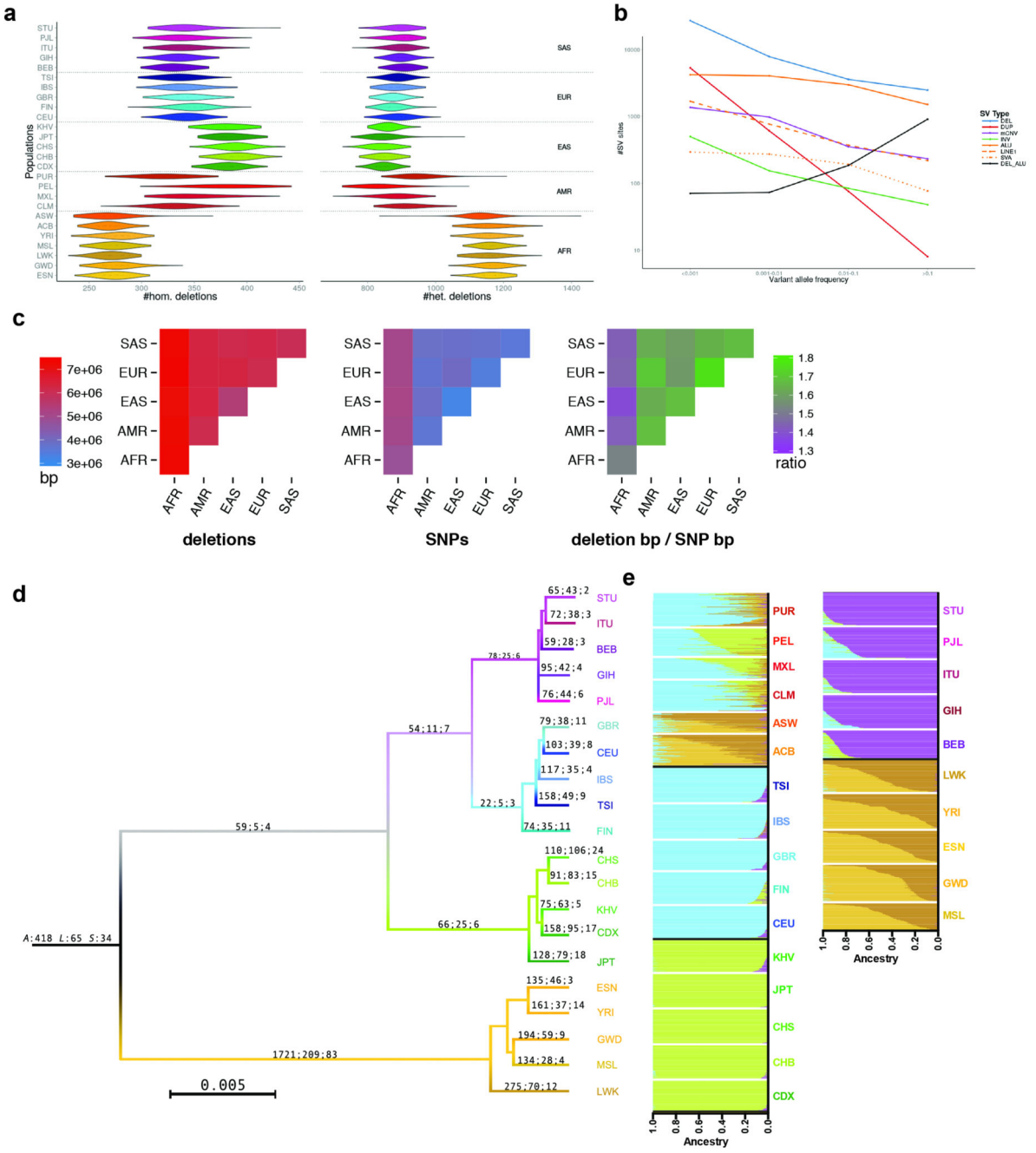
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts



ED Figure 4.

(A) LD properties of deletions, broken down by continental group and shown as a function of VAF. (**B)** LD properties of duplications. (**C)** LD properties of *Alu*, L1 and SVA mobile element insertions. (**D)** LD properties of inversions (with breakdown for two independent inversion sets generated with our inversion discovery algorithm Delly; *i.e.*, CINV=one-sided inversions with support for one breakpoint; INV=two-sided inversions with support for both breakpoints; these two sets are combined into the joint phase3 SV group inversion set).



ED Figure 5.

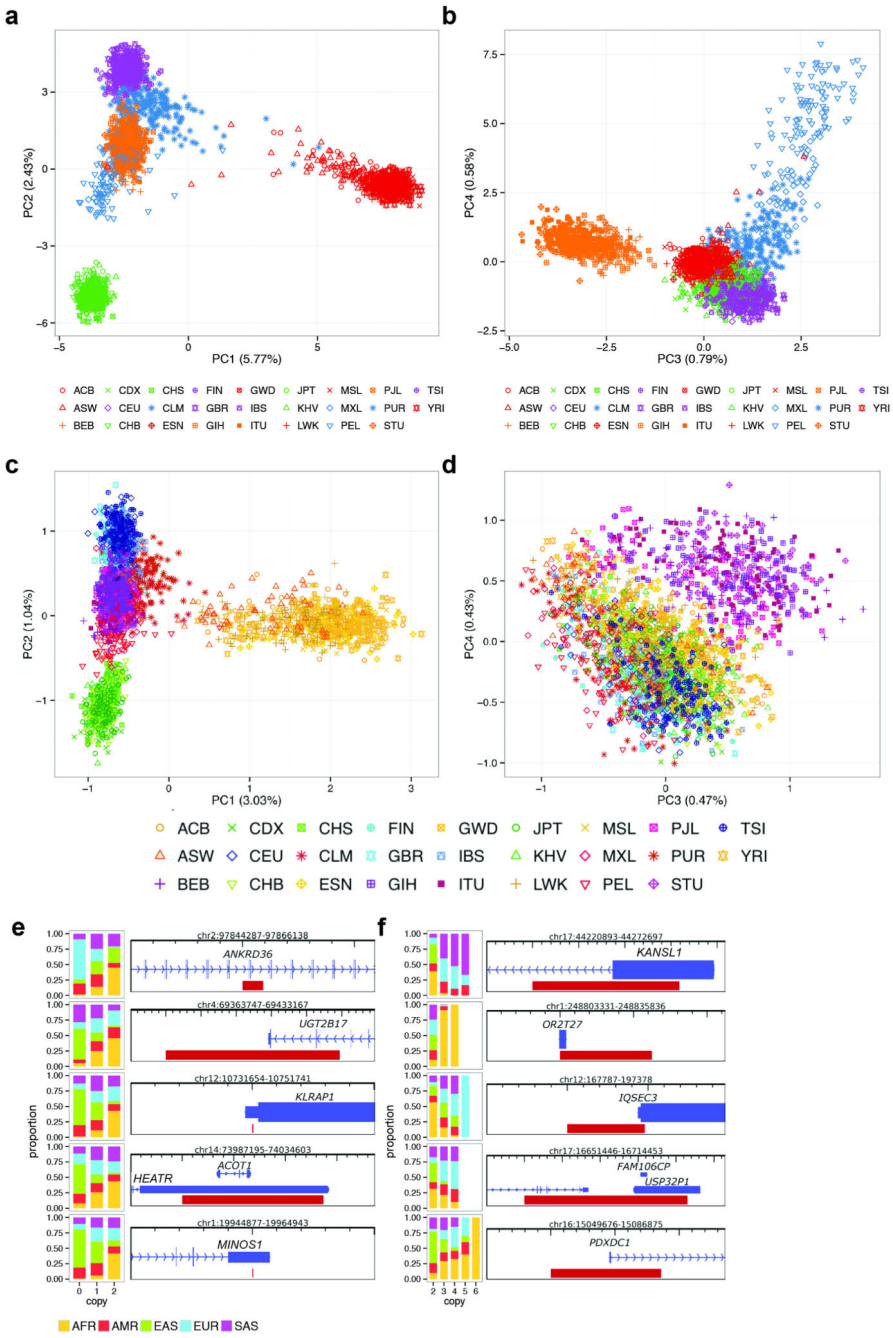
(A) Deletion heterozygosity and homozygosity among human populations for a subset of high-confidence deletions. Populations from the African continental group (AFR) exhibit the highest levels of heterozygosity and thus diversity among humans, but show the overall lowest level of deletion homozygosity among all continental groups. By comparison, East Asian populations exhibited the lowest levels of deletion heterozygosity and the highest levels of homozygosity. (Het., heterozygous. Hom., homozygous.) **(B)** VAF distribution of major SV classes. Bi-allelic duplications represent a notable outlier, showing a striking

Europe PMC Funders Author Manuscripts

depletion of common alleles, which can be explained by the preponderance of genomic sites of duplication to undergo recurrent rearrangement (see main text). As a consequence, most common duplications are classified as multi-allelic variants (*i.e.* mCNVs). (**C)** The number of base pairs (bp) differing among individuals within and between continental groups for deletions (upper panel) and SNPs (middle panel) contrasted with the ratio of deletion bp differences to SNP bp differences (*deletion bp/SNP bp*) among groups (lower panel). Non-African groups exhibit a higher *deletion bp/SNP bp* compared to Africans. (**D)** Neighbor-joining tree of populations constructed from MEIs (homoplasy-free markers) to provide a (simplified) view of population ancestry. The tree is labeled with the number of lineage-specific MEIs (*Alu*:L1:SVA). (**E)** Classification of ancestry in AFR/AMR and AMR admixed populations using homoplasy-free ancestry informative MEI markers. Color usage follows the same scheme as in Fig. 1d, except in the case of AFR individuals, which use both the color in Fig. 1d and another color that is unrelated to any other figure to indicate additional substructure within this group.

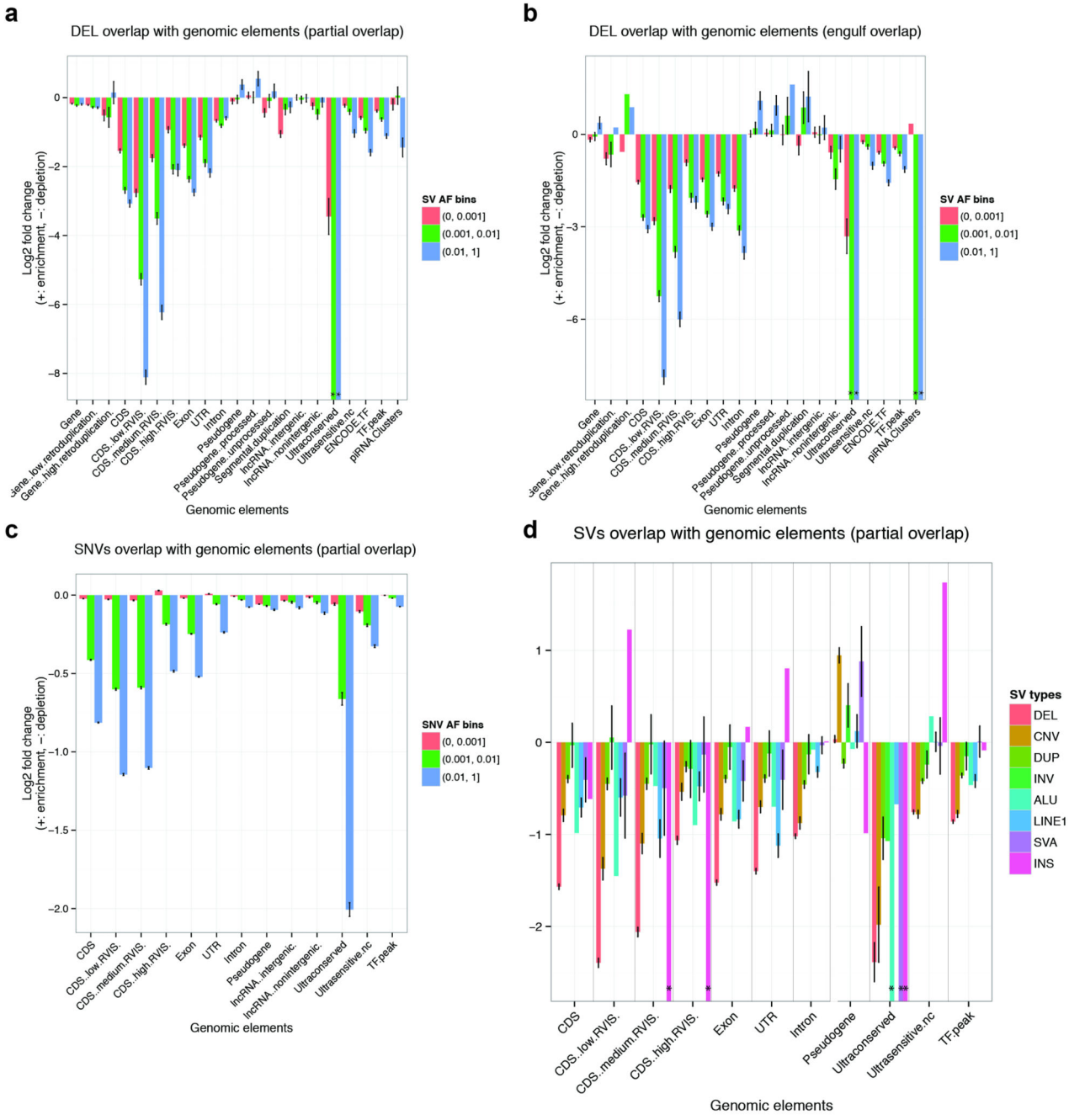
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts



ED Figure 6.

(A) PCA plot of principal components 1 and 2 for deletions. (**B)** PCA plot of principal components 3 and 4 for deletions. (**C)** PCA plot of principal components 1 and 2 for MEIs. (**D)** PCA plot of principal components 3 and 4 for MEIs. (**E)** The five most highly population-stratified deletions intersecting protein-coding genes based on Vst. (**F)** The five most highly population-stratified duplications and multi-allelic copy number variants (mCNVs) intersecting protein-coding genes based on Vst. For abbreviations, see Supplementary Table 1.

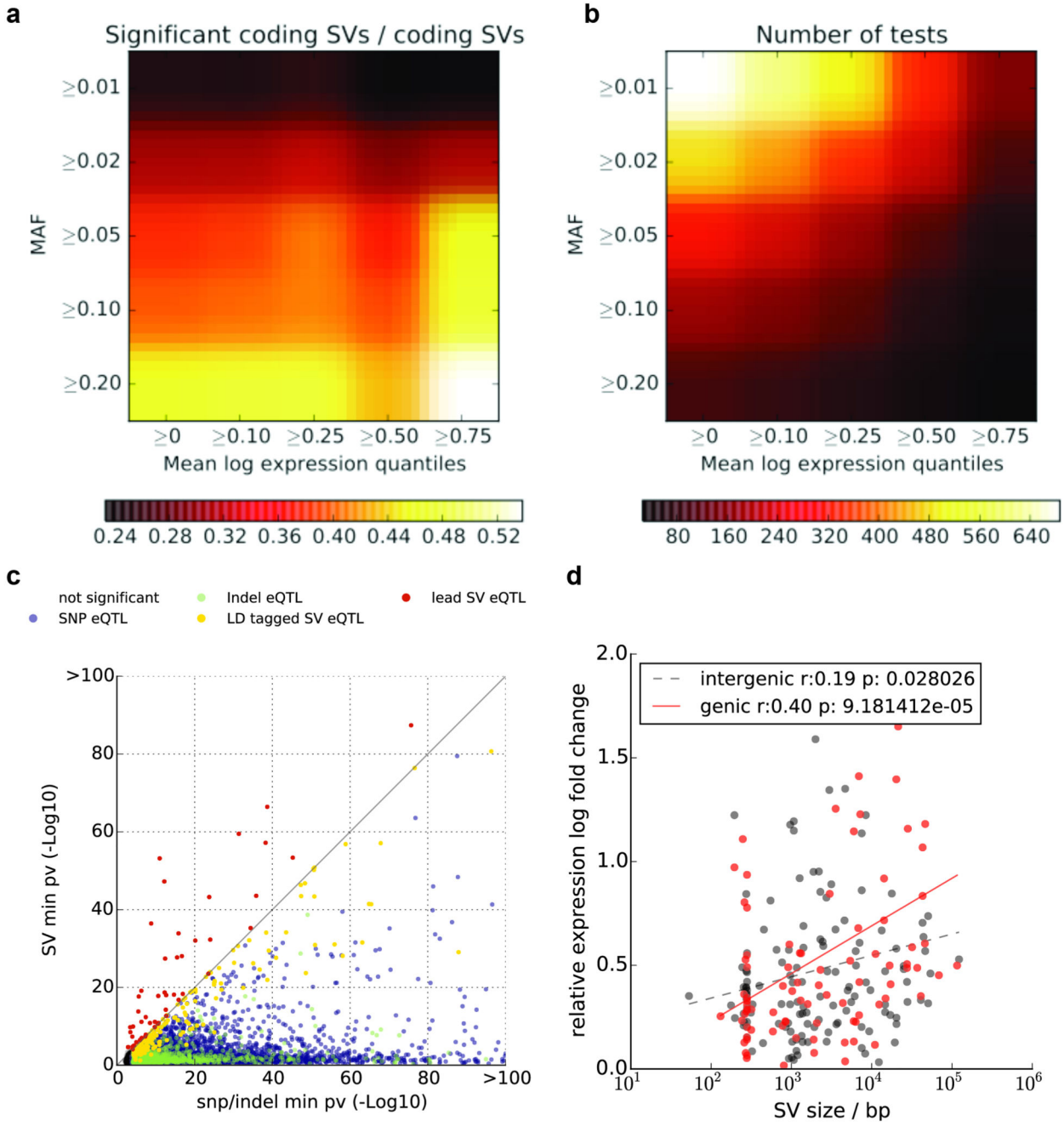


ED Figure 7.

(A) Shadow figure of Figure 2a. Overlap enrichment analysis of deletions (with resolved breakpoints) versus genomic elements, using partial overlap statistic, deletions categorized into VAF bins. (**B)** Similar to (A). The only difference is that engulf overlap statistic is used instead of partial overlap statistic. Engulf overlap statistic is the count of genomic elements (*e.g.* CDS) that are fully imbedded in at least one SV interval (*e.g.* deletions). *no element intersected observed within dataset. (**C)** Similar to (A) and (B), with the enrichment/ depletion analysis pursued for common SNPs as well as more rare single nucleotide

Europe PMC Funders Author Manuscripts

polymorphisms/variants (SNVs). Common SNV alleles show the highest levels of depletion for investigated genomic elements. **(D)** Overlap enrichment analysis of various SV types versus genomic elements, using partial overlap statistic.



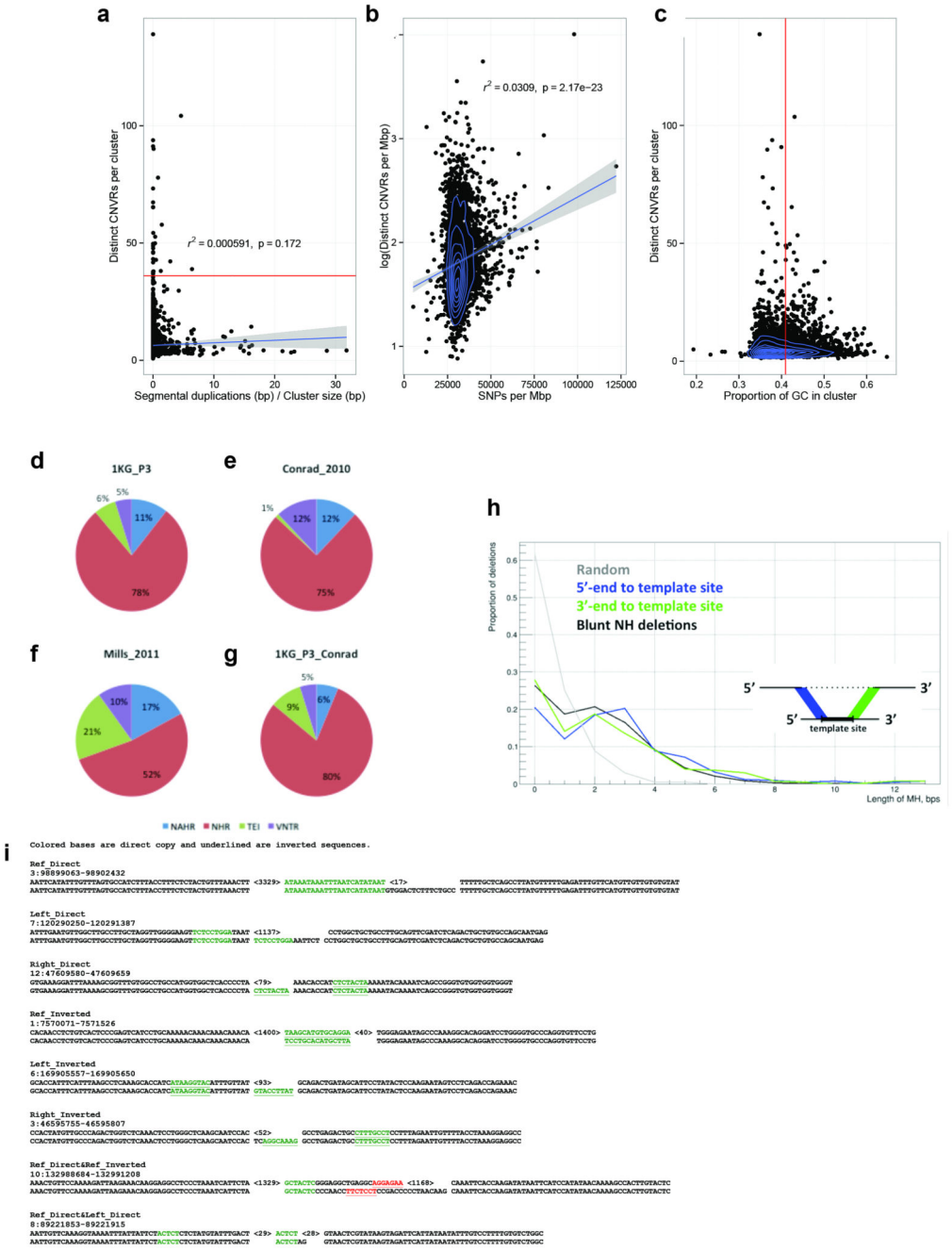
ED Figure 8.

(A) SV-centric eQTL analysis of coding SVs. Shown is the proportion of coding SVs that are eQTLs as a function of the minimum VAF and the expression quartile. (**B)** Total number of coding SVs for corresponding filters. Common SVs (VAF>0.2) in highly expressed genes

Europe PMC Funders Author Manuscripts

(>75% quantile) are very likely to correspond to SV-eQTLs (54%, see also Supplementary Table 8). **(C)** For all genes with significant eQTLs (FDR<10%), shown are raw p-values considering only SNPs (x-axes) or only SVs (y-axes). Genes with (strict lead) SV-eQTLs are shown in red. Genes with a SNP lead eQTL that is in linkage with an SV (r^2>0.5) are shown in orange. SNP lead eQTLs without an SV in LD are shown in blue. (**D)** Relative eQTL effect sizes for genetic and intergenic SV eQTLs (*N*=239) either with an SV-eQTL or an LD tagged SV (in log abundance scale). Shown are regression trends for both genic and intergenic SV eQTls. For genetic eQTLs, a clear relationship between SV effect size is found. For example, genic SVs >10kb have 3-fold larger effect sizes compared to genic SVs < 1kb; P=0.004; t-test.

Europe PMC Funders Author Manuscripts



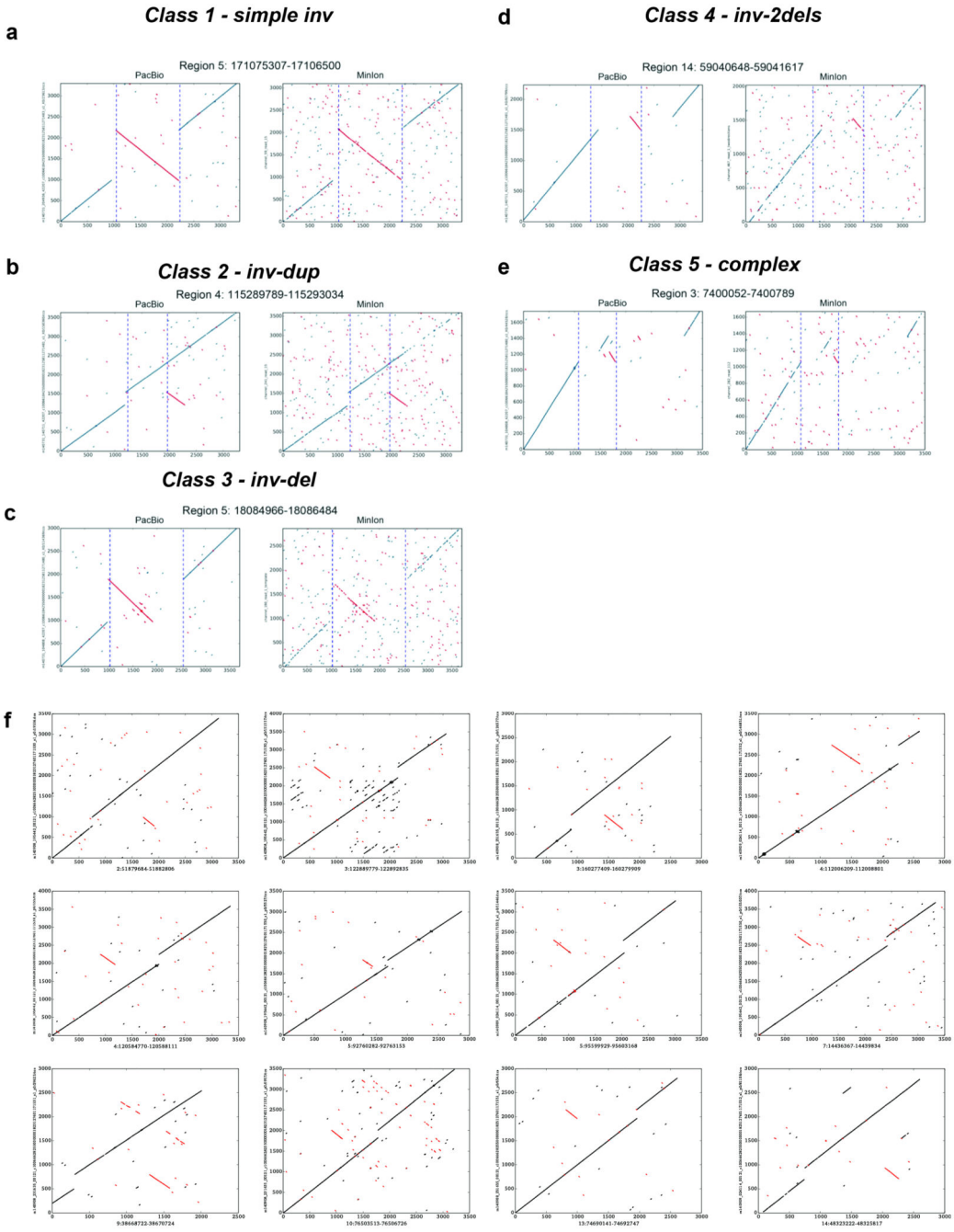
ED Figure 9.

Extensive clustering of recurrent SVs into CNVRs appears unrelated to the extent of segmental duplications (**A**) and is only partially correlating with SNP diversity (**B**) and GC content (**C**). Breakdown of SV mechanism classifications based on criteria from two earlier studies (Conrad et al. and Mills et al.). Shown are results for deletions with nucleotide resolved breakpoints. BreakSeq was used for mechanism inference. (**D**) 1KG_P3: Breakdown for our 1000GP phase 3 SV callset using classification criteria from Mills et al. (**E**) Conrad_2010: summary of mechanism classification results published in Conrad et al.

Europe PMC Funders Author Manuscripts

(**F**) Mills_2011: summary of mechanism classification results published in Mills et al. (**G**) 1KG_P3_Conrad: Breakdown for our 1000GP phase 3 SV callset using classification criteria from Conrad et al. Mechanism classification was pursued using four different categories: Blue=non-allelic homologous recombination (NAHR); green=mobile elements inserted into the reference genomes (appearing deleted in this analysis); red=nonhomology-based rearrangement mechanisms (NHR), such as NHEJ, microhomology-mediated end-joining and microhomology-mediated break-induced replication (involving blunt-ended deletion breakpoints or breakpoints with microhomoloy); purple=expansion or shrinkage of variable numbers of tandem repeats (VNTRs). TEI, transposable element insertion (equivalent with MEI). (**H)** Distribution of lengths of micro-homology (MH) for complex SVs, measured between deletion and corresponding template sites boundaries. Simple deletions, which based on BreakSeq were inferred to be formed by a nonhomology-based SV formation mechanism, such as NHEJ and microhomology-mediated break-induced replication (Supplementary Table 3), are shown as an additional control (here denoted “blunt NH deletions”). **(I)** Origins of inserted sequences in complex deletions inferred by split read analysis. This figure depicts examples for each class shown in Supplementary Table 13.

Europe PMC Funders Author Manuscripts



ED Figure 10.

Examples for five classifications of inversions verified using PacBio and Minion reads: Simple Inversion (**A**), inv-dup (**B**), inv-del (**C**), MultiDel with Inv (here abbreviated as inv-2dels) (**D**) and complex (**E**). (**F**) Several further examples of inverted duplications (inv-dup), the most common form of inversion-associated SV identified in the phase 3 release set. The figure is depicting DNA sequence alignment dotplots (same arrangement as in Figure 3), with the Y-axis referring to PacBio DNA single molecule sequencing reads and the X-axis referring to the reference genome assembly (hg19). Inverted sequences are highlighted

Europe PMC Funders Author Manuscripts

in red. Sequence analysis suggests that these inverted duplications are not typically associated with retrotransposition.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Peter H. Sudmant[#1], Tobias Rausch[#2], Eugene J. Gardner[#3], Robert E. Handsaker[#4,5], Alexej Abyzov[#6], John Huddleston[#1,7], Yan Zhang[#8,9], Kai Ye[#10,11], Goo Jun[12,13], Markus Hsi-Yang Fritz[2], Miriam K. Konkel[14], Ankit Malhotra[15], Adrian M. Stütz[2], Xinghua Shi[16], Francesco Paolo Casale[17], Jieming Chen[8,18], Fereydoun Hormozdiari[1], Gargi Dayama[19], Ken Chen[20], Maika Malig[1], Mark J.P. Chaisson[1], Klaudia Walter[21], Sascha Meiers[2], Seva Kashin[4,5], Erik Garrison[22], Adam Auton[23], Hugo Y. K. Lam[24], Xinmeng Jasmine Mu[8,25], Can Alkan[26], Danny Antaki[27], Taejeong Bae[6], Eliza Cerveira[15], Peter Chines[28], Zechen Chong[20], Laura Clarke[17], Elif Dal[26], Li Ding[10,11,29,30], Sarah Emery[31], Xian Fan[20], Madhusudan Gujral[27], Fatma Kahveci[26], Jeffrey M. Kidd[12,31], Yu Kong[23], Eric-Wubbo Lameijer[32], Shane McCarthy[21], Paul Flicek[17], Richard A. Gibbs[33], Gabor Marth[22], Christopher E. Mason[34,35], Androniki Menelaou[36,37], Donna M. Muzny[38], Bradley J. Nelson[1], Amina Noor[27], Nicholas F. Parrish[39], Matthew Pendleton[38], Andrew Quitadamo[16], Benjamin Raeder[2], Eric E. Schadt[38], Mallory Romanovitch[15], Andreas Schlattl[2], Robert Sebra[38], Andrey A. Shabalin[40], Andreas Untergasser[2,41], Jerilyn A. Walker[14], Min Wang[33], Fuli Yu[33], Chengsheng Zhang[15], Jing Zhang[8,9], Xiangqun Zheng-Bradley[17], Wanding Zhou[20], Thomas Zichner[2], Jonathan Sebat[27], Mark A. Batzer[14], Steven A. McCarroll[4,5], The 1000 Genomes Project Consortium[†], Ryan E. Mills[19,31,*], Mark B. Gerstein[8,9,42,*], Ali Bashir[38,*], Oliver Stegle[17,*], Scott E. Devine[3,*], Charles Lee[15,43,*], Evan E. Eichler[1,7,*,@], and Jan O. Korbel[2,17,*,@]

Affiliations

[1]Department of Genome Sciences, University of Washington, 3720 15th Ave NE, Seattle, WA 98195-5065, USA [2]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Meyerhofstr. 1, 69117 Heidelberg, Germany [3]Institute for Genome Sciences, University of Maryland School of Medicine, 801 W Baltimore Street, Baltimore, MD 21201, USA [4]Department of Genetics, Harvard Medical School, Boston, 25 Shattuck Street, Boston, MA 02115, USA [5]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA [6]Department of Health Sciences Research, Center for Individualized Medicine, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA [7]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA [8]Program in Computational Biology and Bioinformatics, Yale University, BASS 432&437, 266 Whitney Avenue, New Haven, CT 06520, USA [9]Department of Molecular Biophysics and Biochemistry, School of Medicine, Yale University, 266 Whitney Ave, New Haven, CT 06520, USA [10]The Genome Institute, Washington University School of Medicine, 4444 Forest Park Ave, St. Louis, MO 63108, USA

Europe PMC Funders Author Manuscripts

[11]Department of Genetics, Washington University in St. Louis, 4444 Forest Park Ave, St. Louis, MO 63108, USA [12]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, USA [13]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, 1200 Pressler St., Houston, TX 77030, USA [14]Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA [15]The Jackson Laboratory for Genomic Medicine, 10 Discovery 263 Farmington Ave, Farmington, CT 06030, USA [16]Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, 9201 University City Blvd., Charlotte, NC 28223, USA [17]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom [18]Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, CT 06520, USA [19]Department of Computational Medicine & Bioinformatics, University of Michigan, 500 S. State Street, Ann Arbor, MI 48109, USA [20]The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030, USA [21]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK [22]Department of Biology, Boston College, 355 Higgins Hall, 140 Commonwealth Ave, Chestnut Hill, MA 02467, USA [23]Department of Genetics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Bronx, NY 10461, USA. [24]Bina Technologies, Roche Sequencing, 555 Twin Dolphin Drive, Redwood City, CA 94065, USA [25]Cancer Program, Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA [26]Department of Computer Engineering, Bilkent University, 06800 Ankara, Turkey [27]University of California San Diego (UCSD), 9500 Gilman Drive, La Jolla, CA 92093, USA [28]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892 USA [29]Department of Medicine, Washington University in St. Louis, 4444 Forest Park Ave, St. Louis, MO 63108, USA [30]Siteman Cancer Center, 660 South Euclid Ave, St. Louis, MO 63110, USA [31]Department of Human Genetics, University of Michigan, 1241 Catherine Street, Ann Arbor, MI 48109, USA [32]Molecular Epidemiology, Leiden University Medical Center, Leiden 2300RA, The Netherlands [33]Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA [34]The Department of Physiology and Biophysics and the HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, 1305 York Avenue, Weill Cornell Medical College, New York, New York 10065, USA [35]The Feil Family Brain and Mind Research Institute, 413 East 69th St, Weill Cornell Medical College, New York, New York 10065, USA [36]University of Oxford, 1 South Parks Road, Oxford OX3 9DS, UK [37]Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, 3584 CG, The Netherlands [38]Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, NY School of Natural Sciences, 1428 Madison Ave, New York, NY 10029, USA [39]Institute for Virus Research, Kyoto University, 53 Shogoin Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan [40]Center for Biomarker Research and Precision Medicine, Virginia

Commonwealth University, 1112 East Clay Street, McGuire Hall, Richmond, VA 23298-0581, USA [41]Zentrum für Molekulare Biologie, University of Heidelberg, Im Neuenheimer Feld 282, 69120 Heidelberg, Germany [42]Department of Computer Science, Yale University, 51 Prospect Street, New Haven, CT 06511, USA [43]Department of Graduate Studies – Life Sciences, Ewha Womans University, Ewhayeodae-gil, Seodaemun-gu, Seoul, South Korea 120-750

Acknowledgements

We thank Matthew Hurles, Richard Durbin and David Reich for valuable comments during the preparation of this work, Stephen Scherer for providing PCR-based inversion genotyping data for the initial calibration of our inversion caller, Ben Nelson and Vladimir Benes for technical assistance, and Tonia Brown and Nina Habermann for critical review of the manuscript. The following people are acknowledged for contributing to PacBio sequencing or analysis: Ekta Patel, Sandra Lee, Harsha Doddapaneni, Lora Lewis, Robert Ruth, Qingchang Meng, Vanesa Vee, Yi Han, Joy Jayaseelan, Adam English, Jonas Korlach, Mike Hunkapiller, Bruno Hüttel and Richard Reinhardt. We acknowledge the Yale University Biomedical High-Performance Computing Center and high-performance compute infrastructure made available through the EMBL and EMBL-EBI IT facilities. We thank the people generously contributing samples to the 1000 Genomes Project. Funding for this research project came from the following grants: NIH U41HG007497 (to C.L., E.E.E., J.O.K., M.A.B., M.G., S.A.M., R.E.M. and J.S.), RO1GM59290 (M.A.B.), R01HG002898 (S.E.D.) and R01CA166661 (S.E.D.), P01HG007497 (to E.E.E.), R01HG007068 (to R.E.M.), RR19895 and RR029676-01 (to M.B.G.), Wellcome Trust WT085532/Z/08/Z and WT104947/Z/14/Z (to P.F.), an Emmy Noether Grant from the German Research Foundation (KO4037/1-1, to J.O.K.) and the European Molecular Biology Laboratory. C.L. is on the scientific advisory board (SAB) of BioNano Genomics. E.E.E. is on the scientific advisory board (SAB) of DNAnexus, Inc. and is a consultant for Kunming University of Science and Technology (KUST) as part of the 1000 China Talent Program. P.F. is on the SAB of Omicia, Inc. C.L. is an Ewha Womans University Distinguished Professor. E.E.E. is an investigator of the Howard Hughes Medical Institute. J.O.K. is a European Research Council investigator.

References

1. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013; 14:125–138. doi:10.1038/nrg3373. [PubMed: 23329113]
2. Malhotra D, Sebat J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. Cell. 2012; 148:1223–1241. doi:10.1016/j.cell.2012.02.039. [PubMed: 22424231]
3. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009; 10:551–564. doi:nrg2593 [pii]10.1038/nrg2593. [PubMed: 19597530]
4. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011; 12:363–376. doi:nrg2958 [pii]10.1038/nrg2958. [PubMed: 21358748]
5. Wellcome-Trust-Case-Control-Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. 2010; 464:713–720. doi:nature08979 [pii]10.1038/nature08979. [PubMed: 20360734]
6. Mills RE, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011; 470:59–65. doi:nature09708 [pii]10.1038/nature09708. [PubMed: 21293372]
7. Sudmant PH, et al. Diversity of human copy number variation and multicopy genes. Science. 2010; 330:641–646. doi:330/6004/641 [pii]10.1126/science.1197005. [PubMed: 21030649]
8. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. doi:10.1038/nature11632. [PubMed: 23128226]
9. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. doi:nature09534 [pii]10.1038/nature09534. [PubMed: 20981092]
10. Conrad DF, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010; 464:704–712. doi:nature08516 [pii]10.1038/nature08516. [PubMed: 19812545]

Europe PMC Funders Author Manuscripts

11. Kidd JM, et al. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell. 2010; 143:837–847. doi:S0092-8674(10)01197-9 [pii]10.1016/j.cell.2010.10.027. [PubMed: 21111241]
12. Korbel JO, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science. 2007; 318:420–426. doi:1149504 [pii]10.1126/science.1149504. [PubMed: 17901297]
13. Pang AW, et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 2010; 11:R52. doi:gb-2010-11-5-r52 [pii]10.1186/gb-2010-11-5-r52. [PubMed: 20482838]
14. Chaisson MJ, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2014 doi:10.1038/nature13907.
15. Teague B, et al. High-resolution human genome structure by single-molecule analysis. Proc Natl Acad Sci U S A. 2010; 107:10848–10853. doi:10.1073/pnas.0914638107. [PubMed: 20534489]
16. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015 10.1038/nature15393.
17. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010; 26:589–595. doi:10.1093/bioinformatics/btp698. [PubMed: 20080505]
18. Hach F, et al. mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. Nucleic Acids Res. 2014; 42:W494–500. doi:10.1093/nar/gku370. [PubMed: 24810850]
19. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014; 42:D986–992. doi:10.1093/nar/gkt958. [PubMed: 24174537]
20. Stewart C, et al. A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet. 2011; 7:e1002236. doi:10.1371/journal.pgen.1002236. [PubMed: 21876680]
21. Martinez-Fundichely A, et al. InvFEST, a database integrating information of polymorphic inversions in the human genome. Nucleic Acids Res. 2014; 42:D1027–1032. doi:10.1093/nar/gkt1122. [PubMed: 24253300]
22. Pendleton M, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods. 2015 doi:10.1038/nmeth.3454.
23. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009; 25:2865–2871. doi:btp394 [pii]10.1093/bioinformatics/btp394. [PubMed: 19561018]
24. Kloosterman WP, et al. Characteristics of de novo structural changes in the human genome. Genome Res. 2015; 25:792–801. doi:10.1101/gr.185041.114. [PubMed: 25883321]
25. McCarroll SA, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. 2008; 40:1166–1174. doi:ng.238 [pii]10.1038/ng.238. [PubMed: 18776908]
26. Locke DP, et al. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet. 2006; 79:275–290. doi:S0002-9297(07)63135-8 [pii]10.1086/505653. [PubMed: 16826518]
27. Handsaker RE, et al. Large multiallelic copy number variations in humans. Nat Genet. 2015 doi:10.1038/ng.3200.
28. Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is insensitive to recent population history. Nat Genet. 2014; 46:220–224. doi:10.1038/ng.2896. [PubMed: 24509481]
29. Redon R, et al. Global variation in copy number in the human genome. Nature. 2006; 444:444–454. doi:nature05329 [pii]10.1038/nature05329. [PubMed: 17122850]
30. Stefansson H, et al. A common inversion under selection in Europeans. Nat Genet. 2005; 37:129–137. doi:ng1508 [pii]10.1038/ng1508. [PubMed: 15654335]
31. Encode-Project-Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. doi:10.1038/nature11247. [PubMed: 22955616]
32. McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 2009; 5:e1000471. doi:10.1371/journal.pgen.1000471. [PubMed: 19424416]

33. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 2013; 9:e1003709. doi:10.1371/journal.pgen.1003709. [PubMed: 23990802]
34. Stranger BE, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007; 315:848–853. doi:315/5813/848 [pii]10.1126/science.1136678. [PubMed: 17289997]
35. Schlattl A, Anders S, Waszak SM, Huber W, Korbel JO. Relating CNVs to transcriptome data at fine resolution: Assessment of the effect of variant size, type, and overlap with functional regions. Genome Res. 2011; 21:2004–2013. doi:gr.122614.111 [pii]10.1101/gr.122614.111. [PubMed: 21862627]
36. Lappalainen T, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501:506–511. doi:10.1038/nature12531. [PubMed: 24037378]
37. Moore T, Dveksler GS. Pregnancy-specific glycoproteins: complex gene families regulating maternal-fetal interactions. The International journal of developmental biology. 2014; 58:273–280. doi:10.1387/ijdb.130329gd. [PubMed: 25023693]
38. Girirajan S, et al. Relative burden of large CNVs on a range of neurodevelopmental phenotypes. PLoS Genet. 2011; 7:e1002334. doi:10.1371/journal.pgen.1002334. [PubMed: 22102821]
39. International-HapMap-Consortium. A haplotype map of the human genome. Nature. 2005; 437:1299–1320. doi:10.1038/nature04226. [PubMed: 16255080]
40. Conrad DF, et al. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. Nature Genet. 2010; 42:385–391. [PubMed: 20364136]

Europe PMC Funders Author Manuscripts

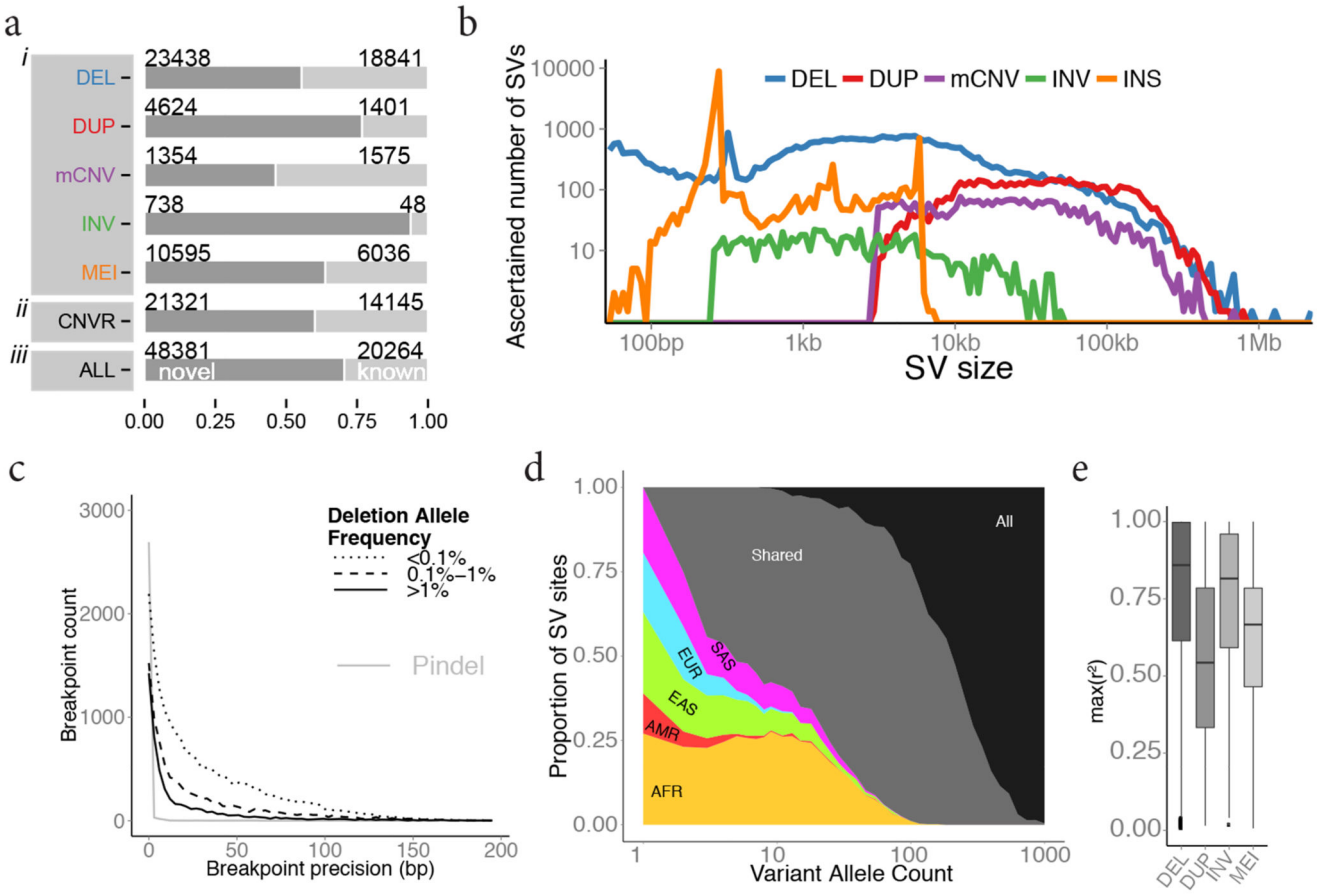


Figure 1. Phase 3 integrated SV callset

a. Novelty based on overlap of our SV set with DGV[19] (upper panel *i*, broken down by SV class), of collapsed CNVRs with earlier 1000 Genomes Project releases[6,8] (*ii*) and of our SV set with refs[6,8] (*iii*). **b.** Size distribution of ascertained SVs (bin width is uniform in log-scale). DEL, biallelic deletion, DUP, biallelic duplication, INV, inversion, INS, non-reference insertion (including MEIs and NUMTs). **c.** Breakpoint precision of assembled deletions stratified by VAF (split-read caller Pindel[23] shown separately). **d.** SV allele sharing across continental groups. **f.** LD properties of biallelic SV classes.

Europe PMC Funders Author Manuscripts

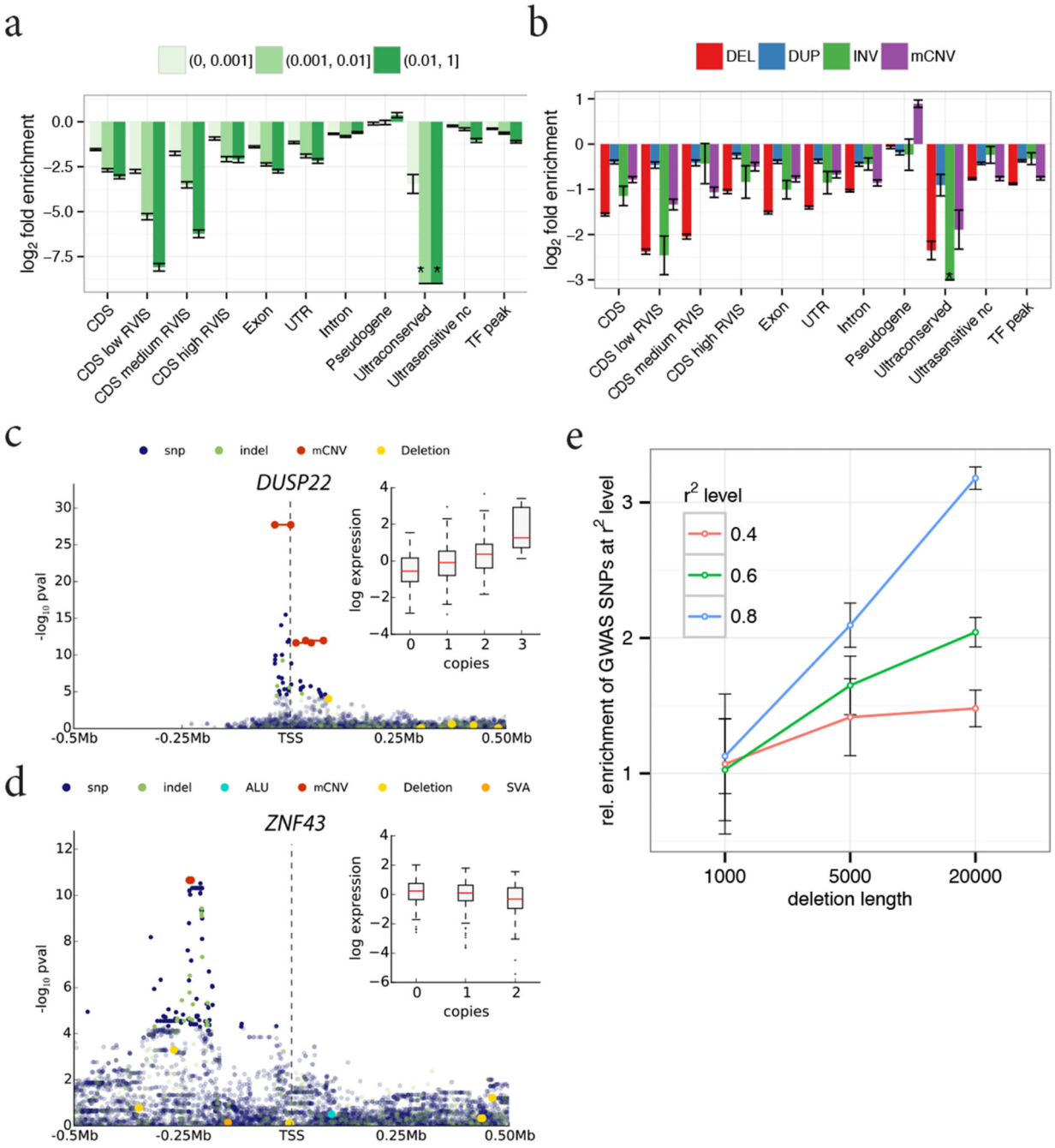


Figure 2. SV functional impact

a. Relative enrichment or depletion of genomic elements within breakpoint-resolved deletions binned by VAF. TF, transcription factor binding site; nc, noncoding. RVIS range from 0–100 (low <20, medium 20–50, high ≥50). *no element intersected. **b.** Enrichment/depletion of genomic elements within different SV classes, compared with breakpoint-resolved deletions. **c.** Manhattan plot of *DUSP22*-eQTL. Inset: boxplots of association between copy-number genotype and expression. **d.** Manhattan plot of *ZNF43*-eQTL. **e.**

Europe PMC Funders Author Manuscripts

Enrichment of SV-containing haplotypes at previously reported GWAS hits (error bars show s.e.m.).

Europe PMC Funders Author Manuscripts

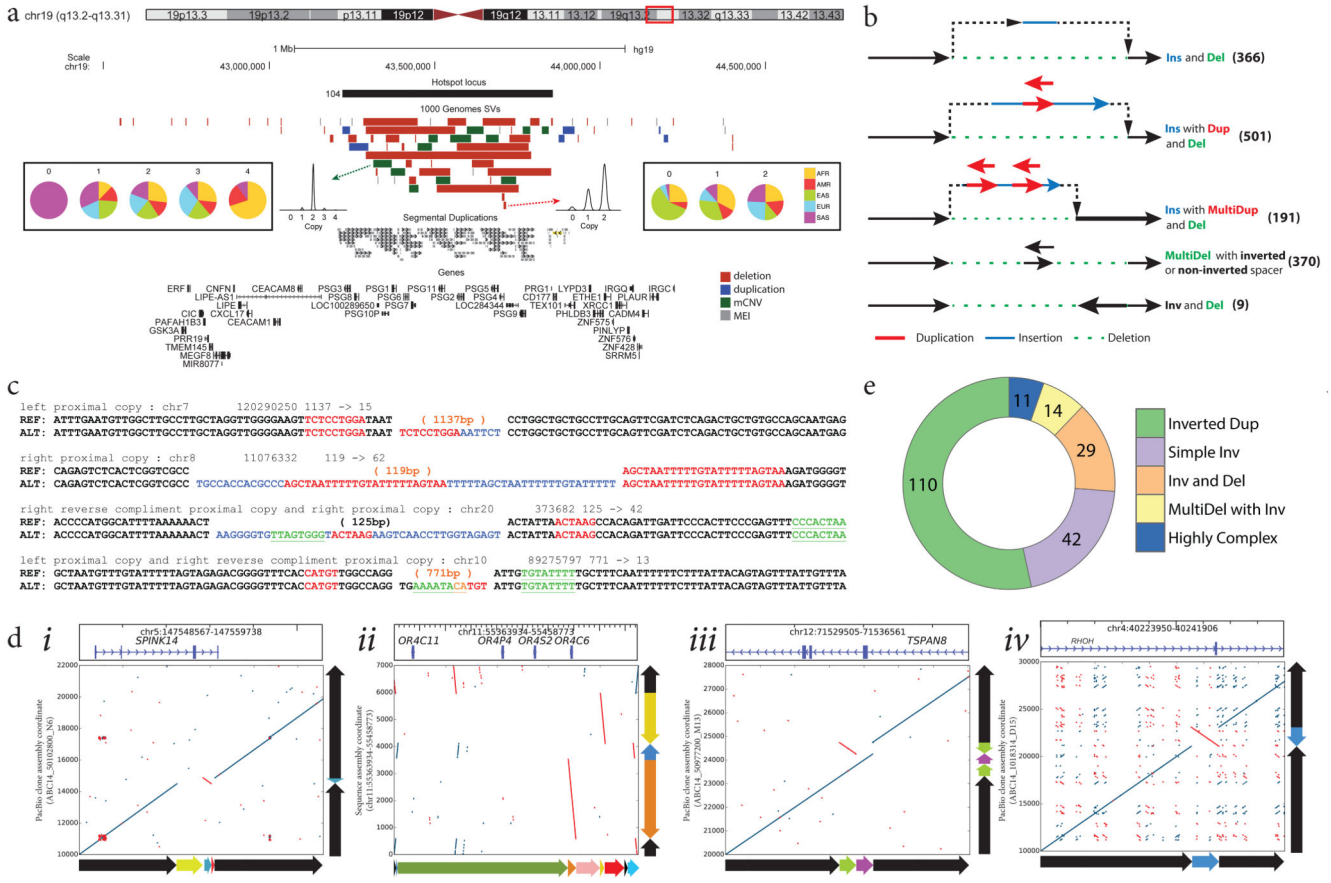


Figure 3. SV complexity at different scales

a. *PSG* locus with clustered SVs. Population copy-number state histograms are shown for two example SVs. **b.** Schemes depicting assembled complex deletions. **c.** Smaller-scale complex deletions identified with Pindel[23]. Flanking sequences are shown for reference (REF) and alternate (ALT) alleles, further to insertions at the breakpoints. Proximal stretches matching the insertion are labeled in red (forward) and green (reverse complement). Blue: insertions lacking nearby matches. **d.** Alignment dotplots depicting inversions (inverted sequences are in red within each dotplot). Adjacent schemes depict allelic structures for REF and ALT. **e.** Inversion complexity summarized.

Europe PMC Funders Author Manuscripts

Table 1

Phase 3 extended SV release

FDR estimates are based on intensity rank-sum testing[8] using Affymetrix SNP6[A] and Omni 2.5 arrays[O], PCR[P], as well as long-read[L], PCR-free (250 bp-read)[D] and CG[C] sequencing (CG-based estimates used reciprocal overlaps of 50% and 20% for deletions and duplications, respectively). [#,V]estimate by comparing MEIs to all calls[#] or all PCR-validated calls[V] from[20] (estimates for individual MEI classes are in Supplementary Table 4). *NA**, no previous data available. Differences in deletion and duplication counts are driven by size-cutoffs and classification of common duplications as mCNVs[27]. [RR]ascertained using read-pairs or read-depth. [SR]ascertained with split-reads[23]. [T]estimated for tandem duplications. [†]estimated for inversions with paired-end support from both breakpoints.

SV class	No. sites	Median size of SV sites (bp)	Median kbp per Individual	Median Alleles per individual	Site FDR	Biallelic site breakpoint precision (bp)	Genotype concordance (non-ref.)	Sensitivity estimates
Deletion (biallelic)	42,279	2,455	5,615	2,788	2% [A] - 4% [O]	15 (±50)[RR] 0.7 (±9.5)[SR]	98% [C]	88% [C]
Duplication (biallelic)	6,025	35,890	518	17	1% [A] - 4% [O]	683 (±1350) [T]	94% [C]	65% [C]
mCNV	2,929	19,466	11,346	340	1% [A] - 4% [O]	-	*NA**	*NA**
Inversion	786	1,697	78	37	17% [L] (9%) [P, †]	32 (±47) [†]	96% [L]	32%
MEI	16,631	297	691	1,218	4% [P]	0.95 (±5.93)	98% [D]	83 [#] - 96% [V]
NUMT	168	157	3	5.3	10% [P]	0.25 (±0.43) [N]	86.1% [P]	*NA**