Human Genomics

PRIMARY RESEARCH

Open Access

Genetic-variant hotspots and hotspot clusters in the human genome facilitating adaptation while increasing instability

Xi Long[1,2] and Hong Xue[1,2,3*]

Abstract

Background: Genetic variants, underlining phenotypic diversity, are known to distribute unevenly in the human genome. A comprehensive understanding of the distributions of different genetic variants is important for insights into genetic functions and disorders.

Methods: Herein, a sliding-window scan of regional densities of eight kinds of germline genetic variants, including single-nucleotide-polymorphisms (SNPs) and four size-classes of copy-number-variations (CNVs) in the human genome has been performed.

Results: The study has identified 44,379 hotspots with high genetic-variant densities, and 1135 hotspot clusters comprising more than one type of hotspots, accounting for 3.1% and 0.2% of the genome respectively. The hotspots and clusters are found to co-localize with different functional genomic features, as exemplified by the associations of hotspots of middle-size CNVs with histone-modification sites, work with balancing and positive selections to meet the need for diversity in immune proteins, and facilitate the development of sensory-perception and neuroactive ligand-receptor interaction pathways in the function-sparse late-replicating genomic sequences. Genetic variants of different lengths co-localize with retrotransposons of different ages on a "long-with-young" and "short-with-all" basis. Hotspots and clusters are highly associated with tumor suppressor genes and oncogenes ($p < 10^{-10}$), and enriched with somatic tumor CNVs and the trait- and disease-associated SNPs identified by genome-wise association studies, exceeding tenfold enrichment in clusters comprising SNPs and extra-long CNVs.

Conclusions: In conclusion, the genetic-variant hotspots and clusters represent two-edged swords that spearhead both positive and negative genomic changes. Their strong associations with complex traits and diseases also open up a potential "Common Disease-Hotspot Variant" approach to the missing heritability problem.

Keywords: Frequency independent, Genetic diversity, Missing heritability, Retrotransposon, Recombination-selection co-saturation, Replication timing

* Correspondence: hxue@ust.hk
[1] Division of Life Science and Applied Genomics Centre, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China
[2] HKUST Shenzhen Research Institute, 9 Yuexing First Road, Nanshan, Shenzhen, China
Full list of author information is available at the end of the article

Background

Genetic variants (GVs) are essential contributors to population diversity, providing an important basis for the investigation of genetic effects on complex traits and disease susceptibility. A wide spectrum of GVs exist in the human genome ranging from point alterations, viz. single-nucleotide polymorphisms (SNPs), to structural variations including copy-number-variations (CNVs). It was found that a typical human genome differed from the reference genome at a median of 4.31 million base pairs, 84.5% of which consisted of SNPs, the most common form of genetic variants in human genomes [1]. Although occurring at lower frequencies, structural variations contribute significantly to genetic diversity among individuals and populations by causing sequence alternations ranging from several to millions of base pairs.

GVs have been analyzed in genome-wide association studies (GWAS), revealing hundreds of thousands of associations of GVs with complex diseases and traits [2]. SNPs are often employed as genetic markers in the “Common Disease-Common Variant” hypothesis-based GWAS [3–5] with the rationale that common diseases may be expectedly attributable to common genetic variants [6–9]. However, so far, only a small fraction of SNP-disease or SNP-trait associations have been discovered, representing a potential component of the missing heritability regarding complex disorders and traits [7, 10, 11]. Other contributing factors to missing heritability would include the poor detection of rare variants strongly associated with complex diseases or traits, giving rise to the “Common Disease-Rare Variant” hypothesis [8, 12] which has been confirmed by discoveries of disease- or trait-associated rare structural variants such as CNVs [13–15]. Additionally, a “hypothesis-driven” approach is found to be useful in examples such as the identification of autism-associated SNPs where shift of codon usage is postulated to alter protein translation efficiency [16, 17]. Models also have been developed to delineate the effects of demographic history and genetic forces on the patterns and maintenance of genetic variants underlining complex traits [18–20]. Nonetheless, a substantial portion of the heritability remains unexplained, giving rise to the need for a more comprehensive understanding of the distribution, formation mechanisms and functional effects of different kinds of genetic variants as a possible gateway to a further reduction of this unexplained portion.

Previous studies have indicated that genetic variants are not evenly distributed in the human genome [21, 22], with density enhancements giving rise to genetic-variant hotspots in some genomic regions, and identified homologous recombination as one of the major mechanisms for the formation of GV hotspots [23–25]. Since transposable elements constitute highly abundant repeat sequences in the genome, they can serve as homologous templates in recombination events that produce GVs [26–28]. Rapid accumulation of sequence variations has been observed in the body of young Alu elements [29], erasing the sequence similarity between homologous Alu pairs, and reducing thereby their capability to serve as templates for homologous recombination. It follows that the age of transposable elements could be associated with different types and frequencies of GVs. Genetic variants could also originate from mechanisms such as background mutations, non-homology repairs, and replicative errors [30–32]. In addition, the choice of DNA double-strand break repair pathways has been found to be cell cycle-dependant [33, 34], making replication timing an important factor in shaping the distribution of GVs.

Human CNV hotspots have been found to overlap with CNVs in the chimpanzee and macaque genomes, pointing to the maintenance of these hotspots by non-neutral evolutionary forces [24]. Such forces could underlie the differential distribution of different kinds of genomic features and genetic variants among the three types of sequence zones in the human genome, viz. the Genic (gene-rich), Proximal (gene-proximal), and Distal (gene-distal) zones [35]. Accordingly, the present study is directed to establishing a high-resolution comprehensive landscape of genetic-variant hotspots and hotspot clusters in the three types of sequence zones, so that an improved understanding may be obtained regarding their formation mechanisms, and constraint forces, their adaptive or destabilizing roles in the genome, and their possible relevance to the problem of missing heritability.

Materials and methods

Data sources

Germline genetic variants employed in genetic variant hotspots detection include biallelic SNPs (~ 77 million entries) and small indels (SIDs, ~ 3 million) from the Phase III 1000 Genomes Project [1] (accessed 18 Febuary 2015), CNVs (~ 1.6 million) from dbVar database [36] (accessed 18 September 2015), microsatellites (MSTs, ~ 40,000) from the “Microsatellite track” [37], and segmental duplication (SDPs, ~ 50,000) from the “Segmental Dups track” [38] of the UCSC Genome Browser database [39] (accessed 20 April 2015). Rare SNPs (minor allele frequency < 1%) located within the exome pull-down target boundaries are excluded from analysis on account of the probability of the rare alleles discovered given disproportionally high coverage of the exons and their flanking regions in the 1000 Genomes Project. The data sources of the 55 functional and structural genomic features analyzed in the present study are given in the “Data source and track detail” column of

Supplementary Table S1A. Information of the germline genetic variants are available in Supplementary Table S1B. The genomic coordinates of gap regions of human genome assembly hg19 are retrieved from the “Gap” track of the UCSC Genome Browser database. Genomic coordinates of immune system gene loci are retrieved from NCBI Gene database [40].

Size-classifications of CNVs

The length distribution of germline CNVs are separated into four size classes using cut points determined as the critical points of the polynomial regression curve fitted to two successive peaks on the length distribution curve as shown in Supplementary Figure S1. The four size classes comprise the short CNVs (5 < length ≤ 61 base pairs [bp]), medium CNVs (61 < length ≤ 952 bp), long CNVs (952 < length ≤ 15,571 bp), and extra-long CNVs (length > 15,571 bp).

Genomic feature quantitation and tripartite genomic zones

For a feature where the number of base pairs is counted, its level in any region is expressed in “Density,” e.g., for the level of CpG islands in a certain region:

$$\text{Density} = \frac{[\text{Total number of bp in CpG}]}{[\text{Total number of bp in region}]}$$

On the other hand, for a feature that is assessed by a numerical score, its level is expressed in “Intensity,” e.g., for the level of methylation in any genomic region:

$$\text{Intensity} = \frac{\sum[\text{Number of bp in methylation entry}] \times [\text{Score of methylation entry}]}{[\text{Total number of bp in region}]}$$

Various features expressed in “Density” or “Intensity” are listed in “Density/Intensity” column of Supplementary Table S1. Genomic features are classified into the Genic, Proximal, and Distal groups based on their co-localization patterns in 500-kb successive and non-overlapping sequence windows in the twenty-two autosomes (Supplementary Figure S2); based on their feature compositions, the 500-kb windows are partitioned into 45.1% Genic, 31.1% Proximal, and 23.8% Distal-zone windows as described by Ng et al. [35] covering 94.0% of total non-gap sequences on the 22 autosomes. Accordingly, autosomal regions in the present study refer to genomic sequences that have been assigned to the Genic, Proximal, or Distal-zone windows on the 22 autosomes.

Density-based genetic-variant hotspots determined by weighted sliding windows

To increase the resolution of hotspot detection, a sliding-window protocol is employed to scan the genome in 1-kb windows sliding by 10-bp steps. As well, to render more precise the genomic boundary of any hotspot as illustrated in Supplementary Figure S3, the number of genetic variant entries in each of the one hundred 10-bp steps (or, n_{step}) in a window is assigned a weight w_{step} that depends on its distance from the window centre, increasing in equal increments from a w_{step} of 0.5 at the extreme margins of the window to 1.0 at the very centre of the window. This gives rise to a symmetrical distribution of step weights within the window, and the weighted density of any genetic variant in each window (D_{win}) is given by:

$$D_{\text{win}} = \sum_{\text{step}=1}^{100} w_{\text{step}} \times n_{\text{step}} / \sum_{\text{step}=1}^{100} w_{\text{step}}$$

The weighted density of the targeted kind of genetic variants is calculated in all 1-kb windows. In the Genic, Proximal, or Distal zones, the top-density windows that cover altogether 5% of the total entries of the targeted genetic variant are identified as hotspots. For example, the hotspot detection procedure for SNPs in the Genic zones (see flowchart in Supplementary Figure S4) consists of (i) each autosome of the human genome is divided into successive 1-kb sliding-windows in 10-bp steps; (ii) D_{win} for SNP entries is determined for all the sliding windows; (iii) those sliding windows that overlap with a 500-kb Genic zone are separated into Part I sliding windows that are entirely located within a Genic zone, and Part II sliding windows that are partially located within a Genic zone; (iv) Part I and Part II windows are ranked separately based on their D_{win} values from the highest to the lowest. The Part I sliding windows with top-ranked D_{win} values are recruited successively from rank 1 down to rank i until their cumulative SNP fraction represents ≤ 5% of total SNP entries in the Genic zones to yield the top SNP windows in Part I (with ranks of 1, 2, 3, … i, red-numbered in the flowchart). The top-ranked Part II sliding windows are recruited from rank 1 down to rank *k* (with ranks of 1, 2, 3, … *k*, red-numbered in flowchart) such that the rank-*k* window displays a weighted SNP density (D_k) greater than or equal to the D_{min} (viz. lowest D_{win}) among the top-ranked SNP windows in Part I. Steps iii and iv are also repeated for the Proximal and Distal zones as described for the Genic zones. (v) Thereupon, Part I and Part II top-ranked windows from the three types of zones are merged together, followed by elimination of any double-counting of windows on account of their zone-crossing to yield altogether a total of 34,487 SNP hotspots in the three types of zones with a mean width of 1994 ± 4528 bp, covering 2.54% of the autosomal regions analyzed. The zone-specific hotspot detection thresholds (D_{min}), and the chromosomal coordinates of different kinds of genetic variant hotspots, are given in Supplementary Table S2 and Supplementary Dataset S1.

Identification of hotspot clusters

When two or more of different kinds of genetic-variant hotspots overlap with one another, the genomic regions occupied by these hotspots are merged into a hotspot cluster. This procedure identifies 1135 clusters in the genome consisting of 2 to 3 kinds of hotspots each, amounting to 0.20% of autosomal sequences and comprising twenty-three kinds of hotspot-compositions, the chromosomal coordinates of which are given in Supplementary Dataset S1.

Determination of natural selection hotspots

To identify positive selection hotspots (PosSel-Hs), the strength of positive selection is evaluated in 1-kb windows in the non-gap regions on 22 autosomes based on (i) average derived-allele frequency (DAF) of all SNPs in each window, (ii) average of the maximum DAF differences (|ΔDAF|) across the two different populations, and (iii) the average of haplotype structure-based statistic $|nS_L|$ [41]. Next, the top-5% windows ranked according to each of these three kinds of positive-selection levels are earmarked as candidate windows under positive selection. All 1-kb windows in the autosomes are classified into ten ranking groups based on their number of informative sites (viz. the SNPs with DAF or nSL statistics). For each candidate window, its level of positive selection is compared with the estimated levels in 10,000 random windows simulated from the corresponding ranking group. The matching of informative site density would accommodate the higher variance in site-based measurement in the windows with low numbers of informative sites. Only windows with significantly higher measurement of positive selection ($p < 0.05$) relative to the random simulations are kept, and successive windows are merged to yield the positive selection hotspots. In this regard, 99,266 DAF hotspots, 89,464 |ΔDAF| hotspots, and 63,060 $|nS_L|$ hotspots are identified, which altogether cover 10.33% of autosomal regions. Likewise, the negative selection hotspots (NegSel-Hs) are identified based on two different measures of purifying selection, viz. (i) sequence conservation score measured using phyloP across 100 species, and (ii) intensity of nucleotide diversity. The windows with top-5% phyloP score or bottom-5% nucleotide diversity are earmarked, and only the windows displaying significantly higher measurements of purifying selection ($p < 0.05$) relative to 10,000 simulation windows of similar levels of informative sites are regarded as NegSel-Hs, which altogether amount to 7.94% of autosomal sequences. The phyloP score is retrieved from the "Conservation" track of UCSC Table Browser [42]. The nucleotide diversity in any window is estimated using VCFtools version 0.1.15 with the "--window-pi" flag [43]. The $|nS_L|$ value for each SNP is the absolute value of the nS_L statistic estimated and normalized across 100 frequency bins in selscan version 1.2.0 with all the default settings [41, 44]. The DAF, |ΔDAF|, $|nS_L|$, and nucleotide diversity used for selection-hotspot determination are based on SNPs and haplotypes from the Phase III 1000 Genomes Project of 2504 individuals. Genomic windows subject to balancing selection retrieved from the supplementary material of Bitarello et al. [45] are merged across African and European populations at different target frequency values, yielding 10, 275 balancing selection hotspots (BalSel-Hs) amounting to 1.71% bp of the autosomal regions. Summaries of selection hotspots are available in Supplementary Table S3A. The chromosomal coordinates and p values of each of the PosSel-Hs and the NegSel-Hs identified are given in Supplementary Dataset S1.

Replication-time segments

Previously, human DNA replication timing has been experimentally assessed in 1-kb sequence windows across the genomes of fifteen different human cell lines by ENCODE project [46, 47], which reveals sequence windows displaying, to different extents, conservation in replication timing across different cell lines. Therefore, it is possible to generate a consent replication-timing profile for the human genome with reference to the fifteen cell lines retrieved from the "UW Repli-seq track" of UCSC Table Browser (viz. GM12878, K562, HeLa-S3, HepG2, HUVEC, IMR90, MCF-7, SK-N-SH, BG02ES, BJ, GM06990, GM12801, GM12812, GM12813, and NHEK). Based on the DNA replication times of 1-kb sequence windows in fifteen cell lines, the human genome is classified into six types of replication-time segments in the present study as follows: (i) for each cell line, record the replication phase(s) that earns the highest score among the six replication phases (G1b, S1, S2, S3, S4, and G2 phases) within every 1-kb window; (ii) for any window, the replication phase that earns the highest score most frequently from the 15 cell lines is chosen as the representative replication phase for all the 15 cell lines. Only windows whose replication time has been assessed in more than eight out of the fifteen cell lines are subject to analysis, yielding 325,635 G1b, 389,304 S1, 395,223 S2, 454,554 S3, 392,600 S4, and 512,129 G2 1-kb windows that fall into the Genic, Proximal, or Distal zonal windows covering altogether 85.7% of the twenty-two autosomes. Windows yielding more than one representative replication phases are not included in the analysis. Successive windows that share the same representative replication phase are merged to yield segments of varied lengths, and the chromosomal coordinates of the six types of replication-time segments are given in Supplementary Dataset S1.

Somatic CNV breakpoints in cluster-containing genes

Coordinates of 23,056 known genes are matched with the locations of hotspot clusters to yield 448 genes that overlap with one or more clusters, and 442 of them display one or more CNVT breakpoint(s). For each of these 442 genes, the cluster-segments and noncluster-segments are compared with respect to their CNVT breakpoint densities, identifying 33 genes where the cluster-segments are significantly enriched in CNVT breakpoints (chi-square test, Bonferroni-corrected), and only one gene where the cluster-segment is significantly depleted in CNVT breakpoints as shown in Supplementary Table S4. In the chi-square tests, the expected number of CNVTs in the cluster-segment of the gene is calculated by multiplying the number of total CNVTs in the gene and the fraction of base pairs covered by the cluster-segments in the gene. Gene coordinates were retrieved from the R package “TxDb.Hsapiens.UCSC.hg19.knownGene” version 3.2.2 using “genes” function in “GenomicFeatures” package version 1.26.4 [48].

Statistical analysis

Empirical p values based on Monte Carlo simulations

In each round of simulation, the locations of target regions are randomly shuffled (by matching the number and sizes of the target regions) within the autosomal regions analyzed. The measurement of interest is assessed for each simulation round. The p value is obtained as $(r + 1)/(n + 1)$ [49], where n is the round of simulations and r is the number of simulations that produce a measurement greater than or equal to the actual level in the case of significant enrichment, or less than or equal to the actual level in the case of significant depletion. Unless otherwise specified, 5,000,000 rounds of simulations are performed.

Enrichment of GWAS-identified SNPs in hotspots and clusters

To estimate using the Monte Carlo method the level of significance regarding the enrichment of GWAS-identified SNPs in the genetic-variant hotspots and hotspot clusters, random regions are simulated in sequence windows with matching levels of average minor-allele frequencies (MAF) in order to accommodate the dependency of GWAS on allele frequency. For this purpose, autosomal sequence windows are classified into ten equal groups according to their individual average MAF values, and the hotspots or clusters are also classified into the same ten groups. The Monte Carlo simulations are conducted by simulating sequence windows from the corresponding MAF group of hotspot or cluster. The simulation results for each group of hotspots and clusters are shown in Supplementary Figure S5.

Comparison of population differentiation between hotspot clusters and non-hotspot regions

To compare the positive-selection strength estimated by population differentiation among the cluster and non-hotspot (viz. the genomic regions outside of hotspots and clusters) regions, the DAF difference between any two populations, viz. $|\Delta\mathrm{DAF}|^{POP1\text{-}POP2}$, is estimated for all the clusters and non-hotspot regions in each of the ten population pairs among African, American, South Asian, European, and East Asian. Let $DAF^{POP}_{\mathrm{Region}}$ represents the average DAF of all SNPs in a specific cluster or non-hotspot region in a population. The $|\Delta\mathrm{DAF}|$ between any two populations, for example African and European, viz. $|\Delta\mathrm{DAF}|^{\mathrm{African\text{-}European}}$, pertaining to the i^{th} cluster and j^{th} non-hotspot region is given by $|\Delta DAF|^{\mathrm{African\ -\ European}}_{\mathrm{Cluster}_i} = |DAF^{\mathrm{African}}_{\mathrm{Cluster}_i} - DAF^{\mathrm{European}}_{\mathrm{Cluster}_i}|$ and $|\Delta DAF|^{\mathrm{African\ -\ European}}_{\mathrm{nonhotspot}_j} = |DAF^{\mathrm{African}}_{\mathrm{nonhotspot}_j} - DAF^{\mathrm{European}}_{\mathrm{nonhotspot}_j}|$. Thereupon, in the case of $\boldsymbol{n}$ total clusters and $\boldsymbol{m}$ total non-hotspot regions, the $\boldsymbol{n}$ $|\Delta DAF|^{\mathrm{African\ -\ European}}_{\mathrm{Cluster}}$ values are compared with the $\boldsymbol{m}$ $|\Delta DAF|^{\mathrm{African\ -\ European}}_{\mathrm{nonhotspot}}$ values by unpaired one-tailed t tests. The procedure is likewise repeated for the other nine population pairs.

Software employed in data processing and visualization

Custom R scripts are employed in hotspot and cluster detections and analyses under R environment version 3.3. Disease- or trait-associated SNPs identified by GWAS at $p < 10^{-8}$ are lifted over from GRCh38.p5 to hg19 using the “liftOver” function in R package “rtracklayer” version 1.34.2 [50]. Processing and quantification of genetic variants and genomic features are conducted using R package “GenomicRanges” version 1.26.4 [48]. Figures are drawn using Cytoscape version 3.6.1 [51, 52], as well as the R packages “ggplot2” version 3.0.0 [53], and “quantsmooth” version 1.40.0.

Results

Density-based genetic variant hotspots and clusters

The germline genetic variants retrieved from different public databases are analyzed including SNPs, SIDs, MSTs, CNVs, as well as SDPs which originated from fixed CNVs [54] (see “Data source” in “Materials and methods” section). The three cuts on the length distribution of germline CNVs in Fig. 1a separate the CNVs into four size classes comprising the short CNVs (SCNVs), medium CNVs (MCNVs), long CNVs (LCNVs), and extra-long CNVs (ECNVs). Notably, the presence of all four different size-classes of CNVs is evident in all three types of genomic zones (Fig. 1a, panels 2–4).

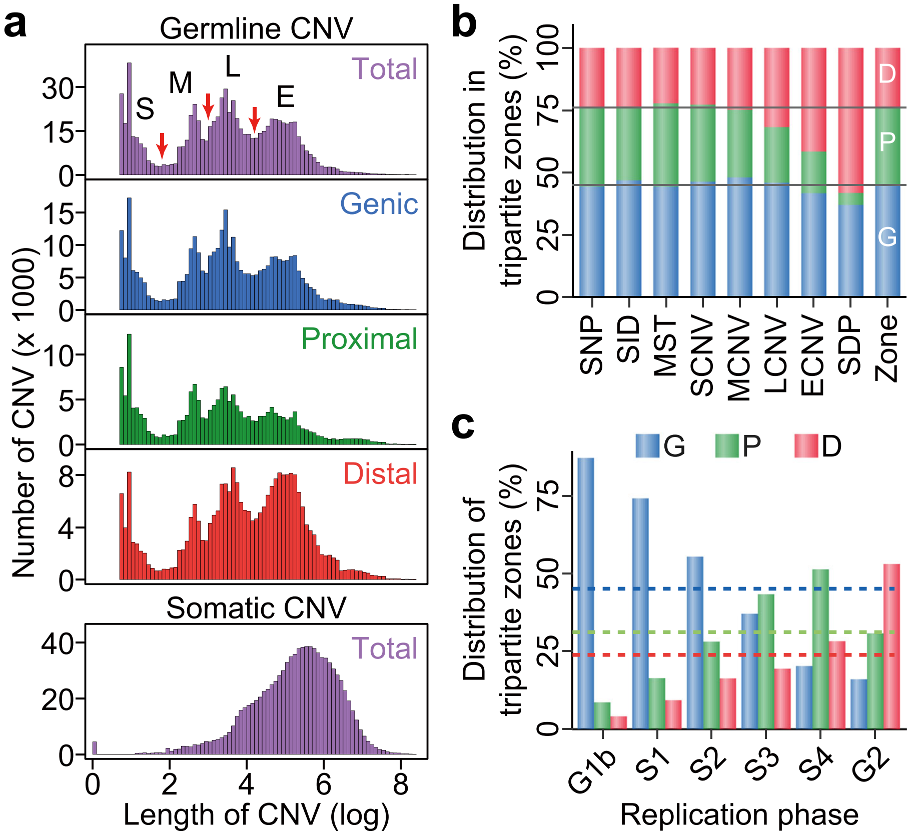

Fig. 1 Genetic variants and replication phases in different genomic zones. **a** Length distributions of CNVs. Upper panel: distribution of germline CNVs in total and in the Genic, Proximal, and Distal genomic zones. Germline CNVs are separated into short (S), medium (M), long (L), and extra-long (E) CNVs based on the three cuts indicated by the red arrows (see "Size-classifications of CNVs" in "Methods" section). Lower panel: distribution of total somatic CNVs in tumors from COSMIC database. **b** Percentile distributions of eight kinds of genetic variants in the Genic (blue), Proximal (green), and Distal (red) zones. The last "Zone" column expresses the percentages of base pairs in the Genic, Proximal and Distal zones amounting to 45.1%, 31.1%, and 23.8% respectively of the total base pairs of the human genome. **c** Percentages of Genic-, Proximal-, and Distal-zone sequences in six different phases of DNA replication ranging from the earliest-replicating G1b to the latest-replicating G2 phases. Colored dashed lines indicate the 45.1% Genic-, 31.1% Proximal-, and 23.8% Distal-zone sequences of the human genome. G, Genic; P, Proximal; D, Distal

The three types of genomic zones display distinct biases in terms of their constituent GVs. As shown in Fig. 1b, the longer GVs, including the breakpoints of LCNVs, ECNVs, and SDPs are prominently enriched in the Distal zones largely at the expense of the Proximal zones. In contrast, the shorter GVs including SIDs, SCNVs, and MCNVs are somewhat enriched in the Genic zones. MSTs are enriched slightly in the Proximal zones, whereas SNPs are evenly distributed over the three zones. Because the average frequencies of genetic variants vary between the Genic, Proximal, and Distal zones (Fig. 1b), the identification of genetic-variant hotspots is performed separately for the three types of zones. To enhance resolution and obtain more precise boundary of genetic-variant hotspots, a weighted sliding-window algorithm with windows of 1000 bp in steps of 10 bp is employed. For each type of genomic zones, the weighted density of any targeted genetic variant is measured in the sliding windows, and all the windows are ranked based on their densities from high to low. The top-density windows that cover altogether up to 5% of the total entries of the targeted genetic variant are identified as hotspots of that kind of genetic variant (see "Density-based genetic-variant hotspots determined by weighted sliding windows" in "Materials and methods" section). Altogether 44,379 hotspots with an average size of 1882 bp are identified in the three genomic zones, equivalent to 3.09% of the total base pairs in the twenty-two autosomes (Fig. 2a and Supplementary Table S2). The enrichments of different kinds of GVs in their respective hotspots relative to simulated genomic regions are significant with p values ranging from 4×10^{-4} to 0 (Supplementary Figure S6). In general, the constitutive GVs in the hotspots display higher allele frequencies than genomic average except in the case of the SID hotspots (Supplementary Figure S7). On the basis that the

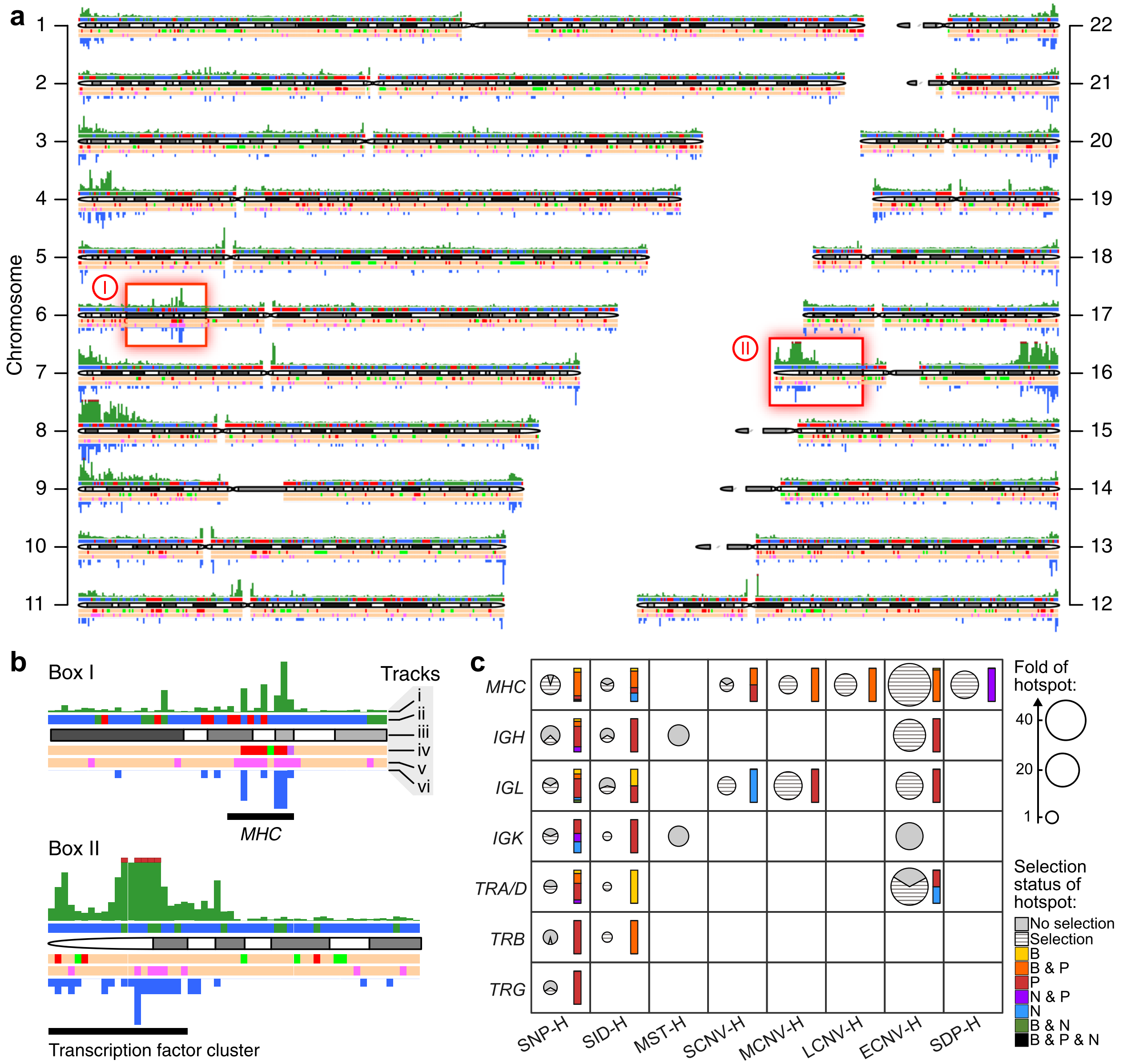

Fig. 2 Landscape of genetic-variant hotspots and hotspot clusters. **a** Tracks on chromosome 1-22 show from top to bottom, in successive and non-overlapping 500-kb windows: (i) coverage of genetic-variant hotspots in window (green bar) capped at 50% by the dark red rectangles in chromosomes 8, 12, and 16; (ii) Genic zones are colored in blue, Proximal in green, and Distal in red; (iii) ideogram; (iv) sequence windows displaying top coverage (top 10%) of positive selection hotspots (red), purifying selection hotspots (green), and both positive and purifying selections (purple); (v) sequence windows displaying top 10% coverage of balancing selection hotspots (pink); (vi) number of hotspots clusters (height of downward blue bars). Beige windows indicate where neither positive nor purifying selection hotspots reach top 10% level in track (iv), or where balancing selection hotspots do not reach top 10% level in track (v). **b** Tracks (i)–(vi) enlarged for red boxes I and II in (**a**) for the locations of *MHC* on chromosome 6, and a cluster of transcription factors on chromosome 16 [55], respectively. **c** Enrichment and selection-status of eight kinds of genetic-variant hotspots (*x*-axis) in immune system gene loci (*y*-axis). Fold of hotspot density inside each immunoprotein gene locus relative to autosomal average is represented by the area of the corresponding pie chart: hashed slices indicating proportion of GV hotspots overlapping, and gray slices indicating non-overlapping, with selection hotspots. Colored bars beside each pie chart show the relative abundance of different kinds of selection hotspots acting on these GV hotspots: "B" stands for GV hotspots overlapping with balancing selection hotspots, "P" with positive selection hotspots, "N" with negative selection hotspots. See Supplementary Table S9 for numerical data. *MHC*, major histocompatibility complex; *IGH*, *IGK*, and *IGL*, immunoglobulin heavy, kappa and lambda loci respectively; *TRA*, *TRB*, *TRD*, and *TRG* are the T cell receptor alpha, beta, delta and gamma loci. Since *TRD* is embedded in *TRA*, they are merged into a *TRA/D* locus. "H" stands for "hotspots"

convergence of two or more kinds of hotspots produces a hotspot cluster (see "Identification of hotspot clusters" in "Materials and methods" section), there are 1135 clusters with an average size of 4816 bp, equivalent to 0.20% of the total base pairs (Fig. 2a and Supplementary Table S2).

Formation mechanisms for hotspots and clusters

Contribution of recombination and retrotransposons

The association of genetic variants with homologous recombination is confirmed by the positive correlation between SNP density and recombination intensity (Fig. 3a), as well as the enhancement of recombination intensity in all GV-containing windows except for the SDP-containing ones (Fig. 3b). It has been proposed that retrotransposons invaded the human genome in successive waves at different times [56], and the ages of some retrotransposon subfamilies have been estimated by previous studies (Supplementary Table S5). Since the SVA, and likewise the Alu, subfamilies of similar ages show similar distribution curves in regions flanking various genetic variants (Fig. 4a and Supplementary Figure S8), the SVA subfamilies are grouped together according to age to yield the SVAef (viz. SVA_E and SVA_F), SVAcd (SVA_C and SVA_D), and SVAab (SVA_A and SVA_B) subgroups; and the AluY subfamilies are grouped together to yield the "very young" AluYvy (AluYa5, AluYb8, AluYb9, AluYg6, AluYf4, AluYd8, AluYa8, AluYk11, AluYh9, and AluYk12), and the "young" AluYy (AluY, AluYc, AluYc3, AluYk4, and AluYf5) subgroups. LINE1 elements are likewise combined into the "very young" L1vy, "young" L1y, "middle aged" L1m, and "old" L1o subgroups (Supplementary Table S5). The oldest SVAab, AluJ, and L1o subgroups all show greater enrichment of SCNVs but not MCNVs, LCNVs, or ECNVs in their vicinities, which is not the case with the youngest SVAef, AluYvy, and L1vy subgroups (Fig. 4b), suggesting that the age of these retrotransposons constitutes a significant determinant of some of the retrotransposon-GV associations. Such age effects conform to the "long-with-young" (p = 0.018 between younger and older retrotransposons) but "short-with-all" (p = 0.263) modes of association between GV length and retrotransposon age (Fig. 4c). A plausible explanation is that, because the younger retrotransposons (shown in red and orange in Fig. 4b) are less numerous than the older ones (shown in blue), they are more sparsely distributed in the genome. As a result, recombinations between pairs of the more sparsely distributed young

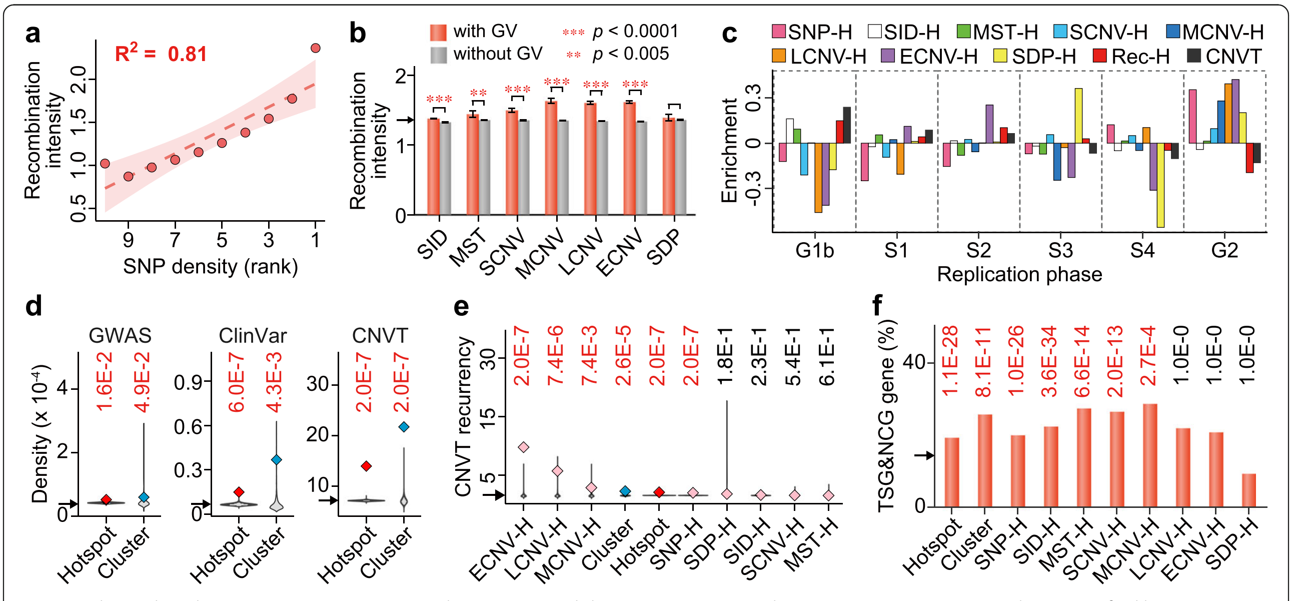

Fig. 3 Relationships between genetic variants and genome instability. **a** Average recombination intensity in 10 equal groups of 1-kb sequence windows ranked according to SNP density from low (left) to high (right). R^2 is the coefficient of determination. **b** Higher recombination intensity in 1-kb windows with GVs compared to windows without GVs (unpaired one-tailed t tests, Bonferroni corrected). **c** Enrichment of eight kinds of genetic-variant hotspots, recombination hotspots (Rec-Hs) and somatic CNVTs in the six DNA replication phases, expressed as density fold-changes relative to autosomal average. **d** Densities of GWAS-identified SNPs (corrected for allele-frequency dependency, see "Statistical analysis" in "Methods" section), breakpoints of variants in ClinVar database, and somatic CNVT breakpoints in total hotspots and total clusters relative to randomly simulated genomic regions represented by the gray violin-plots. All p values smaller than 0.05 are shown in red. **e** Recurrency of somatic CNVT breakpoints in eight kinds of hotspots, in "Hotspot" representing all 44,379 genetic-variant hotspots in the genome, or in "Cluster" representing all 1135 hotspot clusters in the genome, with their p values estimated by the Monte Carlo method. **f** Percentage of hotspot- or cluster-containing genes found in Tumor Suppressor Gene Database (TSG) and Network of Cancer Genes (NCG). The p values estimated using chi-square tests are shown above each bar (Bonferroni corrected). In (**b**, **d** **e**, **f**), arrow on y-axis indicates the autosomal average. Error bars in (**b**), and shaded bands around the curve in (**a**) indicate 95% confidence intervals. "H" stands for "hotspots"

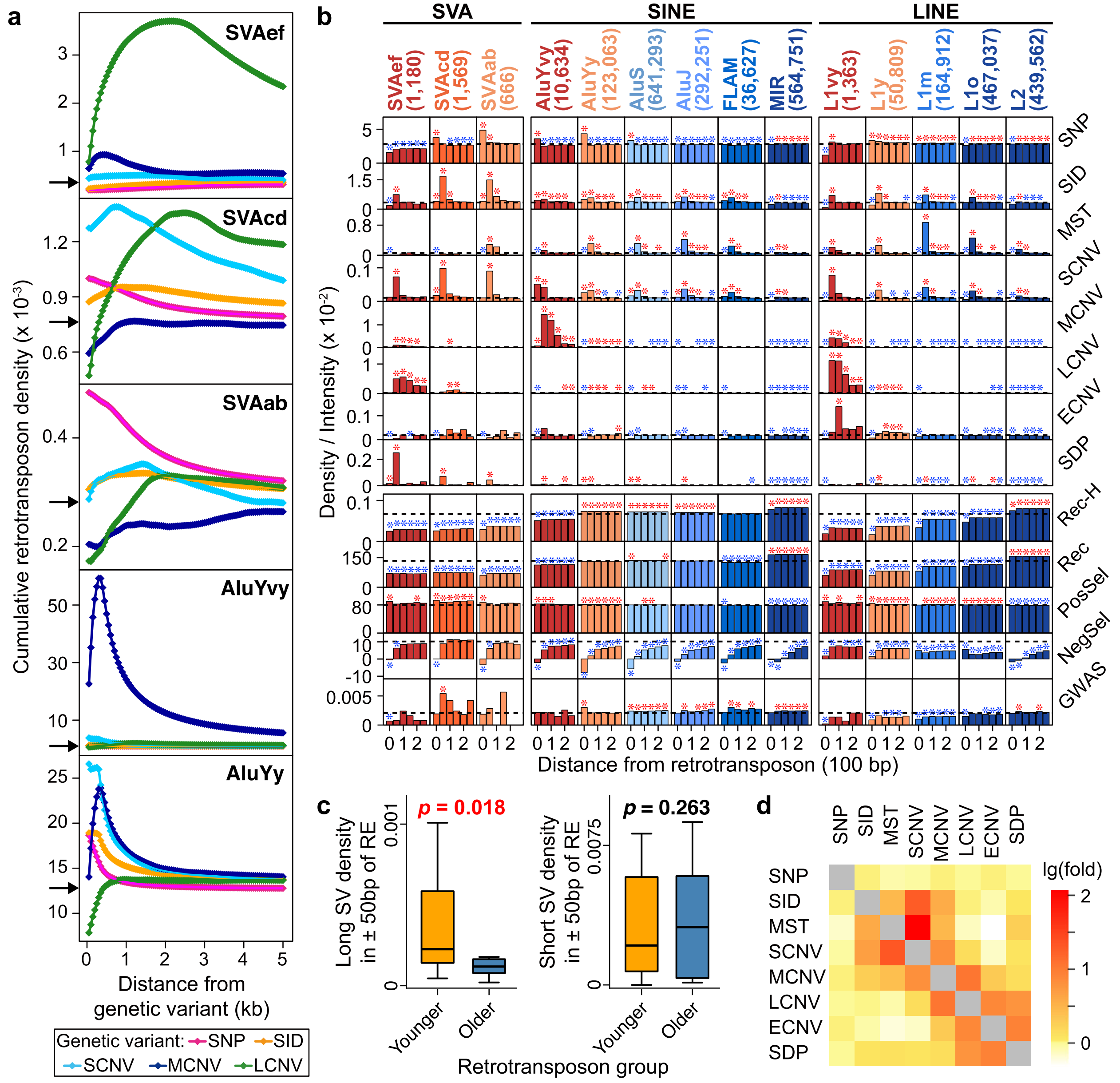

Fig. 4 Relationships between genetic variants and retrotransposons. **a** Density distributions of different retrotransposon groups within (viz. at distance = 0 bp) or near SNPs, SIDs, and breakpoints of SCNVs, MCNVs, and LCNVs. Arrows on *y*-axes indicate autosomal averages of the retrotransposon group. **b** Distributions of genetic variants, recombination hotspots (Rec-H), recombination intensity (Rec), positive selection (PosSel) indicated by average |nSL| values, negative selection (NegSel) indicated by phyloP scores, as well as GWAS-identified SNPs both inside (at distance = 0 bp) and within ± 250 bp of different groups of SVAs, short interspersed nuclear elements (SINE), and long interspersed nuclear elements (LINE). Designations and numbers of the retrotransposon groups indicated at the top of the columns are color-coded based on their relative ages, ranging from red for the youngest to dark blue for the oldest. Colored asterisks mark significant enrichments (red) or depletions (blue) of features based on Monte Carlo simulations (n = 1000; *p < 0.005), and the dashed horizontal lines indicate the respective autosomal averages of each *y*-axis feature. **c** Density of long (MCNV, LCNV, ECNV, SDP) in the left panel, or short (SID, MST, SCNV) structural variations in the right panel, in vicinities of the younger (SVAs, AluYs, L1vy, L1y) or older (AluS, AluJ, FLAM, MIR, L1m, L1o, L2) retrotransposon groups. SV stands for structural variation; RE for retrotransposon; p value < 0.05 is shown in red (unpaired one-tailed t-tests). **d** Enrichment of *y*-axis genetic variants within ± 50 bp of *x*-axis genetic variants expressed by the thermal scale representing the natural-log of the density fold-change relative to the autosomal average

retrotransposons would generate larger structural variations.

As a consequence of the length-age association between genetic variants and retrotransposons, the short variants (SID, MST, SCNV) form a closely co-localized group in Fig. 4d, and the longer variants (MCNV, LCNV, ECNV, and SDP) form a different closely co-localized group. However, although the short and long genetic variants are both associated with the young SVAef, AluYvy, and L1vy in Fig. 4b, the white and light-yellow squares in Fig. 4d indicate that they do not show significant co-localization with one another, which suggests that other mechanisms besides retrotransposon-mediated recombination could also produce genetic variants, and natural selection may also influence the distribution of the genetic variants subsequent to their production.

Effects of natural selection on distribution of genetic variants

In general, the level of positive selection can be assessed based on (i) the allele-frequency spectrum, (ii) population differentiation, and (iii) haplotype structure [57]. First, using the shift of DAF to extreme values as an index of positive selection [58, 59], an elevated percentage of SNPs with DAF > 0.95, points to the enhancement of positive selection in GV hotspots and even more so in hotspot clusters relative to non-hotspot regions (Fig. 5a) throughout the African, American, South Asian, European, and East Asian populations ($p < 10^{-10}$ for clusters in all five populations). Secondly, using population differentiation as an index, the increased values of DAF-difference between population pairs, viz. |ΔDAF| in clusters relative to non-hotspot regions, yielded p values ranging from 10^{-2} down to less than 10^{-13} for eight out of ten population pairs (Fig. 5b), attesting to substantial inter-population differentiation driven by directional selection within hotspot clusters during the evolution of human populations. Thirdly, using haplotype structure-based $|nS_L|$ statistic as an index for both soft and hard positive-selection sweeps [41], the significantly enhanced $|nS_L|$ scores in hotspots and even more so in clusters relative to non-hotspot

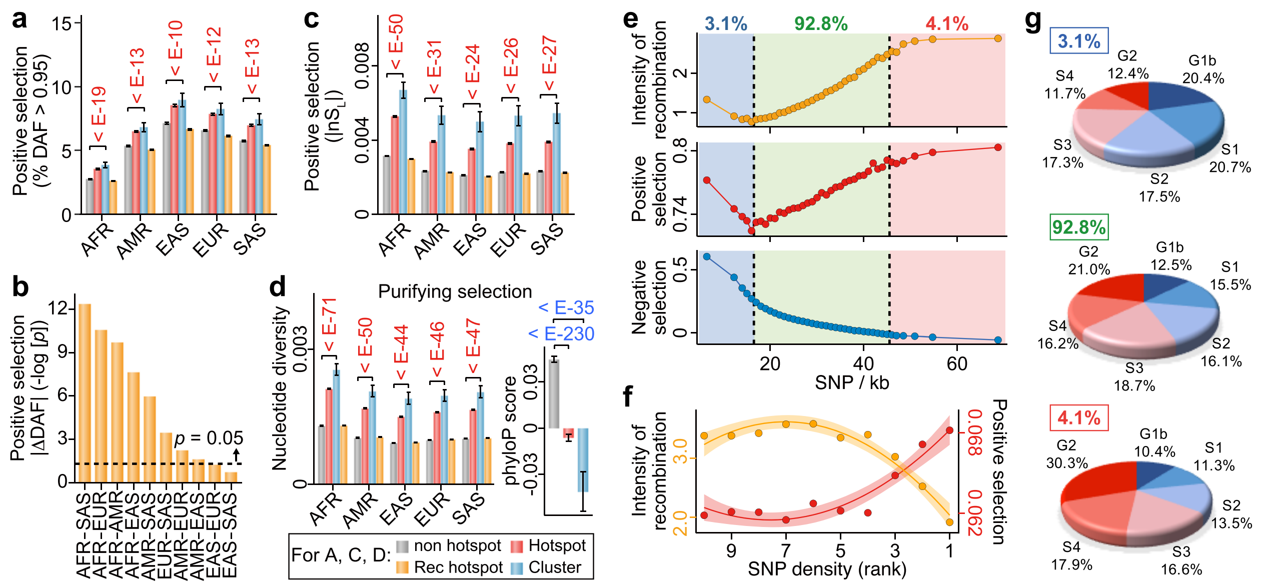

Fig. 5 Relationships of genetic-variant density with recombination, natural selection and replication timing. **a** Strength of positive selection indicated by percentage of SNPs displaying DAF > 0.95 in the African (AFR), American (AMR), South Asian (SAS), European (EUR), and East Asian (EAS) populations within non-hotspot regions, total hotspots, total clusters, and Rec-Hs. **b** Enhancement of positive selection in clusters over non-hotspot regions in terms of |ΔDAF| between ethnic population pairs expressed by negative $\log_{10}$-transformed p values from unpaired one-tailed t tests (see “Statistical analysis” in “Methods” section). **c** Enhanced positive selection in clusters and hotspots relative to the non-hotspot regions indicated by elevated $|nS_L|$ values in five populations. **d** Excess of nucleotide diversity (left panel) or reduction of phyloP score (right panel) in hotspots and clusters. In (**a**, **c**, **d**), bars share the same color-codes, and p values (from unpaired one-tailed t tests) above the bars are shown in red or in blue indicating significantly higher or lower selection in hotspots/clusters relative to the non-hotspot regions. Since no population-specific phyloP data is available, right panel in (**d**) compares the non-hotspot, hotspot, and cluster regions of the human reference genome hg19. **e** Variations of recombination intensity, positive selection (in terms of average $|nS_L|$), and purifying selection (in terms of phyloP score) in 1-kb window groups with low (3.1% of total, blue region), medium (92.8%, green region) and high (4.1%, pink region) SNP densities. **f** Variation of recombination intensity (orange curve) and positive selection in terms of DAF (red curve) among ten groups of SNP hotspots ranked by SNP density from low (rank 10) to high (rank 1). **g** Distribution of the low-, medium-, and high-density windows from part (**e**) in the different DNA replication phases. Error bars in (**a**, **c**, **d**) and shaded bands around the curves in (**f**) indicate 95% confidence intervals

regions (Fig. 5c, $p < 10^{-24}$ for clusters in all five populations) are likewise indicative of the presence of positive selection. A total of 26.4% of GV hotspots and 36.0% of the clusters overlap with sequence windows displaying top-5% levels in one or more of the three different measures of positive selection based on DAF, |ΔDAF|, and $|nS_L|$ (Supplementary Table S3B), significantly more than the simulated regions (p = 0.0001 for both hotspots and clusters, 10,000 simulations). These three different measures of positive selection jointly enable the identification of 200,211 hotspots of positive selection based on the criterion that hotspot sequence windows display top-5% levels in at least one of the three measures (see "Determination of natural selection hotspots" in "Materials and methods" section).

Purifying selection can be assessed based on (i) reduction of genetic diversity, and (ii) cross-species sequence conservation [60]. Significant elevation of genetic diversity, measured by nucleotide diversity [61], is evident in both hotspots and clusters in all five ethnic populations (Fig. 5d left panel) with $p < 10^{-44}$ in all populations for the clusters. Cross-species conservation, measured by sequence conservation score measured using phyloP across 100 species [42], is significantly reduced in both hotspots and clusters relative to non-hotspot regions (Fig. 5d right panel) with $p < 10^{-230}$ for hotspots and 10^{-35} for clusters. Altogether 8.83% of GV hotspots, and 11.1% of the clusters, overlap with sequence windows that display either top-5% levels of phyloP score or bottom 5% of nucleotide diversity, which is significantly less frequent in comparison to simulated regions (p = 0 for both hotspots and clusters, 10,000 simulations). A total of 165,323 purifying or negative selection hotspots are identified as sequence windows that display at once top-5% levels of phyloP score and bottom-5% levels of nucleotide diversity.

It has been proposed that balancing selection represents as an essential adaptive force acting on structural variation hotspots harboring genes involved in immune functions or anthropologically crucial functions [24, 25, 35]. This is confirmed by the present findings that 4.7% of the GV hotspots and 6.0% of the hotspot clusters overlap with the hotspots of balancing selection in the human genome (Supplementary Table S3B). The importance of balancing selection in shaping the distributions of hotspots and clusters is supported further by the significantly higher overlap of balancing selection hotspots with GV hotspots and the clusters relative to simulated non-hotspot regions (p = 0.0001 for both GV hotspots and clusters, 10,000 simulations). The findings that all three different measures of positive selection support the association of genetic-variant hotspots and clusters with positive selection (Fig. 5a–c), and both of the measures of purifying selection reveal a reduction of purifying selection in hotspots and clusters relative to the non-hotspot regions (Fig. 5d), therefore indicate that positive selection and balancing selection are the dominant selection forces acting on genetic-variant hotspots and clusters.

Co-saturation of recombination and selection in genetic variant-enriched regions

The levels of recombination and natural selection are examined at different levels of GV densities in Fig. 5e. In 96.5% of the 1-kb sequence windows in the genome (light green region, Fig. 5e), largely parallel variations of recombination intensity (orange curve) and positive selection (red curve) with SNP density are observed, both running expectedly opposite to the variation of negative selection (blue curve). In the windows with top 2.0% SNP densities (pink region), both the recombination and positive selection curves flatten as they approach saturation. This points to the possible underestimation of recombination rate in the presence of strong positive selection and vice versa. Mechanistically, excessive sequence shuffling by recombination events would limit the effectiveness of positive selection in terms of the attainable DAF, leading to the low percentage of SNPs with high DAF as well as low $|nS_L|$ in over 30,000 previously determined recombination hotspots (Rec-Hs) in the human genome [62, 63] of all the ethnic groups (orange bars, Fig. 5a, c). On the other hand, strong positive selection could eliminate some of the unfit recombinant genotypes, thereby diminishing the footprints of the recombinations and causing an underestimation of their effects. On this basis, the densities of genetic variants, being independent of allele or haplotype frequencies, could represent a more robust measure of genetic diversity than frequency-dependent measures where selection and recombination are highly active. Interestingly, in the SNP hotspots that are 1st or 2nd-ranked in terms of SNP density (Fig. 5f), recombination decreases yet the level of positive selection remains elevated, suggesting that although recombination tends to be underestimated owing to the presence of positive selection, the positive selection is too strong to be suppressed by recombination. Among different GV hotspots, those of SNP, ECNV, and LCNV show the largest percentage of overlap with positive selection hotspots (Supplementary Table S3B), pointing to the presence of exceptional positive selection within these three kinds of hotspots.

In the windows with the lowest-1.5% SNP densities (light blue region, Fig. 5e), low SNP is accompanied by high recombination intensity and positive selection. This finds a possible explanation in the top pie chart of Fig. 5g, which shows that over 50% of these lowest-1.5% SNP-density windows are in fact replicated in the G1b-S2 phases where the fidelity of DNA replication is high [64–

66], suppressing thereby the occurrence of SNPs (see "Replication-time segments" in "Materials and methods" section for the classifications of sequence windows into six replication phases).

Impacts of hotspots and clusters on function and disease

Idiosyncratic association with functional genomic features

It has been observed previously that Alu insertions of different ages enhance SNP occurrences in their neighborhoods differentially [29], suggesting that genomic feature-genetic variant associations vary with the nature of the genetic variant or the genomic feature or both. Accordingly, the co-localizations of eight kinds of GV hotspots with a wide spectrum of genomic features are analyzed in Fig. 6. Some of the co-localizations turn out to be highly idiosyncratic, exemplified by the strong associations of the DNA methylation features with the SNP hotspots; or the large intergenic non-coding RNAs (LINC) with ECNV and SDP hotspots; and histone modification features with MCNV hotspots.

The dissimilar properties of the four size-classes of CNVs are confirmed by the strong associations of histone-modification features such as H3k9me3 with the hotspots of MCNV and SCNV but not those of LCNV and ECNV; the much stronger associations of H2az, H3k79me2, and open chromatin elements (DNase) with

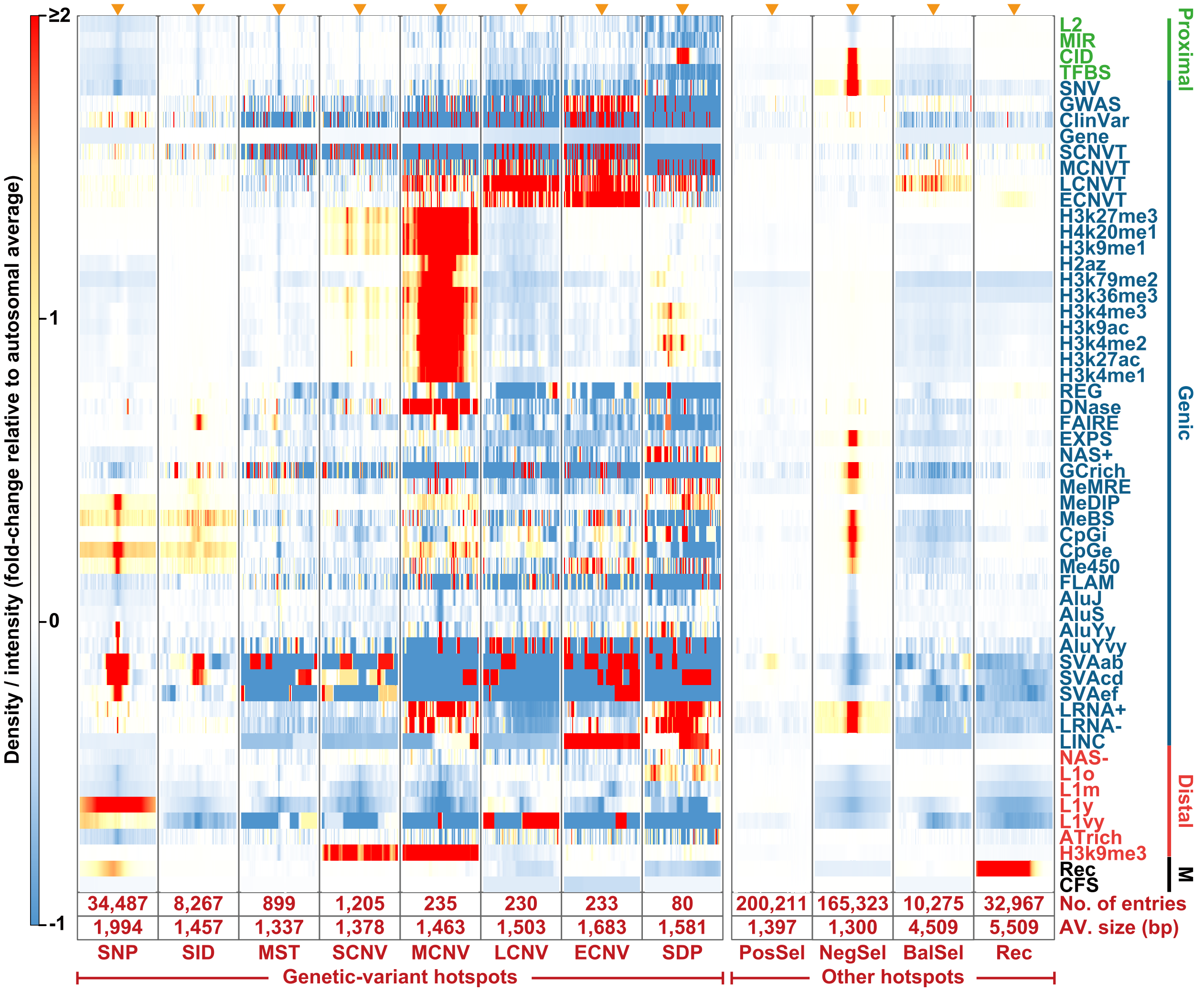

Fig. 6 Genomic features inside and near genetic-variant, selection and recombination hotspots. The density or intensity of 55 types of genomic features occurring between 4-kb upstream (left end) to 4-kb downstream (right end) from the centre (indicated by orange arrowhead at the top of each column) of different kinds of hotspots, expressed in fold-changes relative to the respective autosomal average in accordance to the red-blue thermal scale. The number and average size of different kinds of hotspots are shown in the bottom two rows of the figure. The genomic features indicated on the *y*-axis are grouped into Proximal-zone (green), Genic-zone (blue), and Distal-zone (red) features based on their co-localization relationships (Supplementary Figure S2) [35], along with Marker (M) features (black). The descriptions and abbreviations of different genomic features are given in Supplementary Table S1A

hotspots of MCNV relative to those of SCNV; and the strong association of LINC with the hotspots of ECNV relative to those of MCNV, or even more so those of LCNV or SCNV (Fig. 6 and Supplementary Table S1A). The PosSel-Hs co-localize mildly with the SVAab retrotransposons. The NegSel-Hs co-localize positively with a range of regulatory features, methylation sites, and disease-related sites, while their negative co-localizations with the Alu, SVAab, SVAcd, and L1 retrotransposons support the suggested paucity of purifying selection in transposable elements [67, 68]. The prominent association between histone modifications and MCNV hotspots is further scrutinized in terms of the size of the constituent CNVs of the MCNV hotspots, which reveals striking enrichment of CNVs of 150–200 bp (density peaks in dashed boxed in Fig. 7 and Supplementary Figure S9) in MCNV hotspots with high levels, but not in those with low levels, for all twelve types of histone modifications analyzed.

The 1135 hotspot clusters in the human genome comprised twenty-three different combinations of hotspots, a large majority of which include the hotspots of SNP and hotspots of structural variants. Each kind of hotspot clusters is associated with its own characteristic array of genomic features. The associations of for example the long RNA (LRNA-) with the SNP + MCNV cluster (viz. clusters comprising the SNP and MCNV hotspot) are largely additive, not far from the sum of the individual associations of LRNA- with SNP hotspot and MCNV hotspot; in contrast, the associations of H3k27ac toward the SNP + SCNV cluster are synergistic, far exceeding the sum of the individual associations of H3k27ac with

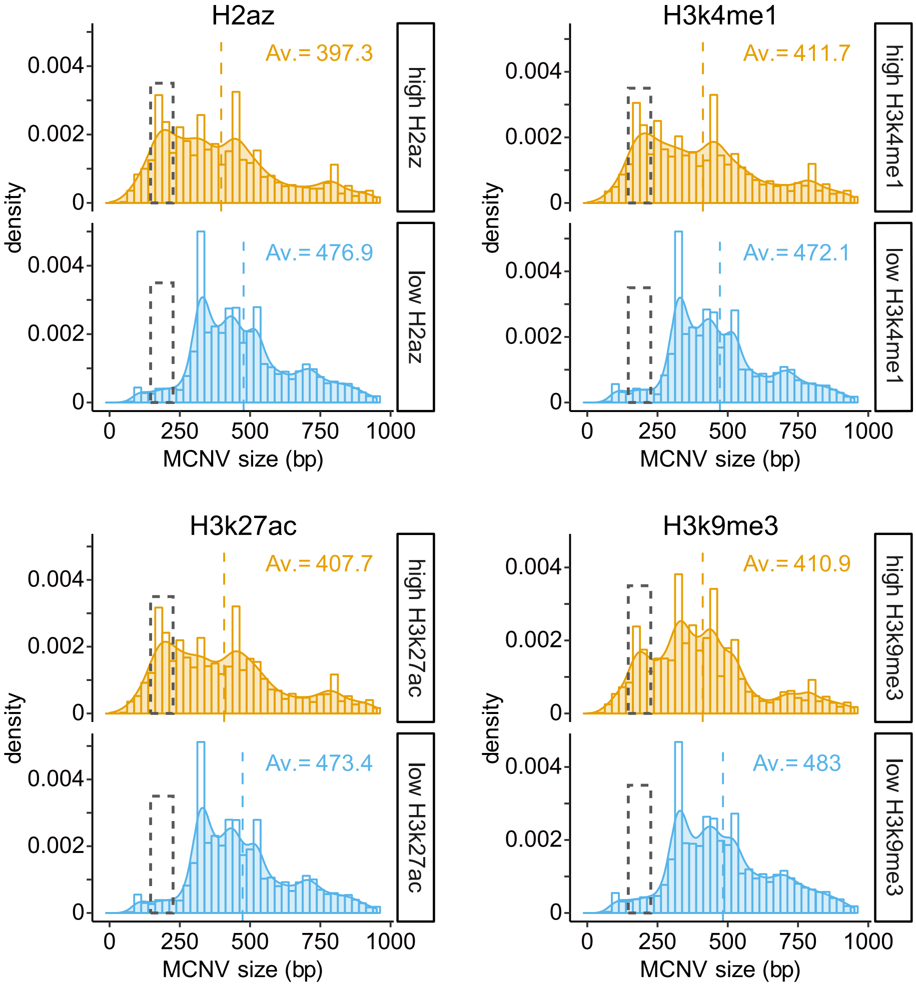

Fig. 7 Length distribution of MCNVs in the MCNV hotspots with high or low levels of histone modifications. Based on the intensity of each type of histone modification, the MCNV hotspots are divided into two groups that show above-average (orange) or below-average (blue) intensities respectively. The contrast between the MCNV hotspots with high levels of histone modification (orange shaded peaks) and those with low levels (blue shaded peaks) is particularly evident in the dashed boxes at 146–226 bp, which corresponds to the length of DNA sequence wrapping around one nucleosome (146 bp) [69] plus the linker DNA (up to 80 bp) [70]. "Av." stands for average CNV length in bp as indicated by the dashed vertical orange or blue line

SNP hotspot and SCNV hotspot (Supplementary Figure S10).

Distributions at immune system gene loci

Although genes are not specially co-localized with GV hotspots or clusters (Fig. 6 and Supplementary Figure S10), the effects of genetic variants and natural selection are often discernable within particular gene clusters (Fig. 2b), as illustrated by the immune system gene loci including *MHC* where protein sequence diversity represents a necessity [21]. Thus the immunoglobulin *IGH*, *IGL*, and *IGK* loci, and the *MHC* locus all overlap with clusters that are enriched with various kinds of hotspots, such as a 40-fold enrichment of ECNV hotspots in the *MHC* locus relative to genomic average (Fig. 2c). The GV hotspots also exhibit abundant co-occurrence of positive in terms of PosSel-H and balancing selection in terms of BalSel-H contents (orange bars, Fig. 2c), but little purifying selection in terms of NegSel-H content (blue bars). These results suggest that the high sequence diversity of the *IG*- and *MHC*-loci has been achieved mainly through positive and balancing selection within the hotspots and clusters. The weaker presence of GV hotspots as indicated by reduced pie size, or selective processes as indicated by hashed slices, in Fig. 2c among the *TRA*, *TRB*, *TRD*, and *TRG* loci of T cell receptors are consistent with the reduced need by these receptors for sequence diversity compared to the *IG*- and *MHC*-loci.

Functional evolution in late-replicating DNA

A well-defined temporal order of DNA replication is characteristic of normal cell division. The periods of genomic DNA synthesis have been classified, from early to late, into the G1b, S1, S2, S3, S4, and G2 phases in fifteen cell types [46, 47], thus allowing the partition of 1-kb autosomal windows into these six phases (see “Replication-time segments” in “Materials and methods” section). Genic zones are enriched in the early-replicating G1b, S1, and S2 phases; Distal zones in the late-replicating S4 and G2 phases; and Proximal zones in the intermediate S3 and late-replicating S4 phases (Fig. 1c). The density of recombination hotspots declines from G1b toward G2 (red bars, Fig. 3c), in accord with the previous finding of a reduction of homologous recombination in late-replicating DNA [33]. High frequencies of GV hotspots are found in G2-phase DNA despite a low level of recombination hotspots.

The DNA sequence windows that undergo early replication in phase G1b are enriched in numerous functional genomic features including the histone-modification sites and gene/regulatory sites throughout the Genic, Proximal, and Distal zones (red rectangles, Fig. 8a top panel), whereas sequence windows in the late-replicating phases S3-G2 are relatively depleted of these functional features (blue rectangles). Nonetheless, within these function-depleted phases, the presence of functional features is more evident in regions where hotspots and clusters locate, especially the histone-modification sites in the Distal-zone hotspots and clusters (red rectangles in middle and bottom panels, Fig. 8a). Six groups of gene pathways are found enriched in the Distal zones (Fig. 8b and Supplementary Figure S11). Among them, the autoimmune thyroid disease, antimicrobial humoral response, natural killer cell, xenobiotic metabolism, epidermal protein disulfide-binding, sensory perception, and neuroactive ligand-receptor interaction pathways are all associated with GV hotspots (red and pink circles, Fig. 8b). Moreover, the latter two pathway groups display the foremost hotspot enrichment, with 46% and 64% of their genes respectively (black slices, Fig. 8c) showing the co-existence of positive, balancing and purifying selections, in contrast to the immune system gene loci which are dominated by positive and balancing selections (Fig. 2c). These findings indicate that the presence of hotspots and clusters could attract various kinds of natural selection forces that can facilitate pathway development even in regions of the genome that usually harbor relatively few functional genomic features.

Enrichment of disease-related variants in hotspots and clusters

The close relationships of genetic-variant hotspots and clusters with different functional genomic features (Fig. 6 and Supplementary Figure S10) suggest that perturbations in the DNA sequences in these regions may result readily in disease-related mutations. In accord with such expectation, the densities of germline SNPs identified from GWAS, the breakpoints of disease-related germline variants in the ClinVar database, and the breakpoints of somatic tumor CNVTs are all significantly enriched within the GV hotspots and clusters relative to simulated regions in the human genome (Fig. 3d). The density of GWAS-identified SNPs within the total GV hotspots is increased to 1.4-fold, to as high as 7.1-fold in ECNV hotspots, or synergistically to 19.0-fold in the SNP + ECNV clusters relative to the autosomal average, far exceeding the sum of their separate enhancements in the SNP and ECNV hotspots (Supplementary Table S6). Instead, clusters formed by SNP and SID does not have this kind of synergistical elevation, but rather a reduced level of disease-associated variants relative to their separate hotspots. This observation suggests that the potential role of SNP + ECNV clusters is pivotal in genomic regions attributable to complex traits and diseases. Furthermore, the GWAS-identified SNPs in the GV hotspots or the clusters are characterized by comparable risk allele frequencies, but higher effect sizes and contributions to heritability, relative to the GWAS-identified SNPs in non-hotspot regions (Supplementary Figure S12). Besides, the GWAS-

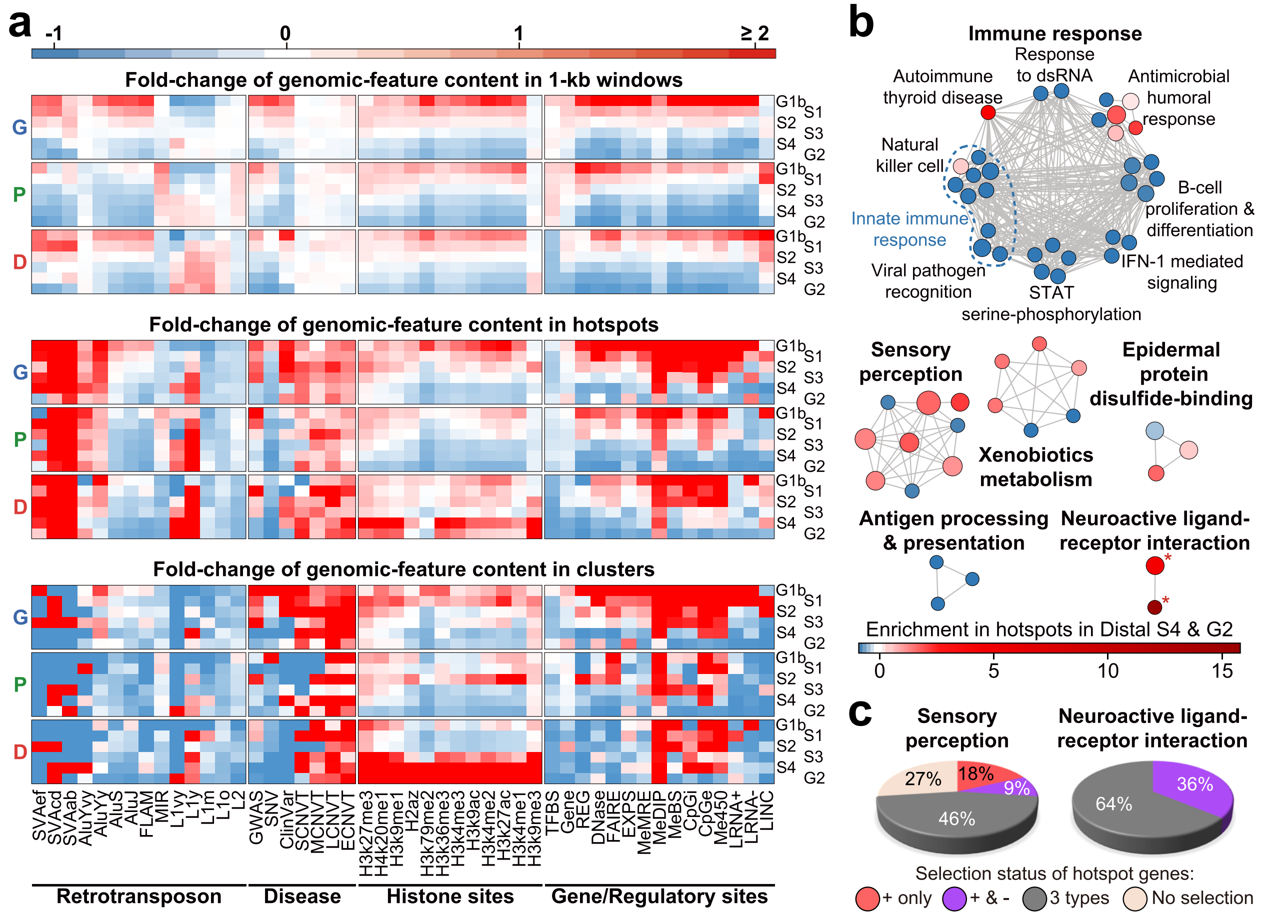

Fig. 8 Distributions of genomic features in different replication phases and genomic zones. **a** Levels of various genomic features (as shown on *x*-axis) within all 1-kb windows (upper panel), all hotspots (middle panel), or all clusters (lower panel) belonging to replication phases G1b-G2 in the Genic (G), Proximal (P), or Distal (D) zones, in accordance to the red-blue thermal scale for fold-change over autosomal average (see Supplementary Table S10 for data on genomic-feature enrichments). **b** Enrichment map for Distal-zone genes annotated using g:Profiler [71] based on Gene Ontology biological process [72] and KEGG [73]. Each node represents a pathway (with 3 to 350 genes) significantly enriched in Distal zones with Benjamini-Hochberg false discovery rate < 0.05, and node size is proportional to the number of Distal-zone genes belonging to the pathway (see pathway IDs in Supplementary Figure S11). Pathways are connected by a gray edge when they share ≥ 50% genes. Node color, represented by the red-blue thermal scale, indicates the fold-change of the fraction of the genes over the fraction of 1-kb sequence windows that overlap with hotspots in the Distal-zone S4 and G2 phase DNA (see Supplementary Table S11). Asterisks mark pathways that overlap with cluster(s). **c** Selection status of pathway genes that overlap with hotspots in the Distal-zone S4 and G2 phases. Fractions of genes overlap only with PosSel-Hs (red slices), with both PosSel-Hs and NegSel-Hs (purple slices), and with all (black slices) or none (beige slices) of PosSel-Hs, NegSel-Hs, and BalSel-Hs

identified SNPs are found significantly enriched inside or near SVAcd, AluYy, AluS, AluJ, FLAM, MIR, and L2 retrotransposons but not the youngest SVAef, AluYvy, and L1vy retrotransposons (Fig. 4b).

The density of CNVTs is likewise increased to 1.9- and 2.9-fold within total hotspots and clusters respectively; or as high as 13.7-fold in the LCNV hotspots and 28.3-fold in the ECNV hotspots (Supplementary Table S6), demonstrating thereby the association of hotspots and clusters with enhanced germline as well as somatic disease-related mutations. Upon division of the somatic CNVTs into the four size-classes of SCNVT, MCNVT, LCNVT, and ECNVT as in the case of the germline CNVs, LCNVTs are most strongly associated with LCNV hotspots, and ECNVTs most strongly with ECNV hotspots (Fig. 6), indicating that LCNV hotspots represent significant sites of LCNVT formation, and ECNV hotspots represent significant sites of ECNVT formation, in keeping with the earlier finding that recurrent germline CNVs provide a useful basis for predicting cancer susceptibility [74]. CNVT recurrency in tumors is particularly strong in the ECNV, LCNV, and MCNV hotspots, in clusters, and in SNP hotspots (Fig. 3e). Such enrichments of disease-related variants in the hotspots and clusters are observed regardless of the genomic zone or replication timing of the DNA sequences (Fig. 8a middle and bottom panels).

There are in the human genome 448 genes that contain one or more clusters, and 24 genes that contain more than two clusters, reaching a maximum of nine clusters in the diabetes-related *PTPRN2* and the potential tumor suppressor *CSMD1* (Supplementary Table S4). Within 33 of these cluster-containing genes, the cluster portions are significantly enriched with somatic CNVTs compared to non-cluster portions (see "Somatic CNV breakpoints in cluster-containing genes" in "Methods" section), which attests to the extraordinary effectiveness of clusters in the generation of CNVT mutations. Among the 448 cluster-containing genes, 113 genes are found in the Network of Cancer Genes 6.0 [75] or the Tumor Suppressor Gene Database 2.0 [76], revealing significant overlaps between cancer genes with hotspot clusters ($p = 8.1 \times 10^{-11}$, Fig. 3f), or with hotspots of SNP, SID, MST, SCNV, or MCNV (Fig. 3f).

Discussion

Formation mechanisms for common and rare variants in hotspots and clusters

SNPs and structural variations in GWAS are widely employed as common and rare genetic markers respectively the "Common Disease-Common Variant" and "Common Disease-Rare Variant" research strategies [5, 12, 13]. In the present study, they are both found to be concentrated in hotspots and clusters that cannot be formed due to background mutations, in a clear departure from the neutral evolution model [77–80]. The presence of numerous hotspot clusters that comprise an SNP hotspot along with the hotspot of some structural variations (Fig. 2 and Supplementary Figure S10) suggests that the formation mechanisms for the SNP hotspots overlap with those for the hotspots of SIDs, MSTs, SDPs, and CNVs of different lengths in the same clusters. It follows that the different structural variations do exhibit overlaps among themselves, as exemplified by the co-localizations between copy number gains and copy number losses, insertions and tandem duplications, as well as duplications and SDPs, supporting that the unstable genomic regions are prone to diversified genetic variants (Supplementary Figure S13). In addition, SNPs, SIDs, MCNVs, LCNVs, and ECNVs are enriched at the telomeric termini of autosomes (Supplementary Figure S14), which is consistent with earlier report of SNP enrichment near the ends of some chromosomes [21]. Such termini-proximal concomitance of GVs might be related to the abundance of terminal repetitive sequences, which could serve as homologous templates for recombination to maintain proper telomeric length and preserve chromosomal stability [81–83]. Evidence suggests that homologous recombination can be mutagenic, producing SNPs due to the error-prone nature of the DNA polymerase involved, or generating structural variations via ectopic homologous recombination events between repeated sequences [30, 32, 84, 85]. As well, the association of homologous recombination with the formation of a range of genetic variants is indicated by the co-occurrence of elevated genetic-variant density and high recombination intensity (Fig. 3a, b). However, only 15.0% of GV hotspots overlap with the previously identified recombination hotspots (Supplementary Table S7), possibly because of underestimation of recombinations on account of the reduction of their footprints by the strong positive selection in the GV hotspots (Fig. 5f), or by other mutagenic mechanisms such as replication-based mechanisms or the non-homologous end joining that contributed to the formation of the GV hotspots and clusters. This is demonstrated by the high concentrations of GV hotspots in G2 phase-replicated DNA despite the relative scarcity of recombination hotspots in such DNA (Fig. 3c), coinciding with the stronger presence of non-homologous end joining compare to homologous recombination in G2-phase DNA [33].

It is noteworthy that nucleotide differences between highly similar sequences (e.g., SDPs) located at different genomic positions may be falsely identified as SNPs due to mismapping of short-reads [25], giving rise to biased associations of SNPs with structural variations. On this basis, false-positive SNPs may be expected to be particularly frequent near SDPs where the repetitive sequences would be conducive to such mismapping. Since SNP densities are much lower in the vicinity of SDPs relative to the vicinities of MCNV, LCNV, and ECNV (Supplementary Figure S13), subtraction of the level detected in the vicinity of SDP from the SNP levels observed from the vicinity of the various CNVs may be considered as a useful correction measure for those SNPs that are near the MCNV, LCNV, and ECNV.

Negative effects of hotspots and clusters

The suggestion that retrotransposons could mediate homologous recombination [26, 27] is supported by the close co-localizations of a range of genetic-variant hotspots and clusters with different retrotransposons, and the positive co-localization of AluYy insertions not only with various hotspots and clusters but also with various recombination features (Figs. 4b and 6). Such findings indicate that GV hotspots and clusters harbor sites of genomic instability which expedite the production of neutral, advantageous, or detrimental variants through recombination. The detrimental effects are evident in the strong associations of hotspots and clusters with cancerous CNVTs, and the enrichment of disease-associated variants in GWAS (Fig. 3d-f). Furthermore, the parallel declines in Rec-H and CNVT in G1b- and G2-replicating DNA (Fig. 3c) support the proposal of homologous recombination as a source of CNVTs [84];

and the strong associations between LCNVT and LCNV hotspots, as well as between ECNVT and ECNV hotspots (Fig. 6) suggest that the selfsame sets of *cis*-acting retrotransposons are utilized as nucleation points for the formation of both germline CNV and somatic CNVT through recombination [86, 87]. Moreover, clusters are also observed for somatic structural variations implying mechanistical linkages of genetic variants in those clusters [88]. However, the germline CNVs and somatic CNVTs are dissimilar in their length distributions, resulting in a merger of the four distinct peaks of germline CNVs into a more even distribution of CNVTs of different lengths (Fig. 1a, bottom panel). A plausible reason may be that, although the occurrences of both types of CNVs are enhanced by recombination events, the frequencies of which vary between the different replication phases (Fig. 3c), the germline CNVs had gone through a process of natural selection to eliminate CNVs disadvantageous to the survival of the host lineage. On the other hand, the somatic CNVTs include newly formed ones that have not been shaped as extensively by natural selection, and hence follow a close-to-normal distribution. The association of CNVTs with GV clusters is particularly strong: there are 33 genes in the genome where CNVT breakpoints are significantly enriched within segments bearing GV clusters relative to other segments of the gene (overall Bonferroni-corrected p value < 0.05). The significance of such enrichment reaches $p < 10^{-58}$ for *CACNA1C* and *SNX29* from the Network of Cancer Genes, and $p < 10^{-61}$ for *WWOX* and *CSMD1* from the Tumor Suppressor Gene Database (Supplementary Table S4). Therefore, the cluster segments of such genes clearly rank among the hotspots of oncogenesis within the genome.

Positive effects of hotspots and clusters

GV hotspots and clusters produce advantageous genetic variants that could be further elevated in frequency by positive selection to expedite functional adaptation. This is illustrated by (i) enrichment of SNPs with DAF > 0.95 in hotspots and clusters in all five ethnic populations ($p < 10^{-10}$, Fig. 5a), accompanied by large overlaps of PosSel-Hs with SNP, LCNV, and ECNV hotspots (Supplementary Table S3B); (ii) rise of the inter-population differentiation $|\Delta DAF|$ owing to positive selection within the clusters (Fig. 5b); and (iii) increase of haplotype homozygosity, measured by $|nS_L|$, on account of positive selection in all five ethnic populations ($p < 10^{-24}$, Fig. 5c). Outstanding examples are provided by the *MHC*- and *IG*-loci which are enriched with SNP, SID, MCNV, LCNV, and ECNV hotspots associated with positive and balancing selections (Fig. 2c), revealing the exceptional hotspot- and cluster-driven evolutionary development of these immune system loci.

Another example of the positive functional effects of hotspots and clusters is provided by the development of the olfactory and taste sensory pathways, and neuroactive ligand-receptor interaction pathways through collaboration of balancing, positive, and purifying selections in the hotspot and cluster regions that amount to only 3.3% of the total base pairs in these otherwise functional deserts of the S4- and G2-replicating Distal zones (Supplementary Table S8), which clearly demonstrates the important advantage of hotspots and clusters as incubators for the accelerated adaptive evolution. The dominance of purifying selection in the protein-coding regions of genes [89] suggests that a majority of the coding sequences in the human genome are already largely optimized by positive selection, and therefore in need of shielding by purifying selection against adverse sequence alterations. On the other hand, the much weaker purifying selection than balancing or positive selection observed in the immunoprotein genes is consistent with the younger age of these genes, where evolutionary development might still be ongoing, propelled by genetic variants working in conjunction with balancing and positive selections. Interestingly, the depletion of LCNV, ECNV, and SDP entries from the Proximal zones relative to the Distal or Genic zones (Fig. 1b) suggests that the presence of lengthy interruptions could be particularly disruptive to tightly organized sequences in the proximal zones such as the sequences flanking transcription factor binding sites, is essential to the function of genomic features in the Proximal zones, which is in accord with the frequent detection of purifying selection in the Proximal zones [35, 67].

The coincidence between hotspots of GVs and positive selection may seem paradoxical, for one might suppose that the selection process would sweep away the GVs. However, a high level of the nS_L statistic could serve as an indicator for the existence of “soft sweep” where genetic variants or multiple variants at a single locus remain standing until the sum of their allele frequencies reaches one [41]. Genetic forces and evolutionary history such as the complex demographic history of human populations may render difficult a rigorous description of positive selection by any single measurement [58, 90]. For example, population bottlenecks can lead to the shift of DAF to extreme values; a population expansion may increase ΔDAF between population pairs; and the nS_L statistic may vary with the demographic model employed. Therefore, the observation of enhancement of positive selection in GV hotspots and clusters based on all three measures for detection of positive selection signals improves the reliability of the observed enhancement.

Missing heritability due to co-occurrence of recombination and positive selection

Genetic components associated with complex traits and disorders could be missed in genomic regions simultaneously subjected to high levels of positive selection and recombination. One of such cases has been well evident in a 3551-bp segment of the schizophrenia-associated *GABRB2* gene that codes for the β2 subunit of $GABA_A$ receptor. The derived alleles of a number of SNPs within this segment are under recent and ongoing positive selection, for they increase through alternative splicing the production of the longer isoform of the β2 subunit and diminish the production of the shorter β2 subunit favored by the ancient alleles [91]. The ancient alleles become less fit than the derived alleles under modern conditions, resulting in positive selection of the derived alleles. The same has been observed in some other genes [92–95]. In the meanwhile, this segment harbors an AluYi6 insertion, which serves as a recombination center to largely enhance recombination rate within the segment [96]. Under such circumstances, the elevated recombination can play positive and negative roles simultaneously. On the one hand, novel sequence variations and haplotypes are brought about by recombination for positive selection to increase the frequencies of the advantageous genotypes. On the other hand, sequence alternations could also lead to functional perturbations that associate with the etiological basis of schizophrenia [97]. The co-occurrence of recombination and positive selection points out a possible genetics mechanism for the development of complex diseases.

A wide range of schizophrenia-like phenotypes displayed by the *GABRB2*-knockout mice, and the reversal of these phenotypic alterations by antipsychotic drug further underline the pivotal role of *GABRB2* in the development of schizophrenia [98]. As well, the association between *GABRB2* and schizophrenia has been validated by multiple genetic studies on different ethnic populations [99–101], and SNPs, haplotypes and CNVs in the 3551-bp segment are found to be associated with schizophrenia taking a candidate gene approach [102, 103]. However, genome-wide association studies report no significant association between any of the SNPs in this 3551-bp segment and schizophrenia. Fine-resolution linkage disequilibrium analysis in this segment reveals much higher recombination in the ancestral allele-containing haplotype groups relative to the derived allele-containing haplotype groups, suggesting active recombination in this region together with intense positive selection on the derived alleles [96]. Such complexity due to the co-occurrence of recombination and positive selection, interacting with each other to affect local genetic diversity landscape, could blur the allele and haplotype signals thereby potentially contributing to the missing heritability problem in GWAS. The depletion of GWAS-identified SNPs near the youngest (viz. SVAef, AluYvy, and L1vy), versus enrichment near the older (viz. AluS, AluJ, MIR, etc.), retrotransposons might point to missing associations stemming from the elevated levels of recombination and positive selection in the youngest retrotransposons (Fig. 4b). Therefore, the genomic regions that are subject to pronounced recombination and positive selection, such as the genetic-variant hotspots and clusters or the vicinity of young retrotransposons, would merit close investigation, more reliance also may be accorded to genetic variants, such as CNVs, that are more resistant to the obscuring effects of recombination together with selection.

“Common disease-hotspot variant” hypothesis

The dissimilar associations of SCNV and MCNV hotspots with histone-modification sites, SNP and SID hotspots with methylation-related sites, and LCNV and ECNV hotspots with somatic CNVTs (Fig. 6) are indicative of the specialized deployment of different genetic variants in the genome to meet specific functional needs. Notably, the associations observed between hotspots and genomic features are unlikely stemmed from the similarity in their detection methods, because data on different genomic features and genetic variants were generated by independent projects with different experimental methods. The strength of association of GV hotspots and clusters with function is underlined by the findings that 34.8% of total GV hotspots and 43.9% of total clusters display some balancing, positive, or negative selection. As a result, sequence perturbations in these pivotal sites may be expected to give rise to phenotypical changes including human disorders. It is noteworthy that searches for phenotype-genotype associations are usefully guided by the “Common Disease-Common Variant” and “Common Disease-Rare Variant” strategies. However, the former is limited by the fraction of heritability of complex traits explained by common SNPs, which is highly variable (20–90%) for different traits and different studies [104, 105], and the latter is confronted by the difficulty of finding the rare variants [12, 35], both leaving substantial proportions of heritability unexplained. The “hypothesis-driven” strategies have been found useful in certain cases, but they usually require some measure of prior knowledge about the disease mechanism. It has also been found that incorporation of functional genomic data in association studies also can increases the discovery of trait/disease-associated sites by about 5% [106]. As well, it is important that the full implications of the sequence changes also should not be missed. Chromatin plays important roles in maintaining DNA stability by preventing recombination processes [107, 108]. By remodelling the chromatin, histone modifications might impact on recombination events that lead to genetic variants, as exemplified by the increased level of H3K56 acetylation that disrupts homologous-based DNA damage repair and causing extensive CNVs

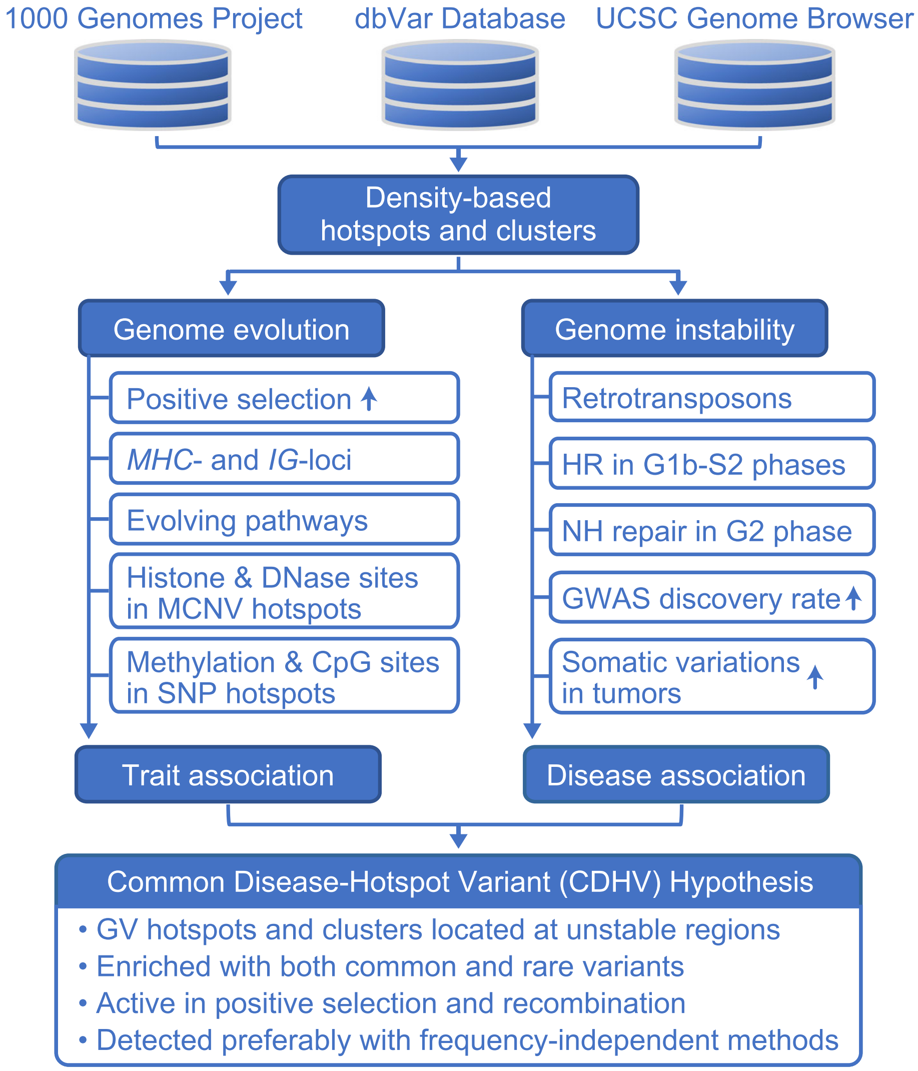
1000 Genomes Project
dbVar Database
UCSC Genome Browser
Density-based hotspots and clusters
Genome evolution
Positive selection ↑
MHC- and *IG*-loci
Evolving pathways
Histone & DNase sites in MCNV hotspots
Methylation & CpG sites in SNP hotspots
Genome instability
Retrotransposons
HR in G1b-S2 phases
NH repair in G2 phase
GWAS discovery rate ↑
Somatic variations in tumors ↑
Trait association
Disease association
Common Disease-Hotspot Variant (CDHV) Hypothesis
• GV hotspots and clusters located at unstable regions
• Enriched with both common and rare variants
• Active in positive selection and recombination
• Detected preferably with frequency-independent methods

Fig. 9 Schematic summary of present findings. Public genetic-variant databases are employed to identify density-based hotspots and clusters, which are found to be dynamic centres with accentuated positive and/or negative changes in the genome. On the one hand, they work in conjunction with positive selection to enhance sequence diversity, evolution of specific genes and pathways, and development of functional genomic features. On the other hand, they are co-localized with destabilizing retrotransposon elements, and occurrences of homologous recombination (HR) in the early DNA replication phases, and non-homologous (NH) repair in the late DNA replication phases. Their associations with high GWAS discovery rates and somatic variants signal the importance of associations with complex traits and diseases, and therefore their utility as the basis for a potential "Common Disease-Hotspot Variant" strategy in a search for the missing heritability in association studies

in the yeast genome [108–110]. This is in accord with our observation of strong co-localization of histone modification sites with CNV hotspots, especially medium size MCNVs but not other sizes of CNVs, in the human genome (Fig. 6). The selective association with histone modification sites is particularly evident in MCNVs comparable in length with DNA sequences that wrap nucleosomes (Fig. 7 and Supplementary Figure S9), suggesting that some of the sequence alterations involved in chromatin remodelling during DNA replication could be related to the nucleosome units. Therefore, it would be of interest to examine whether MCNVs might play some role in the structure and function of nucleosomes. Accordingly, when an investigator finds that some disease displays correlation with an MCNV alteration, he would be encouraged to investigate in depth any histone modification sites in the vicinity, thereby reducing the chances of missing some important functional implication of the MCNV observed. Accordingly, in view of the evident elevations of GWAS discovery rates in GV hotspots and clusters, and the enriched presence of both LCNVTs in LCNV hotspots and ECNV Ts in ECNV hotspots, intensive efforts directed to the discovery of common or rare genetic variants in hotspots and clusters, and characterization of their potential association with complex disease and traits, appear warranted in order to search for some of the missing heritability.

Conclusion

Overall, in the present study, a high-resolution examination has been conducted throughout the Genic, Proximal, and Distal genomic zones in the human genome on eight kinds of germline genetic variants retrieved from public databases. Identification of genetic-variant hotspots and clusters based on regional densities of genetic variants, independent of allele and haplotype frequencies, minimizes underestimation on account of positive selection and recombination saturation. The findings on their distributions, formation mechanisms, associated genomic features, as well as their effects on functional development, complex diseases, and traits underline the pervasive and double-edged sword nature of their impact on the human genome (Fig. 9). They advance functional genome evolution, but they also constitute important foci of cancers and other diseases, both arising from their association with genome instability. The attributions of common and rare variants to common diseases and traits, as formulated in the "Common Disease-Common Variant" and "Common Disease-Rare Variant" hypotheses respectively [5, 12, 13], have brought about important advances in the understanding and detection of the genomic variations related to the phenotypical changes; it is recognized nonetheless that they remain confronted by the problem of missing heritability, where the genomic variants associated with diseases and traits remain largely uncovered [7, 10, 11]. In view of this, and the enrichment of both common and rare genetic variants in GV hotspots and clusters, it is proposed that "Common Disease-Hotspot Variant" hypothesis could provide an additional, complementary approach. The basis of this hypothesis is that common diseases and traits are frequently attributable to genetic variants, common as well as rare types, occurring in regions marked by extraordinary genome instability. Such instability at these unstable regions would enhance the probability of not only germline variants but also disease- or trait-associated germline variants and disease-related somatic variants. Thus, the GV hotspots and clusters represent an unexpected source of disease-related heritability that could easily be missed due to the effects of local recombination at or near saturation levels along with directional selection on allele or haplotype frequencies, and intensified discovery and mutation monitoring of hotspots and clusters could help to uncover some of this missing heritability.

Since the landscape of the unstable regions may be complicated by saturation of recombination and selection, a Common Disease-Hotspot Variant hypothesis-based approach requires experimental and statistical methods that are relatively insensitive to recombination and selection. In this regard, structural variations such as CNVs will serve as valuable markers, for they are, compared to SNPs less prone to mutation saturation arising from increased recombination. GV density-based statistical methods are also advantageous on account of their relative immunity to directional selection as compared to statistical methods of population genetics commonly used in association studies, which are often frequency-based and susceptible to the effects of selection on allelic and haplotype frequencies. Therefore, through additional focus on GV hotspots and clusters, and increased adoption of frequency-independent statistical methods, some of the missing heritability may come to be captured, deepening insight into the genetics of complex disorders and traits, and improving the detection of disease- or trait-associated variants.

Abbreviations

SNP: Single-nucleotide polymorphism; CNV: Copy-number-variation; GV: Genetic variant; GWAS: Genome-wide association study; SID: Small indel; MST: Microsatellite; SDP: Segmental duplication; SCNV: Short germline copy-number-variation; MCNV: Medium germline copy-number-variation; LCNV: Long germline copy-number-variation; ECNV: Extra-long germline copy-number-variation; SCNVT: Short somatic copy-number-variation; MCNV T: Medium somatic copy-number-variation; LCNVT: Long somatic copy-number-variation; ECNVT: Extra-long somatic copy-number-variation; PosSel-H: Positive selection hotspot; NegSel-H: Negative selection hotspot; BalSel-H: Balancing selection hotspot; Rec-H: Recombination hotspot; DAF: Derived-allele frequency; MAF: Minor-allele frequencies; LINC: Large intergenic non-coding RNAs; DNase: Open chromatin elements; LRNA: Long RNAs

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40246-021-00318-3.

Additional file 1: Supplementary FigureS1. Size-classification of germline CNV regions retrieved from dbVar database. **Supplementary Figure S2**. Heat map of pairwise Spearman's correlation coefficients (r_S) among different genomic features and markers in 500-kb non-overlapping autosomal windows. **Supplementary Figure S3**. Refinement of hotpsot boundary by weighted density. **Supplementary Figure S4**. Flowchart of protocol for identifying SNP hotspots. **Supplementary Figure S5**. Density of GWAS-identified SNPs in ten groups of hotspots and clusters ranked based on minor allele frequency from low (group 1) to high (group 10). **Supplementary Figure S6**. Comparison of GV densities in hotspots (red boxplots) and in simulated windows with matching number and size of the hotspots (grey boxplots). **Supplementary Figure S7**. Comparison of GV allele frequencies in hotspots (red bars) and in total autosomal windows (grey bars) binned by the average size of GV hotspots (1,887 bp). **Supplementary Figure S8**. Density distribution of SVA and Alu subfamilies near eight kinds of genetic variants. Cumulative density of 42 SVA and Alu subfamilies within ± 5-kb of SNP, SID, MST, SCNV, MCNV, LCNV, ECNV and SDP entries (in 50-bp increments). **Supplementary Figure S9**. Length distribution of MCNVs in the MCNV hotspots with high or low levels of histone modifications. **Supplementary Figure S10**. Genomic-feature contents of total (first column in upper panel) and 23 kinds of hotspot clusters with distinct hotspot-compositions. **Supplementary Figure S11**. Enrichment Map for Distal-zone-enriched genes annotated using g:Profiler based on Gene Ontology biological process and KEGG. **Supplementary Figure S12**. Comparison of risk allele frequency, odds ratio and heritability of GWAS-identified SNPs located within total autosomal windows (grey bars) relative to those located within hotspots (red bars) or clusters (blue bars). **Supplementary Figure S13**. Co-localization among different kinds of genetic variants. **Supplementary Figure S14**. Distribution of GVs around the telomere ends.

Additional file 2: Supplementary Table S1. Information on genomic features and markers. **Supplementary Table S2**. Properties of genetic-variant hotspots and hotspot clusters. **Supplementary Table S3**. Properties of selection hotspots. **Supplementary Table S4**. List of cluster-containing genes. **Supplementary Table S5**. Estimated ages of retrotransposons. **Supplementary Table S6**. Disease-related variants in GV hotspots and clusters. **Supplementary Table S7**. Overlap of genetic-variant hotspots or clusters with recombination hotspots. **Supplementary Table S8**. Autosomal sequences (in base pairs) in different genomic zones and replication phases that overlap with hotspots or clusters. **Supplementary Table S9**. Enrichment and selection status of hotspots in seven immune system gene loci. **Supplementary Table S10**. Genomic-feature distributions in genomic regions belonging to different replication-phase segments among 500-kb Genic, Proximal and Distal zones. **Supplementary Table S11**. Functional annotation of Distal-zone genes.

Additional file 3: Supplementary Dataset S1. Information of genetic-variant hotspots, hotspot clusters, selection hotspots and replication-phase segments. Genomic coordinates are given in one-based format in human genome assembly hg19.

Acknowledgements

We thank Dr. Taobo Hu and Dr. Siu-Kin Ng for expert advice, and Dr. Zheng-gang Wu for server management.

Authors' contributions

HX conceived and initiated the study, XL collected the data and performed computational analysis, and HX and XL wrote the paper. Both authors read and approved the final manuscript.

Funding

This research was supported by the Innovation Technology Commission of Hong Kong SAR [grant number ITS113/15FP]; and Shenzhen Science and Technology Innovation Commission [grant number JCYJ20170818113656988].

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information files. Custom R scripts are available upon request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Division of Life Science and Applied Genomics Centre, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China. [2]HKUST Shenzhen Research Institute, 9 Yuexing First Road, Nanshan, Shenzhen, China. [3]Centre for Cancer Genomics, School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing, Jiangsu, China.

Received: 3 October 2020 Accepted: 4 March 2021
Published online: 19 March 2021

References

1. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
2. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005–D12.
3. Do R, Willer CJ, Schmidt EM, Sengupta S, Gao C, Peloso GM, et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. Nat Genet. 2013;45(11):1345–52.
4. Lv WQ, Zhang X, Zhang Q, He JY, Liu HM, Xia X, et al. Novel common variants associated with body mass index and coronary artery disease detected using a pleiotropic cFDR method. J Mol Cell Cardiol. 2017;112:1–7.
5. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;51(3):431–44.
6. Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet. 2001;17(9):502–10.
7. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–53.
8. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev. 2009;19(3):212–9.
9. Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2011;13(2):135–45.
10. Maher B. Personal genomes: the case of the missing heritability. Nature. 2008;456(7218):18–21.
11. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11(6):446–50.
12. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11(6):415–25.
13. Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S, et al. Large recurrent microdeletions associated with schizophrenia. Nature. 2008;455(7210):232–6.
14. The International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature. 2008;455(7210):237–41.
15. Kushima I, Aleksic B, Nakatochi M, Shimamura T, Shiino T, Yoshimi A, et al. High-resolution copy number variation analysis of schizophrenia in Japan. Mol Psychiatry. 2017;22(3):430–40.
16. Rogozin IB, Gertz EM, Baranov PV, Poliakov E, Schaffer AA. Genome-wide changes in protein translation efficiency are associated with Autism. Genome Biol Evol. 2018;10(8):1902–19.
17. Alexaki A, Kames J, Holcomb DD, Athey J, Santana-Quintero LV, Lam PVN, et al. Codon and codon-pair usage tables (CoCoPUTs): facilitating genetic variation analyses and recombinant gene design. J Mol Biol. 2019;431(13):2434–41.
18. Lohmueller KE. The impact of population demography and selection on the genetic architecture of complex traits. PLoS Genet. 2014;10(5):e1004379.
19. Sanjak JS, Long AD, Thornton KR. A model of compound heterozygous, loss-of-function alleles is broadly consistent with observations from complex-disease GWAS datasets. PLoS Genet. 2017;13(1):e1006573.
20. Simons YB, Bullaughey K, Hudson RR, Sella G. A population genetic interpretation of GWAS findings for human quantitative traits. PLoS Biol. 2018;16(3):e2002985.
21. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061–73.
22. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16(3):172–83.
23. Perry GH, Tchinda J, McGrath SD, Zhang JJ, Picker SR, Caceres AM, et al. Hotspots for copy number variation in chimpanzees and humans. Proc Natl Acad Sci U S A. 2006;103(21):8006–11.
24. Gokcumen O, Babb PL, Iskow RC, Zhu Q, Shi X, Mills RE, et al. Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. Genome Biol. 2011;12(5):R52.
25. Lin YL, Gokcumen O. Fine-scale characterization of genomic structural variation in the human genome reveals adaptive and biomedically relevant hotspots. Genome Biol Evol. 2019;11(4):1136–51.

26. Sen SK, Han K, Wang J, Lee J, Wang H, Callinan PA, et al. Human genomic deletions mediated by recombination between Alu elements. Am J Hum Genet. 2006;79(1):41–53.
27. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009;10(10):691–703.
28. Song X, Beck CR, Du R, Campbell IM, Coban-Akdemir Z, Gu S, et al. Predicting human genes susceptible to genomic instability associated with Alu/Alu-mediated rearrangements. Genome Res. 2018;28(8):1228–42.
29. Ng SK, Xue H. Alu-associated enhancement of single nucleotide polymorphisms in the human genome. Gene. 2006;368:110–6.
30. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464(7289):704–12.
31. Segurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation in the human germline. Annu Rev Genomics Hum Genet. 2014;15: 47–70.
32. Carvalho CM, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17(4):224–38.
33. Mao Z, Bozzella M, Seluanov A, Gorbunova V. DNA repair by nonhomologous end joining and homologous recombination during cell cycle in human cells. Cell Cycle. 2008;7(18):2902–6.
34. Shrivastav M, De Haro LP, Nickoloff JA. Regulation of DNA double-strand break repair pathway choice. Cell Res. 2008;18(1):134–47.
35. Ng SK, Hu T, Long X, Chan CH, Tsang SY, Xue H. Feature co-localization landscape of the human genome. Sci Rep. 2016;6:20650.
36. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, et al. dbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res. 2013;41(Database issue):D936–41.
37. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80.
38. Bailey JA, Gu ZP, Clark RA, Reinert K, Samonte RV, Schwartz S, et al. Recent segmental duplications in the human genome. Science. 2002;297(5583): 1003–7.
39. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014; 42(Database issue):D764–70.
40. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, et al. Gene: a gene-centered information resource at NCBI. Nucleic Acids Res. 2015; 43(Database issue):D36–42.
41. Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol. 2014;31(5):1275–91.
42. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034–50.
43. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
44. Szpiech ZA, Hernandez RD. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol. 2014;31(10):2824–7.
45. Bitarello BD, de Filippo C, Teixeira JC, Schmidt JM, Kleinert P, Meyer D, et al. Signatures of long-term balancing selection in human genomes. Genome Biol Evol. 2018;10(3):939–55.
46. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci U S A. 2010;107(1):139–44.
47. Thurman RE, Day N, Noble WS, Stamatoyannopoulos JA. Identification of higher-order functional domains in the human ENCODE regions. Genome Res. 2007;17(6):917–27.
48. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118.
49. North BV, Curtis D, Sham PC. A note on the calculation of empirical P values from Monte Carlo procedures. Am J Hum Genet. 2002;71(2):439–41.
50. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009;25(14):1841–2.
51. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
52. Kucera M, Isserlin R, Arkhangorodsky A, Bader GD. AutoAnnotate: a Cytoscape app for summarizing networks with semantic annotations. F1000Res. 2016;5:1717.
53. Wickham H. ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag; 2009.
54. Kim PM, Lam HY, Urban AE, Korbel JO, Affourtit J, Grubert F, et al. Analysis of copy number variants and segmental duplications in the human genome: evidence for a change in the process of formation in recent evolutionary history. Genome Res. 2008;18(12):1865–74.
55. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009;10(4):252–63.
56. Liu GE, Alkan C, Jiang L, Zhao S, Eichler EE. Comparative analysis of Alu repeats in primate genomes. Genome Res. 2009;19(5):876–85.
57. Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreno-Torres A, Pavlidis P, et al. 1000 Genomes selection browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Res. 2014; 42(Database issue):D903–9.
58. Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics. 1995;140(2):783–96.
59. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000;155(3):1405–13.
60. Elyashiv E, Bullaughey K, Sattath S, Rinott Y, Przeworski M, Sella G. Shifts in the intensity of purifying selection: an analysis of genome-wide polymorphism data from two closely related yeast species. Genome Res. 2010;20(11):1558–73.
61. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A. 1979;76(10):5269–73.
62. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. Science. 2005;310(5746):321–4.
63. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449(7164):851–61.
64. Schroering AG, Edelbrock MA, Richards TJ, Williams KJ. The cell cycle and DNA mismatch repair. Exp Cell Res. 2007;313(2):292–304.
65. Carvalho CM, Pehlivan D, Ramocki MB, Fang P, Alleva B, Franco LM, et al. Replicative mechanisms for CNV formation are error prone. Nat Genet. 2013; 45(11):1319–26.
66. Pedersen BS, De S. Loss of heterozygosity preferentially occurs in early replicating regions in cancer genomes. Nucleic Acids Res. 2013;41(16):7615–24.
67. Lunter G, Ponting CP, Hein J. Genome-wide identification of human functional DNA using a neutral indel model. PLoS Comput Biol. 2006;2(1):e5.
68. Ponjavic J, Ponting CP, Lunter G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. 2007; 17(5):556–65.
69. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature. 1997; 389(6648):251–60.
70. Felsenfeld G, Groudine M. Controlling the double helix. Nature. 2003; 421(6921):448–53.
71. Reimand J, Arak T, Vilo J. g:Profiler—a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res. 2011;39(Web Server issue):W307–15.
72. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.
73. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40(Database issue):D109–14.
74. Ding X, Tsang SY, Ng SK, Xue H. Application of machine learning to development of copy number variation-based prediction of cancer risk. Genomics Insights. 2014;7:1–11.
75. Repana D, Nulsen J, Dressler L, Bortolomeazzi M, Venkata SK, Tourna A, et al. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol. 2019;20(1):1.
76. Zhao M, Kim P, Mitra R, Zhao J, Zhao Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic Acids Res. 2016;44(D1):D1023–31.
77. Corbett-Detig RB, Hartl DL, Sackton TB. Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. 2015;13(4):e1002112.
78. Hahn MW. Toward a selection theory of molecular evolution. Evolution. 2008;62(2):255–65.
79. Przeworski M, Hudson RR, Di Rienzo A. Adjusting the focus on human variation. Trends Genet. 2000;16(7):296–302.

80. Aquadro CF, Begun DJ, Kindahl EC. Selection, recombination, and DNA polymorphism in Drosophila. In: Golding B, editor. Non-Neutral Evolution. Boston (MA): Springer; 1994. p. 46–56.
81. Bailey SM, Murnane JP. Telomeres, chromosome instability and cancer. Nucleic Acids Res. 2006;34:2408–17.
82. Dunham MA, Neumann AA, Fasching CL, Reddel RR. Telomere maintenance by recombination in human cells. Nat Genet. 2000;26(4):447–50.
83. Dilley RL, Verma P, Cho NW, Winters HD, Wondisford AR, Greenberg RA. Break-induced telomere synthesis underlies alternative telomere maintenance. Nature. 2016;539(7627):54–8.
84. Kumar Y, Yang J, Hu T, Chen L, Xu Z, Xu L, et al. Massive interstitial copy-neutral loss-of-heterozygosity as evidence for cancer being a disease of the DNA-damage response. BMC Med Genomics. 2015;8:42.
85. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009;10(8):551–64.
86. Coop G, Przeworski M. An evolutionary view of human recombination. Nat Rev Genet. 2007;8(1):23–34.
87. Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet. 2002;3(5):370–9.
88. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020;578(7793):112–21.
89. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. Nat Genet. 2019;51(1):88–95.
90. Torres R, Szpiech ZA, Hernandez RD. Human demographic history has amplified the effects of background selection across the genome. PLoS Genetics. 2018;14(6):e1007387.
91. Zhao C, Xu Z, Chen J, Yu Z, Tong KL, Lo WS, et al. Two isoforms of GABA(A) receptor beta2 subunit with different electrophysiological properties: differential expression and genotypical correlations in schizophrenia. Mol Psychiatry. 2006;11(12):1092–105.
92. Di Rienzo A, Hudson RR. An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet. 2005;21(11):596–601.
93. Vander Molen J, Frisse LM, Fullerton SM, Qian Y, Del Bosque-Plata L, Hudson RR, et al. Population genetics of CAPN10 and GPR35: implications for the evolution of type 2 diabetes variants. Am J Hum Genet. 2005;76(4):548–60.
94. Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, et al. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet. 2000;26(1):76–80.
95. Neel JV. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet. 1962;14:353–62.
96. Ng SK, Lo WS, Pun FW, Zhao C, Yu Z, Chen J, et al. A recombination hotspot in a schizophrenia-associated region of GABRB2. PLoS One. 2010;5(3):e9547.
97. Zhao C, Xu Z, Wang F, Chen J, Ng SK, Wong PW, et al. Alternative-splicing in the exon-10 region of GABA(A) receptor beta(2) subunit gene: relationships between novel isoforms and psychotic disorders. PLoS One. 2009;4(9):e6977.
98. Yeung RK, Xiang ZH, Tsang SY, Li R, Ho TYC, Li Q, et al. Gabrb2-knockout mice displayed schizophrenia-like and comorbid phenotypes with interneuron-astrocyte-microglia dysregulation. Transl Psychiatry. 2018;8(1):128.
99. Yu ZL, Chen JH, Shi HF, Stoeber G, Tsang SY, Xue H. Analysis of GABRB2 association with schizophrenia in German population with DNA sequencing and one-label extension method for SNP genotyping. Clin Biochem. 2006;39(3):210–8.
100. Petryshen TL, Middleton FA, Tahl AR, Rockwell GN, Purcell S, Aldinger KA, et al. Genetic investigation of chromosome 5q GABAA receptor subunit genes in schizophrenia. Mol Psychiatry. 2005;10(12):1074–88 57.
101. Liu J, Shi Y, Tang W, Guo T, Li D, Yang Y, et al. Positive association of the human GABA-A-receptor beta 2 subunit gene haplotype with schizophrenia in the Chinese Han population. Biochem Biophys Res Commun. 2005;334(3):817–23.
102. Lo WS, Lau CF, Xuan Z, Chan CF, Feng GY, He L, et al. Association of SNPs and haplotypes in GABAA receptor beta2 gene with schizophrenia. Mol Psychiatry. 2004;9(6):603–8.
103. Ullah A, Long X, Mat WK, Hu T, Khan MI, Hui L, et al. Highly recurrent copy number variations in GABRB2 associated with schizophrenia and premenstrual dysphoric disorder. Front Psychiatry. 2020;11:572.
104. Golan D, Lander ES, Rosset S. Measuring missing heritability: inferring the contribution of common variants. Proc Natl Acad Sci U S A. 2014;111(49):E5272–81.
105. Shi H, Kichaev G, Pasaniuc B. Contrasting the genetic architecture of 30 complex traits from summary association data. Am J Hum Genet. 2016;99(1):139–53.
106. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet. 2014;94(4):559–73.
107. Szekvolgyi L, Ohta K, Nicolas A. Initiation of meiotic homologous recombination: flexibility, impact of histone modifications, and chromatin remodeling. Cold Spring Harb Perspect Biol. 2015;7(5):a016527.
108. Che J, Smith S, Kim YJ, Shim EY, Myung K, Lee SE. Hyper-acetylation of histone H3K56 limits break-induced replication by inhibiting extensive repair synthesis. PLoS Genet. 2015;11(2):e1004990.
109. Celic I, Verreault A, Boeke JD. Histone H3 K56 hyperacetylation perturbs replisomes and causes DNA damage. Genetics. 2008;179(4):1769–84.
110. Hull RM, Cruz C, Jack CV, Houseley J. Environmental change drives accelerated adaptation through stimulated copy number variation. PLoS Biol. 2017;15(6):e2001333.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.