Supplementary Material

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

Xi Long\textsuperscript{1,2} and Hong Xue\textsuperscript{1,2,3,*}

\textsuperscript{1} Division of Life Science and Applied Genomics Centre, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

\textsuperscript{2} HKUST Shenzhen Research Institute, 9 Yuexing First Road, Nanshan, Shenzhen, China

\textsuperscript{3} Centre for Cancer Genomics, School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing, Jiangsu, China

* Corresponding author:
Email: hxue@ust.hk (HX)

Contents:
- Supplementary Figures S1 to S14
- Captions of Supplementary Tables S1 to S11
- Description of Supplementary Dataset S1
- Supplementary reference
Supplementary Figure S1. Size-classification of germline CNV regions retrieved from dbVar database. Based on the log_{10}-transformed length distribution, CNVs are classified into SCNV, MCNV, LCNV and ECNV based on the critical points (indicated by arrows) of the polynomial regression curves colored in orange, green and red respectively.
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. The pairwise correlation coefficients are expressed in accordance to the red-blue thermal scale with blue square representing positive $r_s$ and red square representing negative $r_s$. The nature and clustering patterns of these genomic features enable their classification into the Genic zone (blue), Proximal zone (green), Distal zone (red) or marker (black) features. Markers including GWAS, CNVT and ClinVar which are allocated to the Genic zones on account of their propensities for positive co-localization with features in these zones.
Supplementary Figure S3. Refinement of hotspot boundary by weighted density. Top panel: local concentration of SCNV breakpoints in three 10-bp bins on chromosome 1. Middle panel: hotspot identified using SCNV density as criterion. Bottom panel: hotspot region identified using weighted SCNV density as criterion in sliding windows. In middle and bottom panels, windows with SCNV densities above threshold (red dashed line) are identified as hotspot (red bins) in sliding windows. With the weighting scheme described in Methods, sliding windows overlapping with the SCNV peaks are assessed with different levels of weighted SCNV density based on the location of SCNV peaks in each sliding window, enabling thereby the refinement of hotspot boundary.
SNP hotspot detection

(i) Divide the genome into 1-kb sliding-windows 10-bp apart

(ii) Window-wise quantification of weighted SNP density ($D_{\text{win}}$)

(iii) Separate windows into two parts

Windows entirely inside the specified type of zone:

Top windows = the top-ranked windows the sum of SNP entries in which reach but not exceed 5% SNPs in the zone. $D_{\text{min}}$ = the minimum $D_{\text{win}}$ of the top windows.

| Rank of window | Weighted SNP density ($w_{\text{win}}$) | Cumulative SNP fraction % |
|----------------|----------------------------------------|----------------------------|
| 1              | 0.9                                    | 0.0001                     |
| 2              | 0.88                                   | 0.0002                     |
| 3              | 0.87                                   | 0.0003                     |
| ...            | ...                                    | ...                        |
| $i$            | $D_{\text{min}}$                       | $\leq 5$                   |
| $i + 1$        | $D_{i+1}$                              | $> 5$                      |
| $N$            | $D_N$                                  | 100                        |

Windows partially inside the specified type of zone:

Top windows = the top-ranked windows the $D_{\text{win}}$ of which equal or exceed $D_{\text{min}}$.

| Rank of window | Weighted SNP density ($w_{\text{win}}$) |
|----------------|----------------------------------------|
| 1              | 0.92                                   |
| 2              | 0.89                                   |
| 3              | 0.87                                   |
| ...            | ...                                    |
| $k$            | $D_k \geq D_{\text{min}}$              |
| $k + 1$        | $D_{k+1} < D_{\text{min}}$             |
| $N'$           | $D_{N'}$                               |

(iv) Part I

Windows entirely inside the specified type of zone

(v) Merge overlapping top windows in red

34,487 SNP hotspots with mean width of 1,994 bp (SD = 4,528 bp), amounting to 2.54% of the autosomal region analyzed

Supplementary Figure S4. Flowchart of protocol for identifying SNP hotspots. The protocol is described in “Density-based genetic-variant hotspots determined by weighted sliding windows” in Materials and Methods.
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). Densities in simulated regions with matching minor allele frequencies are shown by the grey violin plots. A combined result of the ten groups is shown in Fig. 3d.
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). The \( p \)-value and effect size (Cohen’s \( d \)) are shown above each pair of 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). The p-value is shown above each pair of bars. The allele frequencies of SNP and SID is retrieved from 1000 Genomes Project. Recurrency is employed as allele frequency for SCNV, MCNV, LCNV, ECNV and SDP breakpoints. MST lacking allele frequency information is not shown in this figure.
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). The horizontal dashed lines indicate autosomal averages of the retrotransposon subfamilies. The 6 SVA subfamilies and 15 AluY subfamilies shown in the left panel are merged into 5 larger groups in Fig. 4.
Supplementary Figure S9. 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) plus the linker DNA (up to 80 bp). ‘Av.’ stands for average CNV length in bp as indicated by the dashed vertical orange or blue line.
Supplementary Figure S10. Genomic-feature contents of total (first column in upper panel) and 23 kinds of hotspot clusters with distinct hotspot-compositions. The numbers and average sizes of each kind of cluster are indicated in the bottom two rows of each column. Fold-changes of the density/intensity of the genomic features are expressed similar to those in Fig. 6.
Fold-enrichment in hotspots located in Distal S4 & G2

No. genes in hotspots located in Distal S4 & G2

5 genes  118 genes

Supplementary Figure S11. Enrichment Map for Distal-zone-enriched genes annotated using g:Profiler based on Gene Ontology biological process and KEGG. Each circular node represents a pathway (with 3 to 350 genes) significantly enriched in Distal zones with Benjamini-Hochberg false discovery rate < 0.05, and the node size is proportional to the number of Distal-zone genes belonging to the pathway, as illustrated by the two circle signs of ‘5 genes’ and ‘118 genes’. The pathway IDs inside the nodes correspond to those given in the table on the right. Pathways are connected by a grey edge when they share ≥ 50% genes. Enrichment of these pathway genes in the GV hotspots, represented by the red-blue thermal scale and shown in the ‘Fold’ column of the table, is estimated as fold-change of the fraction of the genes relative to the fraction of 1-kb windows that overlap with GV hotspots located in the S4- and G2-phase windows of Distal zones. Supplementary Table S11 gives the description of all the pathways shown in the figure.
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). The \( p \)-value are shown above each pair of bars. The risk allele frequency, odds ratio and heritability of GWAS Catalog SNPs (file date 2020.01.01) are estimated using GWEHS application (López-Cortegano and Caballero 2019) with all default settings and a prevalence value of 1%.
**Supplementary Figure S13. Co-localization among different kinds of genetic variants.**

Enrichment of y-axis GVs at different distances from x-axis GVs, ranging from 0 bp to ± 500 bp in 50-bp increments. Enrichment of each y-axis GV is expressed by its density fold-change relative to its autosomal average density. Small indels (SID) are further separated into small insertions (SINS) and small deletions (SDEL). In addition to the eight kinds of GVs employed in hotspot detection, the variations analyzed in figure also include the ‘calls’ of other eight kinds of structural variations given in dbVar database.
**Supplementary Figure S14. Distribution of GVs around the telomere ends.** GV densities are measured in 5-Mb sequence windows located near the start of the short arm (left) and end of the long arm (right) of 22 autosomes. Autosomal average density of each kind of GV is indicated by the dashed line. Error bar represents 95% CI. Density at each distance is compared to that of all 5-Mb windows in 22 autosomes using one-tailed t-test. Red or blue asterisks represent significant (Benjamini-corrected $p < 0.05$) enrichment or depletion of GV relative to its autosomal average.
Captions of Supplementary Tables
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 retrotansposons.
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.

Descriptions of Supplementary Dataset
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.

Supplementary reference
López-Cortegano E, Caballero A (2019) GWEHS: a genome-wide effect sizes and heritability screener. Genes 10:558 doi:10.3390/genes10080558