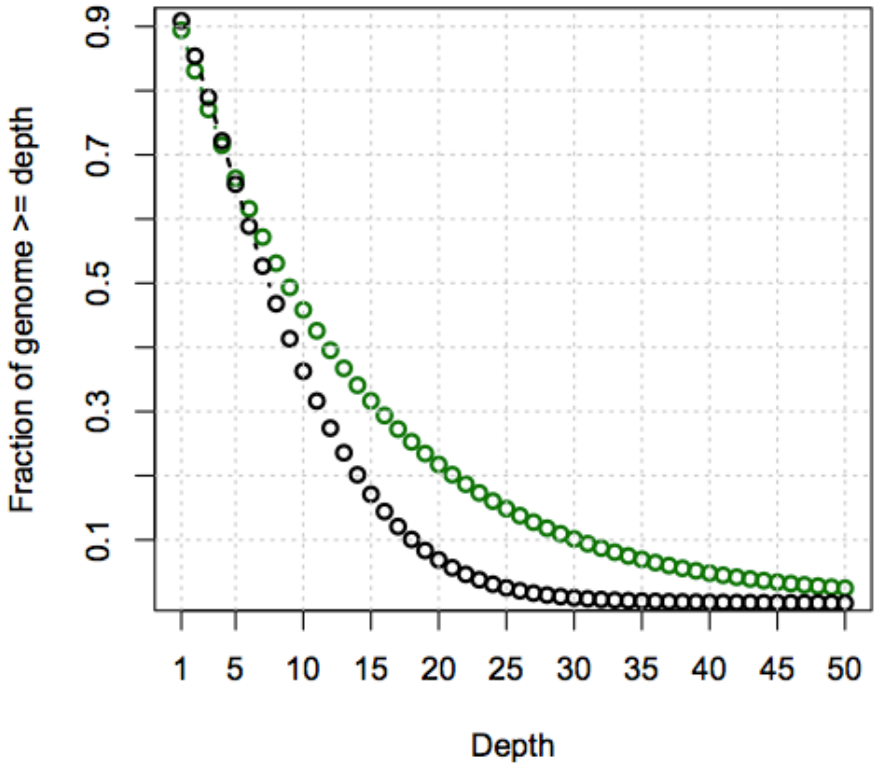
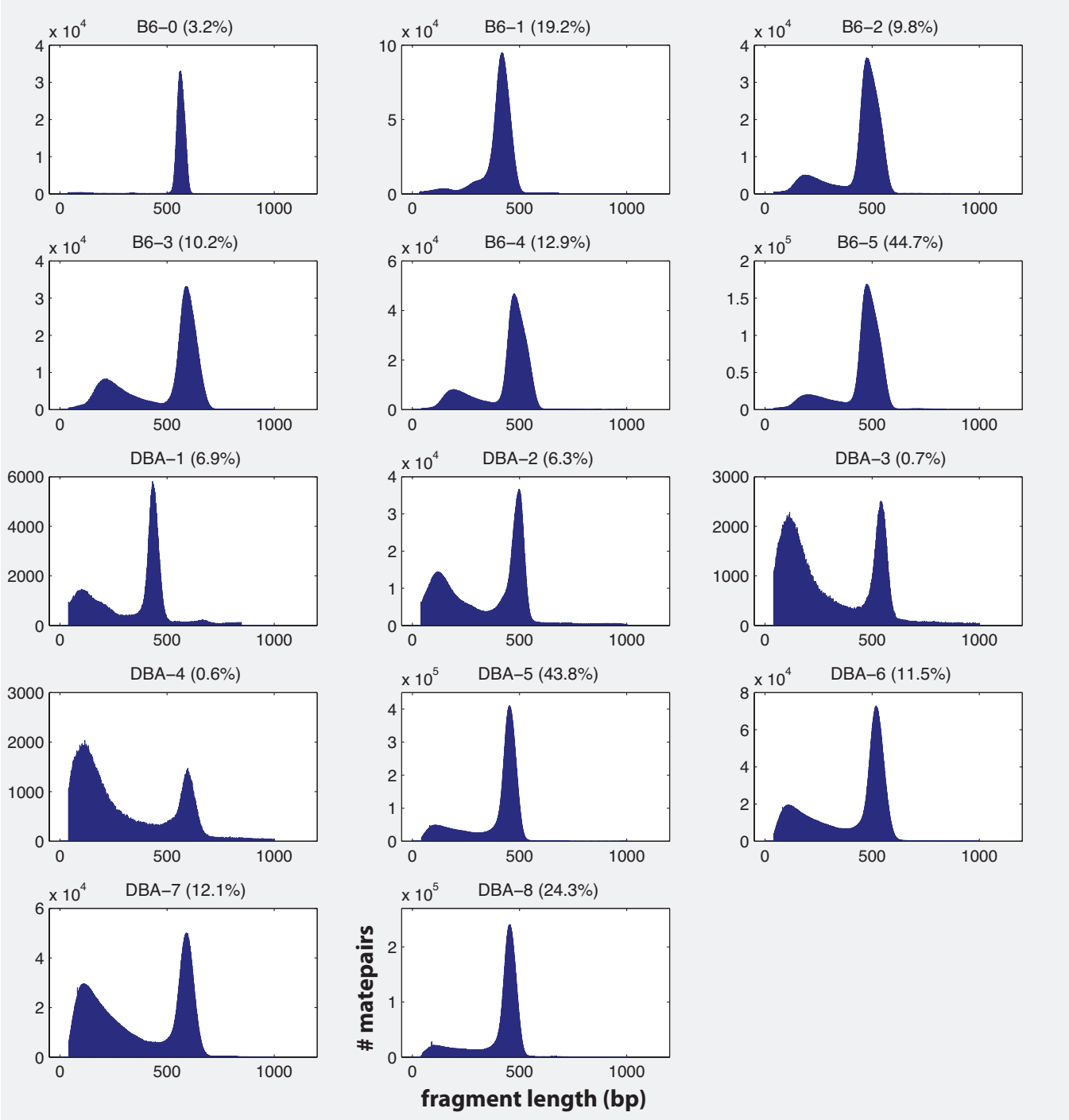
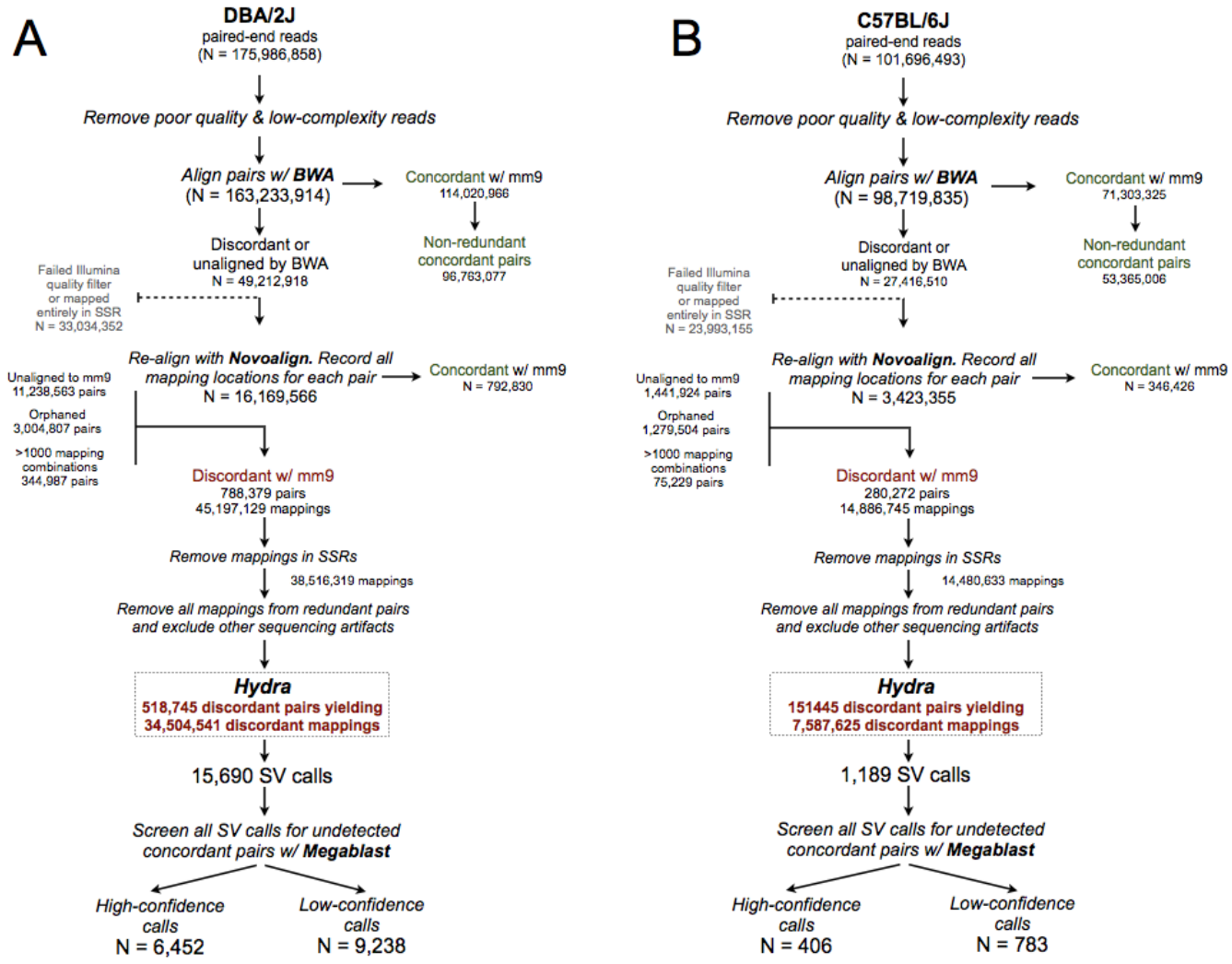
Supplementary Figures



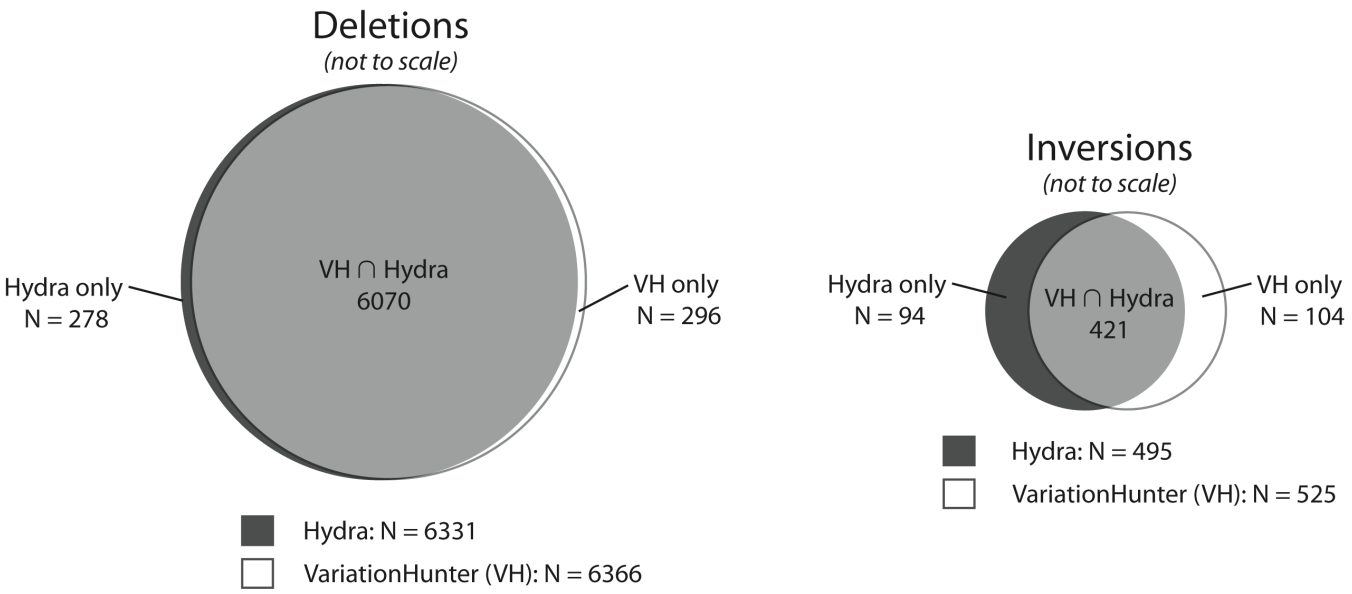
Supplementary Figure 1: Illumina sequence coverage of the B6 and DBA genomes. The cumulative physical coverage of concordantly aligned matepairs is shown for the DBA (green) and B6 (black) genomes. Approximately 85% of each genome was spanned by at least two DNA fragments.



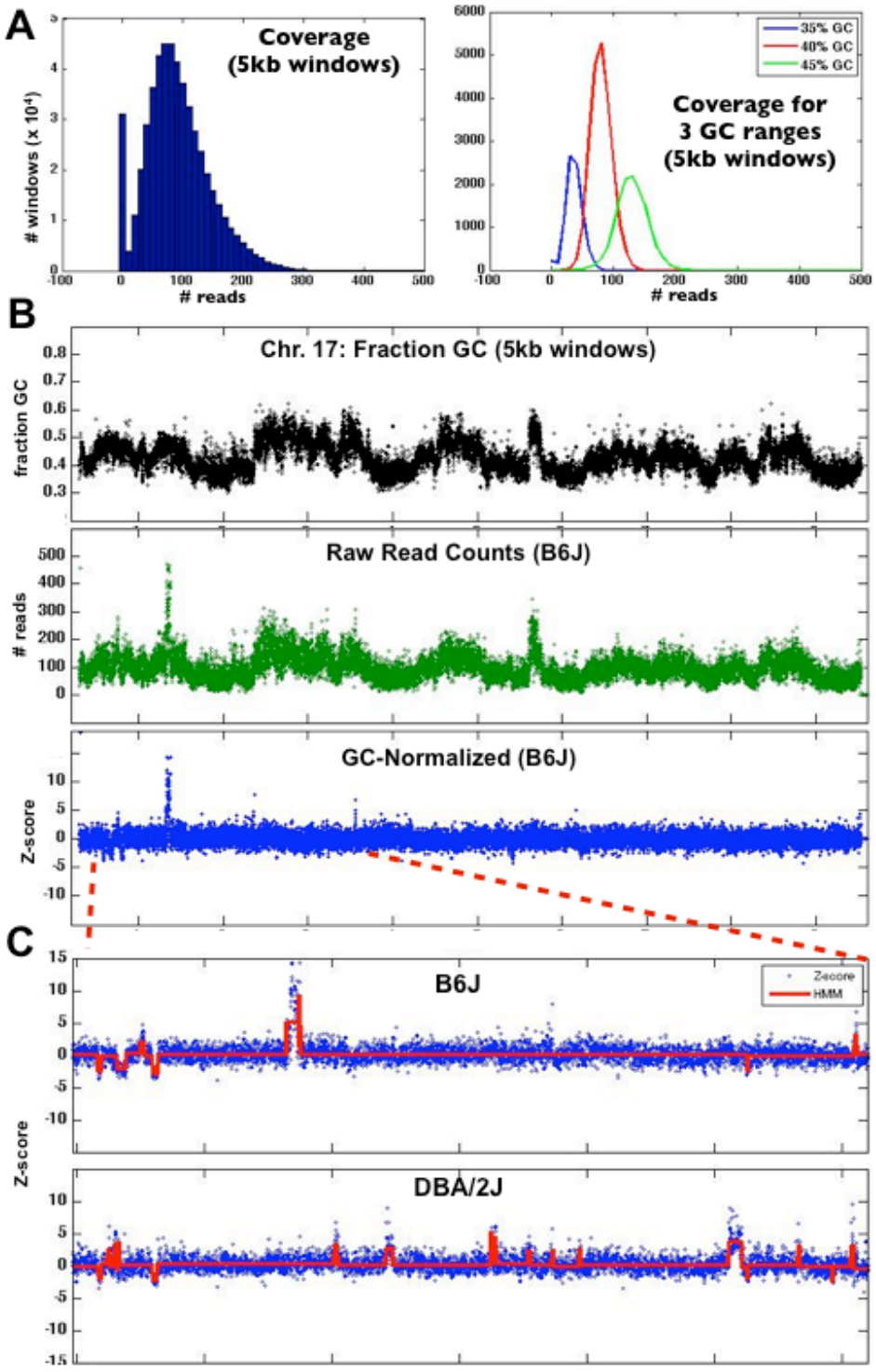
Supplementary Figure 2: DNA fragment length histograms. The distribution of DNA fragment lengths for each sequencing library used in this study is shown. The Y-axis is the length of the fragments in base-pairs and the X-axis is the number of fragments. The library identifier is shown above each histogram, and the percentage of data contributed by each sequencing library to the complete dataset for that strain is shown in parentheses. Note that the libraries with poor size distributions comprise little of the total sequence data.



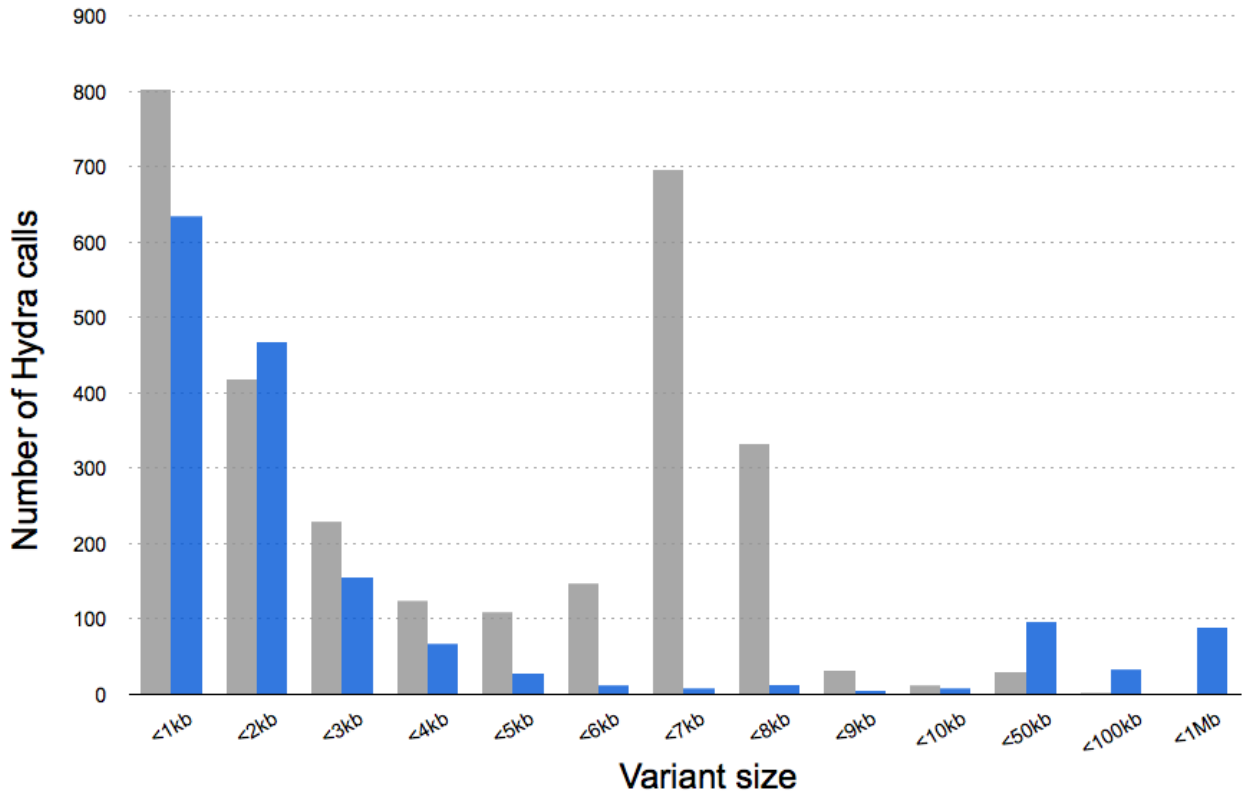
Supplementary Figure 3: SV discovery flowchart. The fates of all matepairs for both the DBA and B6 strains are shown. Matepairs were subjected to quality control and alignment by both BWA and NOVOALIGN. Using the discordant mappings from NOVOALIGN, sequencing artifacts and redundant pairs were excluded from the analysis. All remaining discordant mappings were used for SV breakpoint prediction with HYDRA. All SV calls were screened for missed concordant matepairs with MEGABLAST. High-confidence calls had no such concordant matepairs whereas low-confidence calls had at least one.



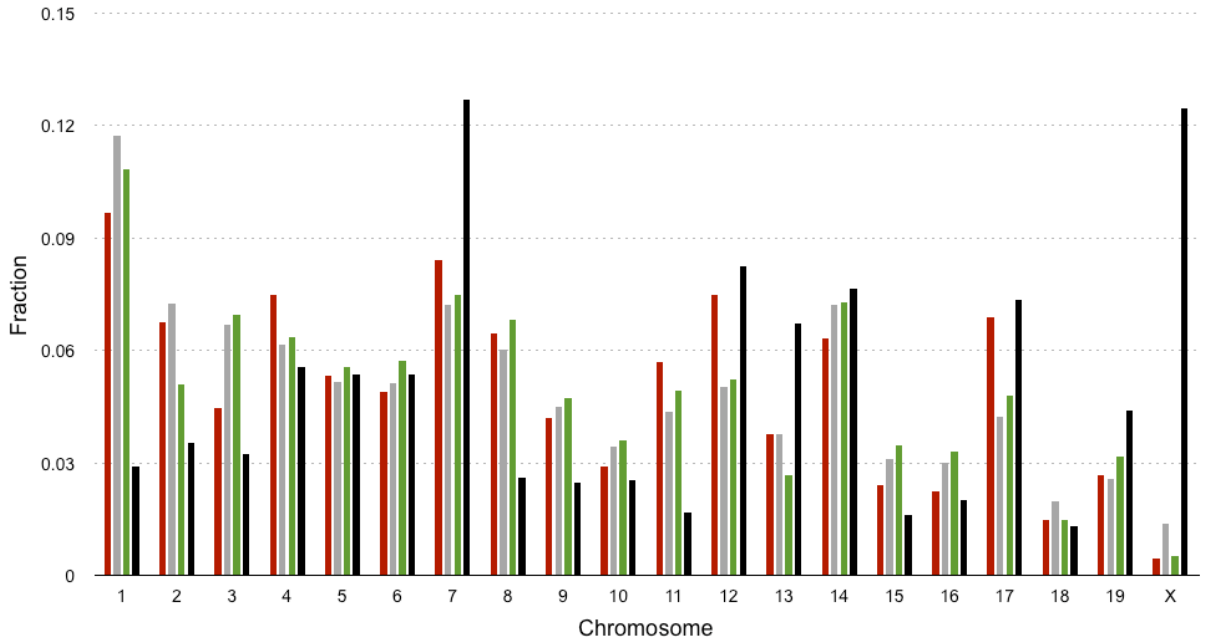
Supplementary Figure 4: Comparison of breakpoints called by HYDRA and VariationHunter. The intersection of deletion and inversion calls made by HYDRA (dark gray) and VariationHunter (VH, white) are shown. The intersection of deletions (6070) and inversions (421) are based on the number of VH calls that were also called by HYDRA. In rare cases, multiple VH calls overlapped the same *single* HYDRA call; consequently for both Venn diagrams, the intersection plus the "HYDRA only" calls is greater than the total number of HYDRA calls while the intersection plus the "VH only" calls matches the total number of VH calls.



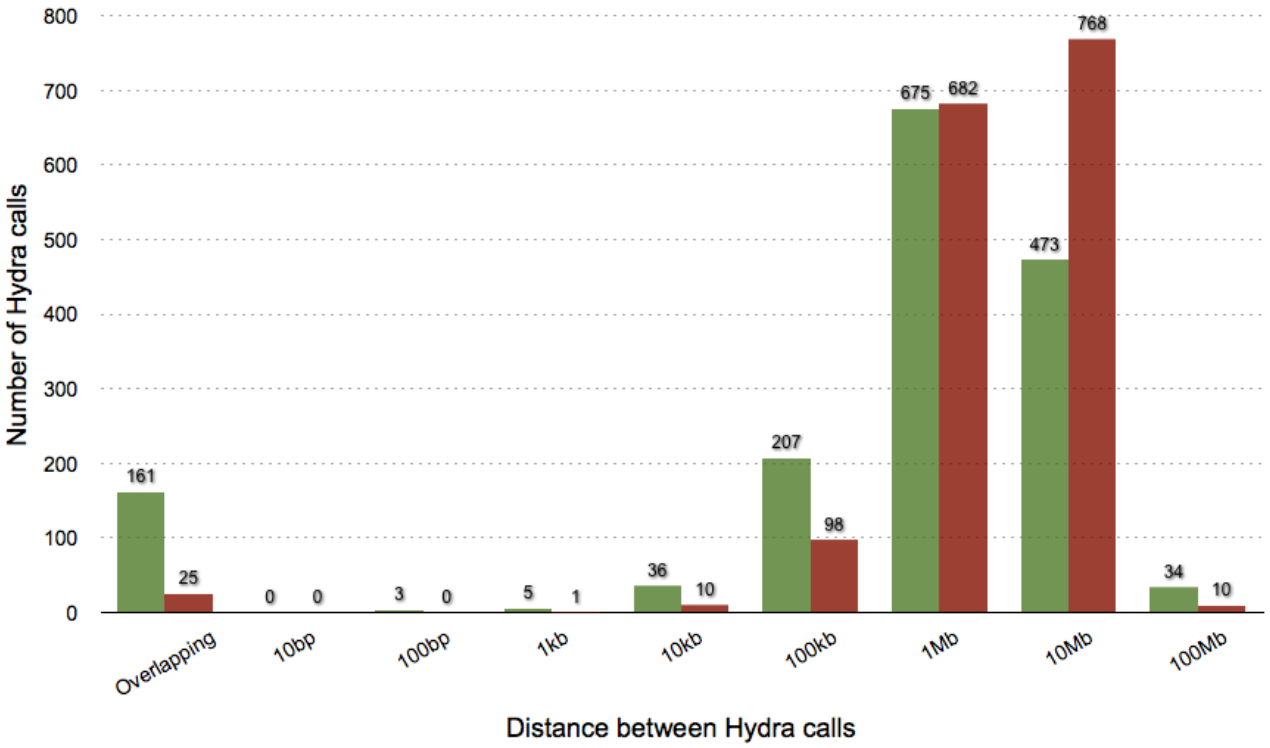
Supplementary Figure 5: Depth of coverage (DOC) analysis. (A) Distribution of read-counts in 5kb windows for B6 (left panel) and read counts for 0.5% GC ranges beginning at the indicated value (right panel). (B) GC fraction, read counts, and GC-normalized data for B6 (Chr. 17). (C) GC-normalized (blue) and HMM-segmented data (solid red lines) for B6 and DBA in a ~35Mb region of chromosome 17, as shown by the red dotted lines linking (B) and (C).



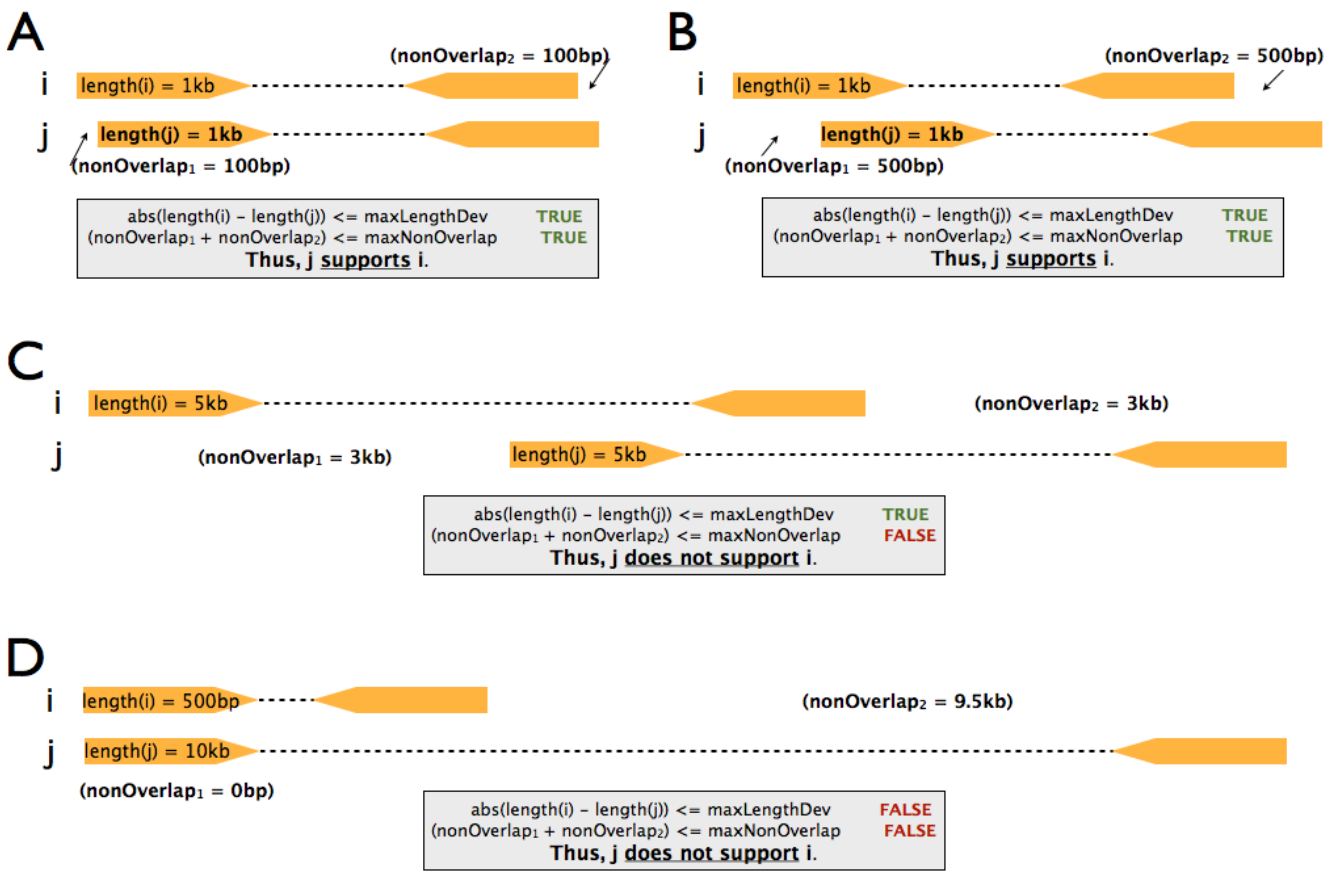
Supplementary Figure 6: Size distribution of HYDRA SV calls. The size distribution of LSV calls (blue) and TEV calls that were detected as insertions in B6 (gray) are shown. The vast majority of structural variation detected in this study is less than 5kb in size. TEV peaks between 6 and 8kb reflect LTR and LINE insertions in B6.



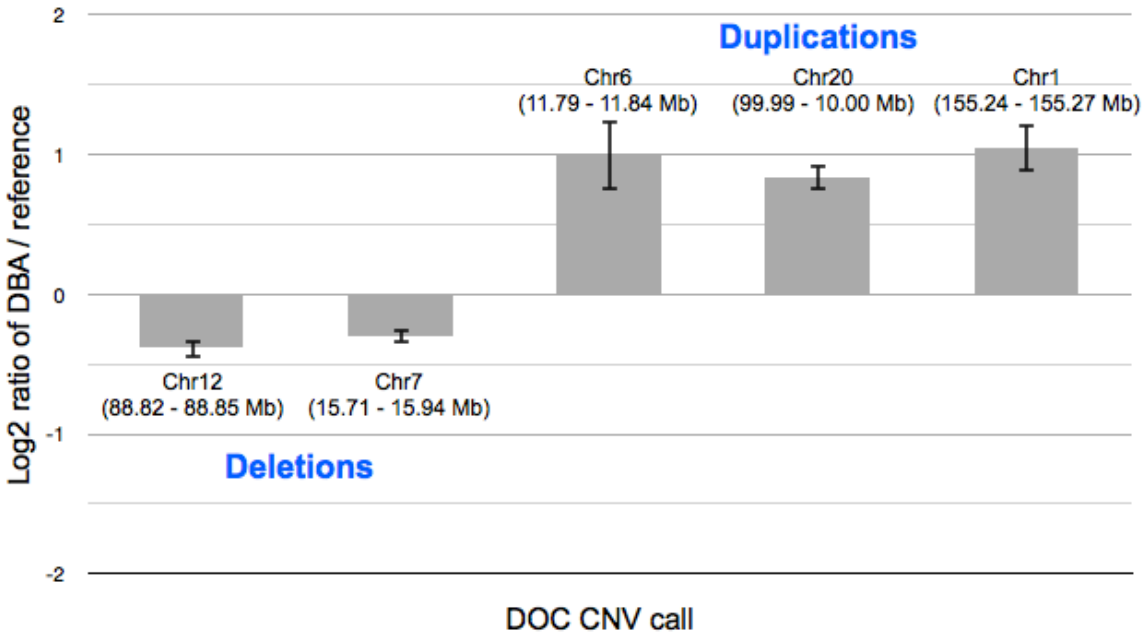
Supplementary Figure 7: Chromosomal distribution of HYDRA SV calls. The proportion of all LSV (red) and TEV (gray) found on chromosomes 1 through X are shown. The fraction of all DBA versus B6 SNPs (green) is also shown as a control for haplotype differences between the two strains. The fraction of each chromosome that is comprised of segmental duplications (black) is also shown.



Supplementary Figure 8: Genomic distance between HYDRA LSV calls. The distribution of observed distance between LSV calls (green) is compared to the expected distribution (red). The expected distribution is based on randomly-permuting the location of LSVs in the genome. Overlapping LSVs are defined as any two LSVs who share at least one bp in common based on the resolution of the HYDRA SV calls. Otherwise, the x-axis represents the genomic distance between LSVs. Distance is computed between *adjacent* LSV calls. Therefore, for each chromosome, if there are N LSV calls, then N-1 distances are computed. Thus the total number of observed distances (1594) is less than the total number of LSV calls (1616).



Supplementary Figure 9: Clustering discordant mappings into HYDRA breakpoint calls. In order to cluster individual discordant mappings into breakpoint calls, HYDRA requires the following: (1) the mappings are derived from distinct matepairs; (2) the mappings span at least 1bp in common; (3) the respective ends of each mapping have the same orientation, with the caveat that HYDRA does have an option to allow +/+ and -/- mappings to support each other for inversion detection; (4) the mappings have similar lengths within the tolerance of the DNA fragment libraries; and (5) the "non-overlap" (i.e., the total portion in each mapping that does not overlap the other mapping) between the mappings does not exceed the tolerance of the DNA libraries. In panel A and B, the lengths of mapping *i* and mapping *j* meet HYDRA's length similarity restriction (*maxLengthDev*). While there is more "non-overlap" between the two mappings in panel B than panel A, it still meets HYDRA's restriction (*maxNonOverlap*). Therefore, the respective mappings in panel A and B are said to support one another and will be clustered into a breakpoint call. In contrast, mappings *i* and *j* in panel C have similar lengths, but the "non-overlap" between the mappings exceeds *maxNonOverlap* and thus HYDRA concludes that they do not predict the same breakpoint. The mappings in panel D exceed both the *maxLengthDev* and *maxNonOverlap* thresholds and are therefore not clustered into a breakpoint call.



Supplementary Figure 10: Validation of CNV calls by quantitative PCR (qPCR). To validate novel copy number variation at randomly selected HMM calls, we performed qPCR. The data is presented as the mean $\log_2$ ratio of the DBA CT value relative to the B6 reference CT value (see Methods). Error bars indicate standard error.

Supplementary Tables

(See SupplementaryTable1.xls online.)

	Observed overlap with SD	Expected overlap with SD	Enrichment	p-Value
All Breaktigs	0.096	0.049	1.95	<1E-3
Simple; N=899	0.100	0.048	2.08	
Complex; N=272	0.081	0.053	1.54	
No Retro-insertions	0.083	0.049	1.72	<1E-3
Simple; N=880	0.091	0.048	1.90	
Complex; N=248	0.060	0.0508	1.19	

Supplementary Table 2: Breakpoint enrichment at segmental duplications (SDs). The genomic positions of all breakpoints were screened for overlap with segmental duplications. Significant enrichment for segmental duplications was observed as compared to the expected enrichment. Expected enrichments were computed based on a 1000 permutations where segmental duplications where randomly placed in the genome.