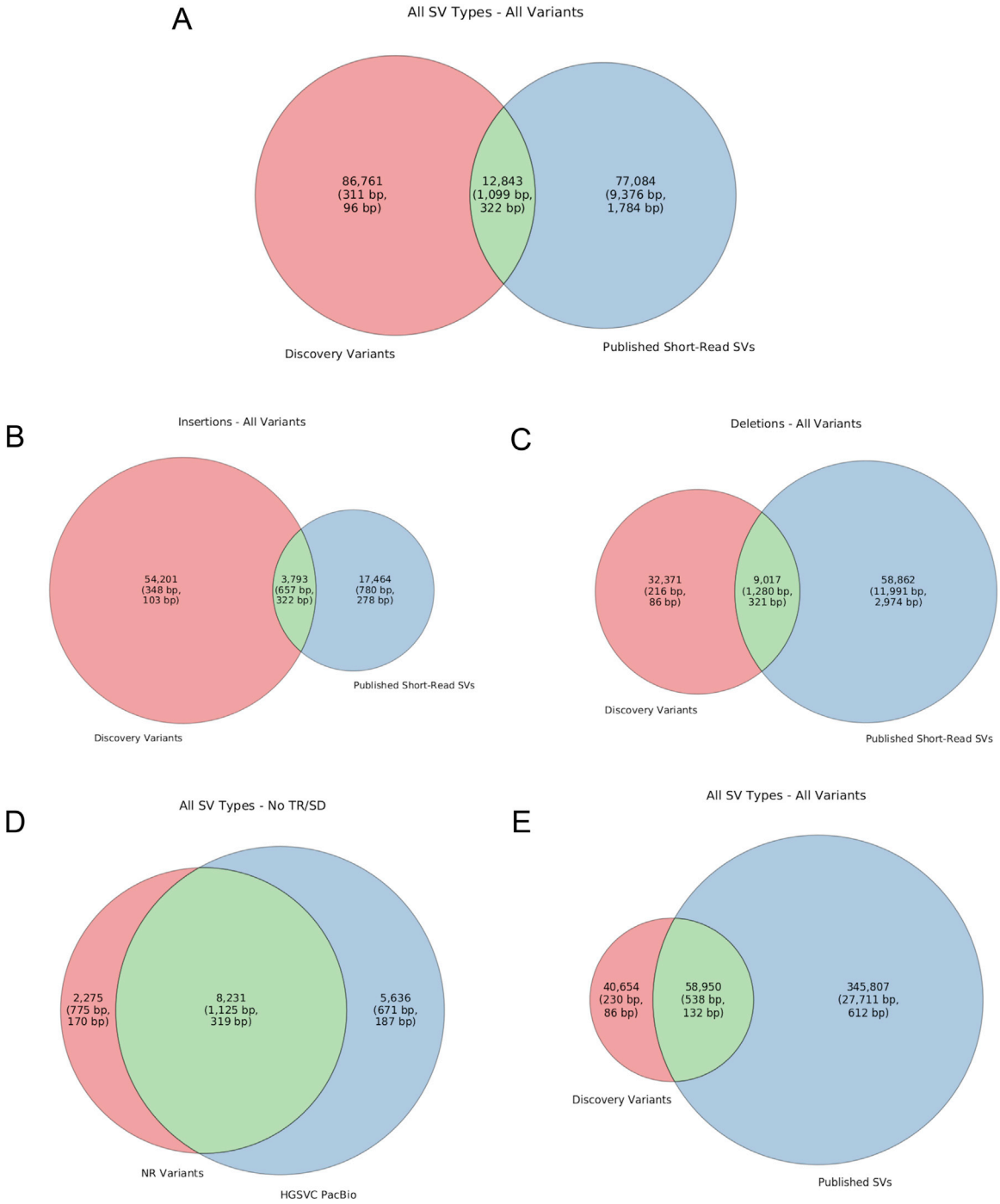

Figure S1. Comparisons with Published Callsets, Related to Figure 1

(A) Nonredundant merged variants (red) versus variants discovered using Illumina sequencing technology (blue) with some variants shared among discovery efforts (green) show a roughly equal number of variants in our 15 discovery samples as more than 3,000 short-read samples.

(B and C) (B) Compared to Illumina technology, a greater increase was observed for insertion SVs than (C) deletion SVs.

(D) Variants in our discovery set for HG00514, HG00733, and NA19240 merged as a nonredundant set ("NR Variants," red) compared with HGSVC PacBio-supported variants for the same biological samples (blue) for variants outside tandem repeats (TRs) and SDs. Our set identifies additional variation likely due to increased read depth; however, the phasing approach employed by HGSVC greatly increases sensitivity.

(E) Comparing the merged discovery set against SVs obtained from many published callsets, including both short- and long-read studies as well as dbVar, shows that almost half of the discovery set in these 15 samples is novel. Many of the variants shared with these studies are now sequence resolved for the first time.

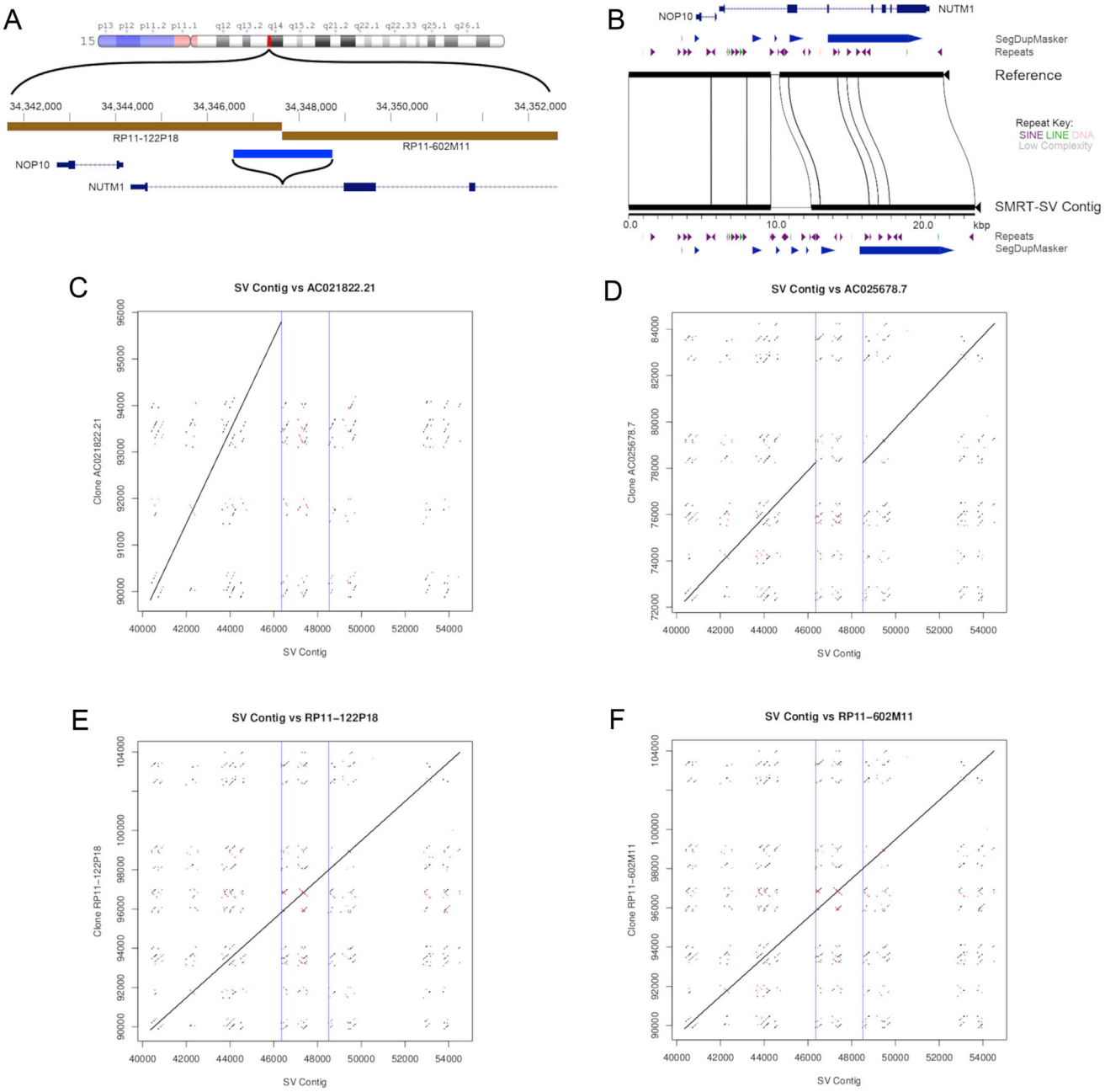

Figure S2. Resolution of a Switch-Point Genome Assembly Error, Related to Figure 5

(A) Two RP11 contigs, RP11-122P18 (AC021822.21) and RP11-602M11 (AC025678.7), were assembled in GRCh38 and a 2.2 kbp insertion (blue bar) was identified precisely at the switch-point of the two RP11 clones (gold bars) in all human genomes.

(B) Miropeats confirms that the reference (top bar) is missing additional sequence represented as an SV insertion (bottom bar) in the CHM1 contig.

(C) A dot plot of the CHM1 assembly (x axis) against RP11-122P18 (AC021822.21, y axis) shows that the contig was truncated at the point of insertion.

(D) A dot plot of the CHM1 assembly (x axis) against RP11-602M11 (AC025678.7, y axis) shows that the contig is continuous over the switch-point, but it does not contain the inserted sequence.

(E) A dot plot of CHM1 (x axis) against a new assembly of RP11-122P18 (y axis).

(F) A dot plot of CHM1 (x axis) against a new assembly of RP11-602M11 (y axis).

Both new assemblies are created by deeply sequencing the clone insert, which corrected both the truncated assembly and the 2.2 kbp of missing sequence.

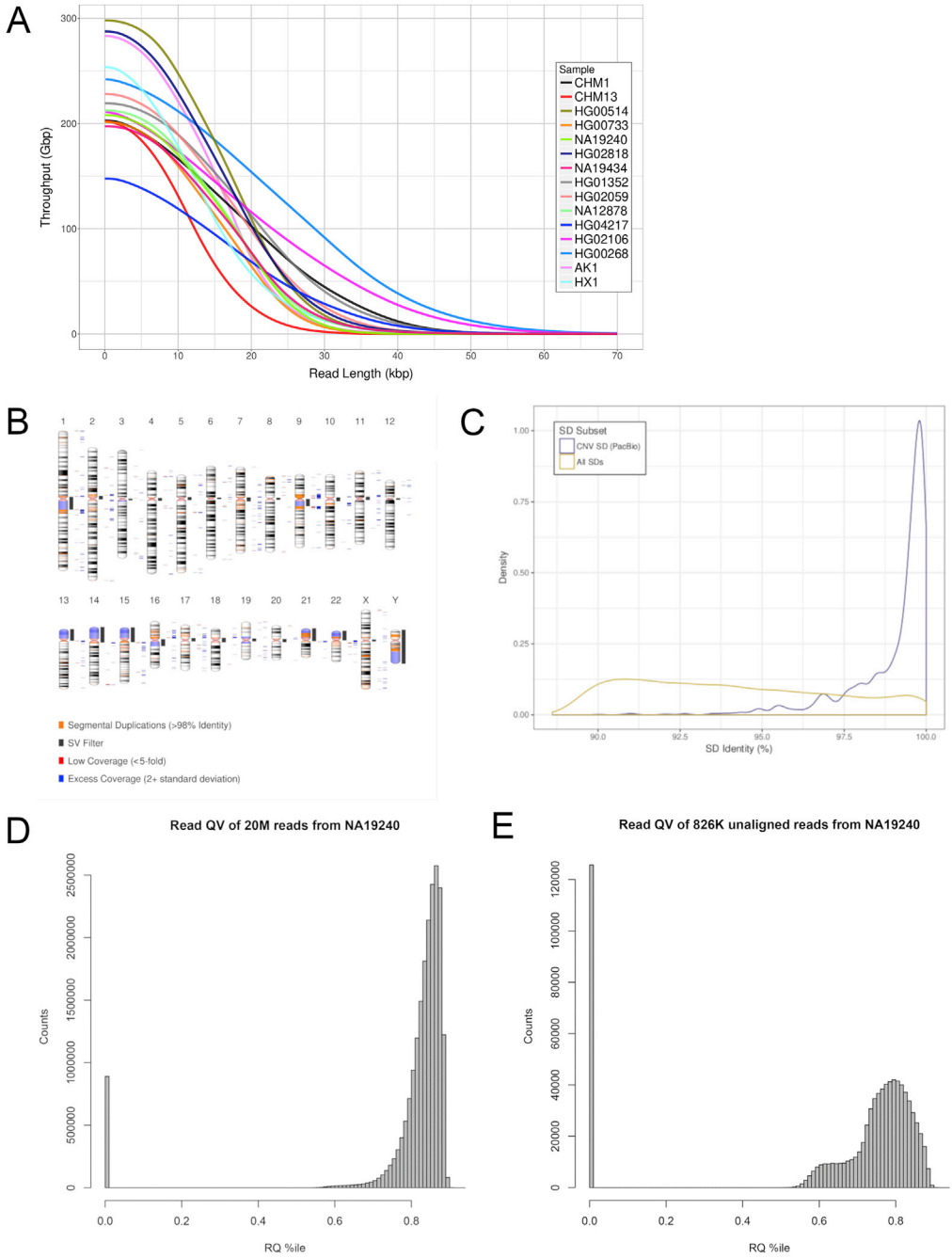

Figure S3. Sequence Data and Alignment Characteristics, Related to STAR Methods

(A) Cumulative distribution of total bases (vertical axis) over read length (horizontal axis) for long-read samples. All samples have at least 119 Gbp (37-fold coverage) in reads 10 kbp+. All but one (CHM13) have at least 57 Gbp in reads 20 kbp+ (17-fold coverage).

(B) Long reads cover a majority of the genome. Locations of SDs (orange, on chromosomes), the centromeric and pericentromeric filter (gray), regions of low mapping coverage (red), and regions of excess mapping coverage (blue). 10+ kbp regions with less than fivefold depth in at least one sample are defined as low coverage (2.2 Mbp). 10+ kbp regions with mapping depth 2+ standard deviations are defined as excess coverage (39.4 Mbp).

(C) Elevated copy number is enriched for high-identity duplications. SD identity with another region in the reference for all SDs (yellow) and SDs intersected by copy number variant (CNV) regions merged from all samples (purple) discovered using PacBio alignments are shown.

(D and E) (D) Quality distribution for all NA19240 reads and (E) all unmapped NA12940 reads. The set of ~20 million PacBio reads had an RQ mean value of 0.8 ± 0.18, while the ~820K unaligned reads have an RQ of 0.65 ± 0.28. These two distributions were significantly different from each other (Wilcoxon rank-sum test $p < 2.2 \times 10^{-16}$).

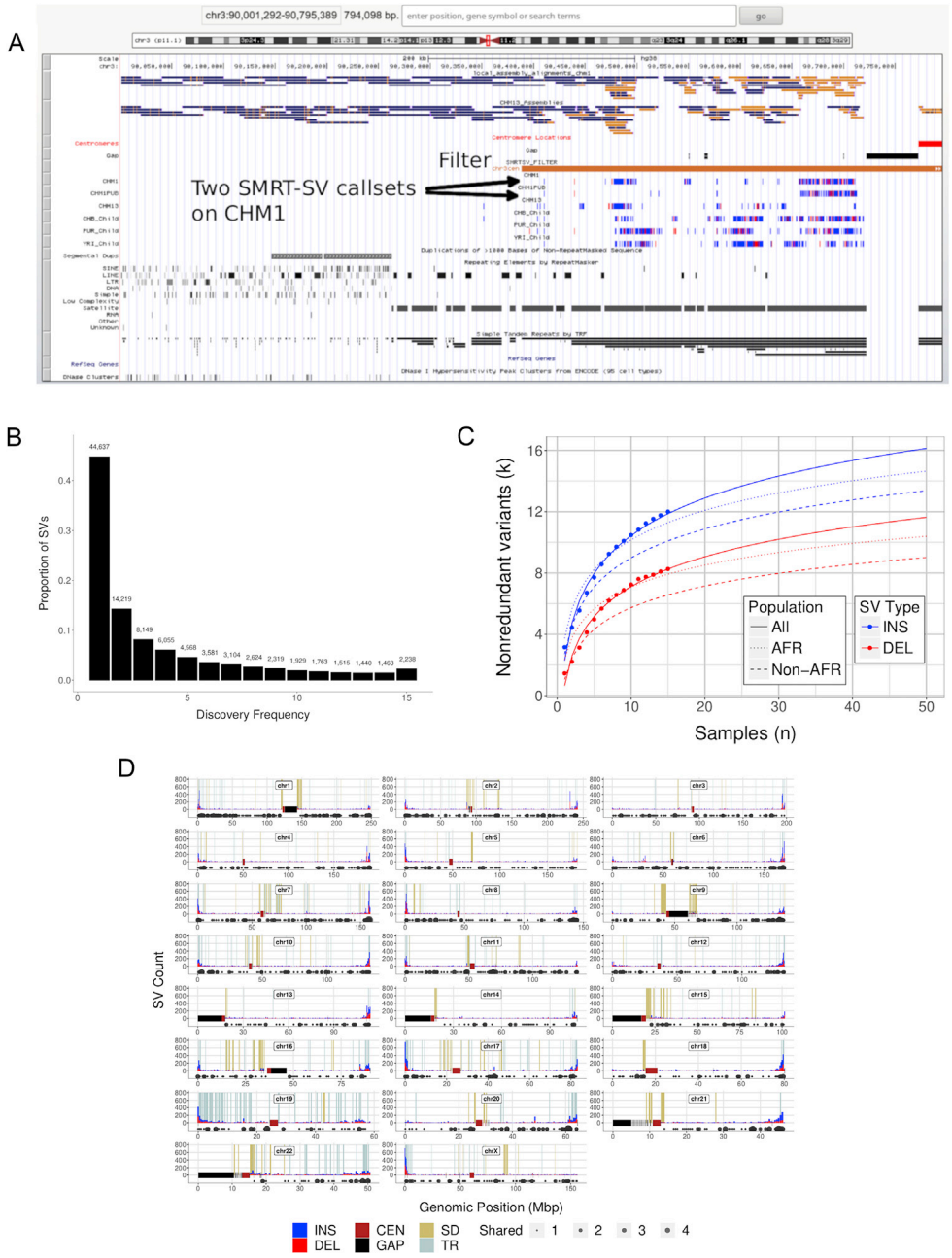

Figure S4. SV Discovery Trends, Related to Figure 1

(A) Low-confidence filter on 3p. Pericentromeric region of chromosome arm 3p with local assemblies for two CHM1 callsets (top; blue bars with yellow indicating alignment mismatches and gaps), GRCh38 modeled centromere (red), gaps (black), and filter (brown). SVs for several callsets (two CHM1, CHM13, HG00514, HG00733, and NA19240) are shown with insertions in blue and deletions in red. SDs, RepeatMasker, and TRF are also shown. From the right, there are no calls in the centromere or gaps. Moving further to the left, both CHM1 callsets show a pattern of very dense SV calls, and they exhibit a clearly differing pattern. These differences are associated with large α-satellite repeats.

(B) The discovery frequency for all SVs in the nonredundant set shows an expected distribution where most variants are rare and fewer variants are found in a large number of samples. There is a noticeable increase for variants in the final bin (shared, discovered in all samples).

(C) Log regression models showing the expected size of the merged SV set (vertical axis) given the number of samples (horizontal axis). For this analysis, we excluded tandem repeats and SDs due to high mutations rates, which leads to identity by state rather than by descent.

(D) SV counts in 500 kbp bins for insertions (blue) and deletions (red) are shown along each human chromosome. Annotations include the centromere (dark red), genome gaps (black), SDs (gold), and tandem repeats (TR, light blue) on the background. Locations with shared variants are shown as bubbles below the chromosome.

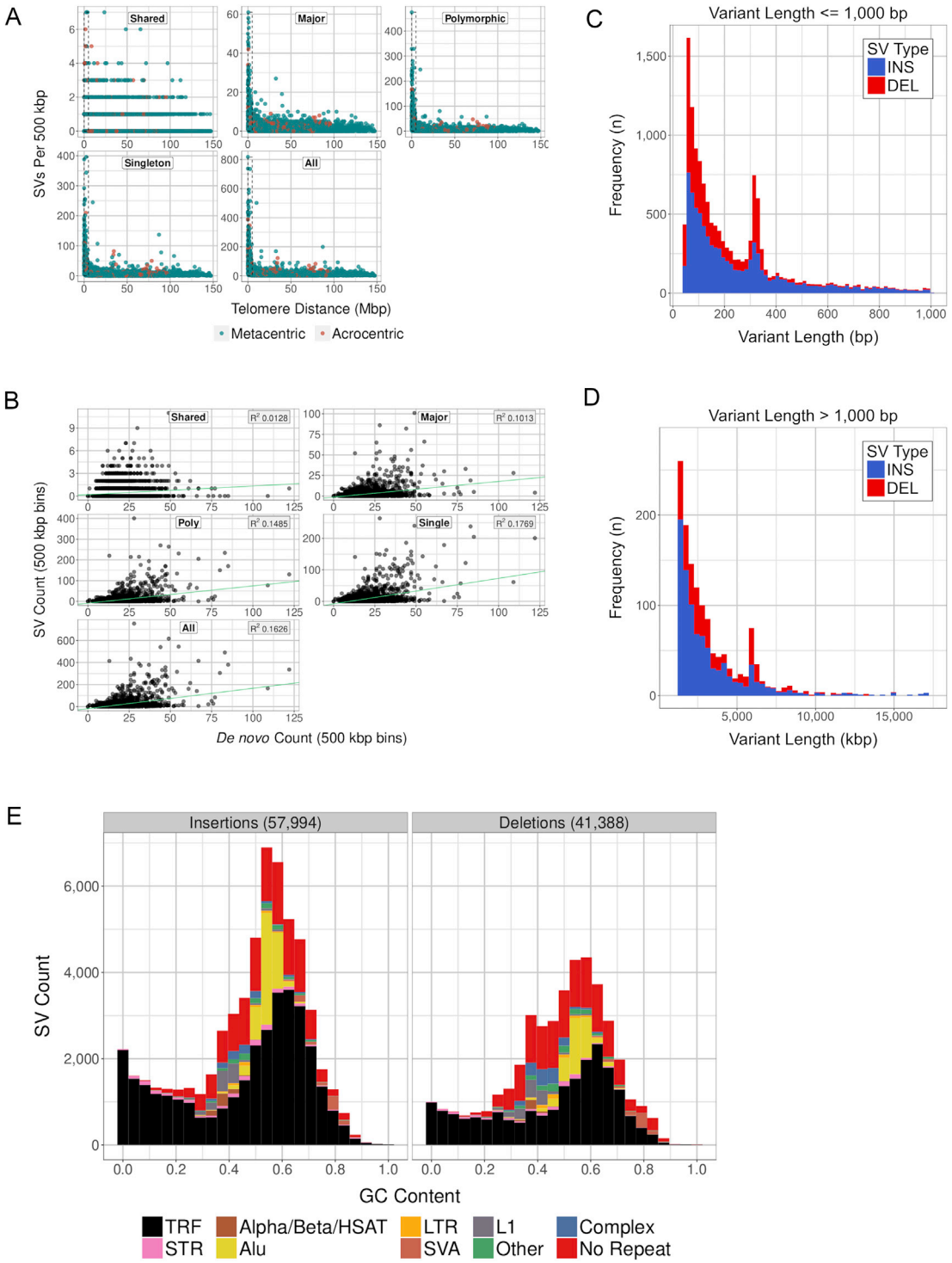

Figure S5. SV Genomic Distributions, Related to Figures 1 and 3

(A) SV discovery is biased toward chromosome ends. The distance from the SV to the end of the chromosome arm is shown for all classes of variants in 500 kbp bins over each chromosome arm. Dashed boxes are drawn around variants within 5 Mbp of the chromosome end.

(B) SV discovery is biased toward sites of *de novo* mutations. Variant counts per 500 kbp bins for *de novo* SNVs (x axis) versus SVs (y axis) show a modest correlation in all discovery classes.

(C) Insertion (blue) and deletion (red) SVs for variants 1 kbp or less decline in frequency with larger SV length with a substantial increase around 300 bp (SINE elements).

(D) Insertion (blue) and deletion (red) SVs for variants larger than 1 kbp also reflect a declining trend with an increase around 6-7 kbp (LINE elements).

(E) In the GC distribution of SVs, there is a clear skew toward low GC content, which is driven by tandem repeats, and it is especially prominent for insertions.

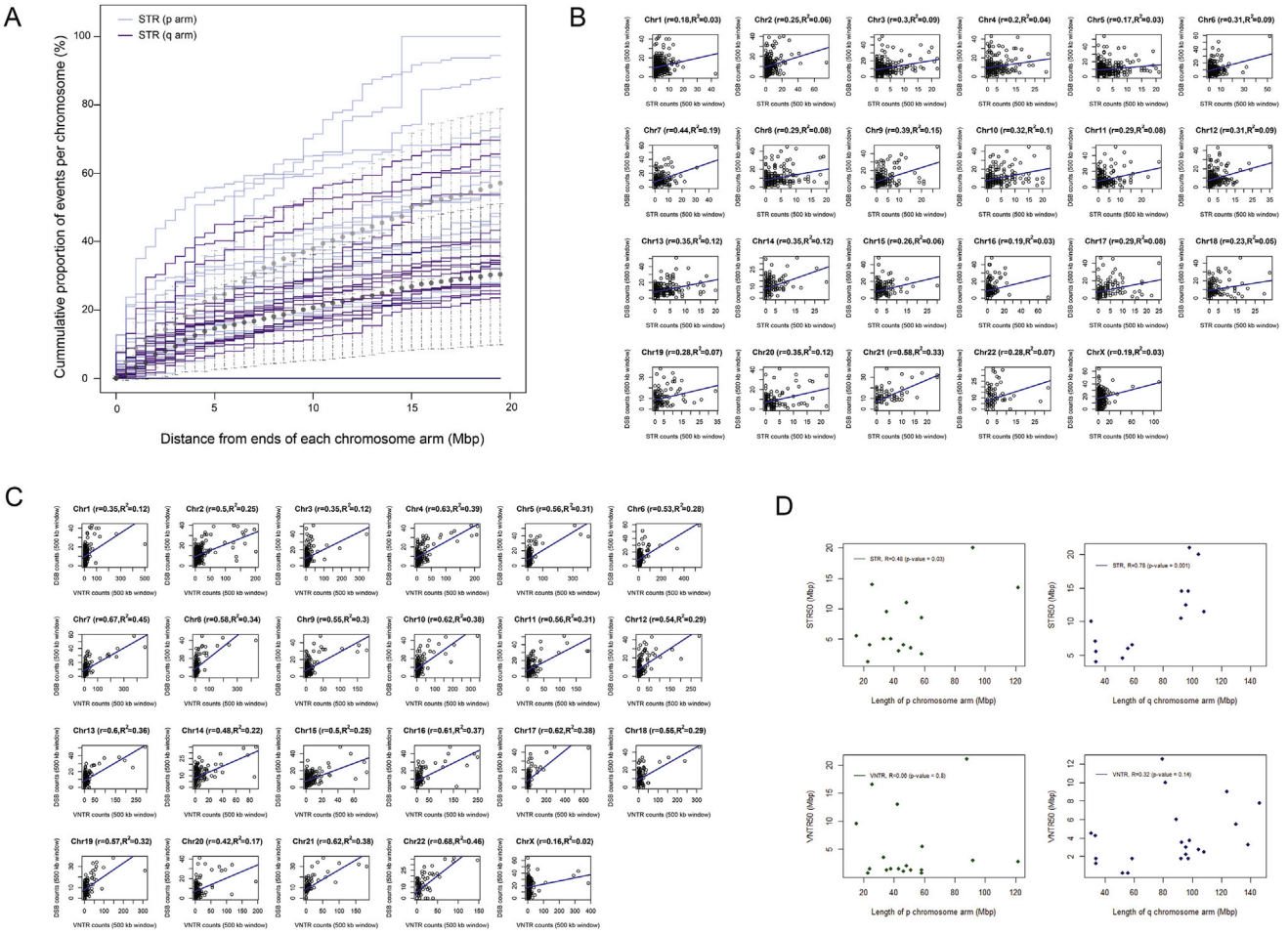

Figure S6. STR and VNTR Distributions, Related to Figure 2

(A) Chromosomal STR distribution. The cumulative abundance for STRs over each chromosome arm is shown. Light and dark colors represent p and q arms, respectively. Chromosomes 1 through X are represented in the plot, with the short arms of acrocentric chromosomes 13, 14, 15, 21, and 22 represented by the horizontal lines at the bottom of the plot. The light- and dark-gray dots represent the genome-wide average across non-acrocentric p and q arms, respectively. Windows of 500 kbp sliding from telomere ends to the centromere were used to count STRs cumulatively. The x axis is truncated at 20 Mbp.

(B) STR enrichment by chromosome. The linear relationship is shown between double-strand breaks (DSB) and STR density, across all chromosomes. The strongest linear relationship between VNTR and DSB density was observed on chromosome 21 ($R^2 = 0.33$). The density value was defined as the total number of events in a 500 kbp window.

(C) VNTR enrichment by chromosome. The linear relationship is shown between DSB and VNTR density, across all chromosomes. The strongest linear relationship between VNTR and DSB density was observed on chromosome 22 ($R^2 = 0.46$). The density value was defined as the total number of events in a 500 kbp window. These correlations were stronger in the case of VNTRs than STRs.

(D) The subtelomeric STR and VNTR abundance as a function of chromosome arm length. We defined STR50 and VNTR50 as the physical position on a chromosome arm that separates evenly all events (i.e., STRs or VNTRs) between the distal and proximal portions of the arm. The Pearson-correlation coefficients between total arm length and STR50 were 0.48 ($p = 0.03$) and 0.78 ($p = 0.001$), while for VNTR50 the coefficients were 0.06 ($p = 0.8$) and 0.32 ($p = 0.14$) for p and q arms, respectively. This indicates that the chromosome arm length differences explain the majority of subtelomeric enrichment for STRs; however, VNTR subtelomeric enrichment is not accounted for by chromosome arm length differences. Interestingly, the p arm seems to have a stronger VNTR subtelomeric enrichment than the q arm, which is not explained by the chromosome arm length differences.

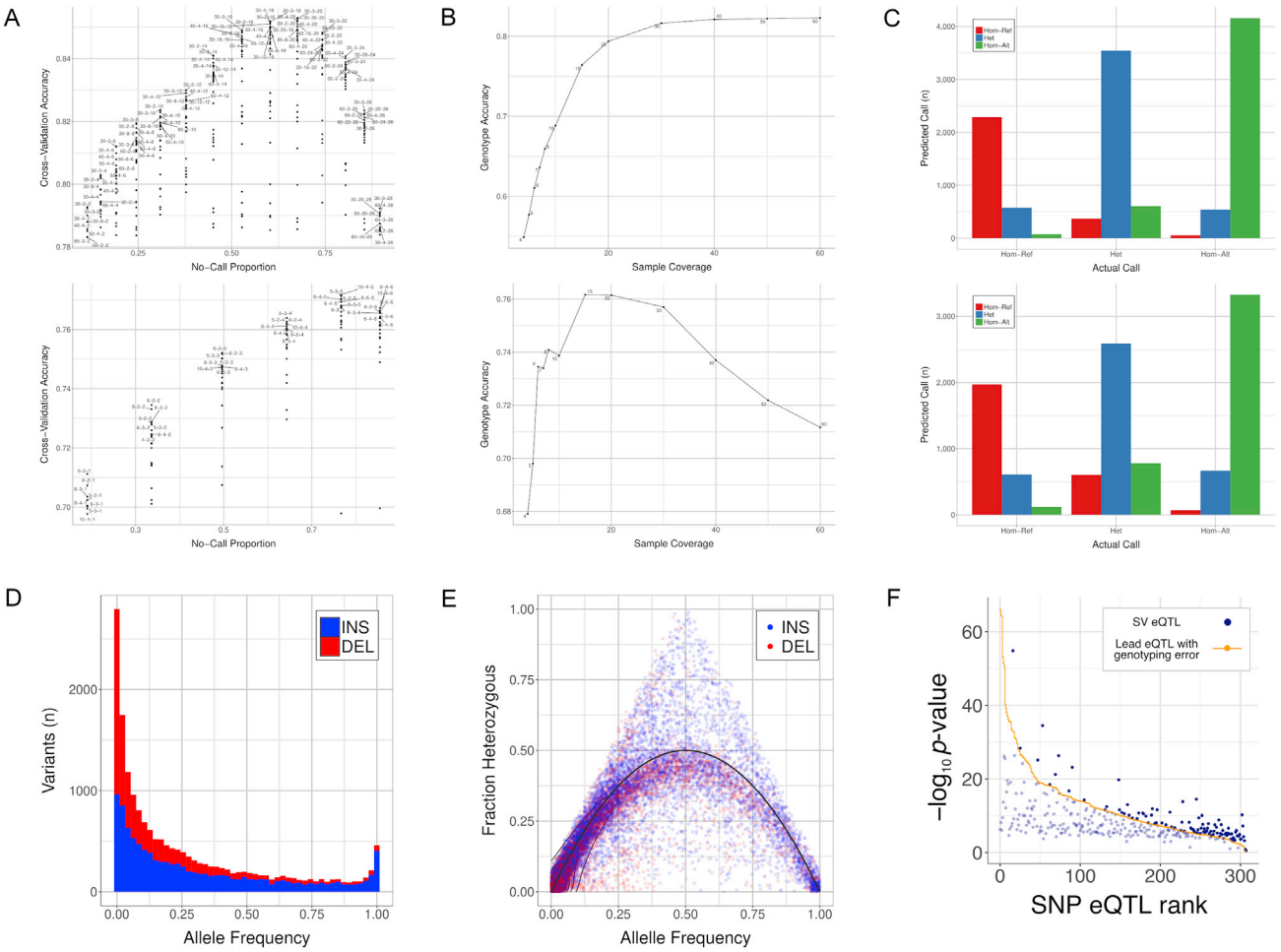

Figure S7. Genotype Model Training and Performance, Related to STAR Methods

(A) To optimize genotyping, several models were trained and tested targeting 30-fold (top) and 6-fold (bottom) samples using 30-fold and 8-fold pseudodiploid samples, respectively. No-call cutoff values tested were 2, 4, 6, …, 30 for 30-fold and 1, 2, 3, …, 6 for 6-fold yielding a proportion of variants that were not called in each test (horizontal axis), and the accuracy was calculated for the remaining variants (vertical axis). The top eight models for each no-call cutoff value are labeled with the training sample coverage (first number), training sample no-call cutoff (second number), and test sample no-call cutoff (third number). The top models (first two numbers) were not greatly affected by the no-call cutoff.

(B) Models selected for 30-fold (top) and 6-fold (bottom) samples were further tested to see how well they generalized when the sequencing coverage varies. We scaled the no-call cutoff with read depth by setting the value at ~25% of the expected read depth (e.g., 15 for the 60-fold sample). For the 30-fold model, the accuracy increases with read depth even though it was trained on 30-fold data. Although the 6-fold model accuracy declines after 15-fold, it is more accurate at lower coverage than the 30-fold model.

(C) We quantified misclassifications for the 30-fold (top) and 6-fold (bottom) models with cross-validation. Known calls (horizontal axis) are categorized by the model's prediction (vertical axis). Genotype error is mostly attributable to miscounting one allele.

(D) The genotype allele frequency distribution confirms that the majority SVs are rare in the human population and that a small number are fixed. Generated using SVs outside tandem repeats and SDs where the callable frequency was 20% or greater.

(E) The fraction of heterozygous variants (vertical axis) and allele frequency (horizontal axis) closely models Hardy-Weinberg equilibrium (red line). SV density is shown with blue lines. Also generated using SVs outside tandem repeats and SDs where the callable frequency was 20% or greater.

(F) Nearly half (186 of 379) of the SV eQTLs we identified were more significantly associated to the expression level of a nearby gene compared to the lead SNP eQTL when considering genotyping errors.