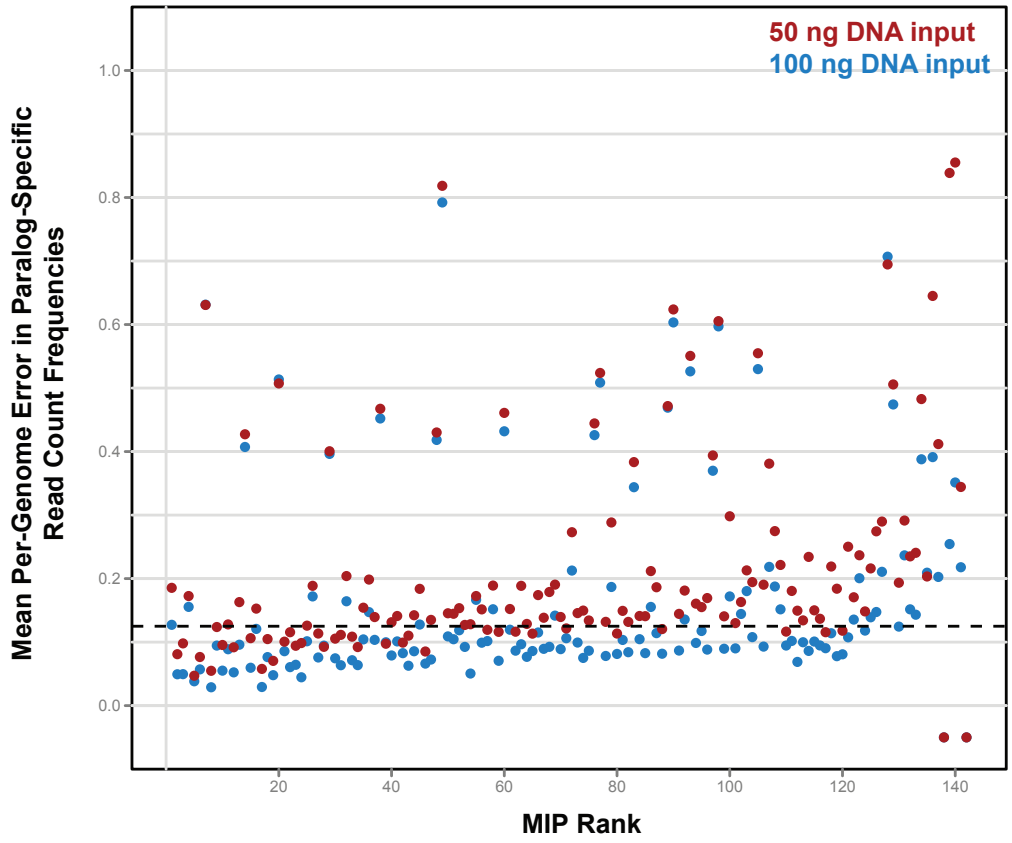
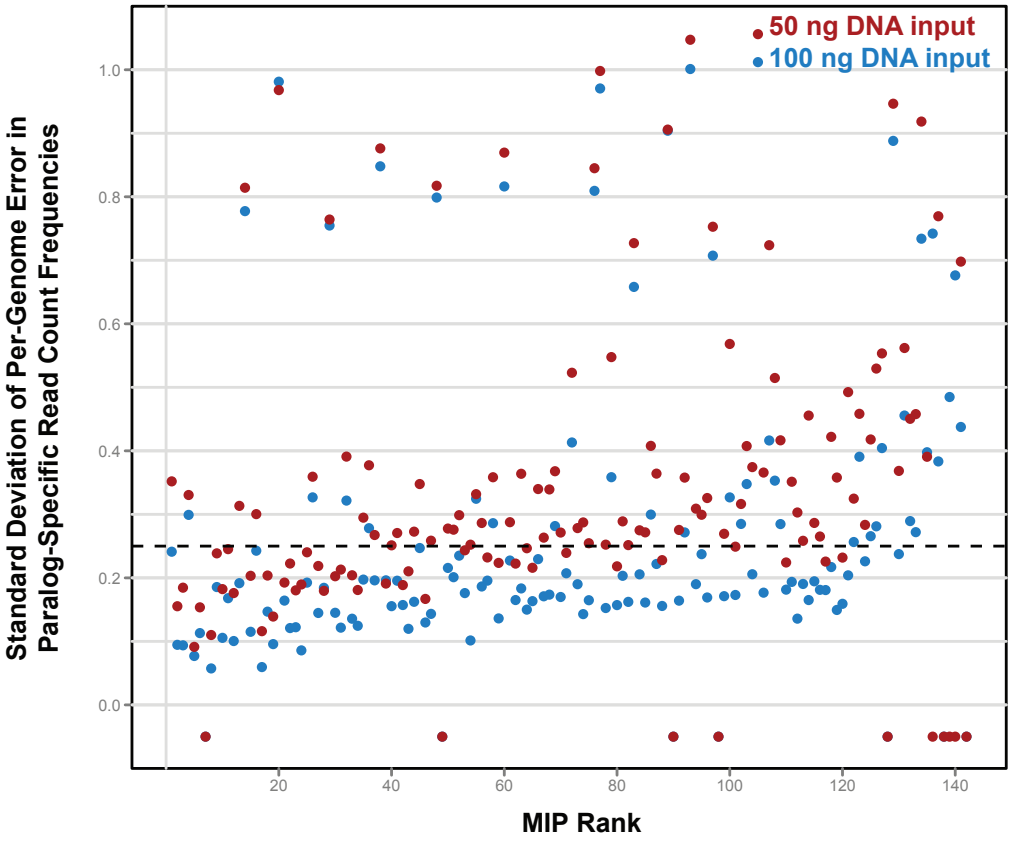
Supplementary Information for:

Rapid and accurate large-scale genotyping of duplicated genes and discovery of novel sites of interlocus gene conversion

Xander Nuttle[1], John Huddleston[1,2], Brian J. O'Roak[1], Francesca Antonacci[1,3], Marco Fichera[4,5], Corrado Romano[4], Jay Shendure[1], and Evan E. Eichler[1,2]

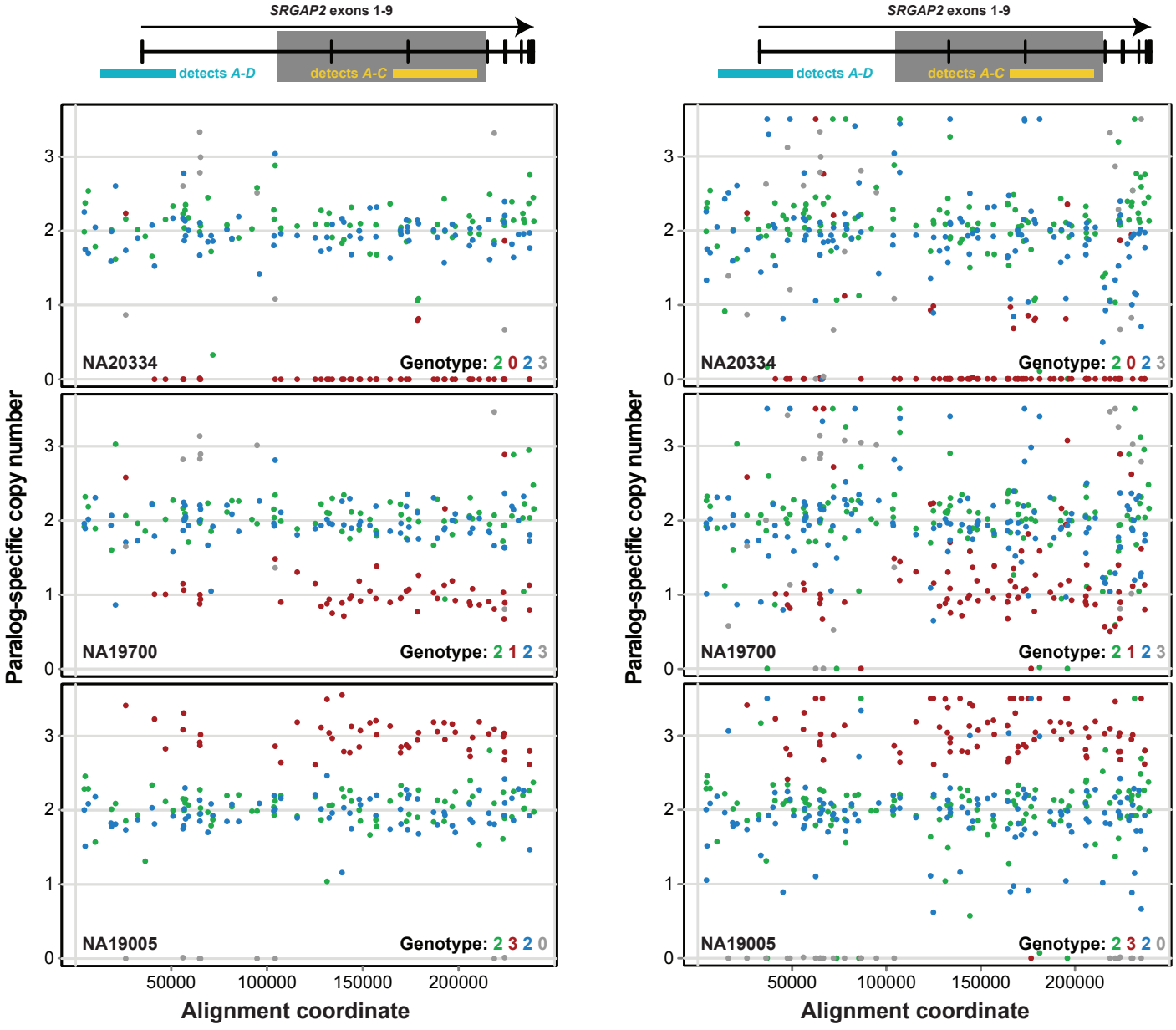
Supplementary Figure 1





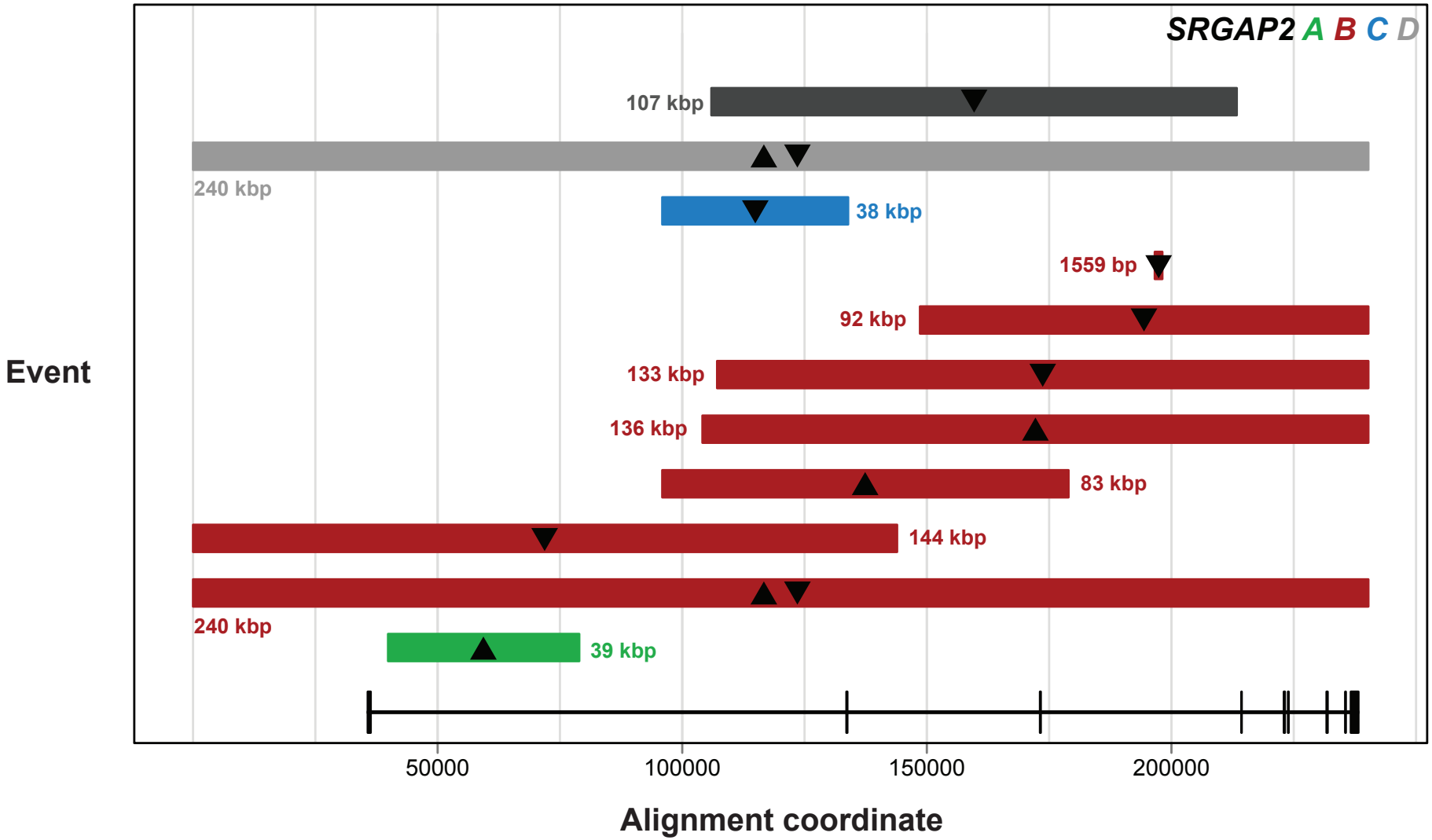
Supplementary Figure 1. Performance assessment of *SRGAP2* copy number genotyping MIPs. For a given genome assayed, error was calculated for each MIP as the sum of the absolute values of the differences between observed and expected mapped read count frequencies for each *SRGAP2* paralog and for a non-paralog-specific category (not all MIPs targeted sequence where all four *SRGAP2* paralogs can be distinguished). The per-genome means (top) and standard deviations (bottom) of these error values are plotted for each MIP using data from 31 individuals assayed in the initial 50 ng (red) and 100 ng (blue) replicate capture experiments. Negative plotted values correspond to mean errors and standard deviations of errors greater than 1.1. MIPs are ranked in the plot by total corresponding mapped read count in the 100 ng capture data for the 31 individuals, with MIPs having the highest such counts on the left. Dashed lines indicate thresholds we imposed in selecting MIPs for inclusion in our final pool. These error data highlight the increase in accuracy attained by using 100 ng of DNA rather than 50 ng of DNA for the capture reactions. Most likely, more independent capture events occur and sampling error accordingly declines with increased DNA input.

Supplementary Figure 2



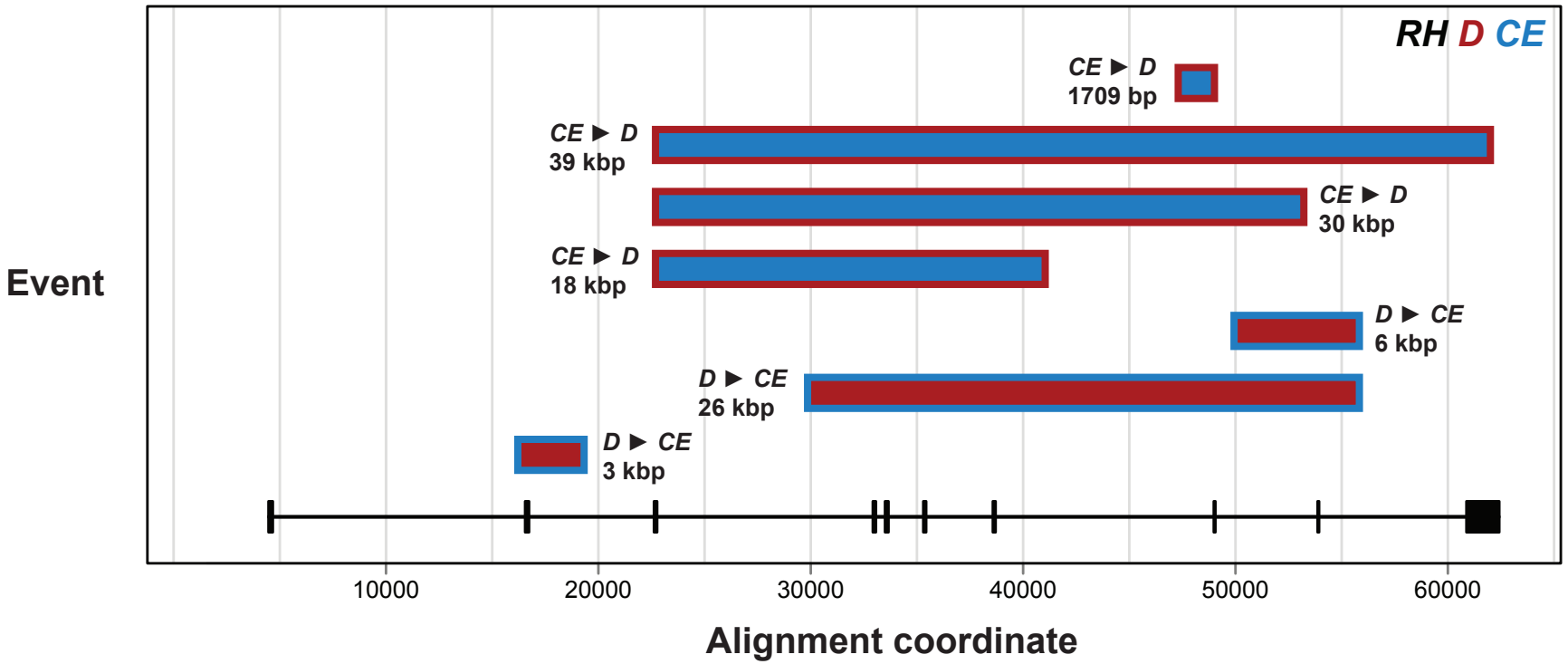
Supplementary Figure 2. Comparison of the full *SRGAP2* MIP set with the final selected set. The left panels show paralog-specific copy number estimates for 90 high-performing MIPs across ~240 kbp of aligned *SRGAP2* genomic sequence, as in Figure 2. The right panels show corresponding data for the full set of 142 MIPs. All values > 3.5 were set to 3.5 for plotting purposes. Even though the right panels show more noise, the same automated genotype call (consistent with FISH) is made regardless of whether data from the final MIP set only or the full MIP set is considered. Extending this analysis, we compared genotype calls made from the same experiment using data from the full *SRGAP2* MIP set to those made using only data from the final selected set. With one exception, the genotypes were identical for 48 individuals tested when comparing the full set with the selected set. Interestingly, for the one discordancy orthogonal data supported the genotyping call from the full set as opposed to the selected set.

Supplementary Figure 3



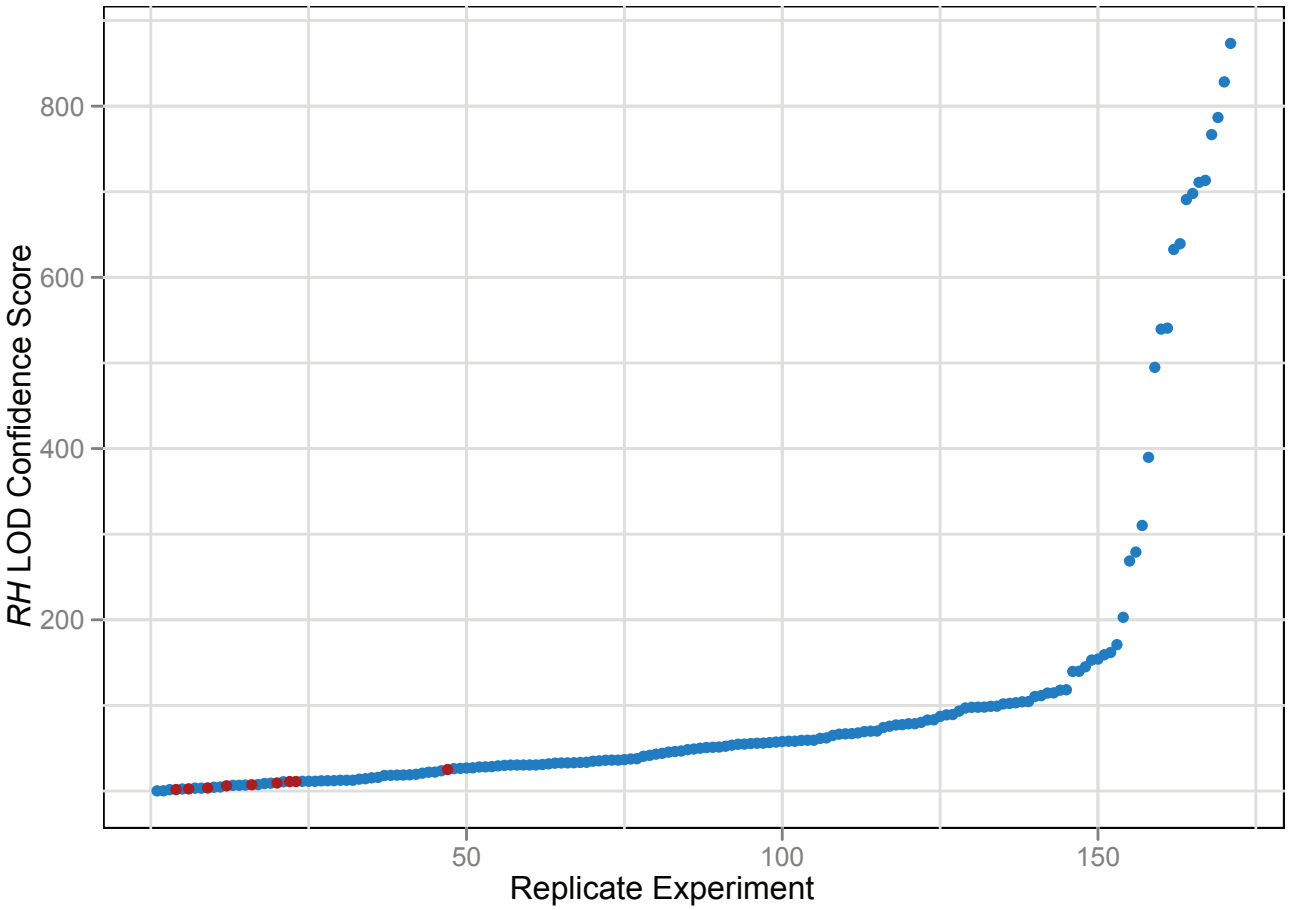
Supplementary Figure 3. Structural variation in *SRGAP2* paralogs. Locations of duplications (depicted by colored boxes with upward-pointing triangles) and deletions (depicted by colored boxes with downward-pointing triangles) identified from MIP-based genotyping of 1,056 HapMap individuals are plotted relative to duplicated *SRGAP2* exons. Dashed lines indicate events that extend beyond the extent of duplicated sequence shared between all four *SRGAP2* paralogs. Reported approximate sizes of all events are minimum estimates, calculated as the number of base pairs between the centers of MIP target sequences for the 5'-most and 3'-most MIPs signaling each event (except for events extending beyond duplicated *SRGAP2* sequence, where *SRGAP2* duplication boundaries are used in this calculation). The precisions of these size estimates are governed by the spacing and paralog-specificity of MIPs targeting surrounding regions, but typically allow for breakpoint resolution within a few kbp to a few tens of kbp. The dark gray box depicts the *SRGAP2D* internal deletion. Its breakpoints are known with very high-precision from clone-based capillary sequencing[13].

Supplementary Figure 4



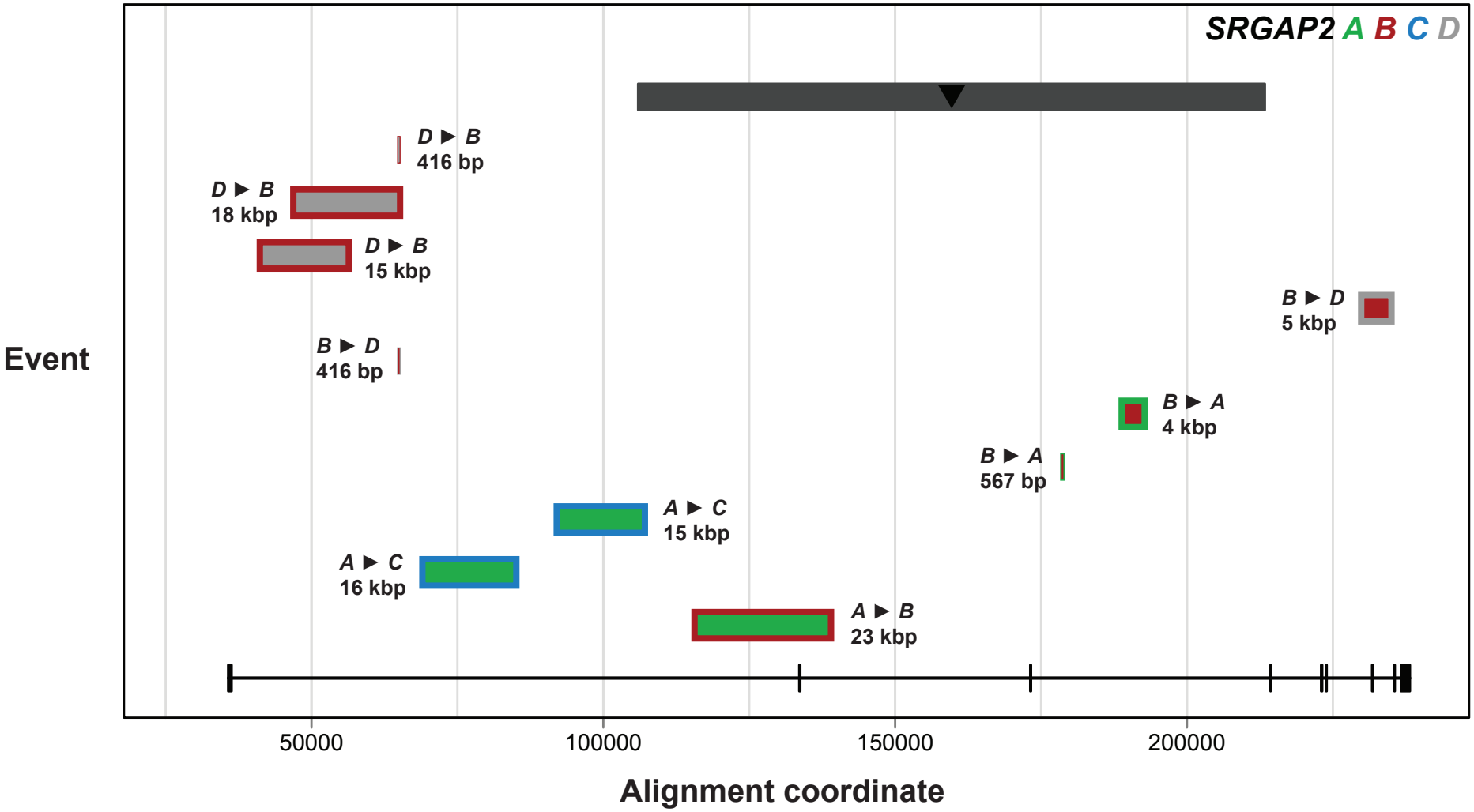
Supplementary Figure 4. Signatures of interlocus gene conversion in *RH* paralogs. Locations of putative *RH* interlocus gene conversion events (depicted by two-colored boxes) identified from MIP-based genotyping of 1,056 HapMap individuals are plotted relative to duplicated *RH* exons (corresponding to *RHD* transcript variant 1). Inner fill colors indicate putative conversion donors, while border colors indicate corresponding putative conversion acceptors. Reported approximate sizes of all events are minimum estimates, calculated as described in the legend to **Supplementary Fig. 3**.

Supplementary Figure 5



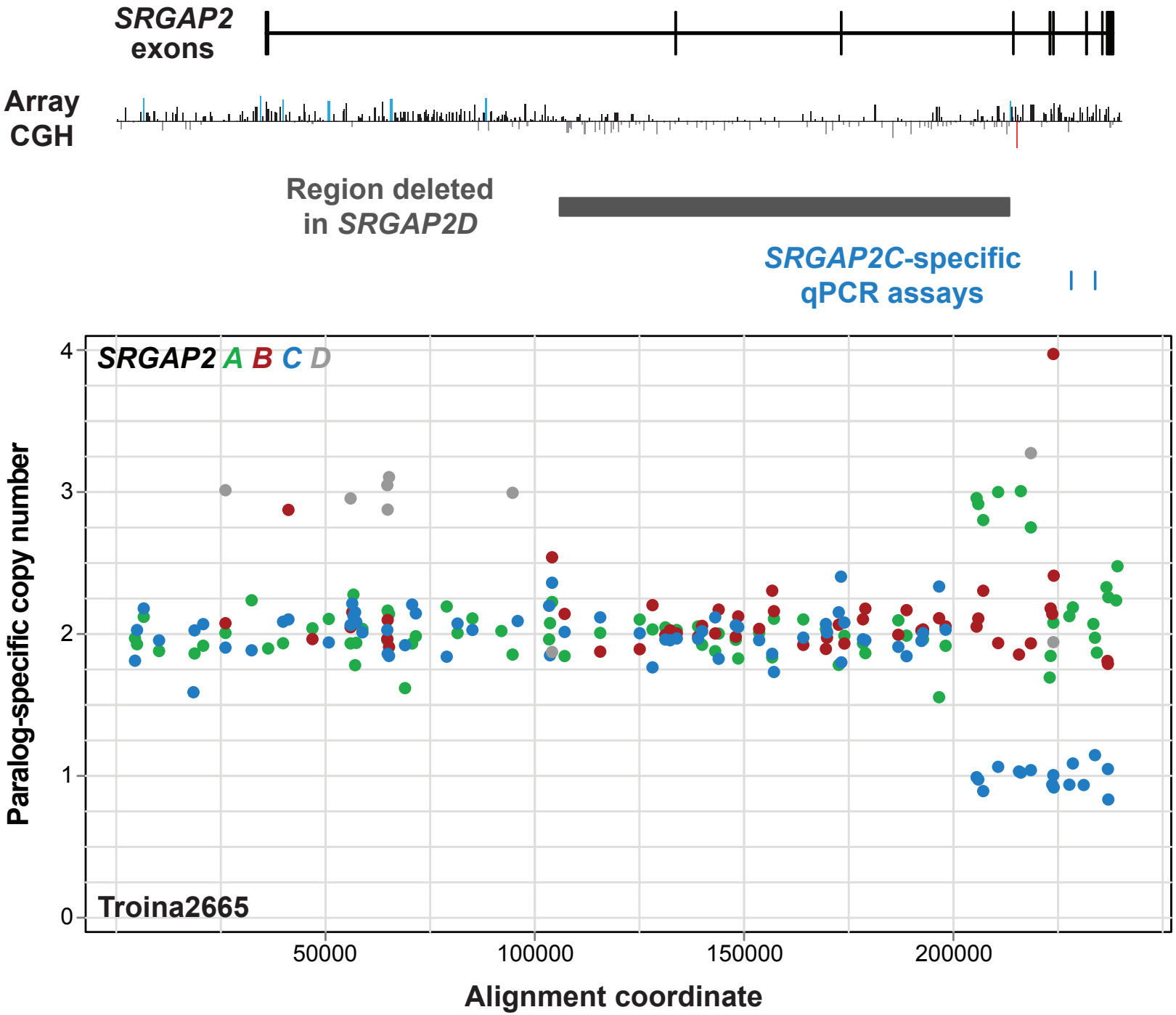
Supplementary Figure 5. Distribution of LOD confidence scores for MIP-based *RH* paralog-specific copy number genotypes from 171 replicate experiments. Discordancies are shown in red. The highest scores correspond to individuals having homozygous deletion of *RHD*. These data allow potential genotyping errors to be readily distinguished from high-confidence genotype calls.

Supplementary Figure 6



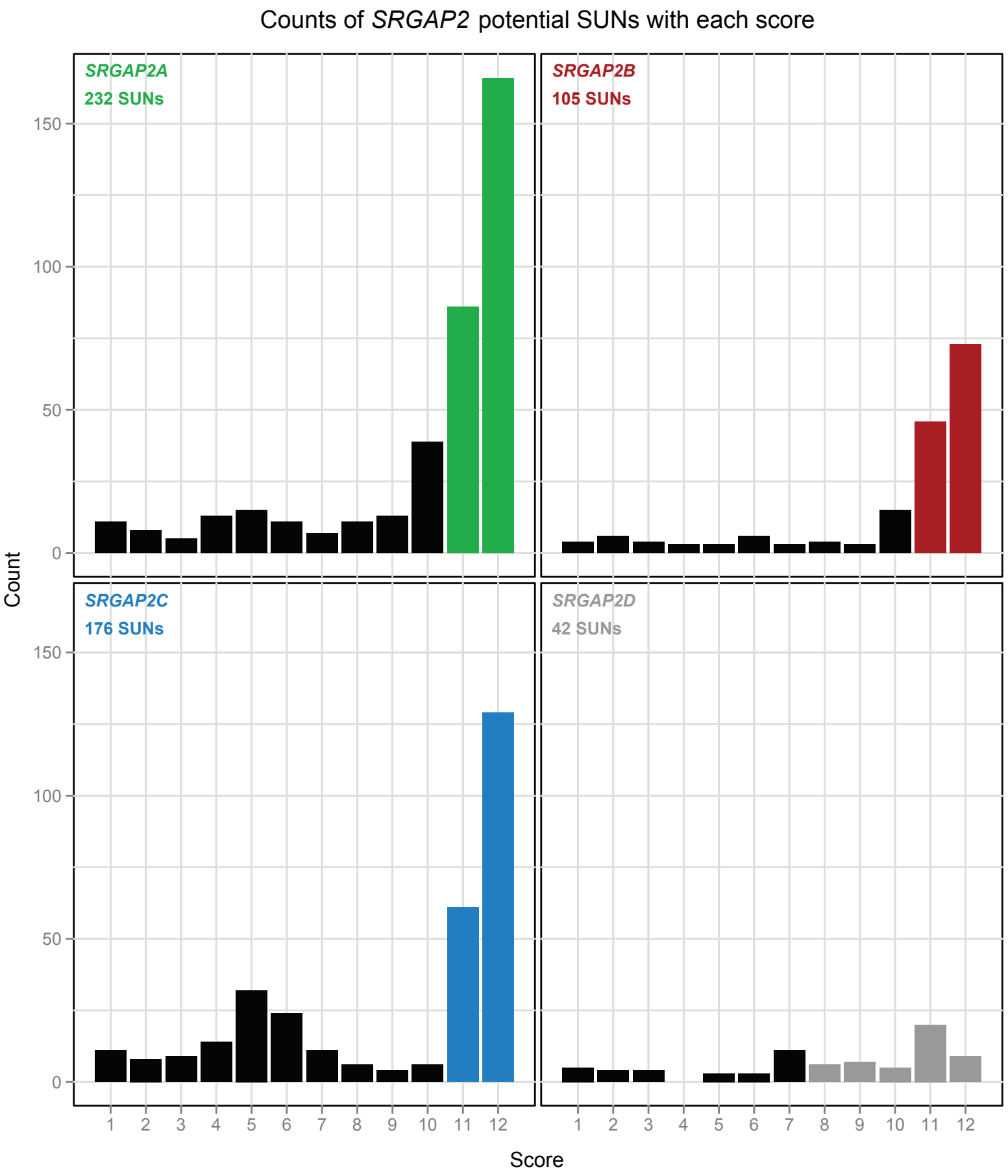
Supplementary Figure 6. Signatures of interlocus gene conversion in *SRGAP2* paralogs. Locations of putative *SRGAP2* interlocus gene conversion events (depicted by two-colored boxes) identified from MIP-based genotyping of 1,056 HapMap individuals are plotted relative to duplicated *SRGAP2* exons. Colors and reported sizes follow the convention described in **Supplementary Fig. 4**. The dark gray box depicts the *SRGAP2D* internal deletion. We note that our power to detect gene conversion events between *SRGAP2B* and *SRGAP2D*, paralogs having ~99.6% sequence identity both located within chromosome 1q21.1, was limited. This limited power largely reflects our prioritization of *SRGAP2A* and *SRGAP2C* in designing MIPs for copy number genotyping.

Supplementary Figure 7



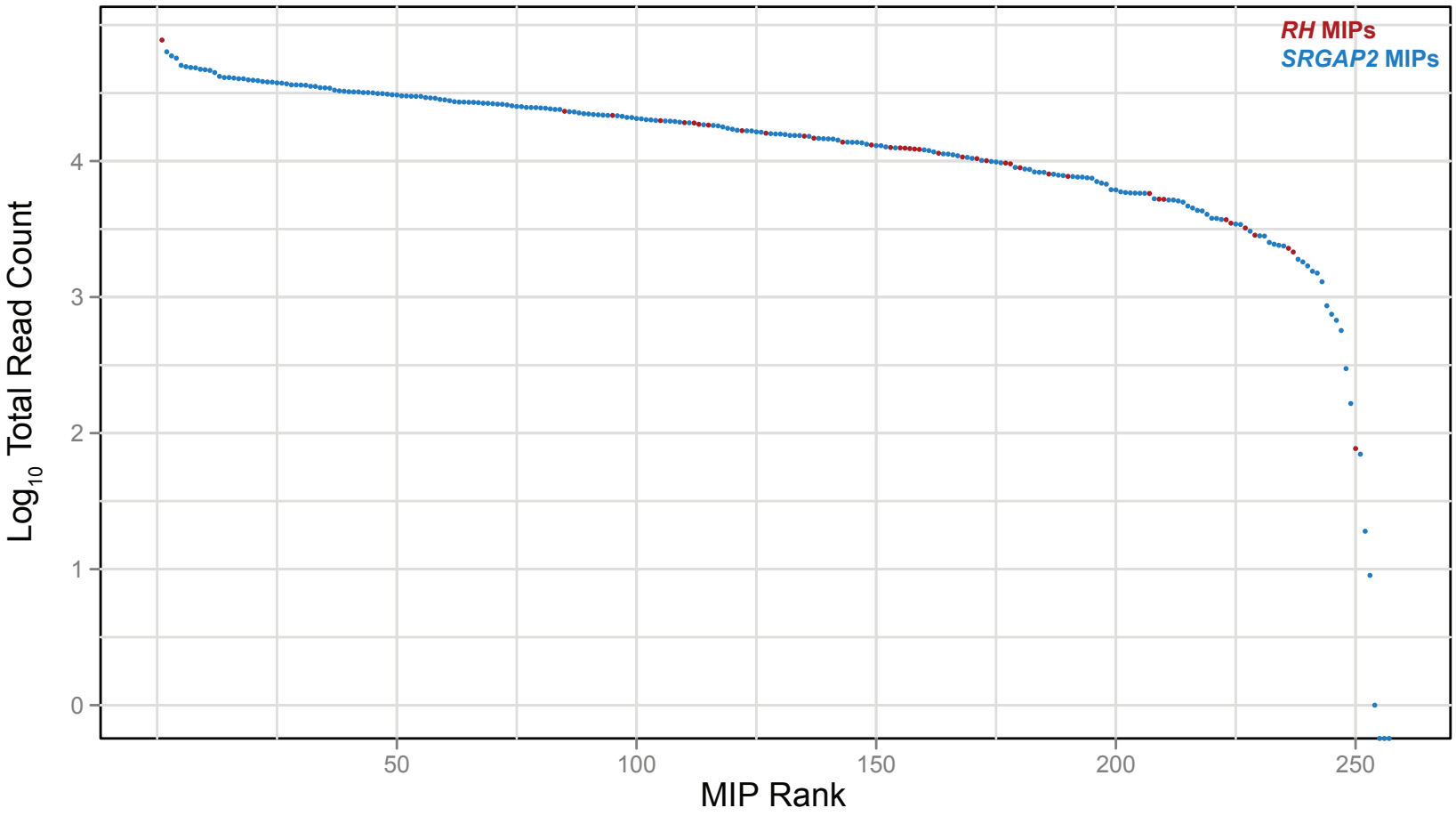
Supplementary Figure 7. Array CGH and qPCR validation of an interlocus gene conversion signature. The array CGH profile for *SRGAP2* loci predicts a gain in an individual with intellectual disability, likely involving *SRGAP2D* because the array signal disappears over the *SRGAP2D* internal deletion region. However, two independent *SRGAP2C*-specific qPCR assays targeting introns 6 and 7 predict a *SRGAP2C* deletion, a result seemingly inconsistent with the array data. MIP genotyping provides further support for the *SRGAP2D* duplication and suggests that gene conversion involving *SRGAP2C* as an acceptor explains the qPCR results. MIP data from this individual show evidence for multiple putative interlocus gene conversion events affecting the last few duplicated *SRGAP2* exons.

Supplementary Figure 8



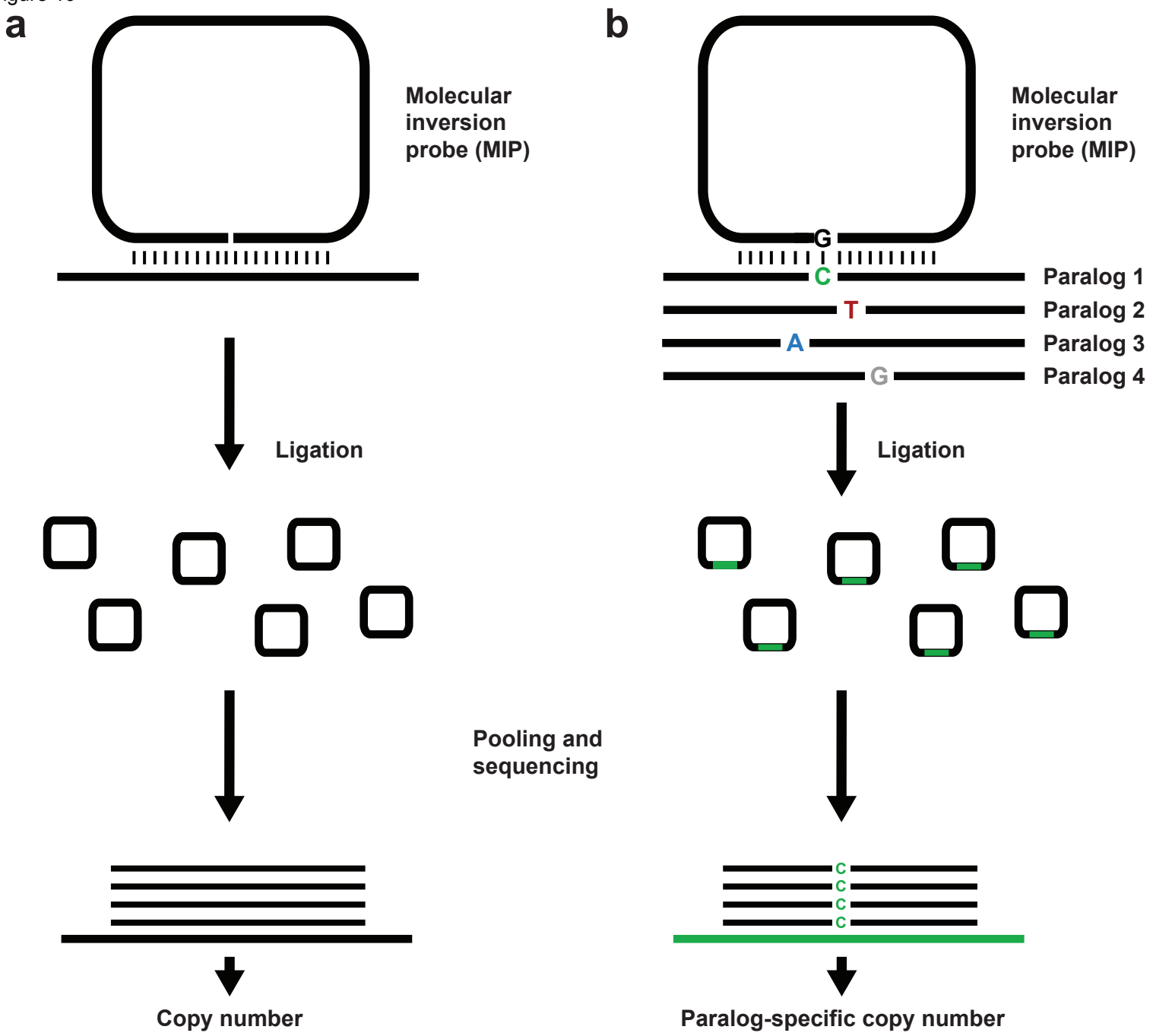
Supplementary Figure 8. Score histograms for *SRGAP2* potential SUNs. All *SRGAP2* potential SUNs having at least a single overlapping 30-mer SUNK were scored on a scale of 0-12. Scores were calculated as the sum over all 30-mer SUNKs overlapping the potential SUN of the number of high-coverage genomes analyzed supporting the SUNK's presence, divided by the total number of 30-mer SUNKs overlapping the potential SUN. This score can thus be interpreted as the average number of high-coverage genomes supporting a potential SUN's presence. Low scores reflect low allele frequency, sequence masking at or near a potential SUN position, or some combination of these factors, while high scores indicate a likely high potential SUN allele frequency and thus high value for copy number genotyping. The histograms show the distributions of potential SUN scores rounded to the nearest integer for all four *SRGAP2* paralogs. Colored bars correspond to potential SUNs defined to be true SUNs. Counts of SUNs scoring < 0.5 are omitted from the plot, as SUNs with these scores may indeed be present in several analyzed high-coverage genomes but could not be assessed due to sequence masking.

Supplementary Figure 9



Supplementary Figure 9. Counts of total reads mapped to each MIP target for the 100 ng 48-individual capture experiment. All reads included in these counts passed all filters described in the copy number genotyping section of the **Online Methods**. All *SRGAP2*-targeting and *RH*-targeting MIPs are ranked in the plot by total corresponding mapped read count, with MIPs having the highest such counts on the left. These data provide insight into the relative capture efficiencies of different MIPs and inform MIP rebalancing. The tight distribution of total corresponding mapped read count values (within 1.5 logs for the 227 MIPs having highest such counts) suggests capture efficiency was fairly uniform between MIPs. MIPs having the fewest corresponding mapped read counts were almost all exon-targeting with the lowest design score (-1), used because no higher-scoring alternative MIPs could be designed that would still target the desired exonic sequence.

Supplementary Figure 10



Supplementary Figure 10. MIP-based multiplex ligation-dependent probe amplification. a) MIP arms could be designed to hybridize to adjacent sequences, such that hybridization followed by ligation results in circularly closed molecules. Barcoding, pooling, and sequencing these molecules, mapping reads to corresponding reference sequence, and quantifying read depth should provide insights into copy number of targeted loci in a manner akin to MLPA. This approach would allow for up to ~2000 sites to be assayed in this manner simultaneously. Furthermore, these probes could be combined with conventional MIP probes in the same reaction. b) If the MLPA-MIP were designed such that the final base of one hybridization arm was complementary to a SUN, this assay might be able to achieve paralog-specificity.