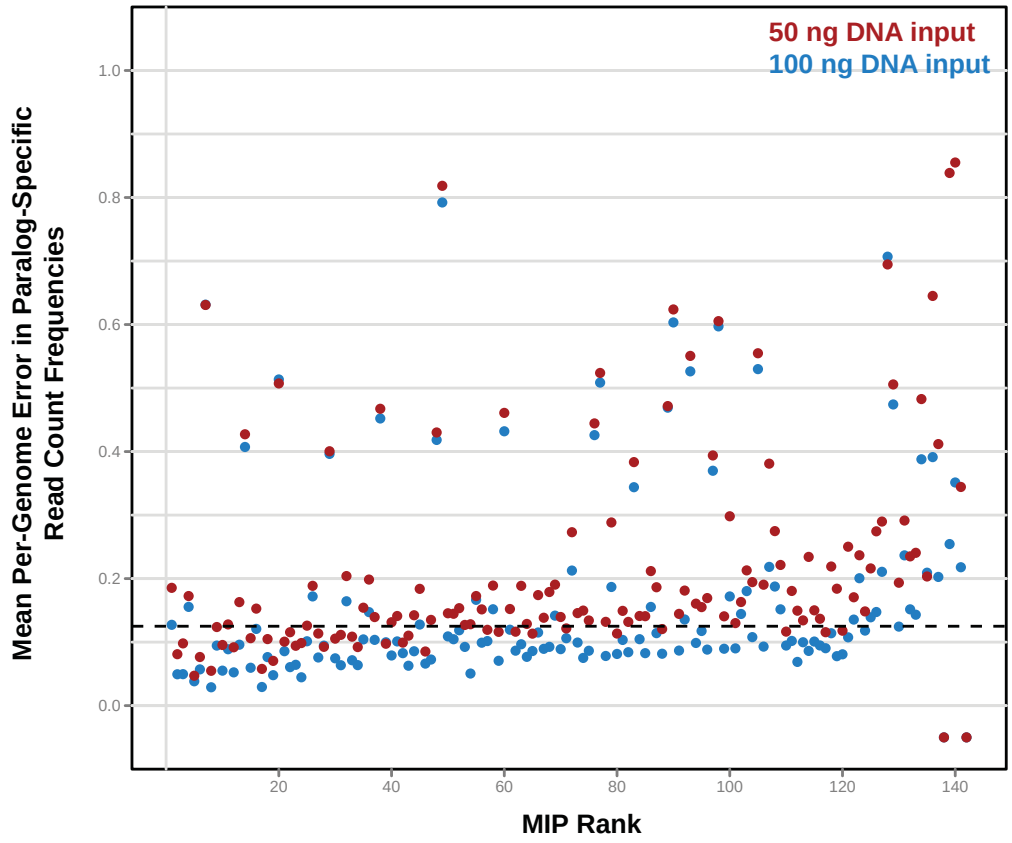
Supplementary Information for:

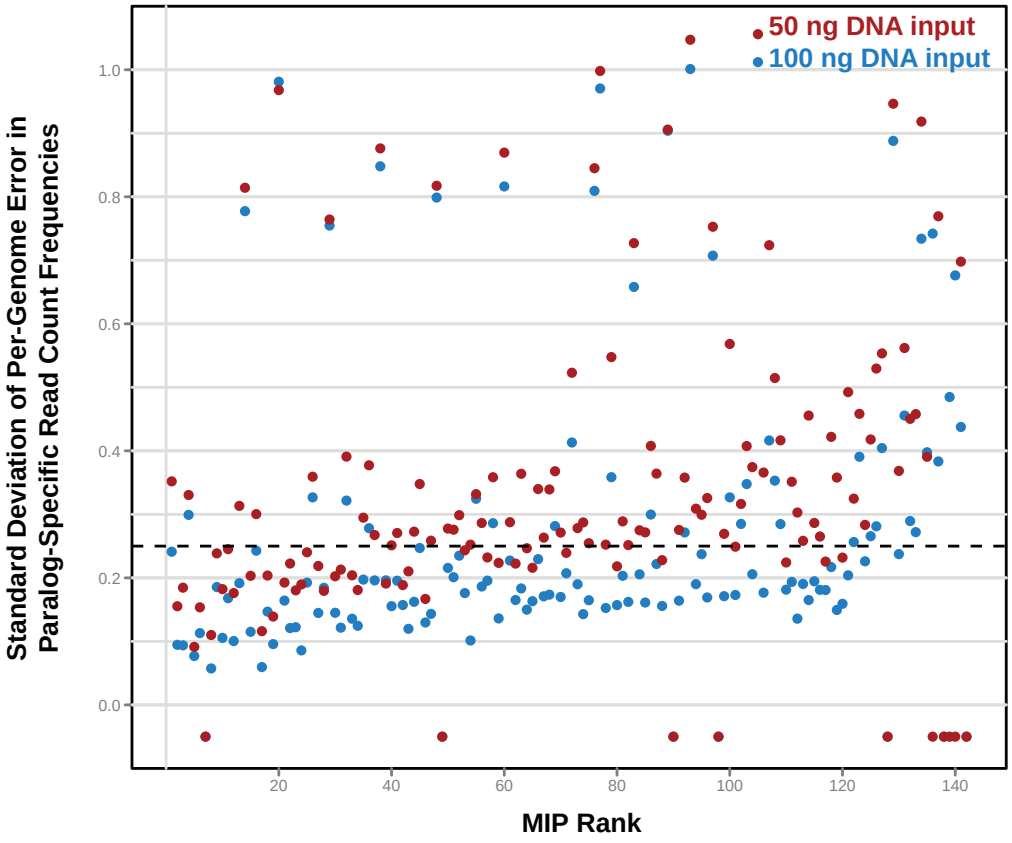
Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions

Xander Nuttle, John Huddleston, Brian J. O'Roak, Francesca Antonacci, Marco Fichera, Corrado Romano, Jay Shendure, and Evan E. Eichler

Nature Methods: doi:10.1038/nmeth.2572

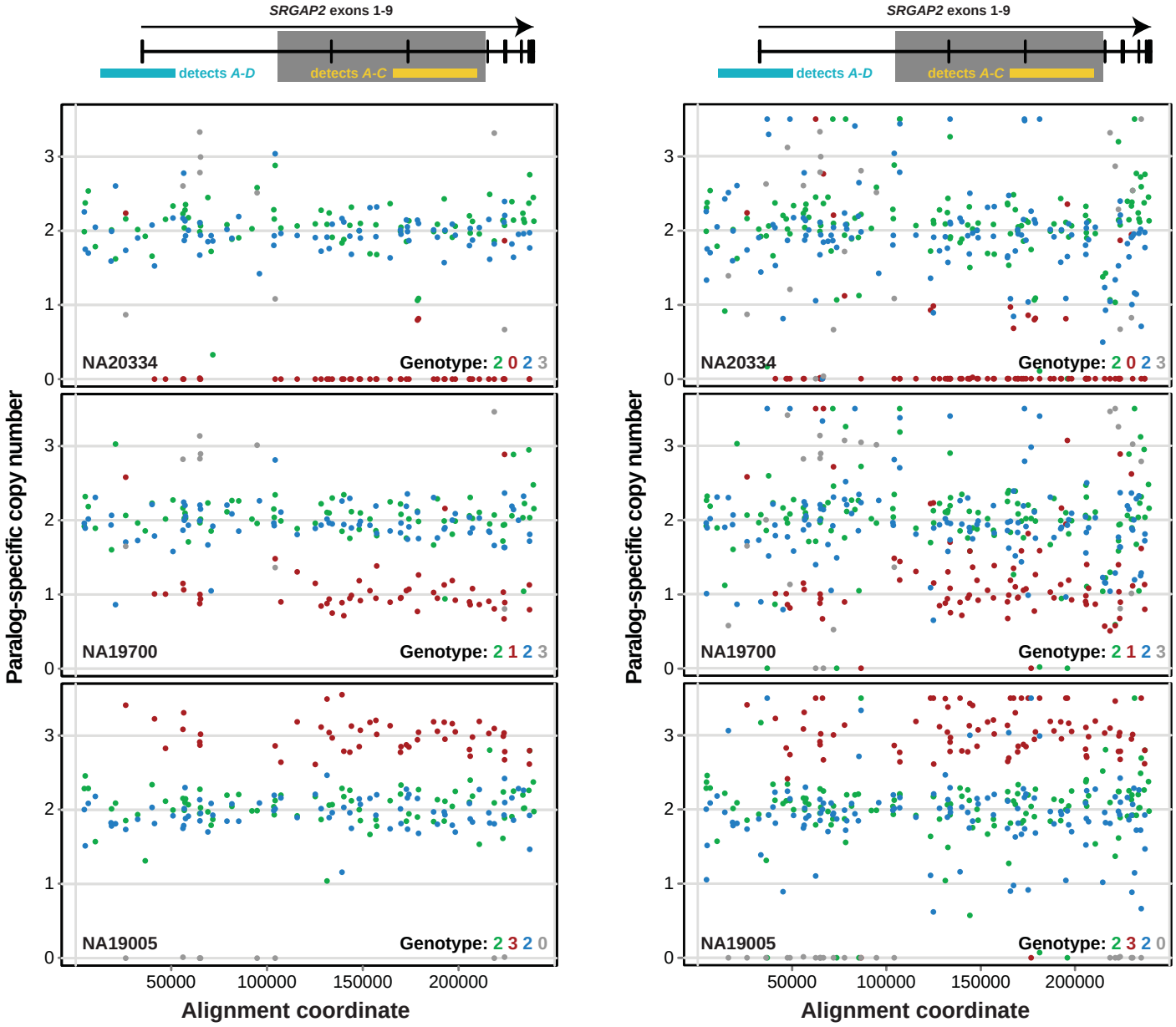
Supplementary Figure 1





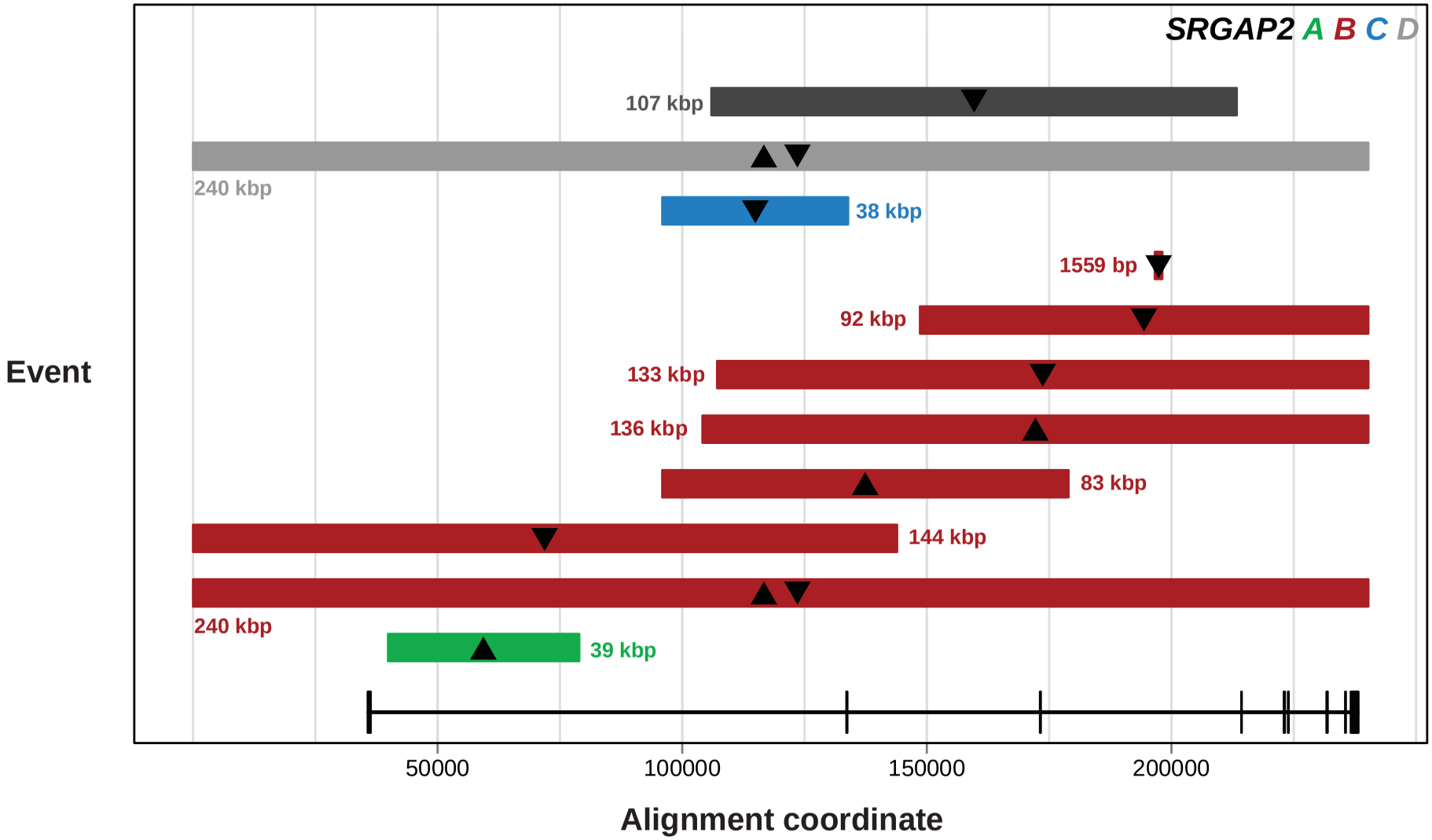
Supplementary Figure 1. Performance assessment of *SRGAP2* copy number genotyping MIPs. For a given genome assayed, error was calculated for each MIP as the sum of the absolute values of the differences between observed and expected mapped read count frequencies for each *SRGAP2* paralog and for a non-paralog-specific 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.

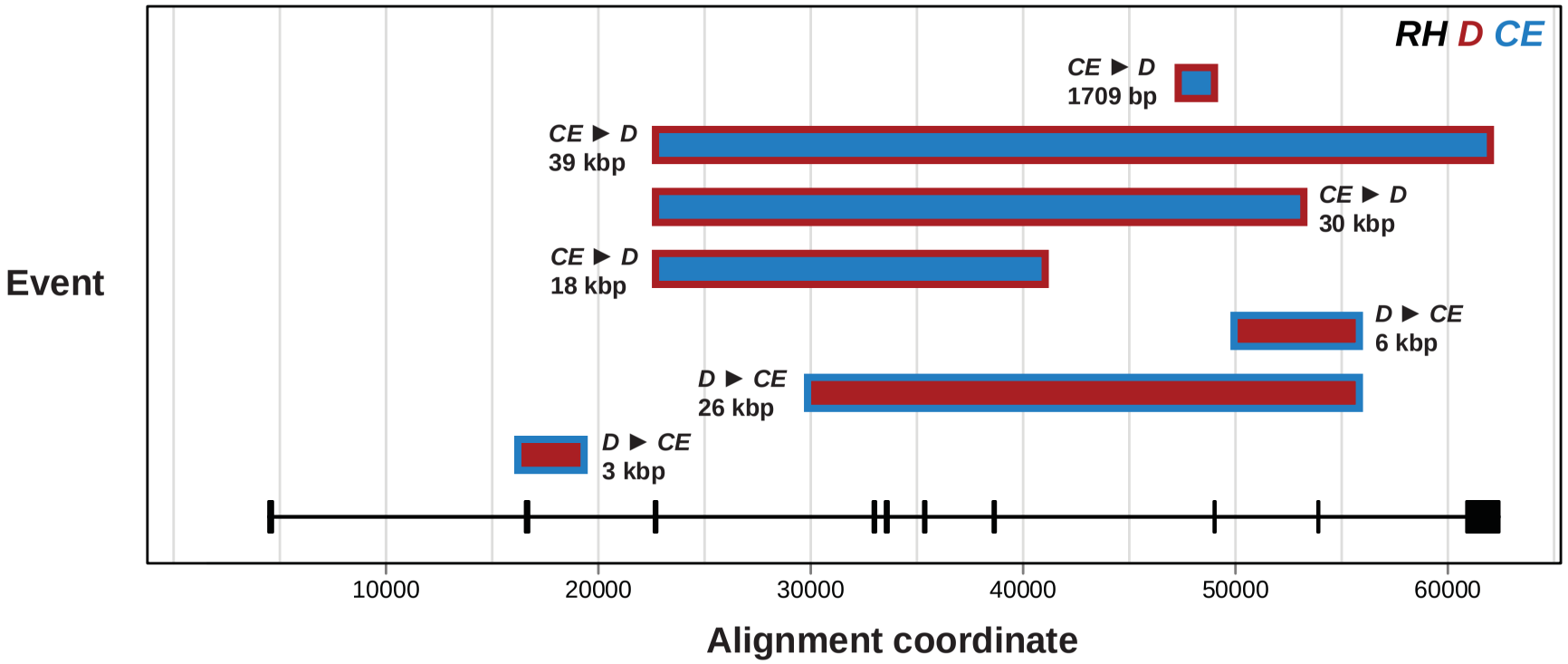
Nature Methods: doi:10.1038/nmeth.2572



Nature Methods: doi:10.1038/nmeth.2572

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].

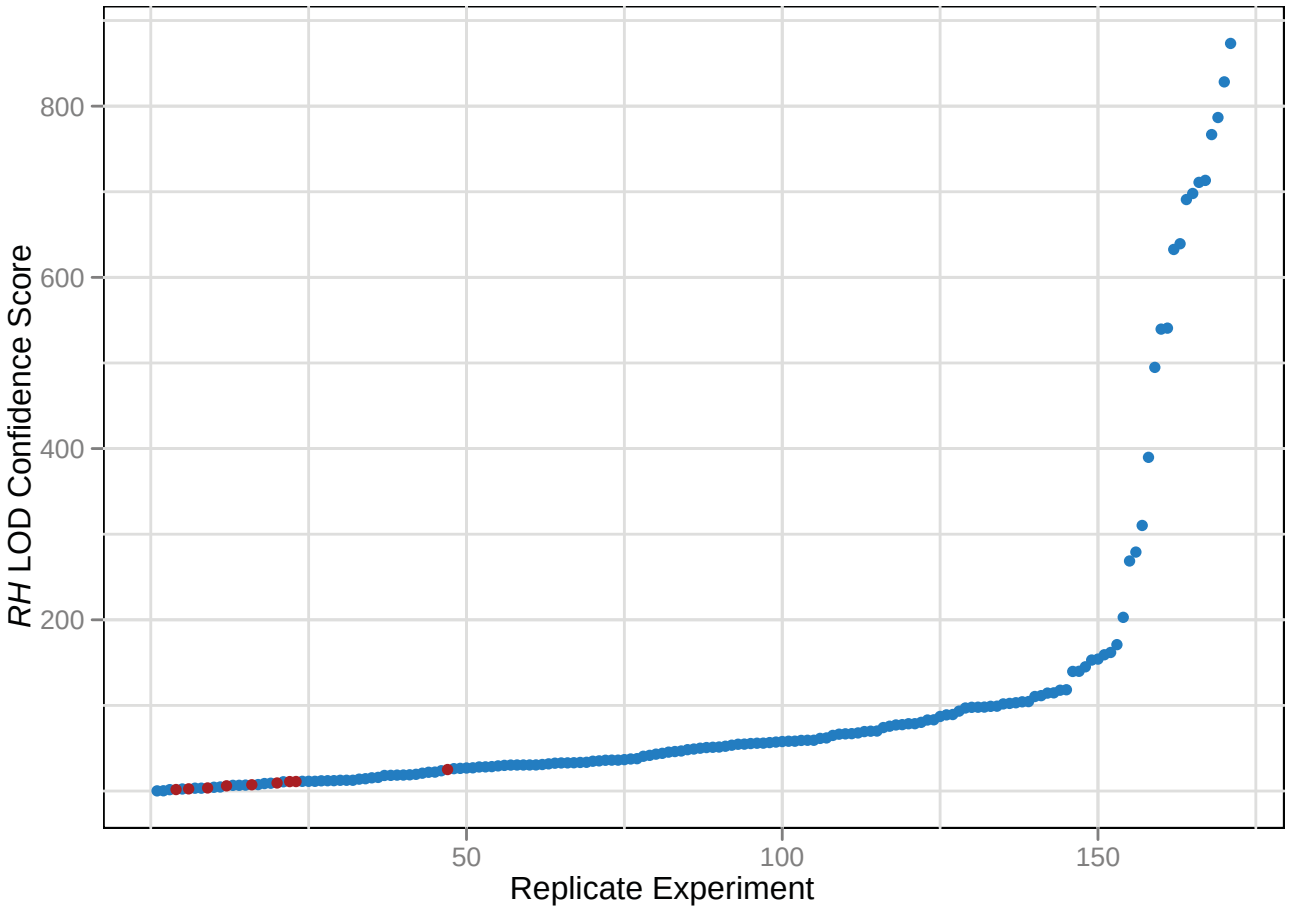
Nature Methods: doi:10.1038/nmeth.2572


RH D CE
Event
CE ▶ D
1709 bp
CE ▶ D
39 kbp
CE ▶ D
30 kbp
CE ▶ D
18 kbp
D ▶ CE
6 kbp
D ▶ CE
26 kbp
D ▶ CE
3 kbp
10000
20000
30000
40000
50000
60000
Alignment coordinate

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**.

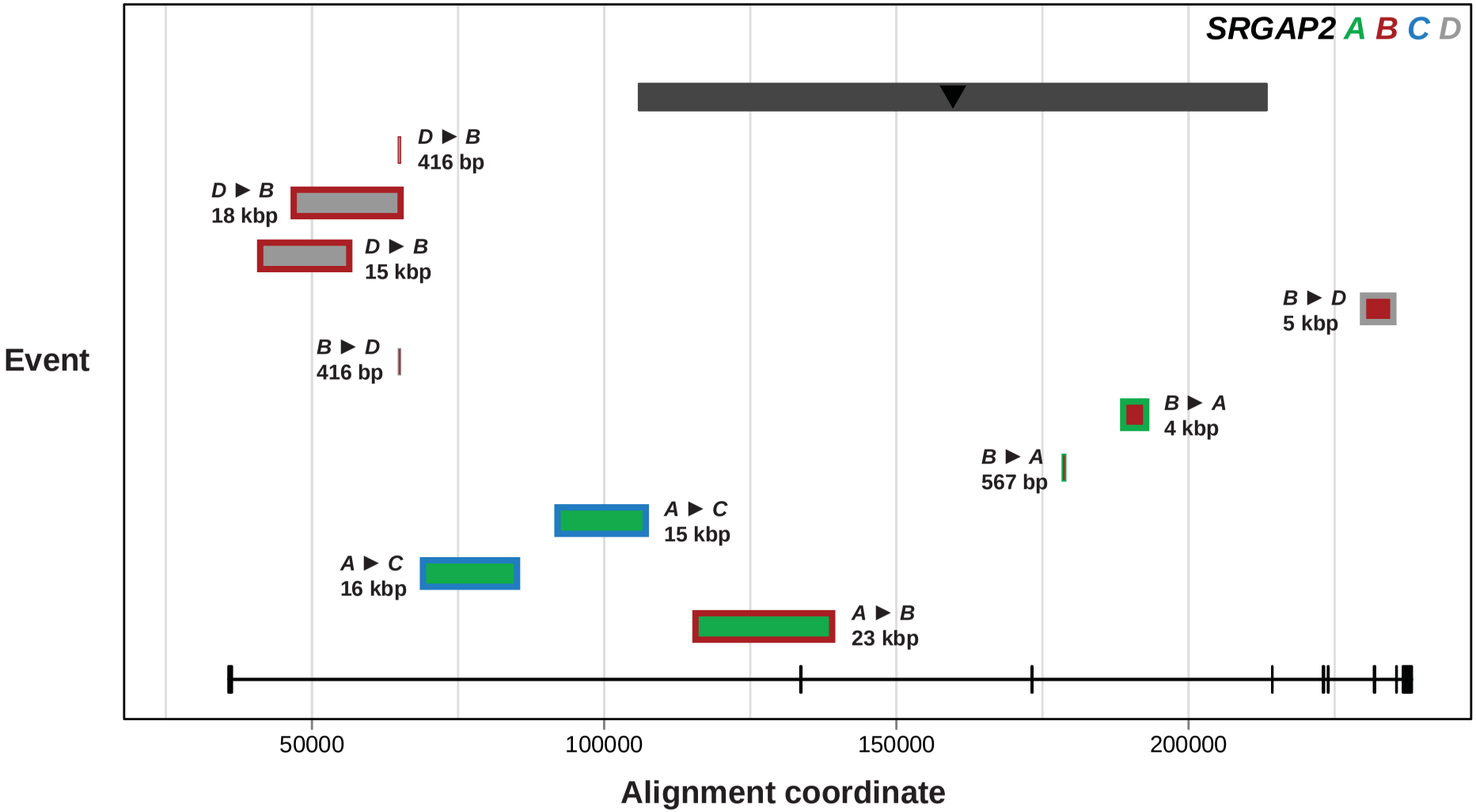
Nature Methods: doi:10.1038/nmeth.2572

Supplementary Figure 5

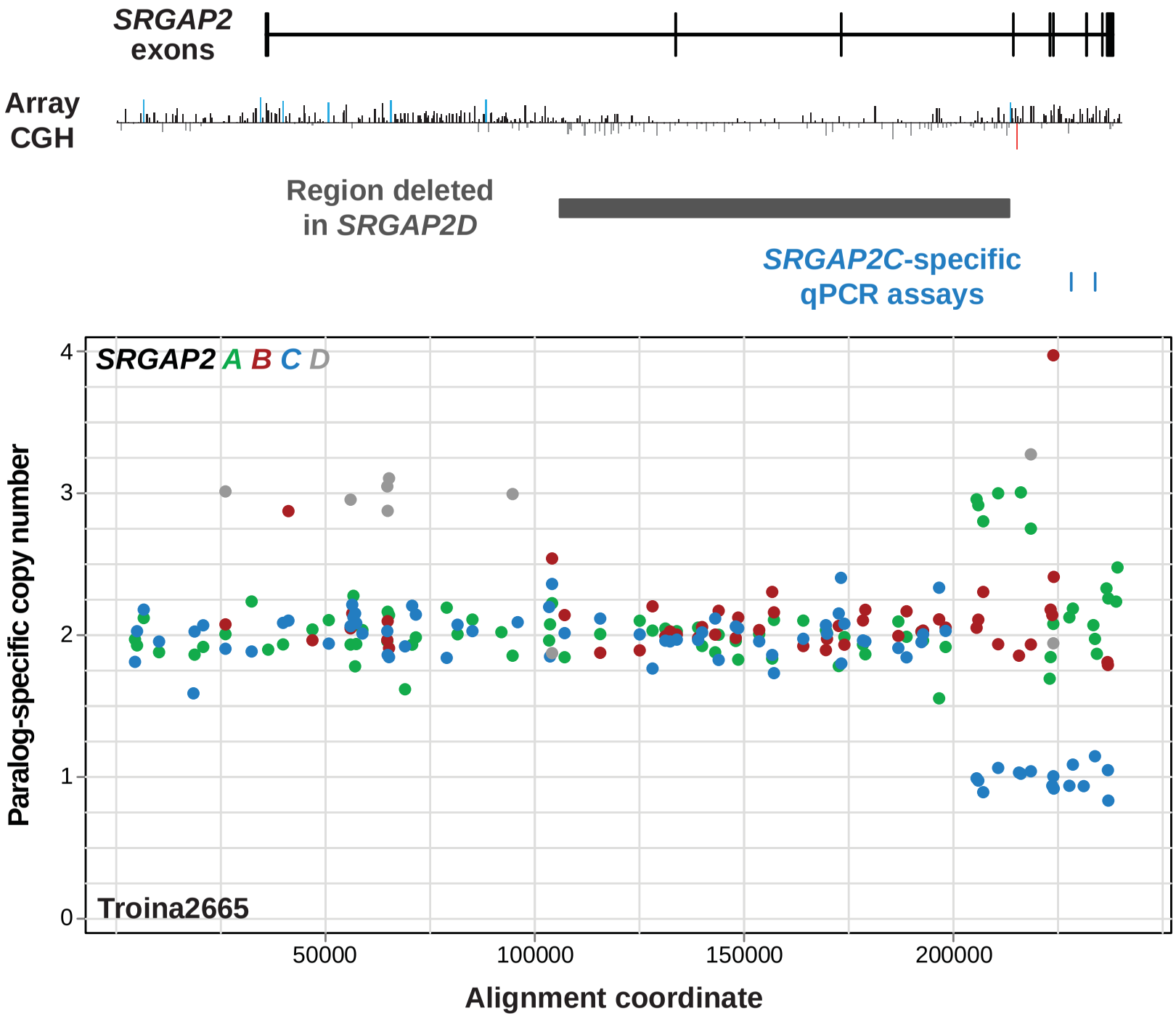


Supplementary Figure 5. Distribution of LOD confidence scores for MIP-based *RH* paralog-specific copy number genotypes from 171 replicate experiments. Discordancies are shown in red. The highest scores correspond to individuals having homozygous deletion of *RHD*. These data allow potential genotyping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

Nature Methods: doi:10.1038/nmeth.2572


SRGAP2 A B C D
Event
D ▶ B
416 bp
D ▶ B
18 kbp
D ▶ B
15 kbp
B ▶ D
5 kbp
B ▶ D
416 bp
B ▶ A
4 kbp
B ▶ A
567 bp
A ▶ C
15 kbp
A ▶ C
16 kbp
A ▶ B
23 kbp
50000
100000
150000
200000
Alignment coordinate

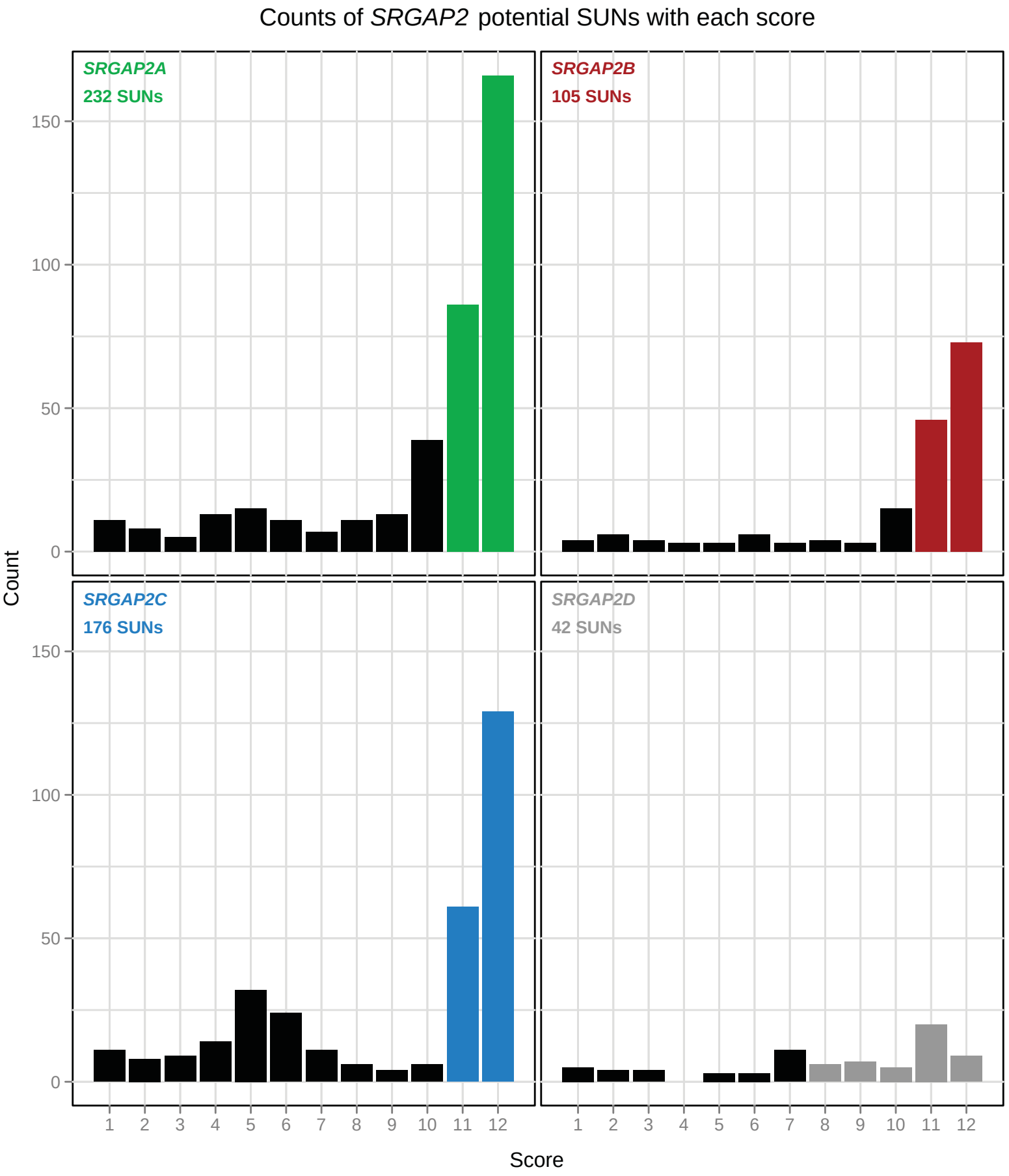
Supplementary Figure 6. Signatures of interlocus gene conversion in *SRGAP2* paralogs. Locations of putative *SRGAP2* interlocus gene conversion events (depicted by two-colored boxes) identified from MIP-based genotyping of 1,056 HapMap individuals are plotted relative to duplicated *SRGAP2* exons. Colors 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.



Nature Methods: doi:10.1038/nmeth.2572

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

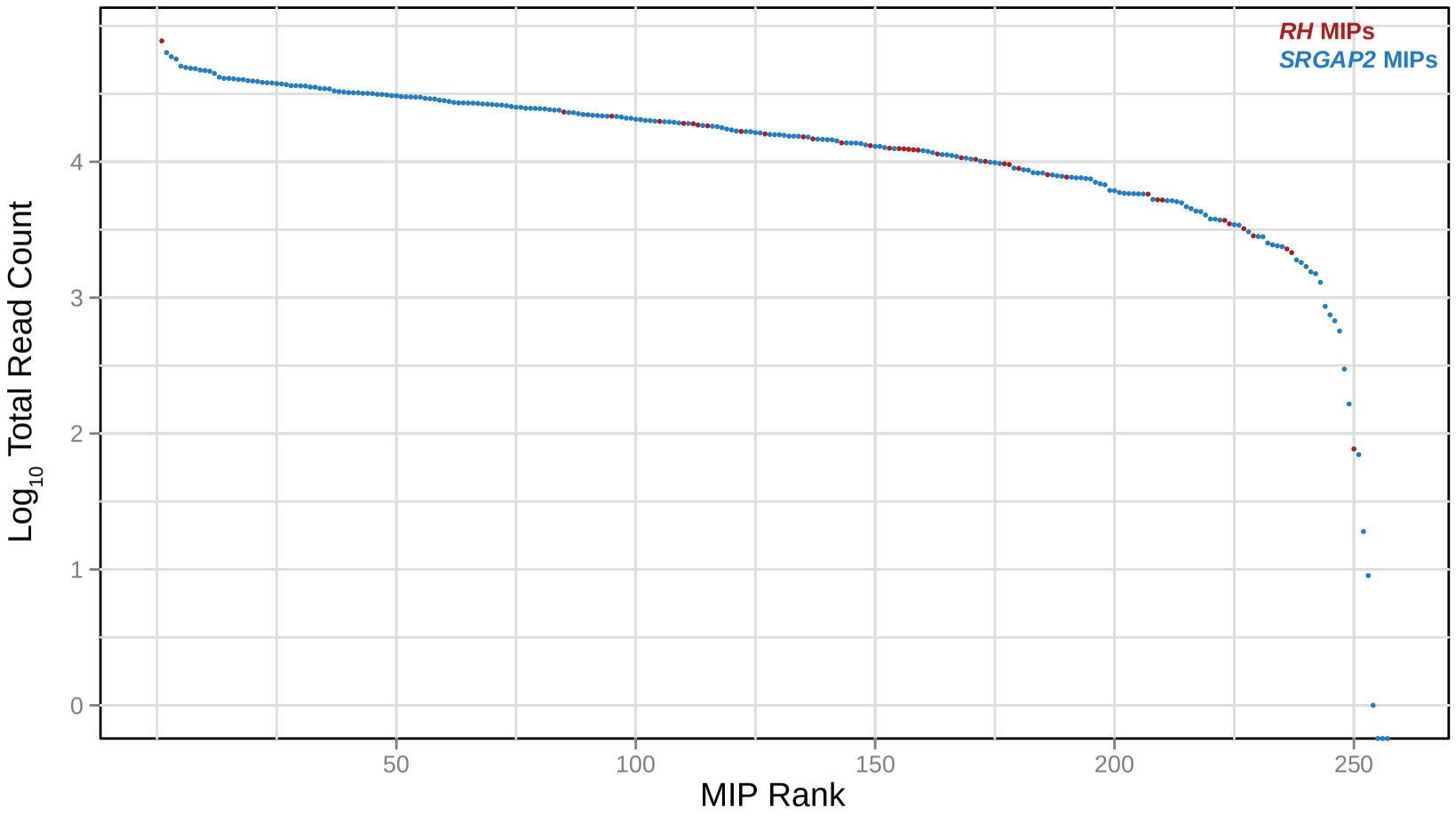


Nature Methods: doi:10.1038/nmeth.2572

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.

Nature Methods: doi:10.1038/nmeth.2572

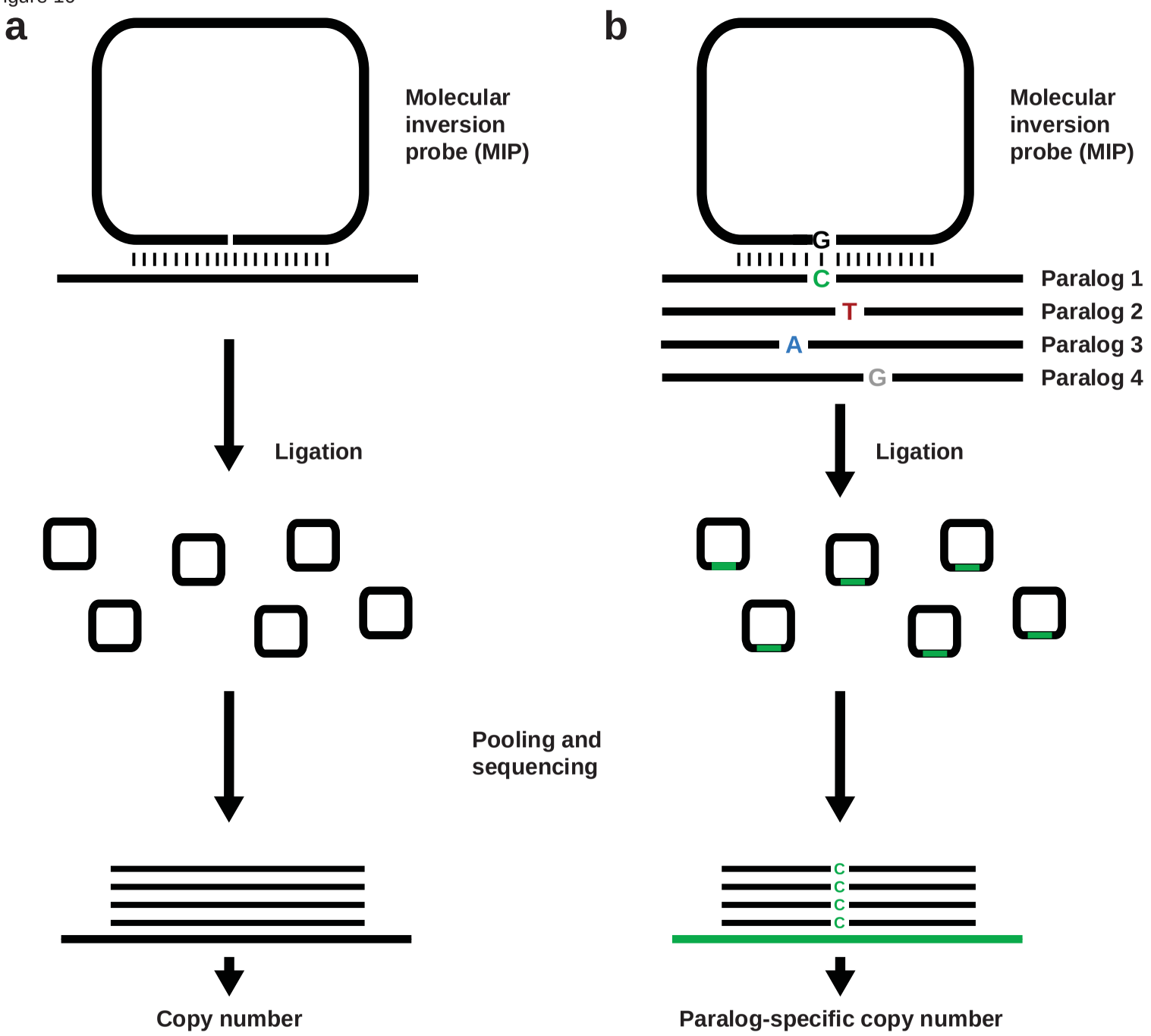
Supplementary Figure 9



Nature Methods: doi:10.1038/nmeth.2572

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



Nature Methods: doi:10.1038/nmeth.2572

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.

Supplementary Table 1. High-coverage genomes used for SUN analysis.

Genome	Ethnicity	Reference
YH-1	CHI	62
NA10851	CEPH	63
NA18507	YRI	35
NA18508	YRI	35
Jay_Flatley	European	
Korean-1-gr	Korean	64
KB1	Khoisan	65
ABT	Bantu	65
NA12891	CEU	66
NA12892	CEU	66
NA19238	YRI	66
NA19239	YRI	66

Supplementary Table 7. LOD confidence scores for MIP-based RH paralog-specific copy number genotypes from 171 replicate experiments.

Individual*	*RHD* copy number	*RHCE* copy number	LOD confidence score
NA21434	1	2	0.183269
NA21434	1	2	0.300784
Troina02753_50	2	2	1.596775
NA20344	2	3	1.743351
NA19789	2	2	2.398601
NA20334	2	3	2.542917
NA20288	1	2	3.320578
NA20344	1	2	3.322344
NA19700_50	2	3	3.611023
NA20127_50	2	2	4.399925
NA20771	1	2	4.732674
HG00319_50	2	3	6.081959
NA20289_50	1	2	6.612153
NA19901_50	1	2	6.683607
NA20288	1	2	6.932922
NA20756	2	3	7.314588
Troina02753_100	2	2	7.538668
NA19700	1	2	8.729696
NA21774	1	2	9.123378
NA19761	3	2	9.408723
SG9906627_50	1	2	10.795023
NA19901_100	2	3	11.008318
NA20543	2	3	11.095718
HG00319_100	1	2	11.327155
NA19190_100	2	2	11.355645
NA20334_50	1	2	11.370777
NA12275	2	2	11.928825
12523.p1_100	1	2	12.039967
NA19065_50	2	2	12.095013
NA20334_100	1	2	12.547751
NA20815	2	2	12.568038
NA12248	1	2	12.573907
NA20356	1	2	13.891574
NA20771_100	1	2	14.39308
NA12878_50	1	2	15.450185
NA21774	1	2	15.920752

NA06991	1	2	18.184914
NA20543_50	1	2	18.314733
NA19901	1	2	18.575463
12523.p1_50	1	2	18.672419
HG00275_50	1	2	18.948416
NA20289_100	1	2	19.525161
NA12878_100	1	2	20.827254
NA20543_100	1	2	21.918577
NA18976_50	2	2	22.098344
NA18609	1	2	23.675737
NA20281	2	3	25.106067
NA19190	2	2	26.124906
NA18976_100	2	2	26.463111
NA20289	1	2	26.887572
NA18548_50	1	3	27.116302
NA19700_100	1	2	28.104513
NA19761_50	2	2	28.209919
NA20281_50	2	2	28.524735
NA19190_50	2	2	29.362947
NA18548_100	1	3	29.948507
NA18553_50	2	2	30.319171
NA19201_50	2	2	30.409405
NA19703	1	2	30.421565
NA12248	1	2	30.459737
NA19703_50	1	2	30.475232
NA18951	2	2	30.990924
SG9906627_100	1	2	31.726145
NA19789_50	2	2	32.50885
NA19625_100	2	2	32.784573
NA19625	2	2	32.871337
NA18633	2	2	33.027036
SG9881737_50	2	2	33.37415
NA19204_50	3	2	33.574044
NA20771_50	1	2	34.806548
NA12878	1	2	35.225563
NA19783	2	2	35.951394
NA20756_50	1	2	36.081647
NA19625_50	2	2	36.096223
NA18951	2	2	36.686763
NA19703_100	1	2	37.419181
NA18522_50	2	2	37.790271

NA19204_100	3	2	40.603231
NA19004_50	2	2	41.663101
NA19005_50	2	2	43.060665
NA20756_100	1	2	44.028124
NA18603	2	2	45.44171
NA18862	2	2	46.110129
NA18986_50	2	2	46.772449
NA20127_100	2	2	48.31181
NA20322	2	2	49.049383
NA20356	1	2	50.018455
NA18989_50	1	2	50.774258
NA18548	1	3	51.106908
NA19005	2	2	51.398237
NA19761_100	2	2	52.307211
NA18933_50	2	2	53.530331
GC21416_50	1	2	54.638637
NA18859	2	2	54.793925
NA20322	2	2	55.453359
NA19001	1	2	55.844486
NA18862	2	2	55.934918
NA20127	2	2	56.580004
HG00421_100	2	2	57.17978
HG00275_100	1	2	57.801162
NA18933	2	2	58.183854
NA19708	2	2	58.223977
NA18989_100	1	2	59.106864
NA18633_50	2	2	59.291233
NA20281_100	2	2	59.310818
HG00475_50	2	2	61.470213
NA19708	2	2	62.014388
NA18522_100	2	2	64.951549
NA20815_50	2	2	66.406696
HG00421_50	2	2	66.790405
NA18599	1	2	67.022838
NA19055	2	2	67.855638
NA12275_50	2	2	69.432692
NA18986_100	2	2	69.894956
NA19004_100	2	2	70.068124
NA19201_100	2	2	74.123443
NA18859	2	2	75.701588
NA18534	2	2	77.164583

HG00326_50	2	2	77.546555
NA18856	2	2	78.5951
NA18599	1	2	78.639253
HG00326_100	2	2	80.21937
NA20814_50	1	2	82.960369
SG9881737_100	2	2	83.188347
NA18594	2	2	87.094854
NA06991	1	2	88.960074
NA18933_100	2	2	89.371968
NA12275_100	2	2	93.343849
GC21416_100	1	2	96.98639
NA18553_100	2	2	97.780269
NA19783_50	2	2	97.902986
NA18633_100	2	2	98.033023
NA18594	2	2	98.935106
NA19005_100	2	2	99.110332
NA19065_100	2	2	101.760569
NA19789_100	2	2	102.344933
HG00475_100	2	2	103.10901
NA20814_100	1	2	104.208546
NA20815_100	2	2	104.479242
NA19001	1	2	110.411256
NA18976	2	2	111.405636
NA19065	2	2	114.399469
NA19055	2	2	114.683915
NA18995	2	2	117.76671
NA18609	1	2	118.301373
NA18995	2	2	139.679885
NA20770	0	2	139.850179
NA18534	2	2	145.082368
NA21575	0	2	153.008411
NA19783_100	2	2	154.102642
NA18603	2	2	159.169288
NA21575	0	2	161.854293
NA18856	2	2	170.977949
NIMH811_50	0	2	202.875425
NA12761_50	0	2	268.702772
NIMH811_100	0	2	279.112721
NA20770_50	0	2	310.207473
HG00261_50	0	2	389.770957
NA12761_100	0	2	494.845954

NA11993	0	2	539.544656
NA20770_100	0	2	540.689146
13398.p1_50	0	2	632.598674
NA06993	0	2	639.267365
NA12156	0	2	690.908492
NA12156	0	2	697.954205
HG00327_50	0	2	710.964006
HG00261_100	0	2	713.274178
NA06993	0	2	766.7774
13398.p1_100	0	2	786.710325
NA11993	0	2	828.339405
HG00327_100	0	2	873.238792

*_100 means the experiment used 100 ng DNA input; _50 means the experiment used 50 ng DNA input; otherwise, the experiment used 200 ng DNA input

discordant genotype

Supplementary Table 8. Regions of GRCb37 missing paralogous sequence.

Paralog-specific copy number of the following regions of GRCb37 was > 2 for > 90 % of 885 individuals (from 1KG) genotyped using reads mapping to singly unique nucleotide k-mers genome-wide. See Supplementary Methods for additional information.

Chromosome	Start	End
chr1	16786166	17125657
chr1	120525184	120697155
chr1	142535435	142731021
chr1	142781023	142967760
chr1	143901186	144095782
chr1	144810725	145401370
chr1	148781362	148954459
chr10	46916792	47161683
chr15	20232225	21217697
chr16	22427520	22722898
chr17	21284461	21364800
chr20	26198816	26319568
chr21	10697897	11188128
chr4	10000	104168
chr4	190440380	190695159
chr5	21476842	21583500
chr6	239986	393297
chr6	57184931	57616325
chr8	2185352	2295948
chr9	66454657	66614194
chr9	67224538	67366295

These start and end coordinates follow the convention for a bed file.

Supplementary Table 9. Summary of MIP capture experiments and sample information.

Experiment name	48-individual experiment	HapMap experiment	Troina2665 MIP capture
Number of individuals*	48	1040	192 (mostly for another study)
Samples analyzed	see Supplementary Table 3	HapMap samples	Troina samples
MIP pool used	Pool 1	Pool 2	Pool 3
Total number of MIPs in pool used	257	207	2361
DNA input	50 ng and 100 ng	200 ng	100 ng
Sequencing platform	Illumina MiSeq	Illumina HiSeq 2000	Illumina HiSeq 2000
Sequencing details	300-cycle v1 reagent kit; read 1—151 cycles; index (barcode) read—8 cycles; read 2—151 cycles	spike in on single lanes (25 % of lane capacity); up to 384 samples per lane	entire single lanes; 192 samples per lane

*some samples were analyzed in both the 48-individual and HapMap experiments (the number of distinct samples analyzed for each sample set is given below)

Samples analyzed
1056 HapMap
3 Signature Genomics
2 Simons Simplex Collection probands
2 Troina
1 NIMH

Sample set	**Description**
HapMap	individuals from International HapMap project populations: African American from the Southwest United States (ASW), Centre d'Etude du Polymorphisme Humain collection (CEU), Han Chinese from Beijing (CHB), Han Chinese South (CHS), Finnish from Finland (FIN), British from England and Scotland (GBR), Japanese from Tokyo, Japan (JPT), Luhya in Webuye, Kenya (LWK), Maasai from Kinyawa, Kenya (MKK), Mexican ancestry in Los Angeles, California (MXL), Toscani in Italia (TSI), and Yoruba from Ibadan, Nigeria (YRI)
Signature Genomics	see table S4 of reference 13
Simons Simplex Collection	see reference 67 and table S4 of reference 13
NIMH	from the National Institute of Mental Health
Troina	individuals with intellectual disability from Troina, Italy

Supplementary Table 10. Automated copy number genotyping heuristics.

Heuristics governing allowed copy-number state transitions**
No more than 2 copy-number state transitions allowed
Copy-number state transitions must either affect a single paralog only or affect 2 paralogs reciprocally, such that aggregate copy number is the same between pre-transition and post-transition copy-number states

Heuristics for calling 1 *SRGAP2* copy-number state transition**
Score of highest-scoring 1-transition path - score of highest-scoring 0-transition path must be > 40
Must have ≥ 5 MIPs in both called *SRGAP2* copy-number states
Score of highest-scoring 2-transition path - score of highest-scoring 1-transition path must be ≤ 40

Heuristics for calling 2 *SRGAP2* copy-number state transitions**
Score of highest-scoring 2-transition path - score of highest-scoring 0-transition path must be > 40
Must have ≥ 5 MIPs in all 3 called *SRGAP2* copy-number states*
Score of highest-scoring 2-transition path - score of highest-scoring 1-transition path must be > 40

*Three copy-number states can mean 3 different copy-number states over the spatial extent of duplicated *SRGAP2* sequence, or instead mean 2 different such states, where the one spatially in the middle differs from the others, which are the same. Having 2 copy-number state transitions divides the spatial extent of duplicated *SRGAP2* sequence into 3 regions, each of which, according to this heuristic, must include at least 5 MIPs.

**These heuristics can of course be adjusted to make automated calling of internal structural variation and gene conversion more aggressive or more conservative and allow for the possibility of calling multiple internal events in an automated fashion. However, they seemed to work well for *SRGAP2*. Visual inspection of paralog-specific count frequency plots is recommended when assessing all automated calls. In addition to providing a sense of whether any given call appears to reflect a real event, this allows multiple internal events to be detected as long as one is called by the automated caller.

Supplementary References

62. Wang J. *et al*. The diploid genome sequence of an Asian individual. *Nature* **456**, 60-65 (2008).
63. Park H. *et al*. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.* **42**, 400-405 (2010).
64. Ahn S.M. *et al*. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res*. **19**, 1622-1629 (2009).
65. Schuster S.C. *et al*. Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943-947 (2010).
66. 1000 Genomes Project Consortium *et al*. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
67. Fischbach G.D. & Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192-195 (2010).