The American Journal of Human Genetics, Volume *93*

Supplemental Data

Reliable Identification of Genomic Variants from RNA-Seq Data

Robert Piskol, Gokul Ramaswami, and Jin Billy Li

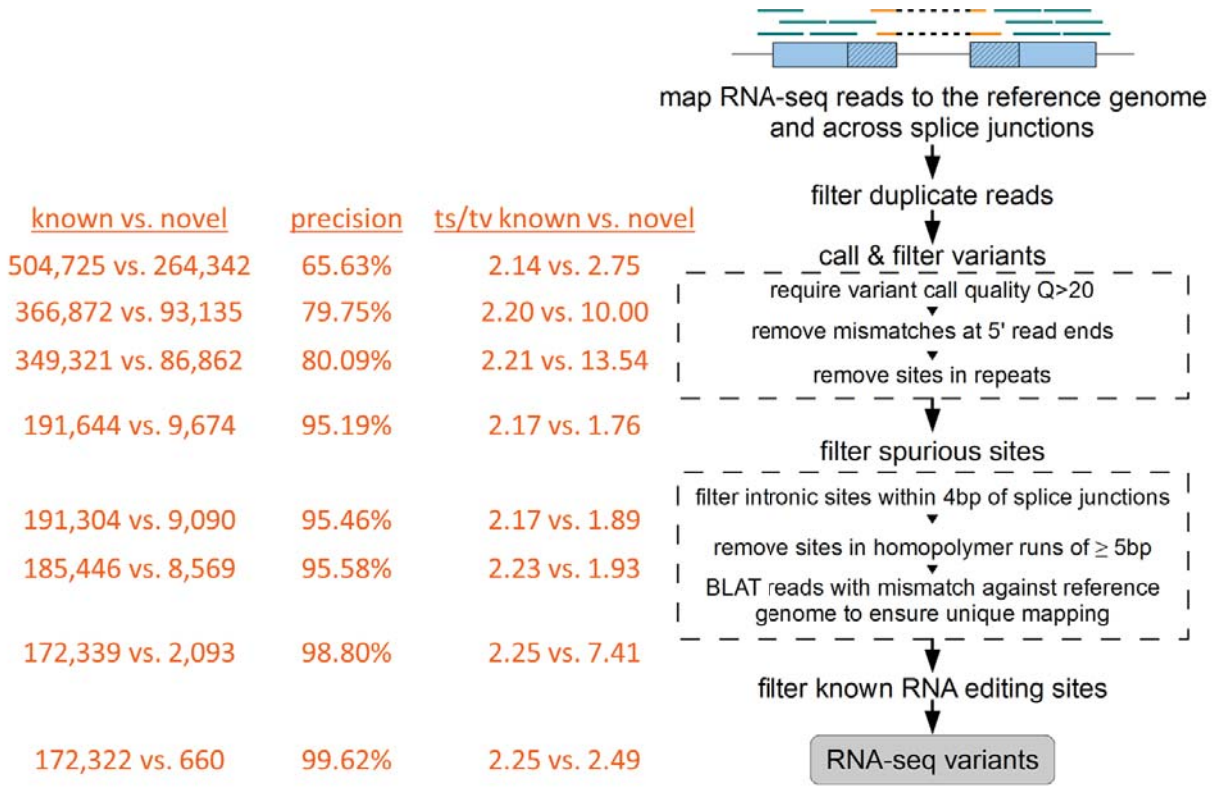


Figure S1. Outline of the filtering procedure applied in our computational pipeline. Orange numbers describe the number of known/novel GM12878 sites that remained after each filtering step, as well as the precision and the ts/tv ratios after each filtering step.

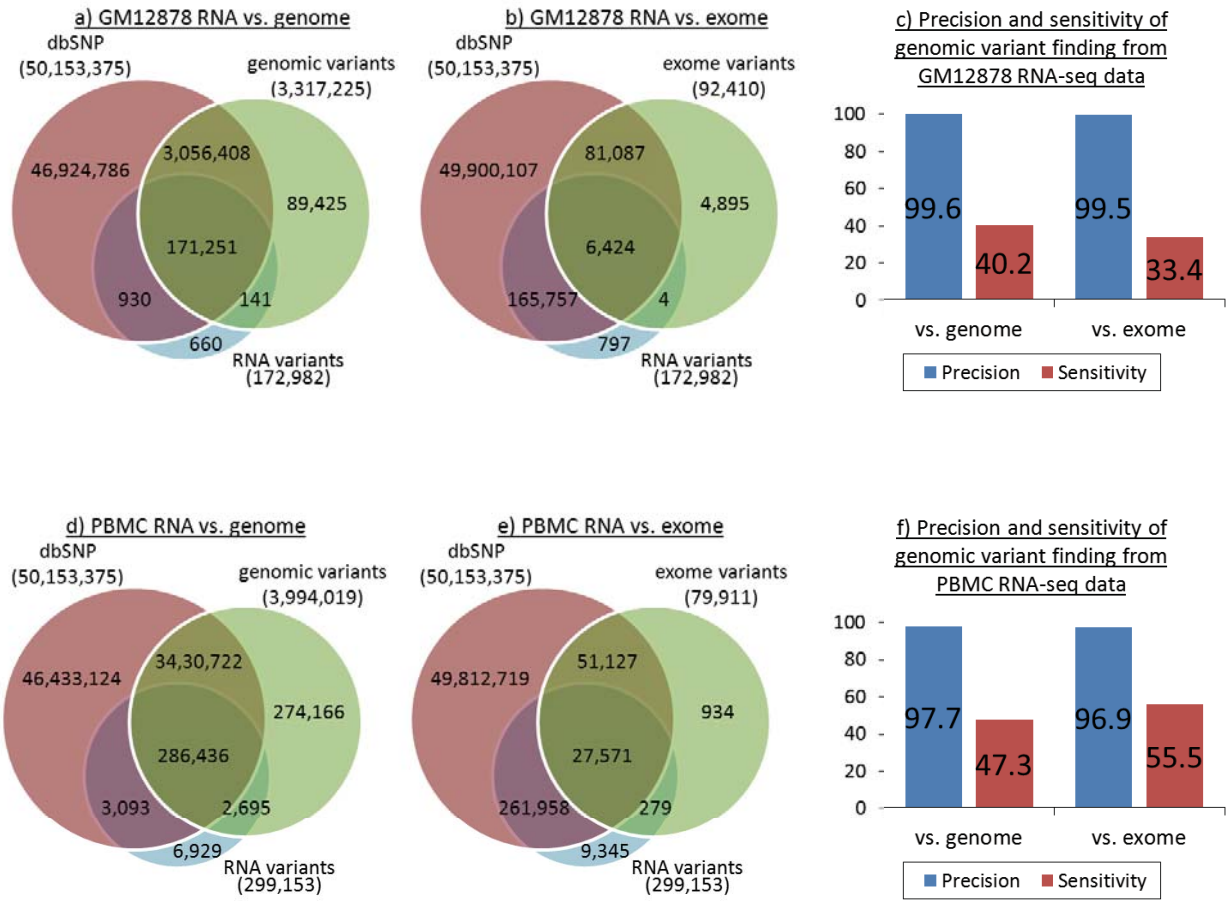


Figure S2. Comparison of SNPs identified via RNA-seq and genome sequencing of GM12878/PBMCs or present in dbSNP. (**a,b**) Overlap of RNA-seq and genome/exome variants of GM12878. (**d,e**) Comparison of RNA-seq variants and genome/exome variants in PBMCs. (**c,f**) Precision and sensitivity of RNA variant finding in GM12878 and PBMCs. Precision was calculated as the percentage of all RNA-seq variants confirmed by WGS/WES or present in dbSNP. Sensitivity describes the percentage of all exonic variants found through WGS/WES of GM12878/PBMCs and recovered with RNA-seq data of the same data set. (Note: circles are schematics and therefore not to scale).

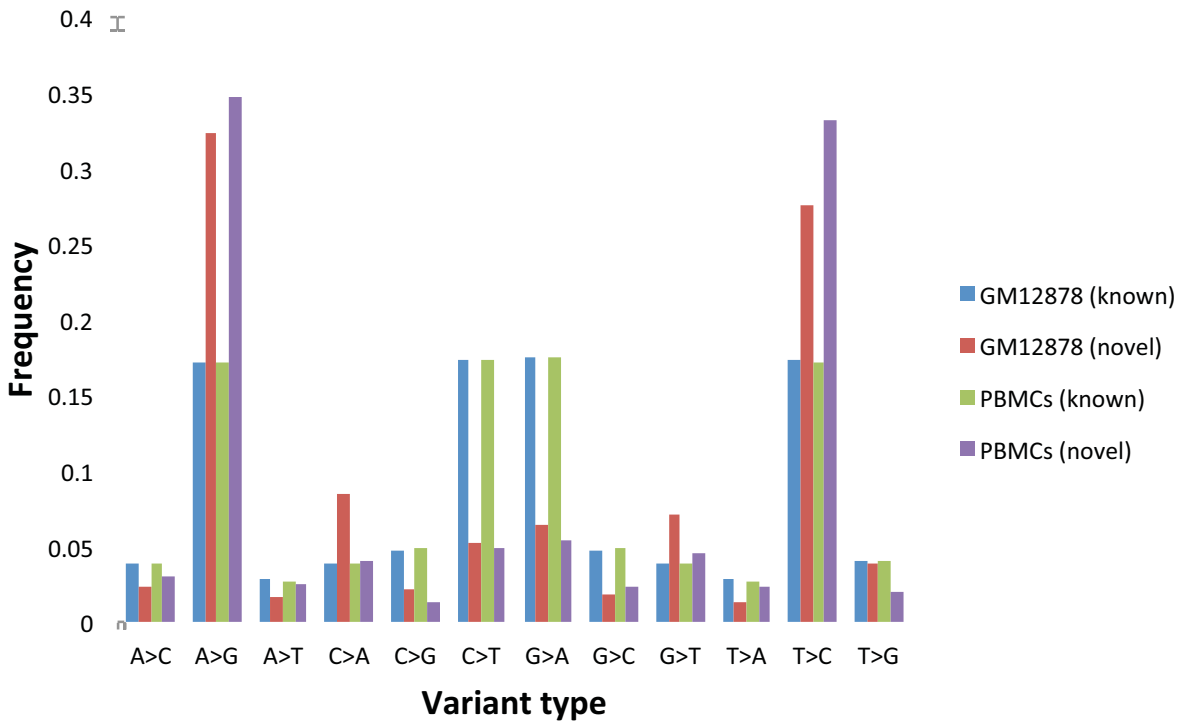


Figure S3. Types of known and novel variants in the GM12878 and PBMC datasets.

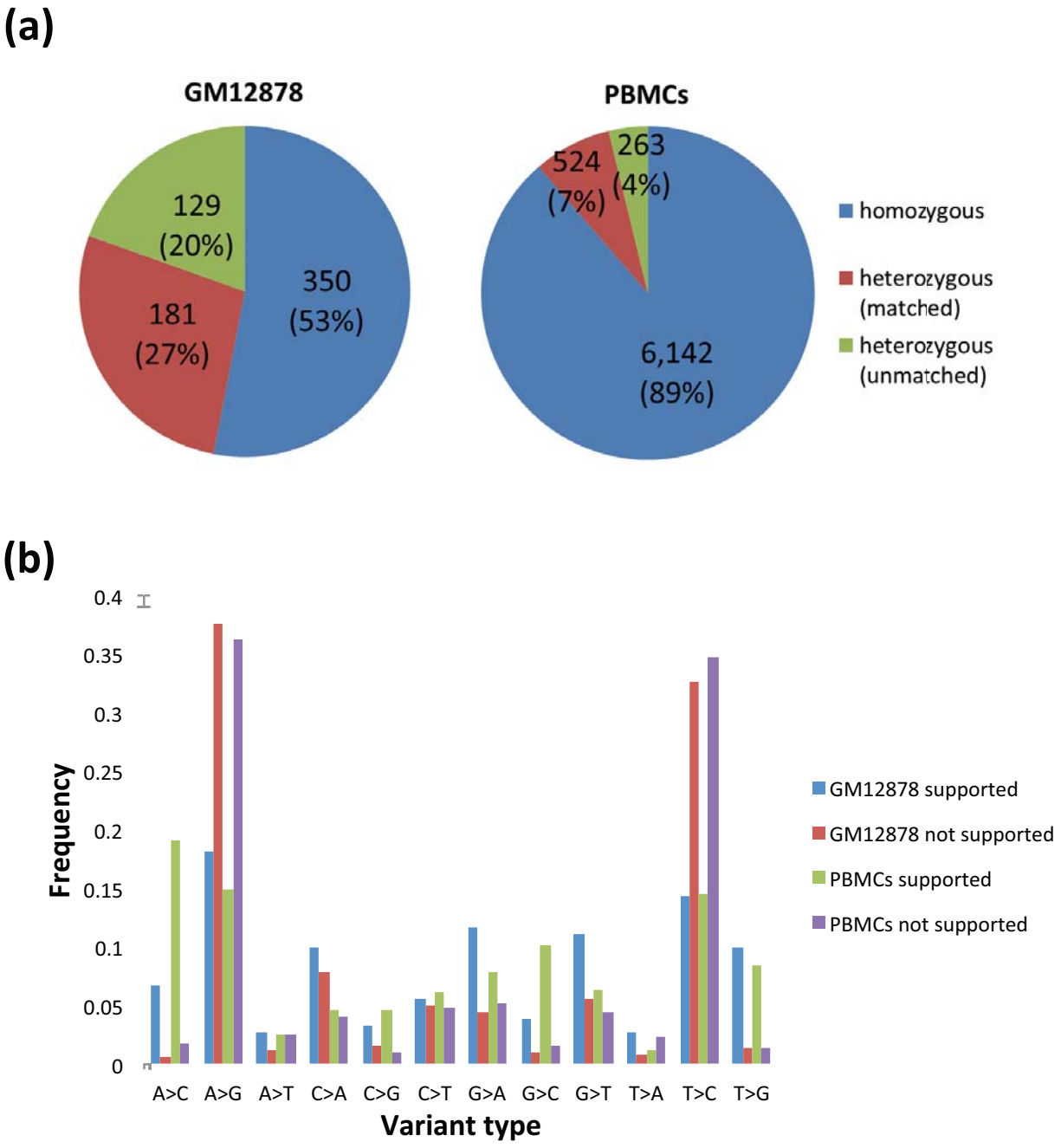


Figure S4. Novel variants in GM12878 and PBMCs. (**a**) Fractions of novel RNA-seq variants with evidence for support from WGS (homozygous: no variation in WGS; heterozygous matched: variation in WGS matches RNA-seq variant; heterozygous unmatched: variation in WGS does not match RNA-seq variant), (**b**) types of WGS-supported (heterozygous matched) and WGS-unsupported (homozygous, heterozygous unmatched) novel variants in (**a**).

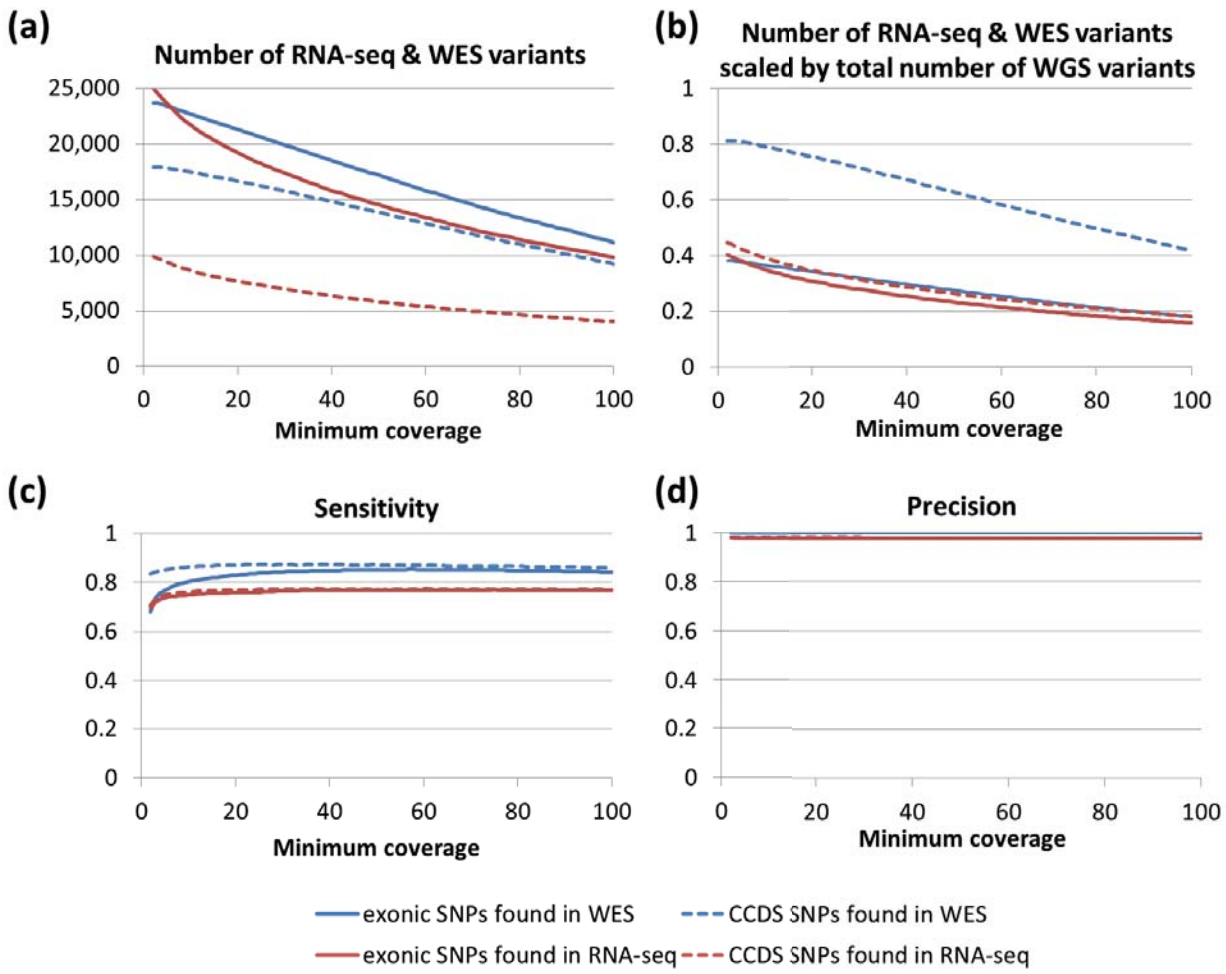


Figure S5. Comparison of sensitivity and precision between variant calling from RNA-seq and exome-seq experiments. We identified all genomic SNPs in (1) exonic regions (coding exons & UTRs) and (2) consensus CDS (CCDS) regions. Subsequently, RNA-seq and WES variants that were found in these regions served to determine (a,b) the number and scaled number of RNA-seq & WES variants, (**c**) the sensitivity and (**d**) the precision depending on the RNA-seq and WES coverage. Sensitivity was determined as the number of RNA-seq/WES variants that had the same type as the genomic variant, divided by the total number of genomic variants with a certain minimum coverage. Precision was calculated as the number of correctly identified genomic variants divided by the total number of detected RNA-seq/WES variants (variants with matching type as well as variants that do not match the WGS type or were not found in WGS).

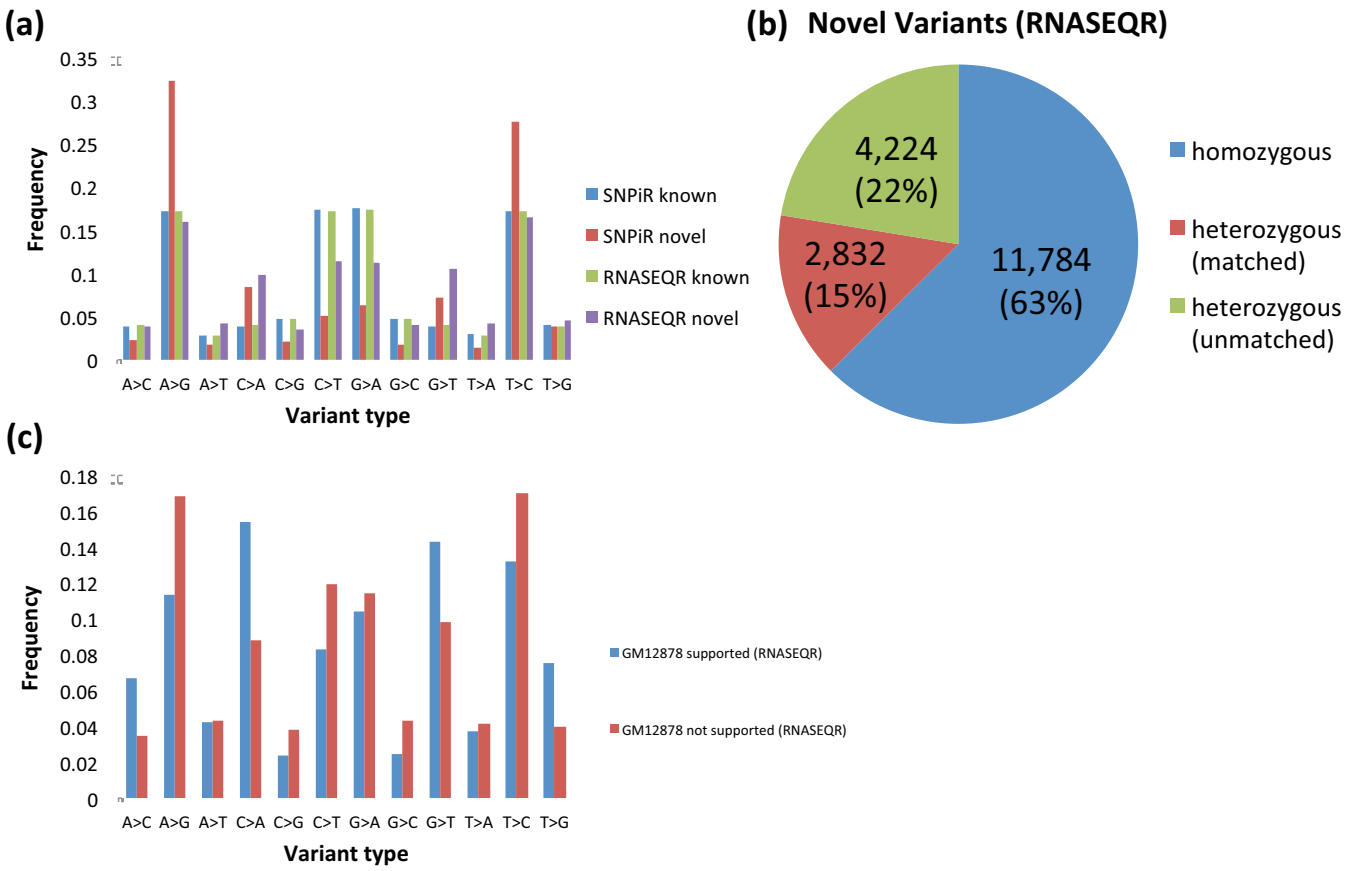
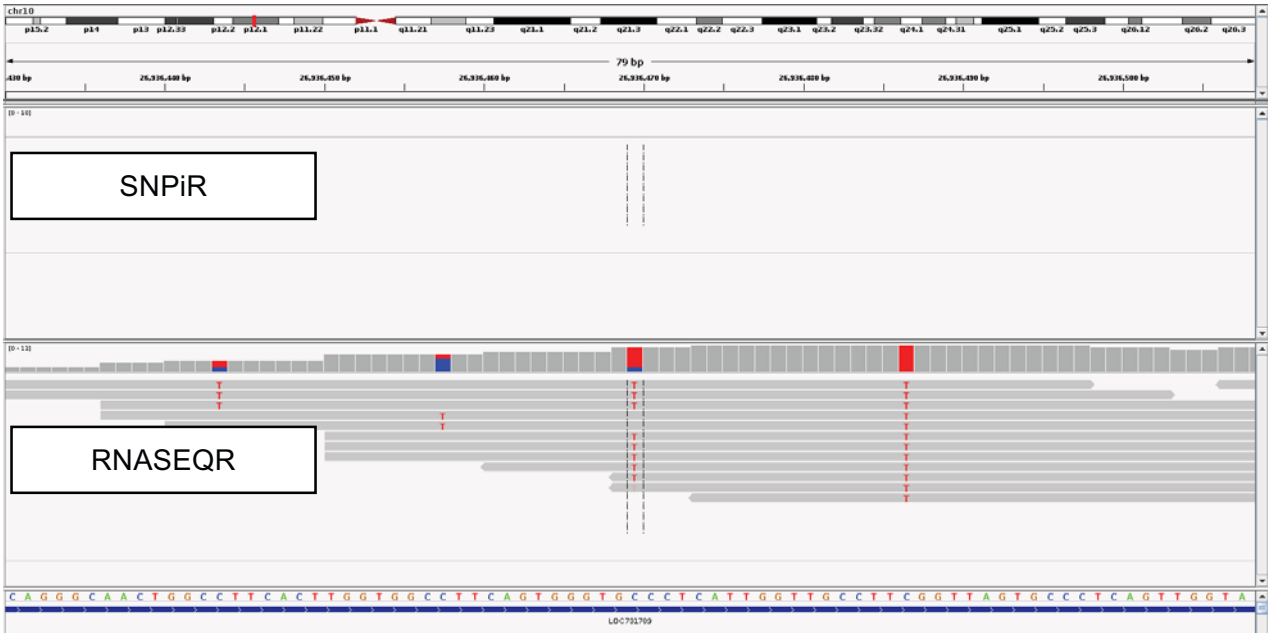


Figure S6. Comparison of SNPiR and RNASEQR variants. (a) Mutational Spectrum for novel and known variants found by our method (SNPiR) and RNASEQR in the GM12878 data set (SNPiR variants are taken from Supplementary Figure 3), **(b)** fractions of novel RNASEQR variants with evidence for support from WGS (homozygous: no variation in WGS; heterozygous matched: variation in WGS matches RNA-seq variant; heterozygous unmatched: variation in WGS does not match RNA-seq variant), and (**c**) types of WGS-supported (heterozygous matched) and WGS-unsupported (homozygous, heterozygous unmatched) novel variants in (**b**).

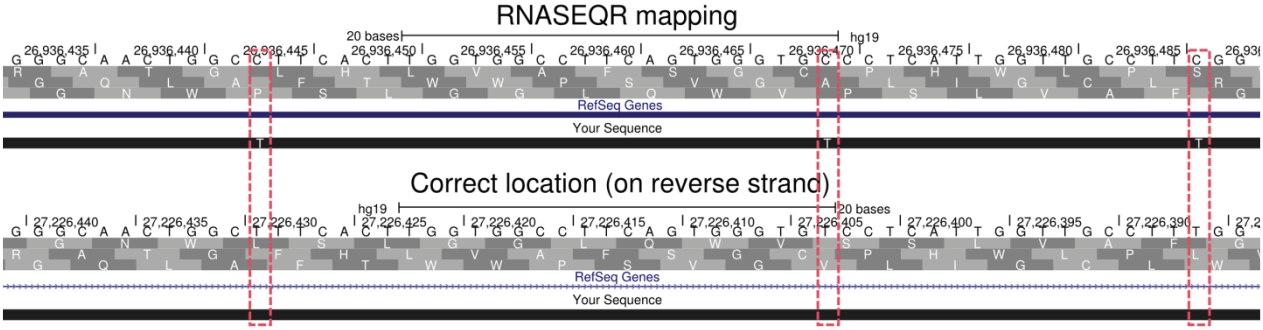
Figure S7. Sequence mapping details

(**a**) Position chr10:26,936,469:

Example for variation in RNASEQR mapping that is caused by not-uniquely mapped reads.

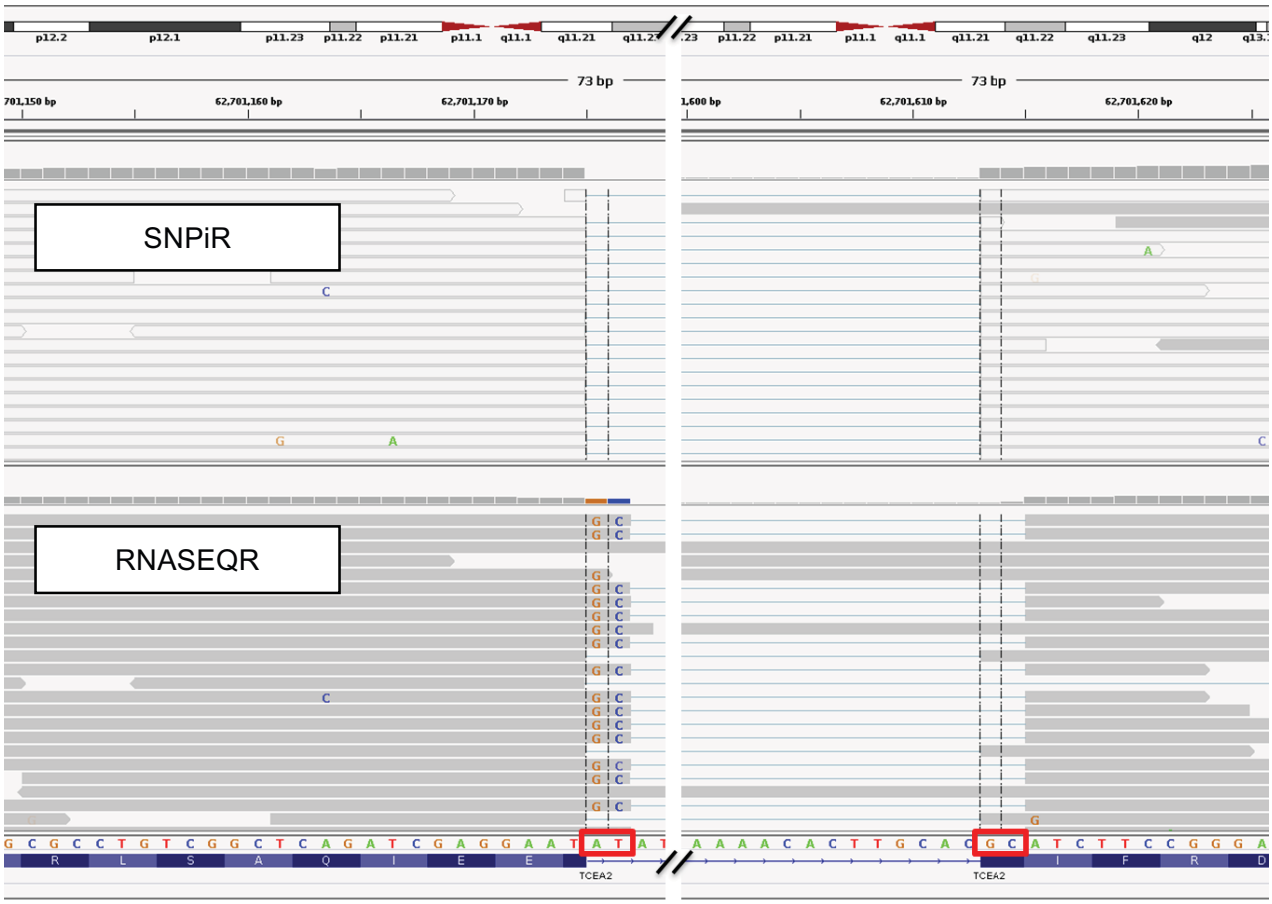


Comparison between the mapping location assigned by RNASEQR (see above) and the original location of a read in a different region of the genome. Mismatches in the RNASEQR mapping correspond exactly to reference nucleotides in the correct location:



(**b**) Position chr20:62701176:

Example for variation in RNASEQR mapping that is caused by incorrectly spliced reads. The two sides of a splicing junction are shown.



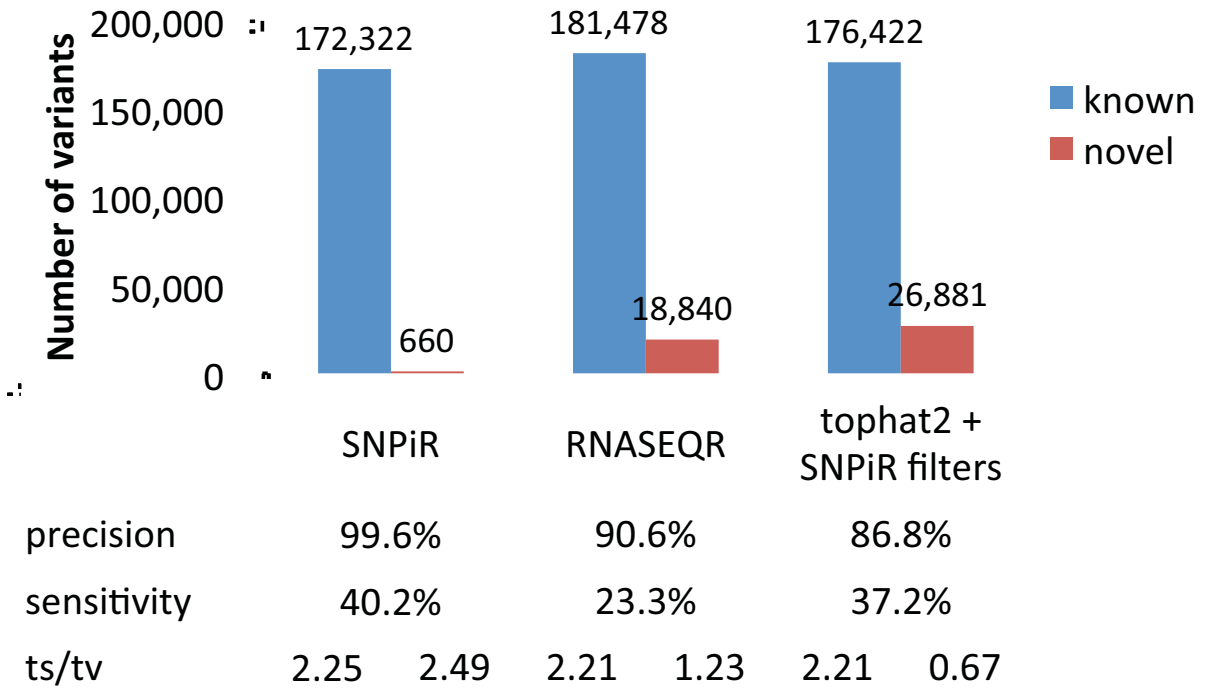


Figure S8. Comparison of total number of RNA-seq variants, precision and sensitivity between SNPiR, RNASEQR and tophat2. For tophat2 all variants were filtered using the SNPiR procedure.

Table S1. Summary of the data used in this work

	GM12878			PBMCs		
	RNA-seq	**WGS**	**WES**	**RNA-seq**	**WGS**	**WES**
Number of reads (million)	499.4	3,218.0	103.1	3,232.0	1,233.9	124.2
Number of filtered reads (million)[1]	110.3	2.580	45.1	500.1	1,166.9	94.1
Read length	76	36 – 76	36 – 101	101	100	101
Number of variants	172,982	3,317,225	92,410	299,153	3,944,019	79,911

[1] after removal of reads that cannot be uniquely mapped by BWA, that map to exactly the same location and have a mapping quality q<10

Number of paired end reads and total discovered nucleotide variants for RNA-seq, WGS and exome-seq (WES) of GM12878 and PBMCs. WES data is based on the Agilent SureSelect kit.

Table S2. Evaluation of SNPiR on random subsamplings of the GM12878 RNA-seq data

			All variants				Exonic variants				
Sample name	sample size	all	known	novel	precision	SD (known)	all	known	novel	**precision**	SD (known)
5_1	5000000	9472	9312	160	0.9831	30.05	2064	2028	23	0.9826	8.33
5_2	5000000	9445	9310	135	0.9857		2068	2033	24	0.9831	
5_3	5000000	9505	9373	132	0.9861		2052	2021	17	0.9849	
10_1	10000000	17321	17072	249	0.9856	107.87	3244	3195	26	0.9849	24.66
10_2	10000000	17141	16910	231	0.9865		3249	3195	34	0.9834	
10_3	10000000	17128	16896	232	0.9865		3204	3155	22	0.9847	
20_1	20000000	33219	32777	442	0.9867	79.57	4894	4809	35	0.9826	5.57
20_2	20000000	33062	32635	427	0.9871		4883	4802	35	0.9834	
20_3	20000000	33163	32754	409	0.9877		4890	4816	32	0.9849	
50_1	50000000	62820	62015	805	0.9872	47.96	6530	6399	43	0.9799	6.08
50_2	50000000	62739	61935	804	0.9872		6529	6402	41	0.9805	
50_3	50000000	62824	62027	797	0.9873		6519	6394	39	0.9809	
100_1	100000000	97017	95846	1171	0.9880	160.13	7618	7470	43	0.9806	20.55
100_2	100000000	96709	95564	1145	0.9882		7621	7473	41	0.9806	
100_3	100000000	96787	95636	1151	0.9881		7584	7435	47	0.9804	
complete set	499900000	172982	172322	660	0.9962		9803	9774	29	0.9970	

Known variants describe all sites that were confirmed through WGS or were present in dbSNP. Novel variants could not be identified in any of these two sources.

Table S3. Comparison of SNPiR and RNASEQR mapping and variant calling results for GM12878 and PBMCs

	GM12878		PBMCs	
	SNPiR	**RNASEQR**	**SNPiR**	**RNASEQR**
Number of reads (million)	499.2		3,232.0	
Read length (base)	76		101	
Number of filtered reads (million)	110.3	176.8	1,166.9	1,400.9
Total number of variants	172,982	200,318	299,153	315,945
Number of known variants	172,322	181,478	292,224	275,013
Number of novel variants	660	18,840	6,929	40,923