Figures & Supplementary materials for: Joint genotyping of public RNA-seq samples to identify tissue specific trans-eQTLs.

Figure 1.

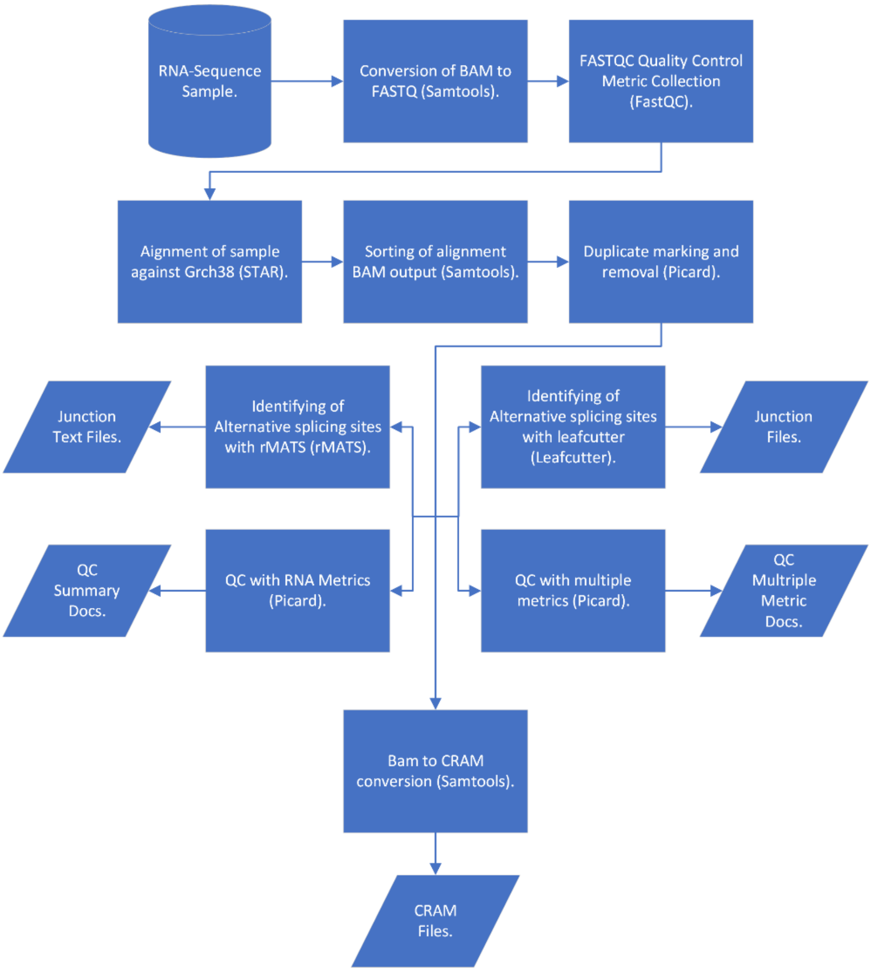
RNA-Sequence Sample.
Conversion of BAM to FASTQ (Samtools).
FASTQC Quality Control Metric Collection (FastQC).
Aignment of sample against Grch38 (STAR).
Sorting of alignment BAM output (Samtools).
Duplicate marking and removal (Picard).
Junction Text Files.
Identifying of Alternative splicing sites with rMATS (rMATS).
Identifying of Alternative splicing sites with leafcutter (Leafcutter).
Junction Files.
QC Summary Docs.
QC with RNA Metrics (Picard).
QC with multiple metrics (Picard).
QC Multriple Metric Docs.
Bam to CRAM conversion (Samtools).
CRAM Files.

Figure 1. Flowchart of RNA-seq alignment pipeline. Showing the order of execution and Nextflow structure of the subprocesses. Starting with conversion and quality control steps and proceeding to the STAR alignment of a sample sequence against the human reference genome. Followed by identification of splicing sites and BAM to CRAM conversion.

Figure 2.

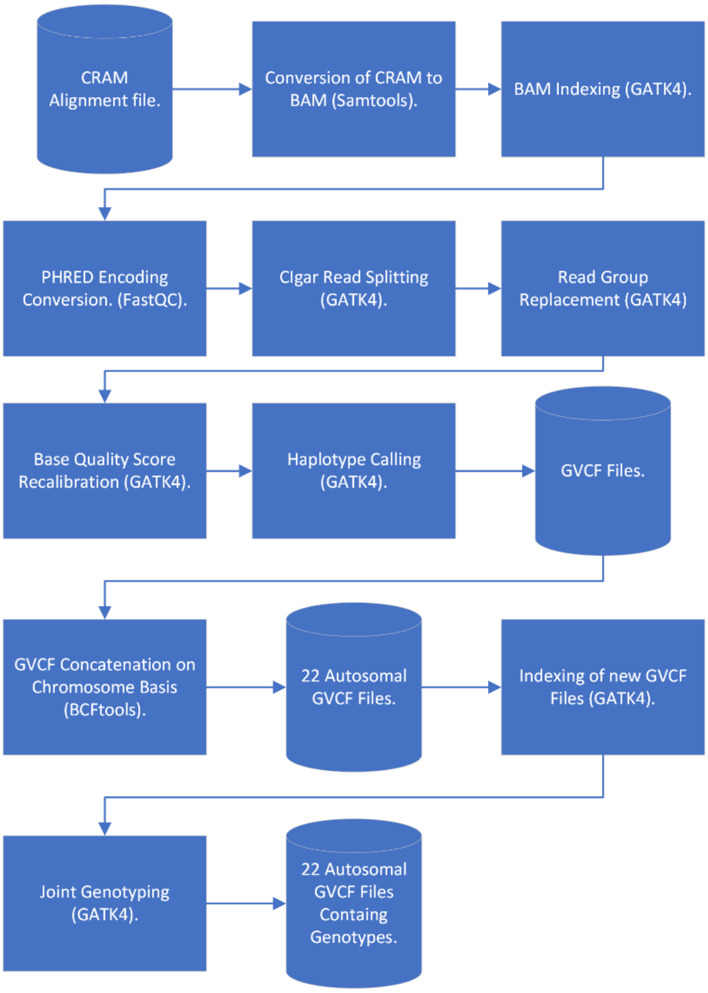
CRAM Alignment file.
Conversion of CRAM to BAM (Samtools).
BAM Indexing (GATK4).
PHRED Encoding Conversion. (FastQC).
CIgar Read Splitting (GATK4).
Read Group Replacement (GATK4)
Base Quality Score Recalibration (GATK4).
Haplotype Calling (GATK4).
GVCF Files.
GVCF Concatenation on Chromosome Basis (BCFtools).
22 Autosomal GVCF Files.
Indexing of new GVCF Files (GATK4).
Joint Genotyping (GATK4).
22 Autosomal GVCF Files Containg Genotypes.

Figure 2. Flowchart of RNA-seq based alignment genotyping pipeline. This starts out with format conversion steps which are required for the splitting for cigar read splitting and base quality score recalibration. After which the haplotypes are called and merged on chromosome basis. The resulting GVCF file is then passed to the joint genotyping step, which produces a list of variants found for the samples and their corresponding genotypes.

Figure 3.

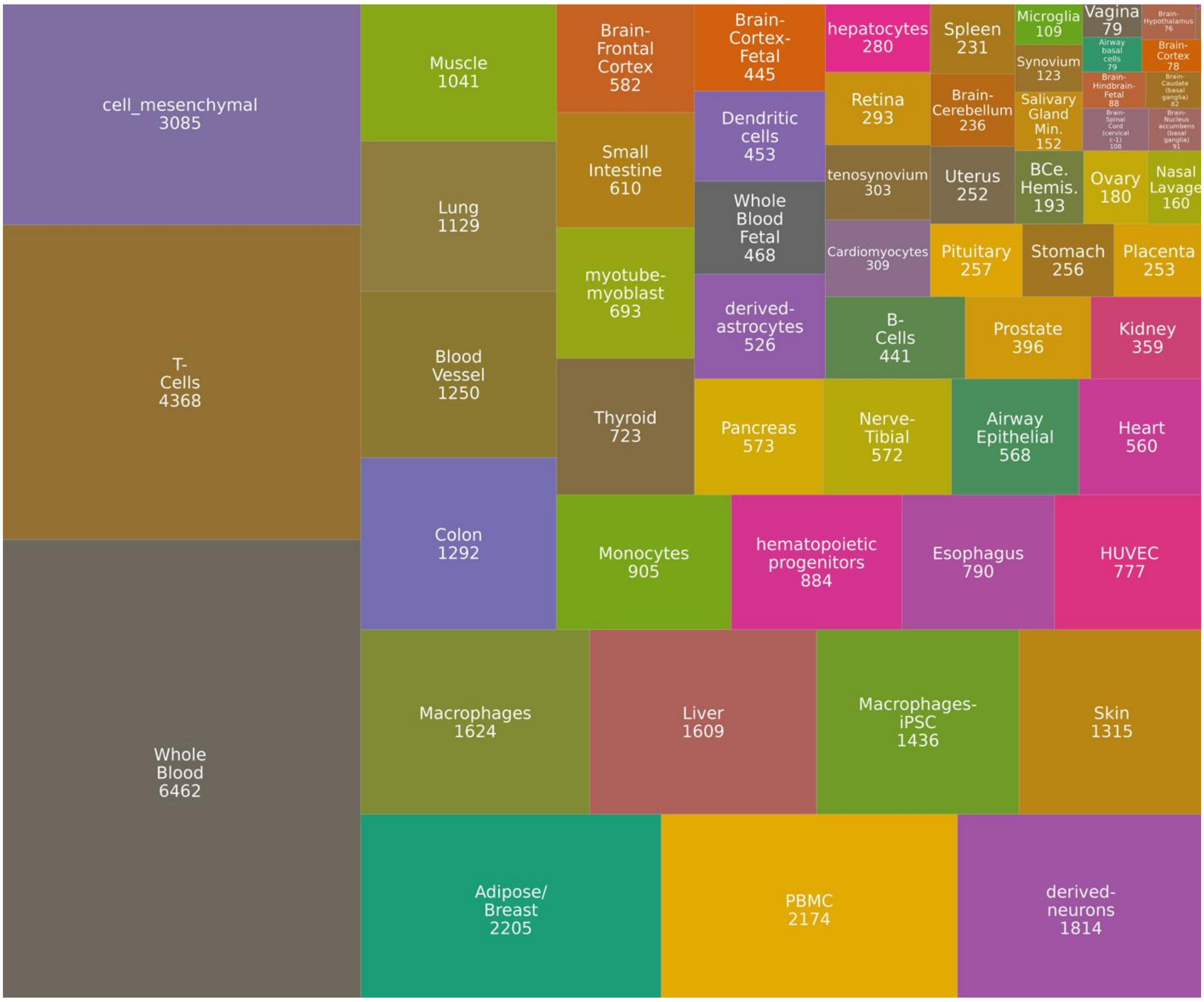

Figure 3. A tissue distribution of all the 46,410 public RNA-seq samples. This figure shows that there are hundreds of samples representing different tissues. This does now show any filtering or thresholds steps. It is a display of the raw metadata state.

Figure 4.

Minimum Threshold	Studies Kept	Studies Discarded	Samples Kept	Samples Discarded
60	146	1,991	29,880	16,530
50	169	1,968	31,118	15,292
40	211	1,926	32,959	13,451
30	269	1,868	34,933	11,477
20	376	1,761	37,501	8,909
10	656	1,481	41,282	5,128

Figure 4. Showing the varying amounts of studies and samples lost at potential thresholds for the minimum number of samples per study. This shows that there are thousands of samples which are part of studies that only contain a couple of samples for a specific tissue.

Figure 5.

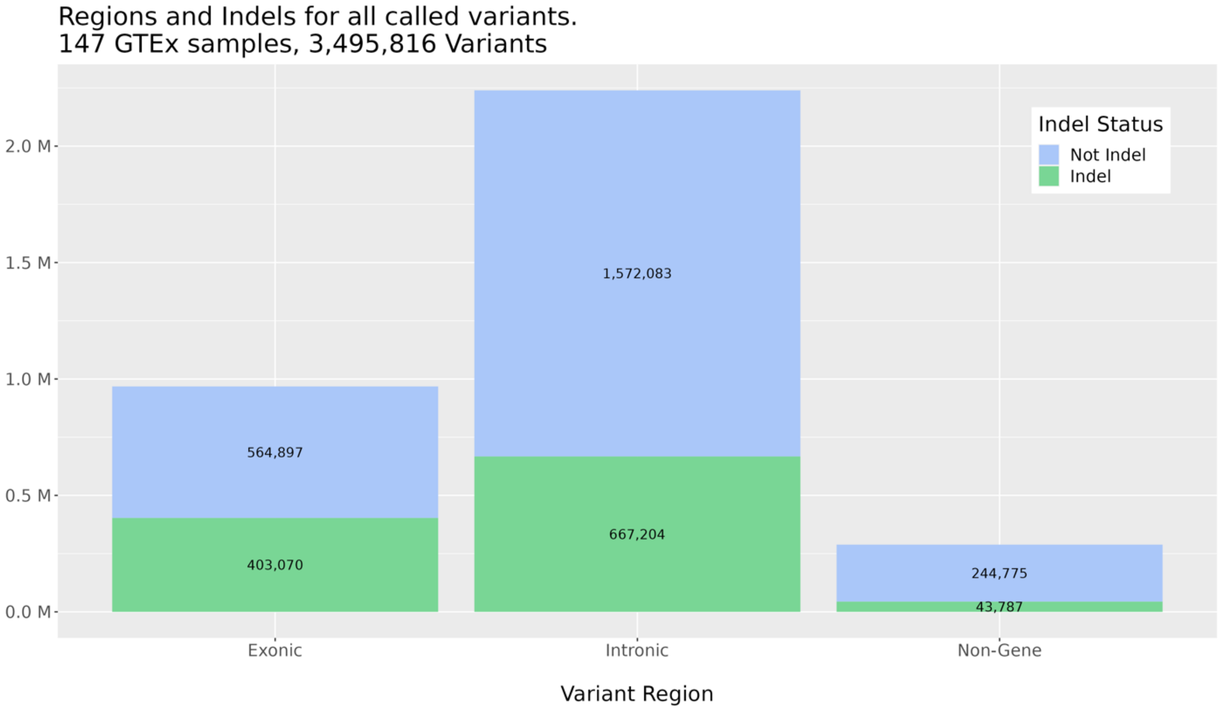

Figure 5. Overview of the variant locations and how much of their respective regions is assigned to insertions/deletions or SNPs. A total of 3,495,816 variants are further divided into groups of the regions they are located in. These variants are called for 147 GTEx whole blood samples.

Figure 6.

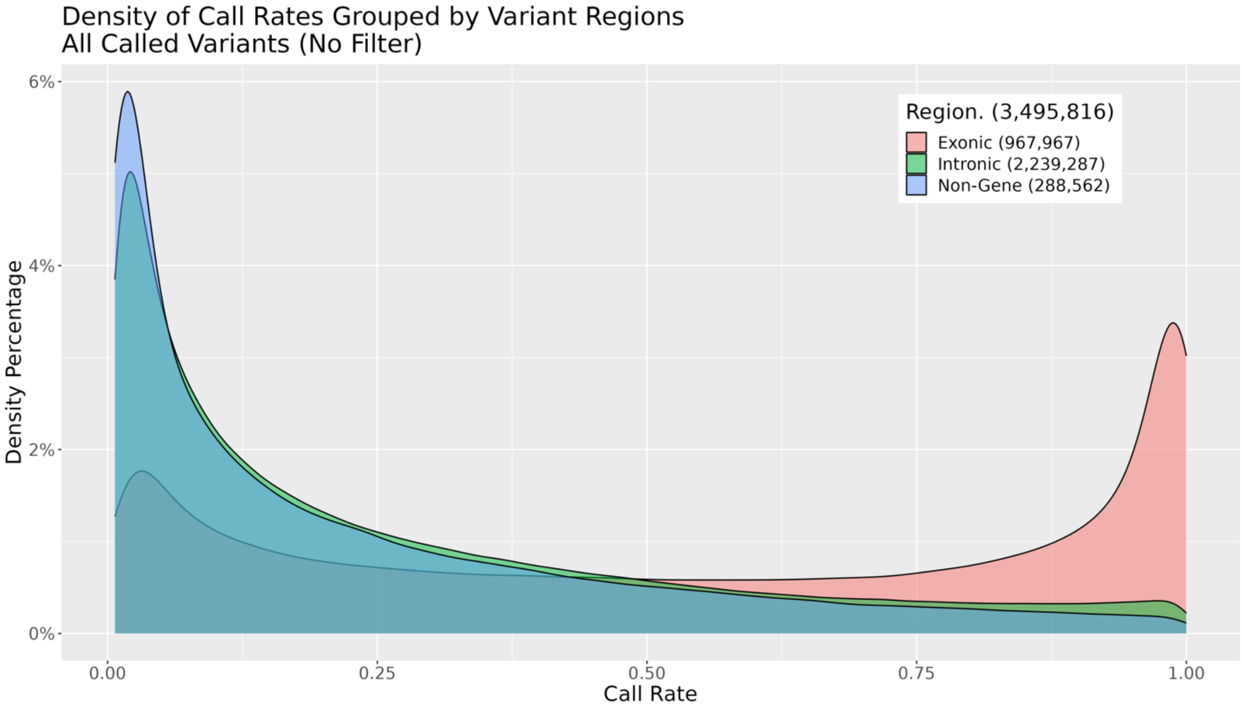

Figure 6. Call rates for all called variants shown for each variant region. This shows a distinct difference in the distribution of call rates depending on the region a variant was found in. This figure also introduces the region distribution for all variants. Variants in exonic regions contains 967,967 variants, while those located in intronic, and non-gene regions encapsulate 2,239,287 and 288,562 variants, respectively. The call rates for the intronic and non-gene variants are notable.

Figure 7.

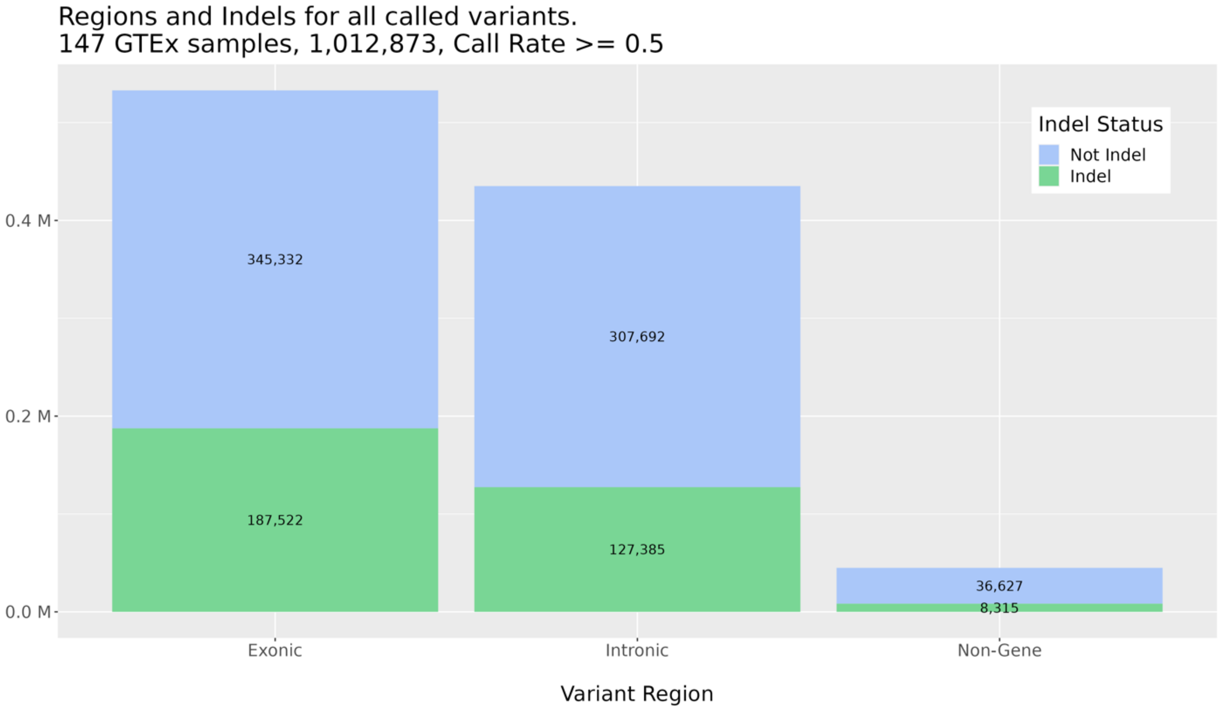

Figure 7. Overview of the variant locations and how much of their respective regions is assigned to insertions/deletions or SNPs after applying a minimum call rate filter of 0.5. This resulted in the removal of 2,482,943 variants (71%) of the original variants.

Figure 8.

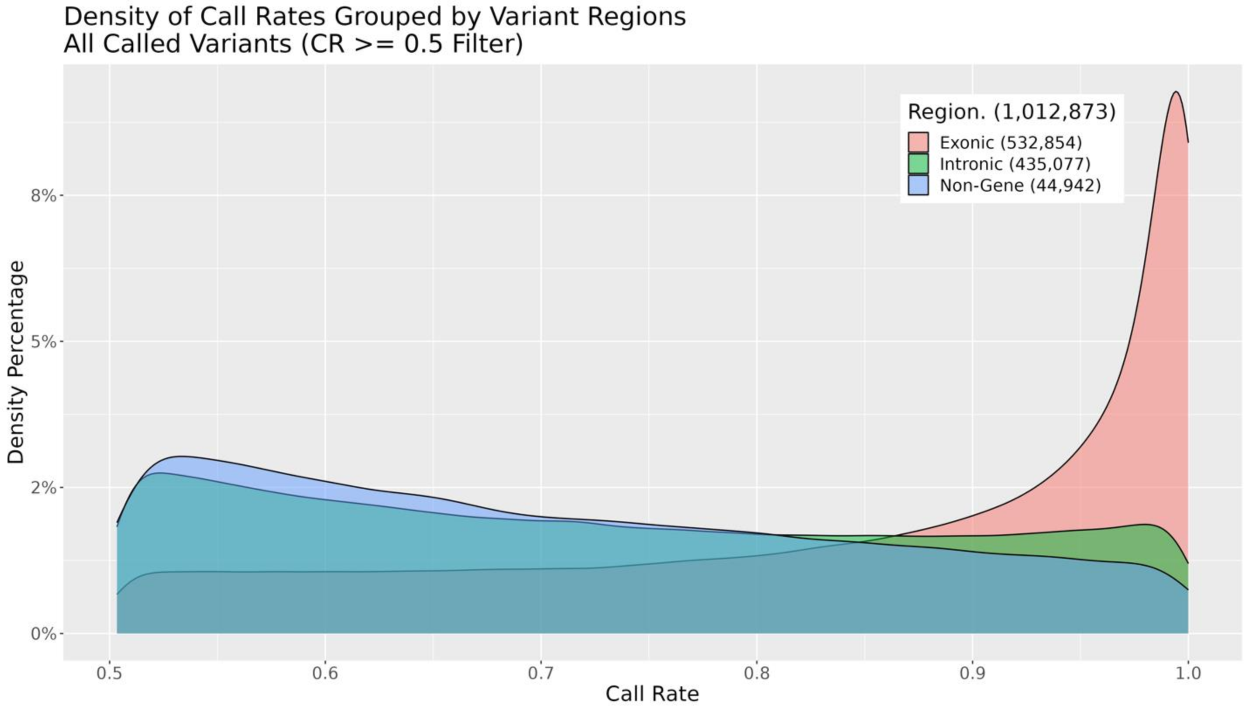

Figure 8. Call rates for all called variants shown for each variant region after a minimum call rate filter of 0.5. This shows a distinct difference in the distribution of call rates depending on the region a variant was found in. The figure shows a substantial difference between the initial distributions and the post filter one. 1,012,873 Variants are shown here, for which 532,854 (52%) variants are now located in exonic regions, while the intronic variants now only make up 43% of the data. The variants outside genes only make up 44, 942 (5%) variants of the data.

Figure 9.

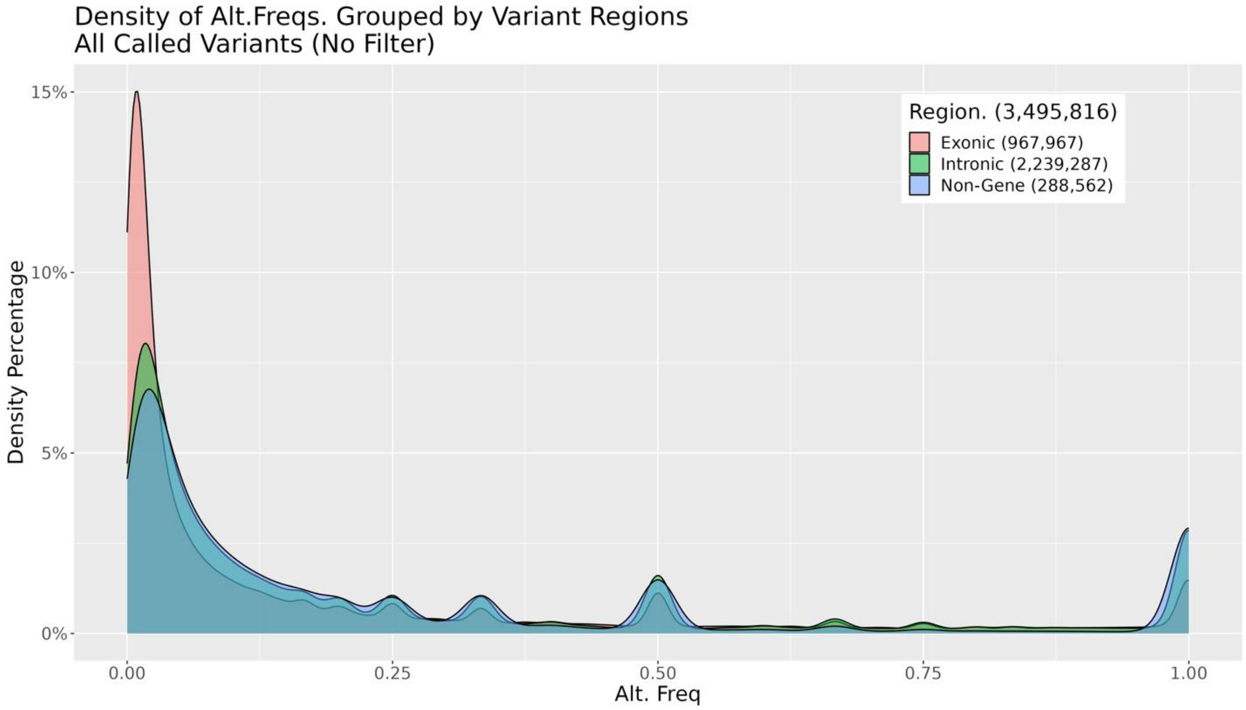

Figure 9. Alternative allele Frequencies for all called variants shown for each variant region. For the 3,495,816 variants, it shows that most of the data is of a lower call rate. This means that many of the variants are quite rare.

Figure 10.

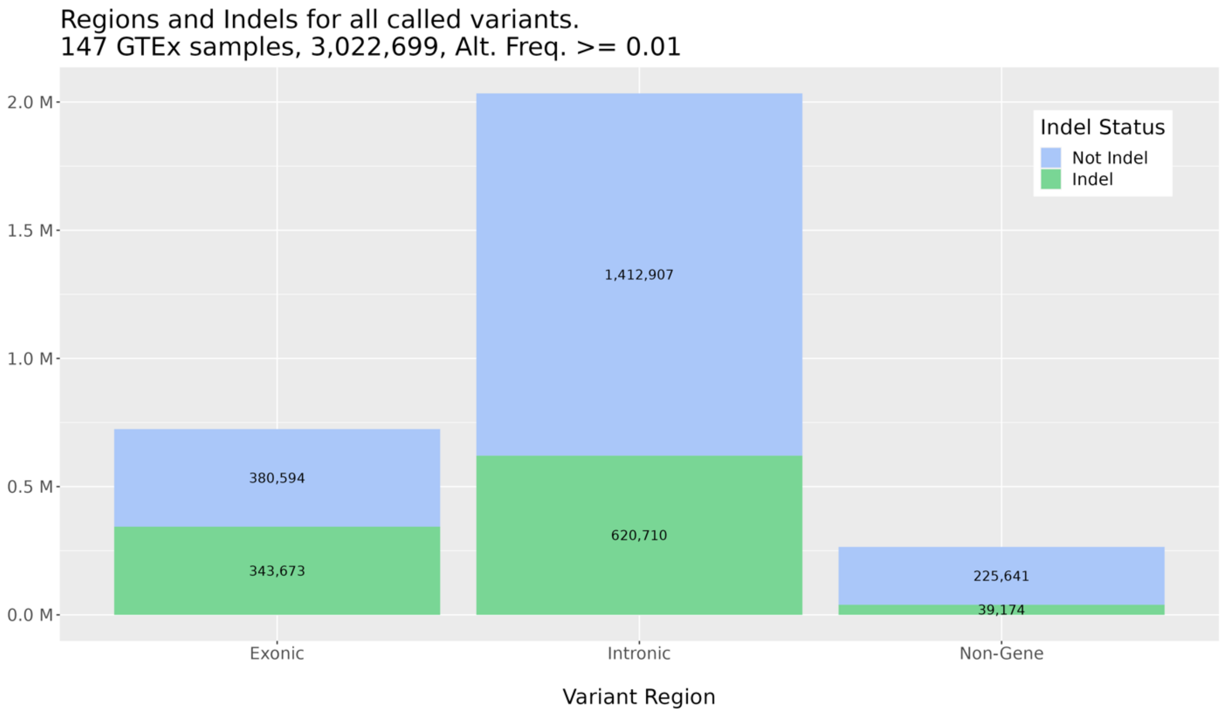

Figure 10. Overview of the variant locations and how much of their respective regions is assigned to indels or SNPs after applying a minimum alt. freq. filter of 0.01. This resulted in the removal of 473,117 variants (14%) of the original variants.

Figure 11.

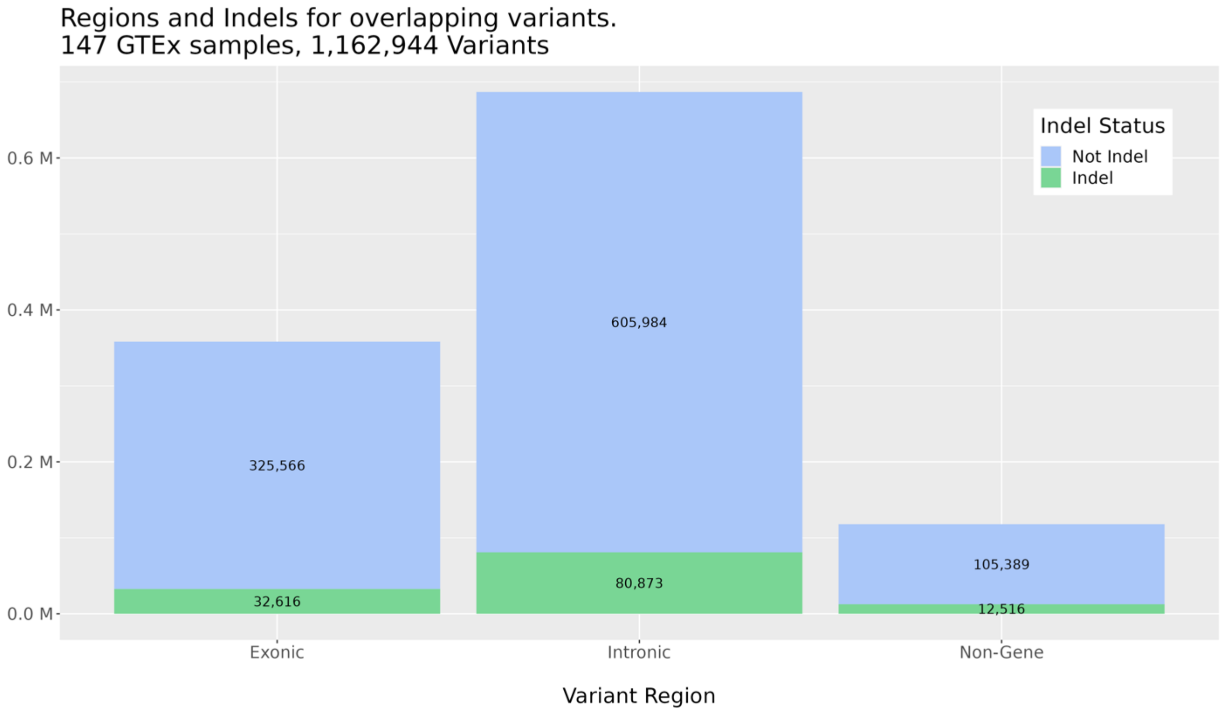

Figure 11. Overview of the variant locations and how much of their respective regions is assigned to insertions/deletions or SNPs, for all variants that are present in both our pipeline output and WGS validation genotypes. 1,162,944 variants are present in both variant call files. This figure shows that most overlapping variants are SNPs, not insertions or deletions.

Figure 12.

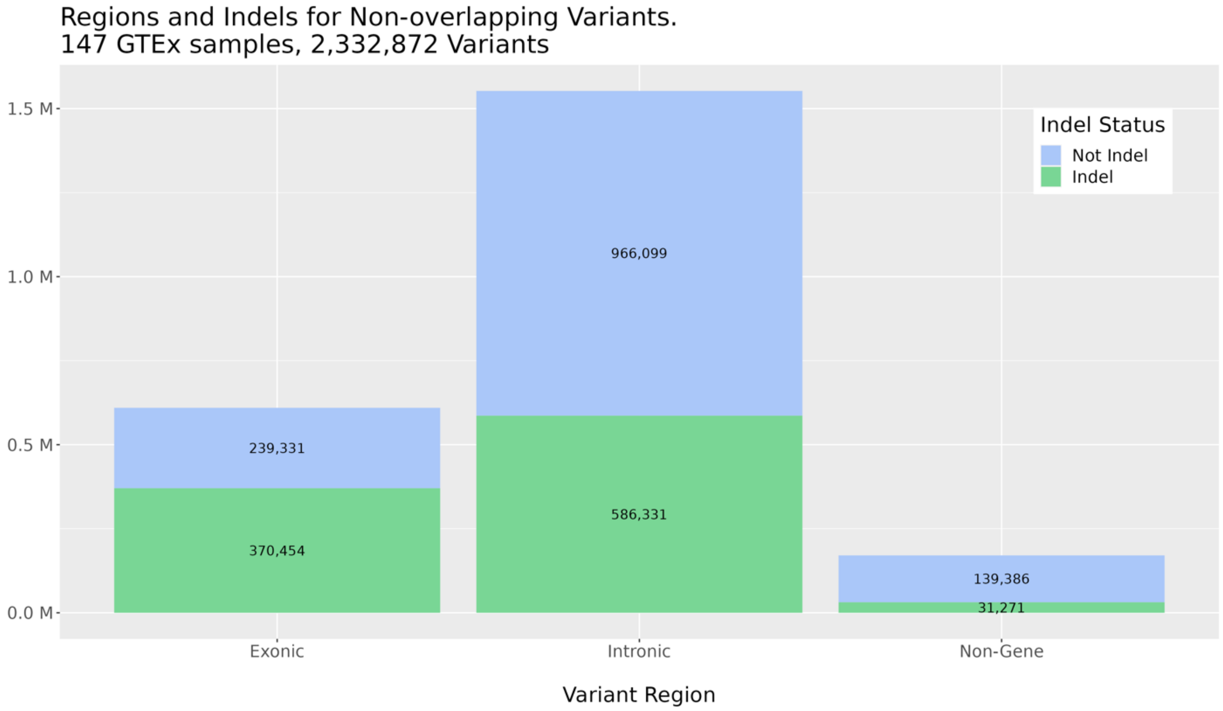

Figure 12. Overview of the variant locations and how much of their respective regions is assigned to insertions/deletions or SNPs, for all variants that are absent in the WGS variant call file but present in the RNA-seq based variant call file. 2,332,872 variants do not overlap between both files.

Figure 13.

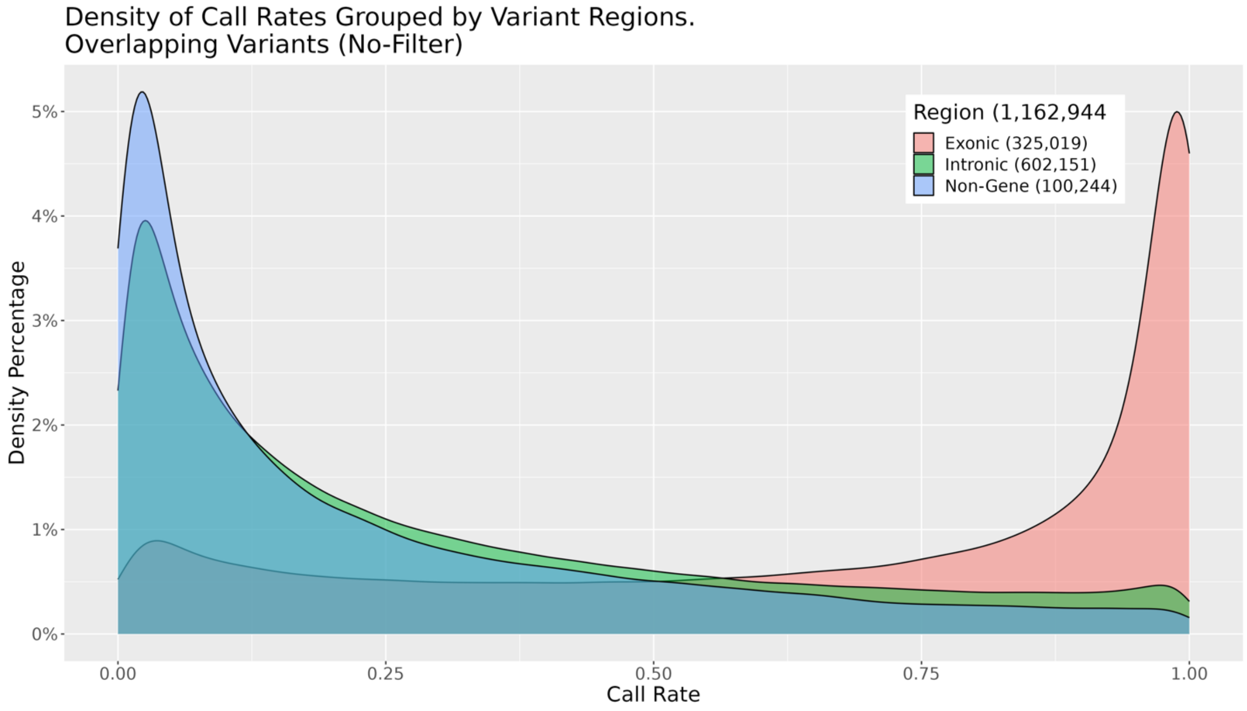

Figure 13. Call rates for all overlapping variants shown for each variant region. This once again shows that there is a great divide between the call rate distribution when looking at the different variant regions.

Figure 14.

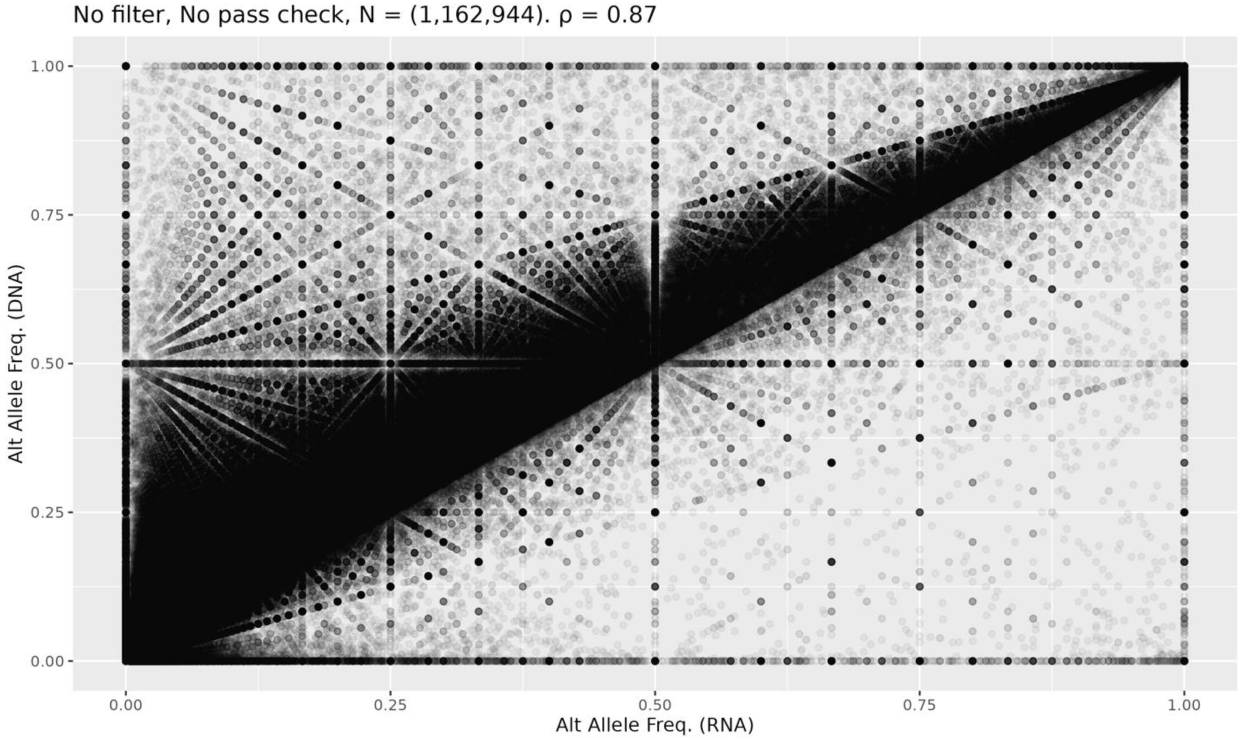

Figure 14. *Alternative allele frequencies for the RNA-seq derived genotypes against the WGS derived genotypes. Showing a correlation coefficient of 0.87, which is accompanied by a great amount of noise.*

Figure 15.

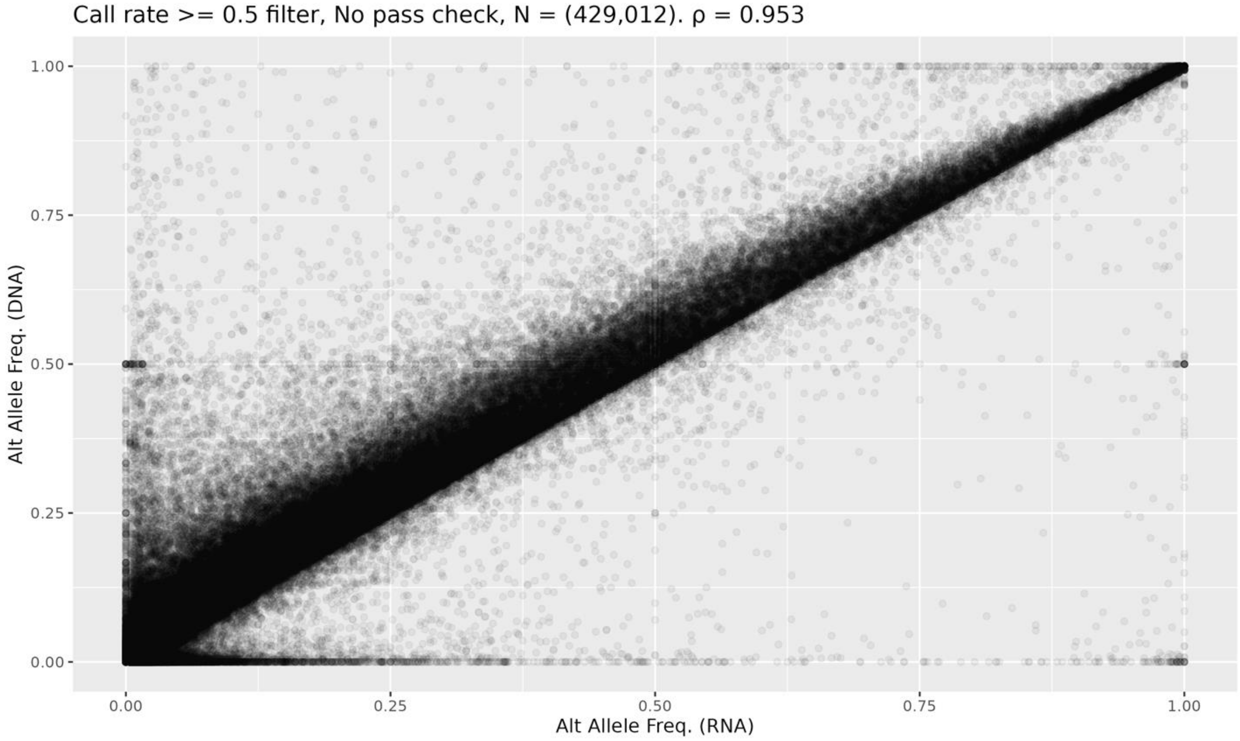

Figure 15. *Alternative allele frequencies for the RNA-seq derived genotypes against the WGS derived genotypes with a minimum call rate of 0.5. Showing a correlation coefficient of 0.953, which is accompanied by slight noise.*

Figure 16.

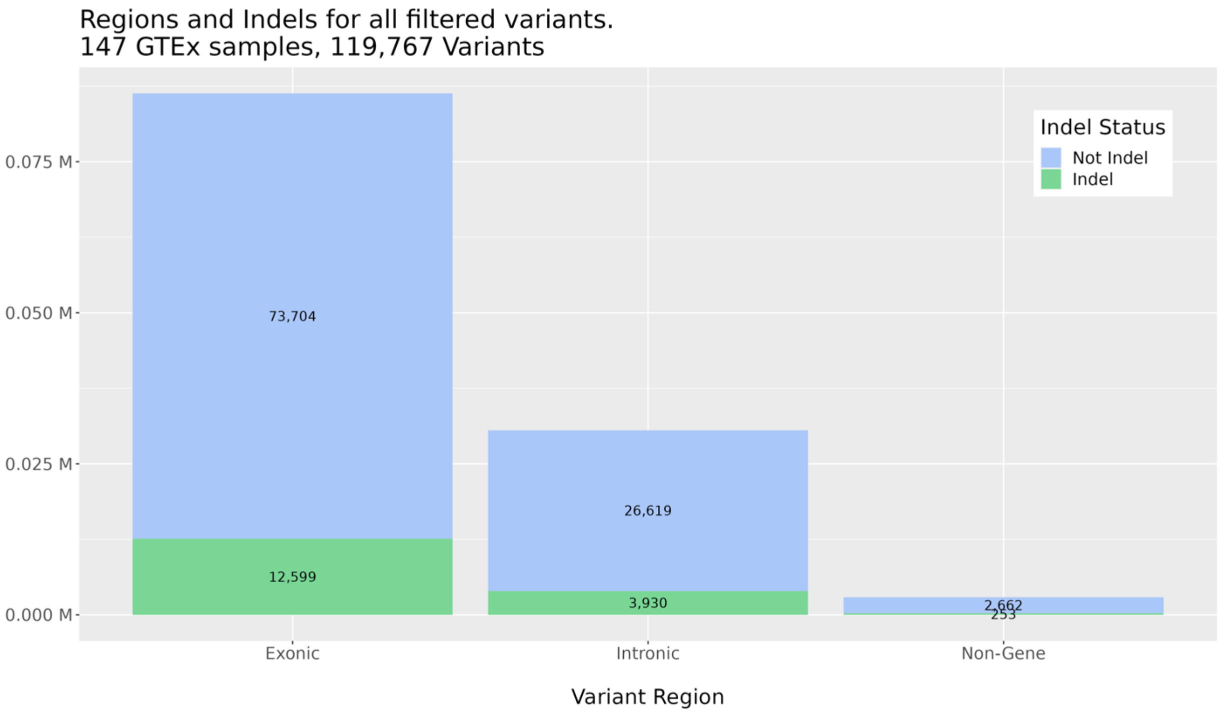

Figure 16. *Summary of variants which pass the custom VCF filter script with the following thresholds: A minimum call rate of 0.5, allelic depth of 5 and genotype quality of 10.*

Figure 17.

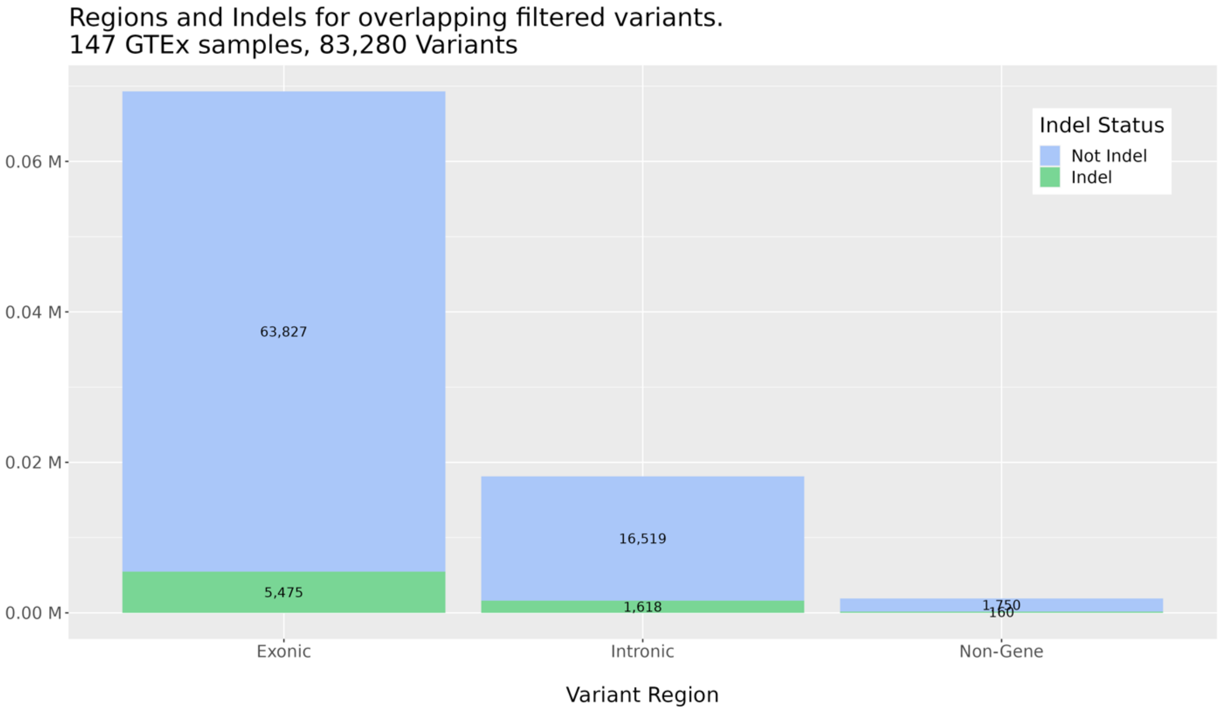

Figure 17. *Summary of overlapping variants which pass the custom VCF filter script with the following thresholds: A minimum call rate of 0.5, allelic depth of 5 and genotype quality of 10.*

Figure 18.

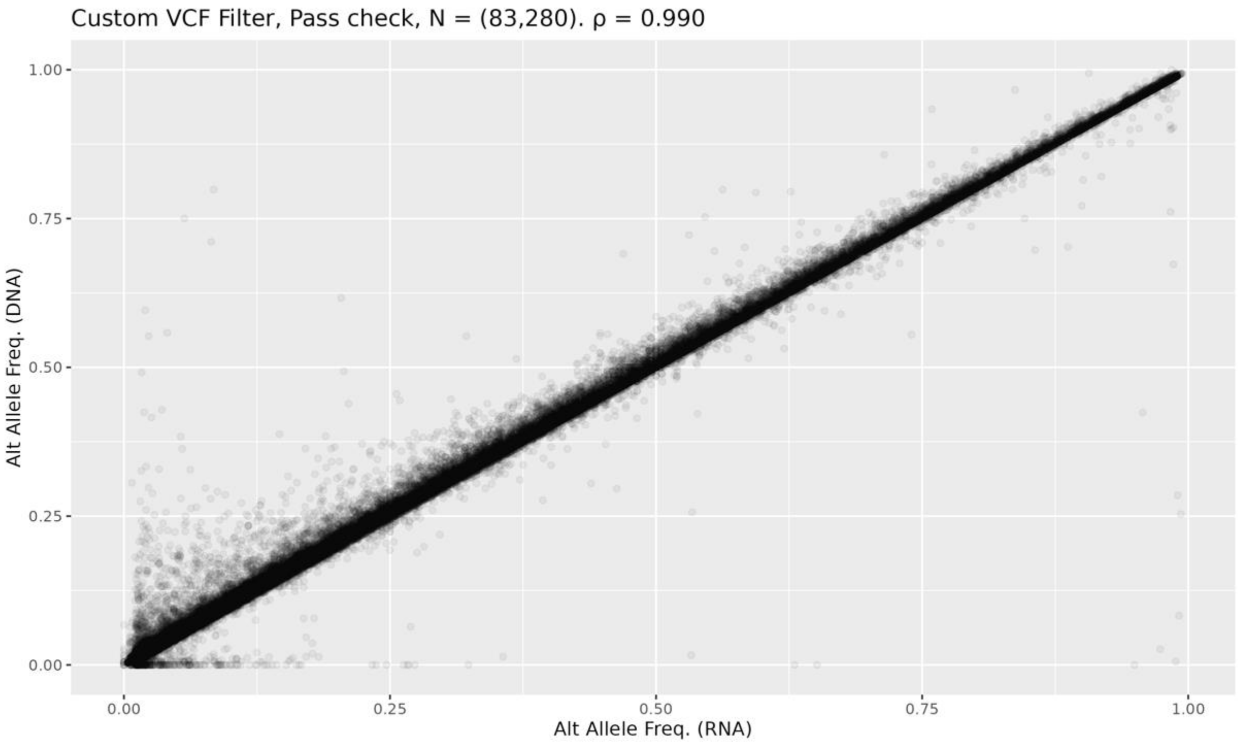

Figure 18. *Alternative allele frequencies plotted against each other for the overlapping and filtered variants. Correlation coefficient is 0.99.*

Figure 19.

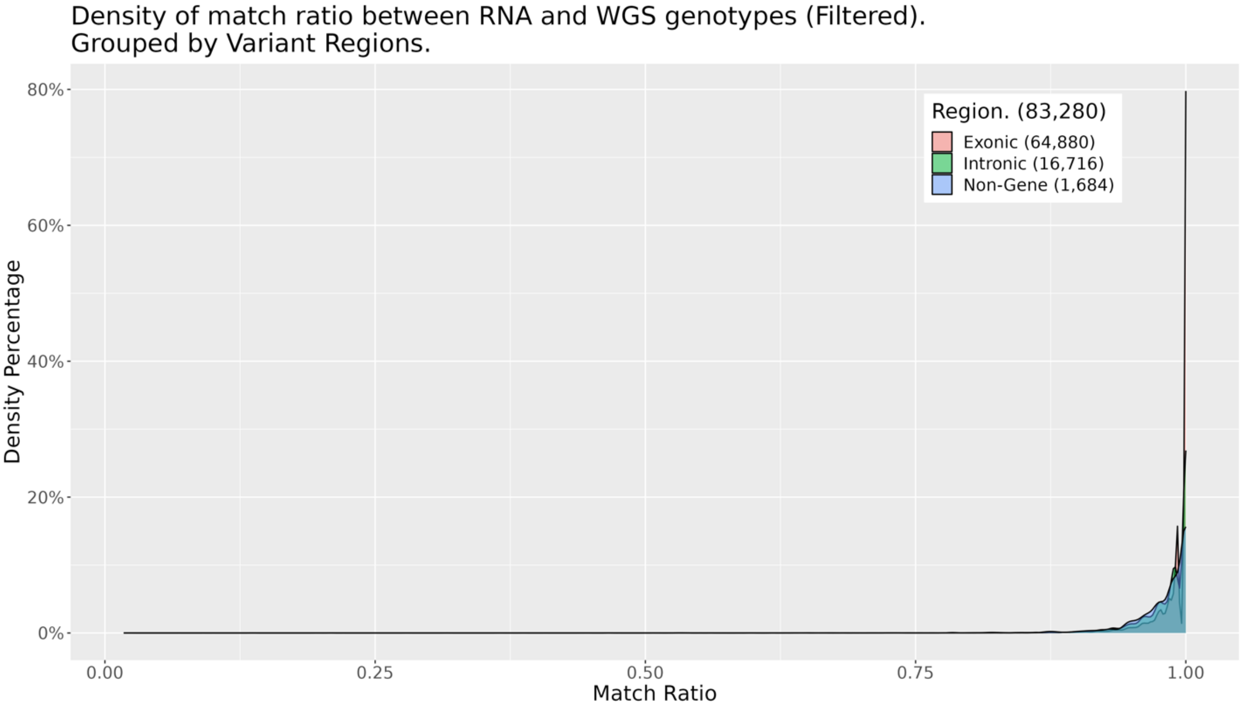

Figure 19. This density plot shows the distribution of the ratio which represents exactly how many of the genotypes for a variant are the same for their respective samples between both files.

Figure 20.

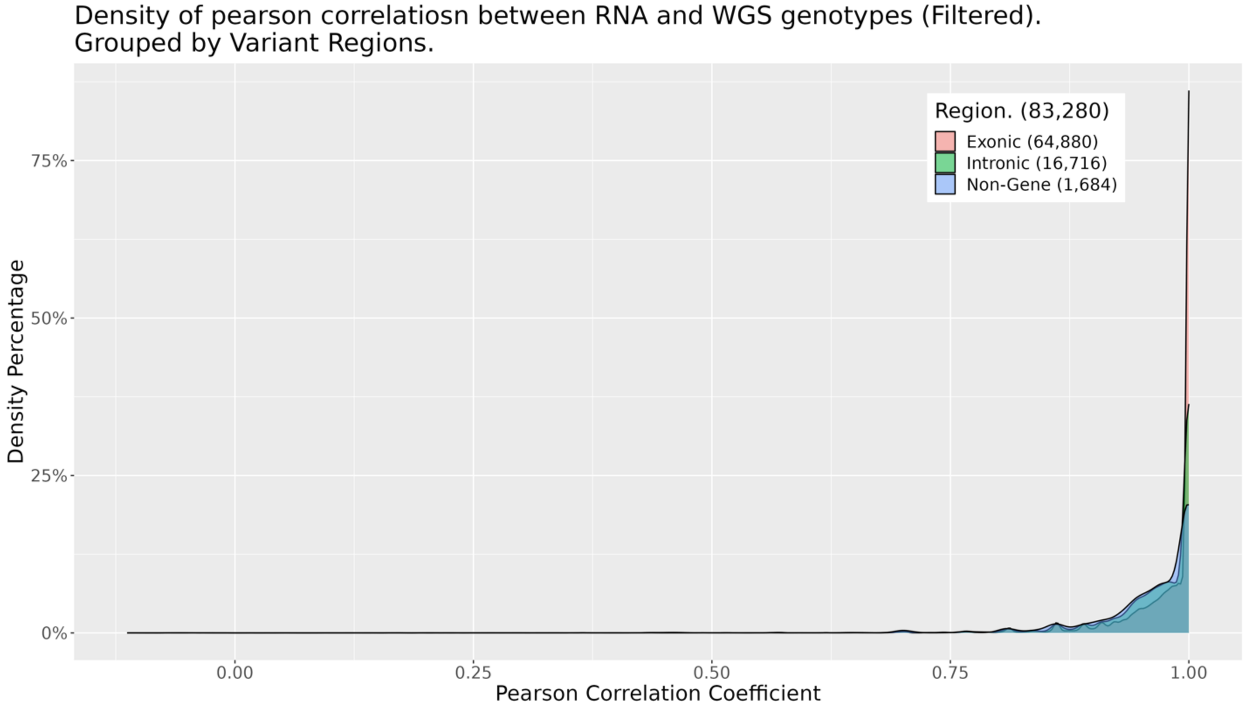

Figure 20. This density plot shows the distribution of the Pearson correlation coefficients which represents how many of the genotypes for a variant are the same for their respective samples between both files.