https://rgd.mcw.edu

Variant analysis in an inbred rat population – A lesson from the Hybrid Rat Diversity Panel

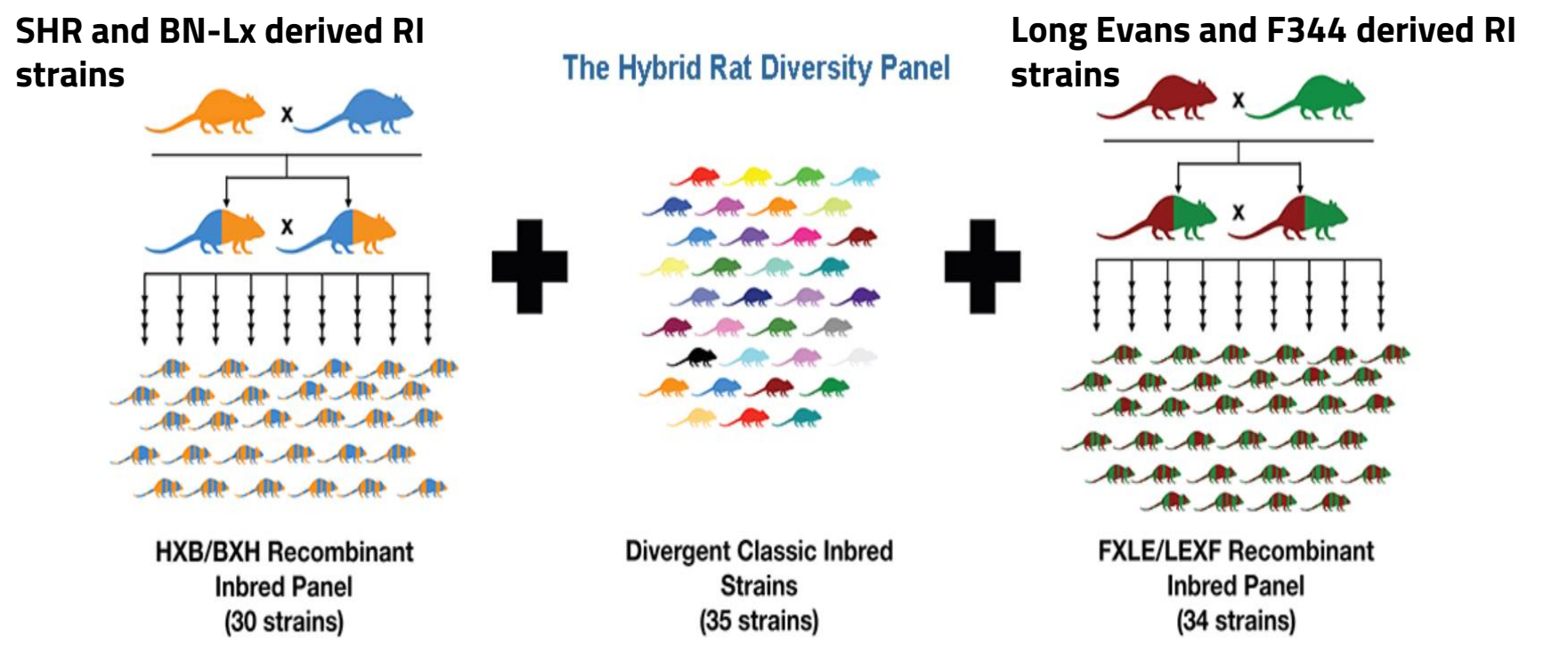
Monika Tutaj[1], Akiko Takizawa[1], Lynn Malloy[1], Rebecca Schilling[1], Kent C Brodie[2], Jeffrey L De Pons[1], Wendy M Demos[1], G Thomas Hayman[1], Mary L Kaldunski[1], Stanley JF Laulederkind[1], Jennifer R Smith[1], Marek A Tutaj[1], Mahima Vedi[1], Shur-Jen Wang[1], Anne E Kwitek[1], Melinda R Dwinell[1]

[1]Department of Physiology, Medical College of Wisconsin, Milwaukee, WI,
[2]Clinical and Translational Science Institute, Medical College of Wisconsin, Milwaukee, WI

Abstract

Currently, many human genetic studies identify disease-associated loci and variants using whole exome sequencing (WES)/whole genome sequencing (WGS) methods. However, the associated variants are often either Variants of Unknown Significance (VUS) and/or not even within gene coding regions, making identifying causal variants in known genes challenging. Therefore, researchers utilize multiple model organisms, different genetic backgrounds, and environmental stressors to link the variants to orthologous genes, pathways, molecular networks and eventually disease phenotypes. The Rat Genome Database provides information on genomic variants across laboratory rat strains for all rat genome assemblies (RGSC_v3.4, Rnor_5.0, Rnor_6.0, mRatBN7.2), and integrates it with strain phenotype data to aid in interpretation of the variants. 96 strains from the Hybrid Rat Diversity Panel (HRDP) were sequenced and analyzed at MCW as part of the Hybrid Rat Diversity Program, a resource to rederive classic and recombinant inbred rat strains, sequence their genomes to identify variations, and make the data and the strains accessible to the research community. In 2020, we used mRatBN7 and Rnor_6.0 to assess the detection rate of homozygous, heterozygous, genic, and intergenic single nucleotide variants (SNVs) and short indels from WGS of 47 HRDP strains. We also performed variant discovery using non-reference rat genomes assembled and released in 2022 (SHRSP/BbbUtx, WKY/Bbb, SHR/Utx). All datasets were analyzed with the Genome Analysis Toolkit from the Broad Institute, designed and optimized for human data. There are no existing recommendations for variant discovery in non-human, inbred populations like rat with low effective sizes. In addition, the systematic comparisons of variant callers were not conducted in rat populations, so we lack the truth and training datasets to evaluate variant call accuracy and therefore to efficiently exclude false positive calls. We address these variant analysis challenges and propose strategies for selecting and prioritizing candidate variants for disease model studies. We differentiate variants present in likely misassembled genomic regions in the BN reference, in repetitive regions, and in regions with accumulations of heterozygous variants. Thus, scientists can identify potential disease-causing mutations and distinguish them from low confidence ones in the context of RGD's integrated multi-species data, have comparative insight in data alignment and distribution (JBrowse2, VCMap and VariantVisualizer) and prioritize variants for validation in the available HRDP strains. Finally, they can confirm if the elected SNVs and indels occur in QTLs and genes associated with disease phenotypes for which the strains were selected, and have impact on transcripts (SnpEff, VEP) and proteins (PolyPhen2) shared between strains representing the same disease model.

What is the Hybrid Rat Diversity Panel ?


SHR and BN-Lx derived RI strains
The Hybrid Rat Diversity Panel
Long Evans and F344 derived RI strains
x
HXB/BXH Recombinant Inbred Panel (30 strains)
Divergent Classic Inbred Strains (35 strains)
FXLE/LEXF Recombinant Inbred Panel (34 strains)

- blood pressure regulation
- glucose and fatty acid metabolism
- taste aversion
- herpes simplex encephalitis
- metabolic syndrome
- alcohol consumption

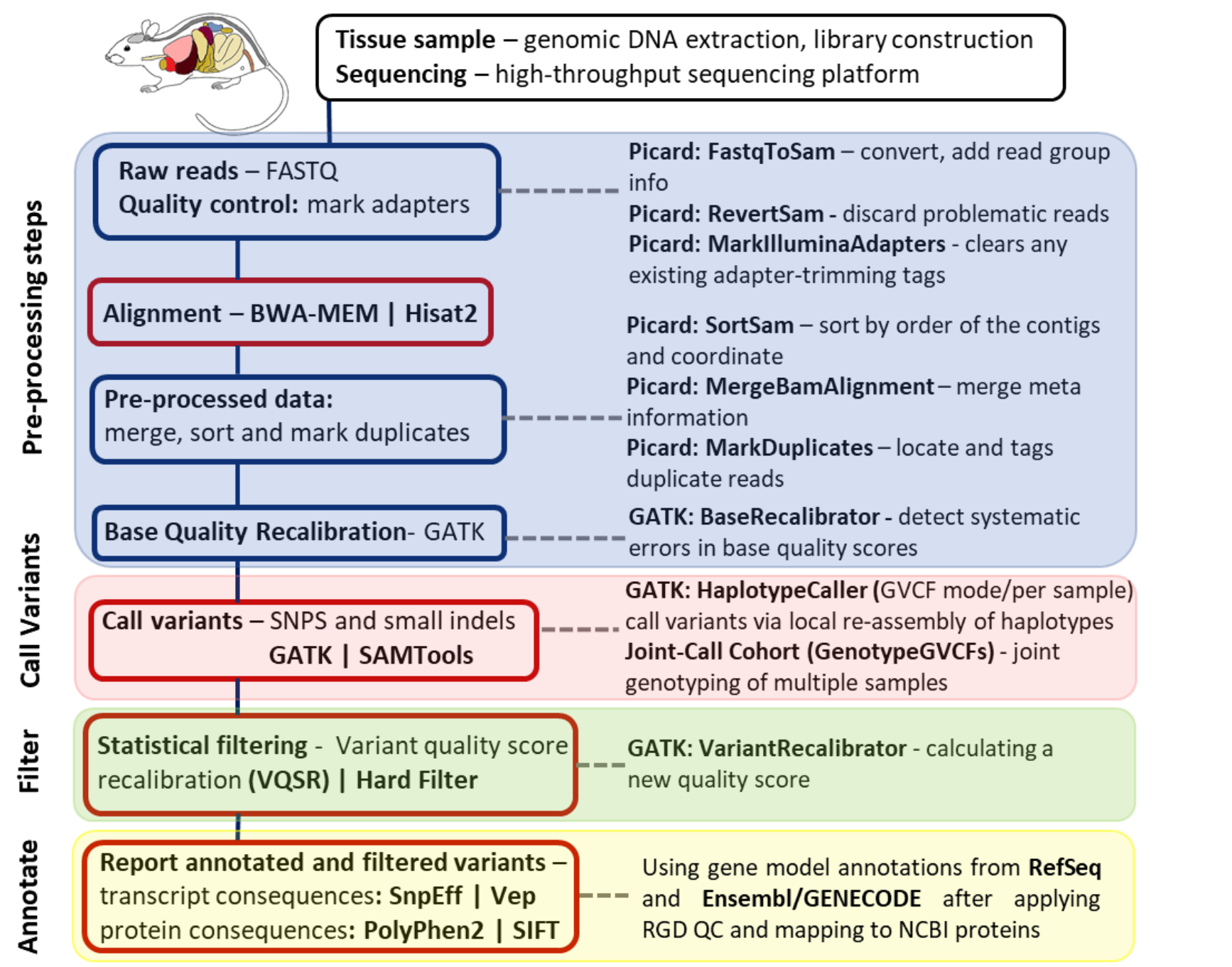
- cardiovascular diseases
- renal diseases
- obesity
- diabetes
- pulmonary diseases
- behavioral variation
- anxiety
- autoimmunity

- seizures
- epilepsy
- male reproductive disorders
- lymphoma and leukemia

Hybrid Rat Diversity Panel was selected to:

1. Provide stable genetic and phenotypic strains to allow researchers to conduct reproducible experiments
2. Maximize the genetic diversity among strains and to maximize power to detect specific genetic loci associated with a complex trait (QTL mapping resolution)
3. Extend the whole genome sequencing to all HRDP inbred rat strains with susceptibility to different complex diseases
4. Facilitate the translation of disease-related genetics and genomics research to pre-clinical and clinical studies

Analysis Workflow


Tissue sample – genomic DNA extraction, library construction
Sequencing – high-throughput sequencing platform

Pre-processing steps

Raw reads – FASTQ
Quality control: mark adapters

Picard: FastqToSam – convert, add read group info
Picard: RevertSam - discard problematic reads
Picard: MarkIlluminaAdapters - clears any existing adapter-trimming tags

Alignment – BWA-MEM | Hisat2

Picard: SortSam – sort by order of the contigs and coordinate
Picard: MergeBamAlignment – merge meta information
Picard: MarkDuplicates – locate and tags duplicate reads

Pre-processed data:
merge, sort and mark duplicates

Base Quality Recalibration- GATK

GATK: BaseRecalibrator - detect systematic errors in base quality scores

Call Variants

Call variants – SNPS and small indels
GATK | SAMTools

GATK: HaplotypeCaller (GVCF mode/per sample) call variants via local re-assembly of haplotypes
Joint-Call Cohort (GenotypeGVCFs) - joint genotyping of multiple samples

Filter

Statistical filtering - Variant quality score recalibration (VQSR) | Hard Filter

GATK: VariantRecalibrator - calculating a new quality score

Annotate

Report annotated and filtered variants – transcript consequences: SnpEff | Vep
protein consequences: PolyPhen2 | SIFT

Using gene model annotations from RefSeq and Ensembl/GENECODE after applying RGD QC and mapping to NCBI proteins

1. Pre-processing steps involve marking adapter sequences, alignment to the rat genome reference **mRatBN7.2** and marking duplicates.
2. Base Quality Scores Recalibration corrects biases introduced by sequencing platforms and assigns scores empirically determined from the read data using validated variants.
3. Variant calling is accomplished by running the GATK HaplotypeCaller that simultaneously detects SNVs and Indels via local de-novo assembly of haplotypes (method to increase accuracy of the variant call comparing with position-based algorithm).
4. In the filtering process we remove less reliable variant calls: variants with low coverage, low quality, strand biased, located in SNV clusters, and supported by low-confidence read alignment.

Variants Discovery Challenges

Short read mapping, variant detection and variant interpretation limitations

1. The increase in sequencing capacity has identified large numbers of potentially pathogenic variants however supporting functional evidence is often sparse or inconclusive.
2. Genome sequencing is increasingly being adopted in the clinic to provide genetic diagnoses for patients with rare diseases. The general approach is to prioritize variants disrupting gene function and ignore small variants that can generate splicing errors or disrupt poorly annotated regulatory elements.
3. The copy number, repeat and structural variation, non-coding and splice site variants that may have a larger impact on the genome than SNVs and small insertion/deletions (indels) are harder to detect with short read sequencing technologies.
4. Functional inference prediction tools have often high false positive rates, as missense mutations found by genome sequencing, identified as deleterious have little impact on the clinical phenotype of individual cases.
5. Incomplete annotations and genome references together with the error-prone experimental and computational steps impede discovery of true variants data from DNA or RNA sequences.
6. Variant calling algorithms are imperfect. It is particularly difficult to identify variations in the low-complexity and homologous genomic regions, which contain noncoding elements.
7. Variant calling can be thwarted by sample heterogeneity and contaminations.

Pursue High Confidence Variants

Variants identified in different rat populations

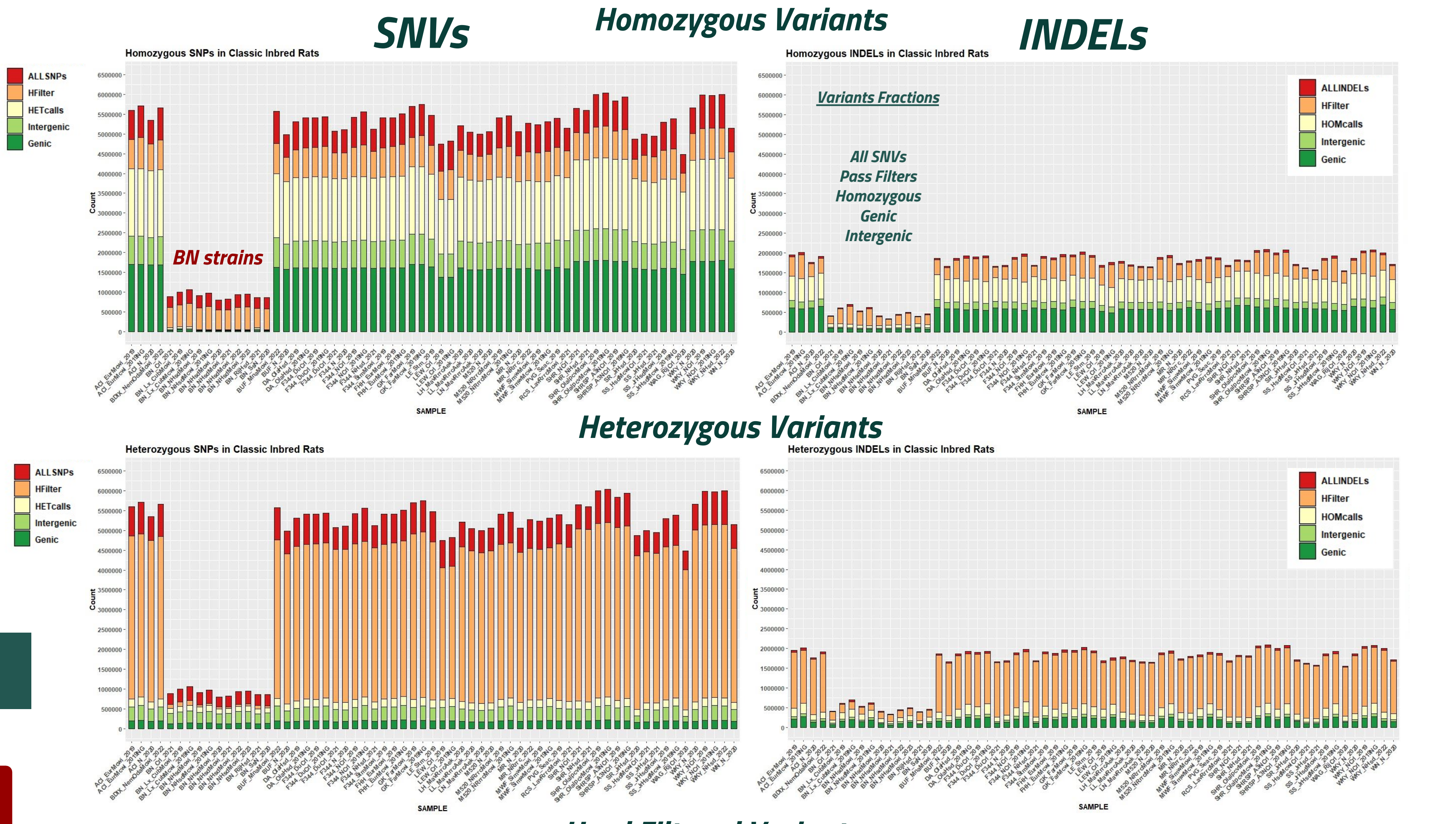
Genome assembly version	RGSCv3.4	Rnor5.0	Rnor6.0	mRatBN7	mRatBN7
Rat strains	40	64	67	88	107
Rat samples	52	74	90	108	131
SNPs	13,059,423	12,521,159	18,140,193	13,205,288	13,668,613
Insertion	1,466,176	2,431,570	1,902,501	3,818,073	4,030,074
Deletion	782,581	2,868,674	2,408,148	4,555,354	4,794,013
dbSnp version	136	138	146	EVAv2	EVAv3

Hard Filter - 6 parameters PASS

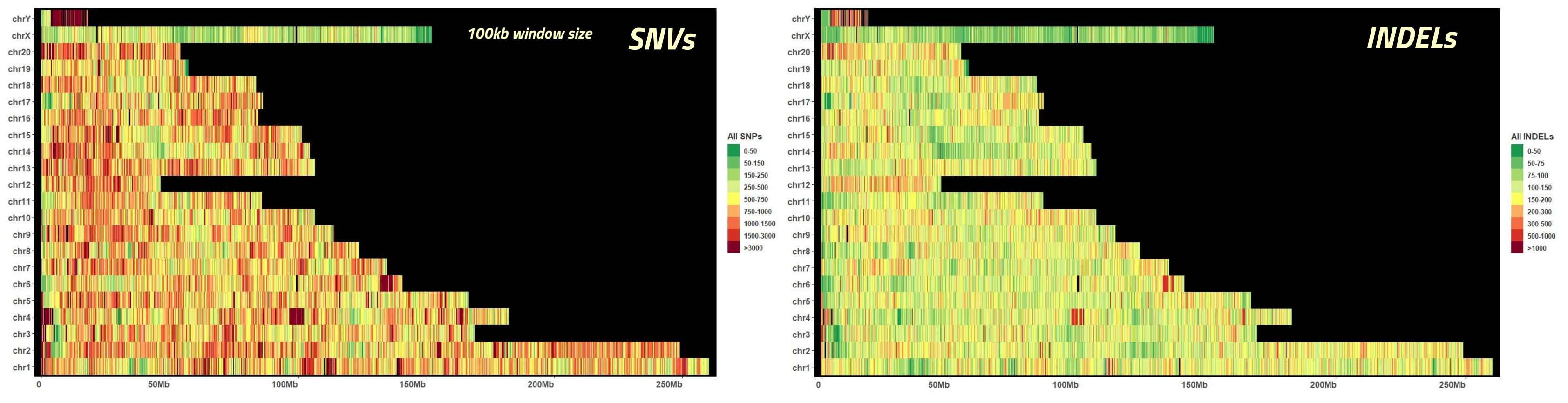
1. *QualByDepth (QD)*
2. *FisherStrand (FS)*
3. *StrandOddsRatio (SOR)*
4. *RMSMappingQuality (MQ)*
5. *MappingQualityRankSumTest (MQRankSum)*
6. *ReadPosRankSumTest (ReadPosRankSum)*

High confidence variants

Number of variants decrease after applying hard filtering parameters and only variants supported by more than 8 reads were selected as high confidence set.

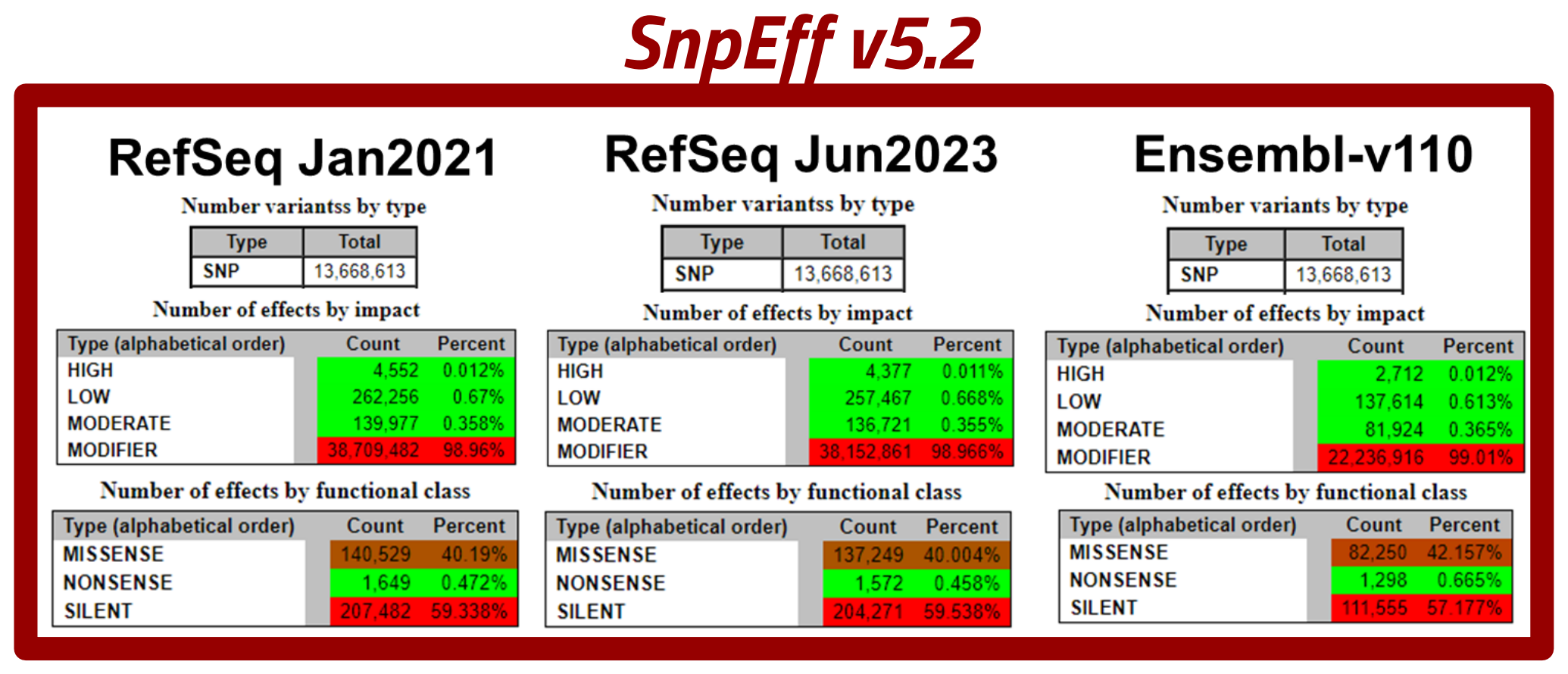


Hard Filtered Variants



Variants Functional Consequences

SnpEff v5.2


RefSeq Jan2021

Number variantss by type

Type	Total
SNP	13,668,613

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	4,552	0.012%
LOW	262,256	0.67%
MODERATE	139,977	0.358%
MODIFIER	38,709,482	98.96%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	140,529	40.19%
NONSENSE	1,649	0.472%
SILENT	207,482	59.338%

RefSeq Jun2023

Number variantss by type

Type	Total
SNP	13,668,613

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	4,377	0.011%
LOW	257,467	0.668%
MODERATE	136,721	0.355%
MODIFIER	38,152,861	98.966%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	137,249	40.004%
NONSENSE	1,572	0.458%
SILENT	204,271	59.538%

Ensembl-v110

Number variants by type

Type	Total
SNP	13,668,613

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	2,712	0.012%
LOW	137,614	0.613%
MODERATE	81,924	0.365%
MODIFIER	22,236,916	99.01%

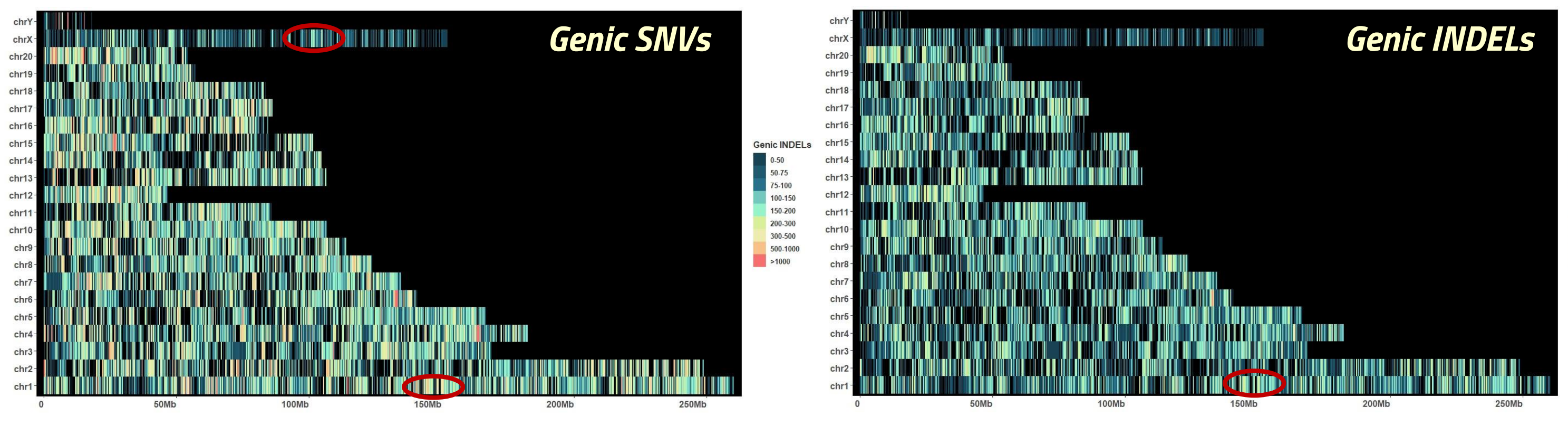
Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	82,250	42.157%
NONSENSE	1,298	0.665%
SILENT	111,555	57.177%

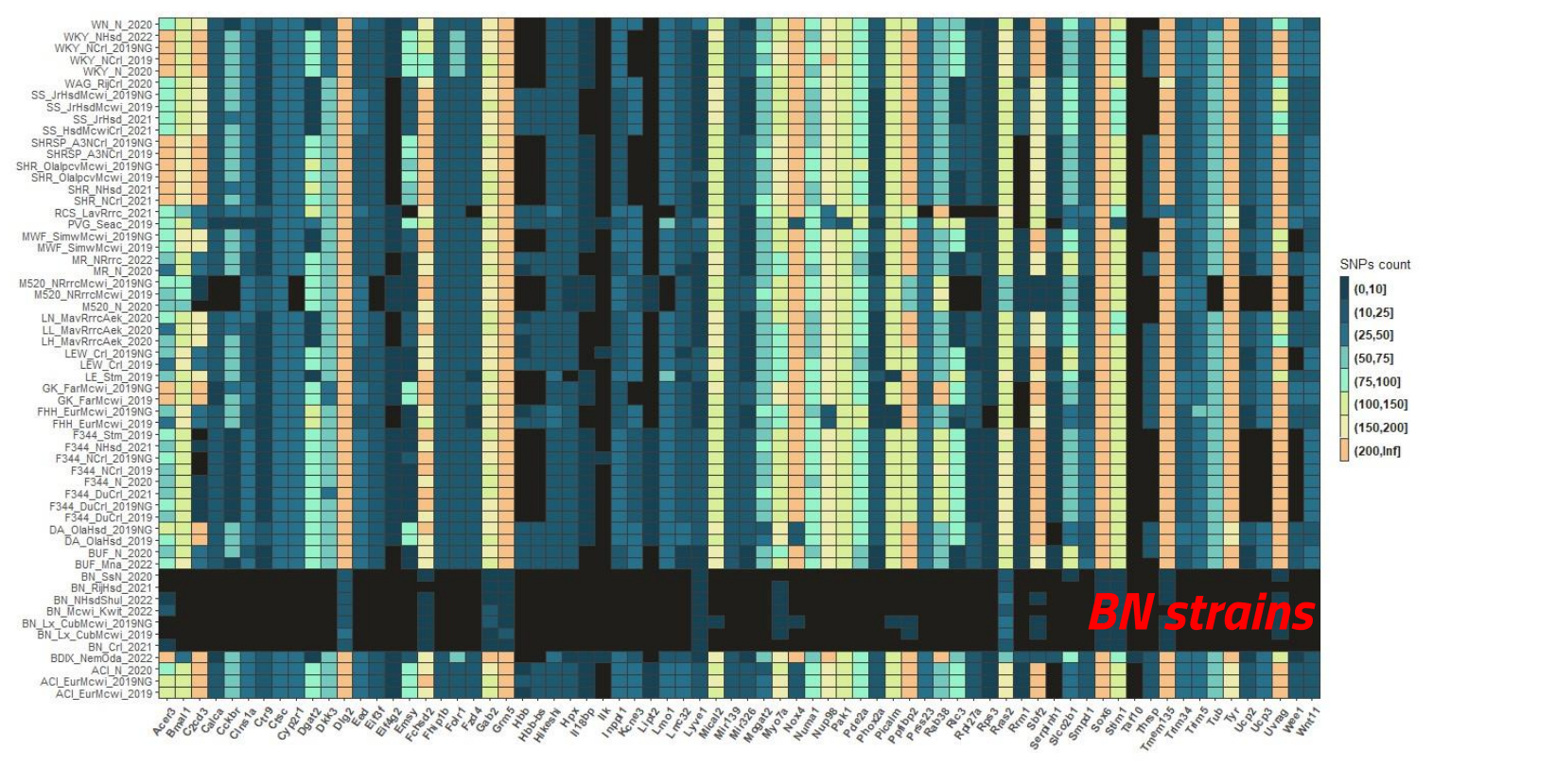
Additional variants filtering by high impact on transcript level shows region with higher number of variants for further verification. Predicted consequences depend on the choice of the gene model annotations.

Variants of HIGH impact Sequence Ontology terms

1. *Transcript ablation*
2. *Splice acceptor variant*
3. *Splice donor variant*
4. *Stop gained*
5. *Frameshift variant*
6. *Stop lost*
7. *Start lost*
8. *Transcript amplification*
9. *Feature elongation*
10. *Feature truncation*



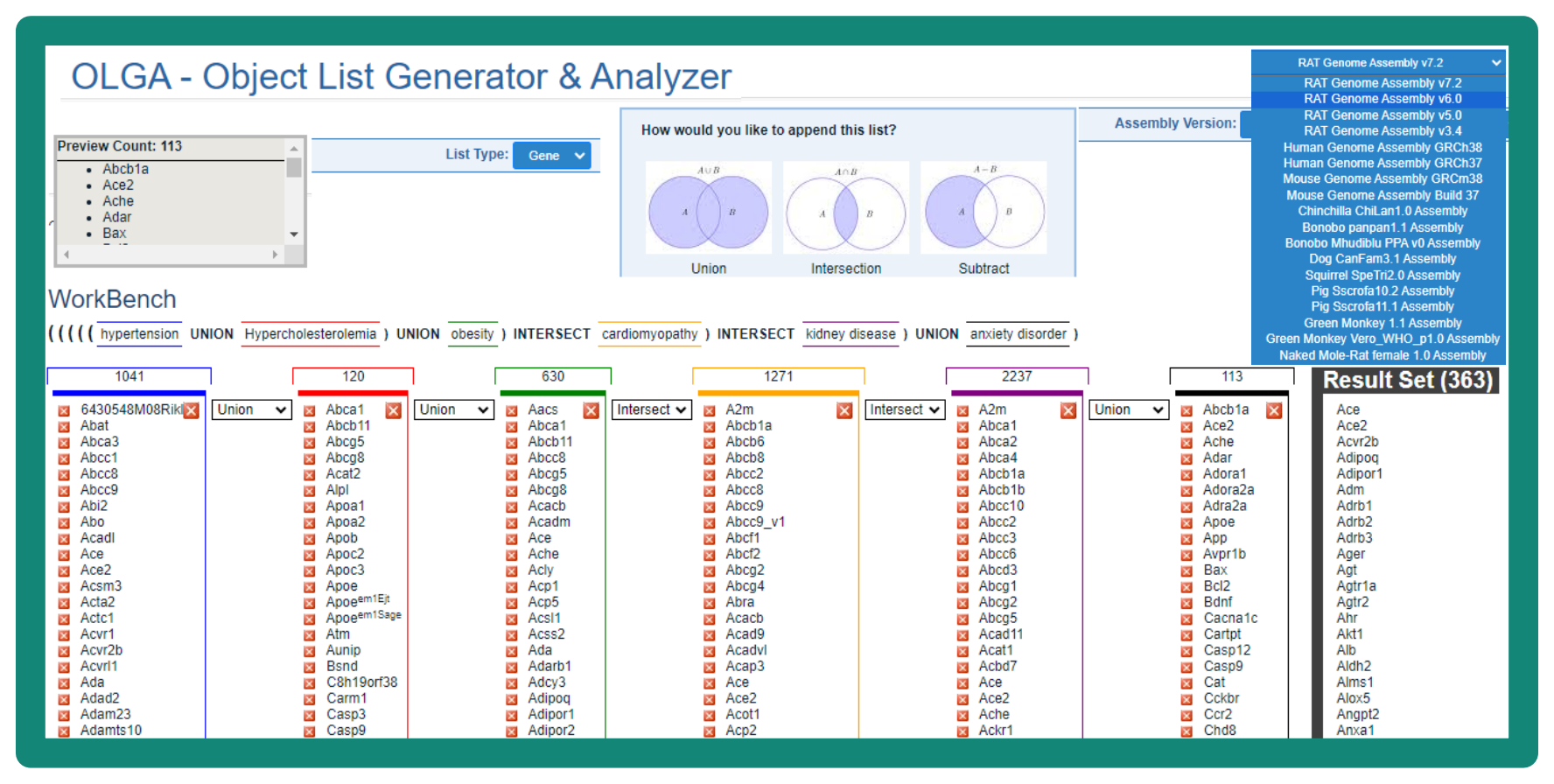
Homozygous Genic Variants



Selected Genes from chr1-30Mb region

Disease association

DO terms



(((((hypertension UNION Hypercholesterolemia) UNION obesity) INTERSECT cardiomyopathy) INTERSECT kidney disease) UNION anxiety disorder)

Emerging clusters of variants present in a few strains and particular genes can potentially relate to individual phenotypes

Selected 363 genes

Assembled genomes are not complete. Analyses of genome variants, but also synteny, duplications, relatedness depend on having the correct sequence.

Classic Inbred Rats are models in

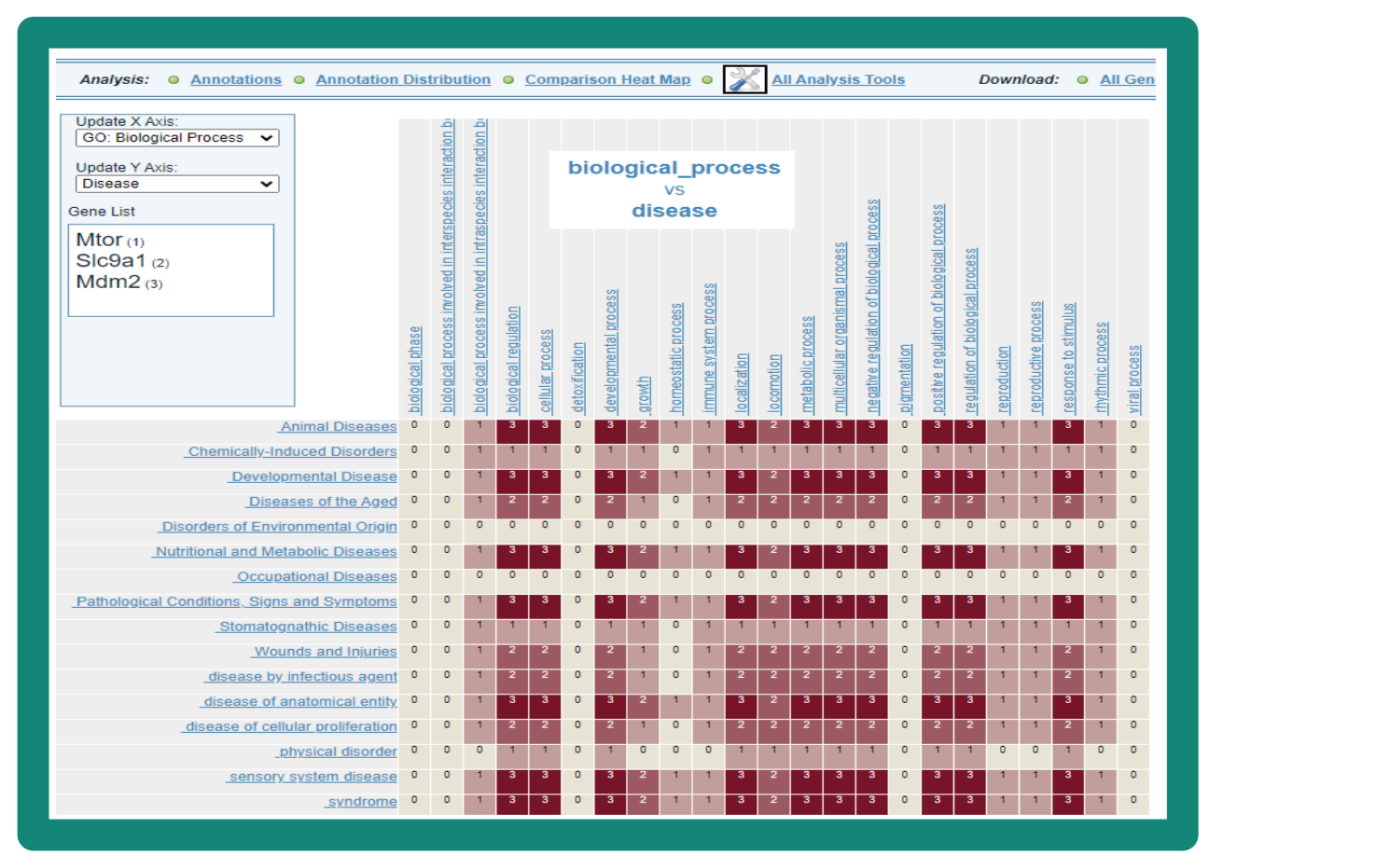
- *cardiovascular diseases*
- *renal diseases*
- *obesity*
- *diabetes*
- *pulmonary diseases*
- *behavioral variation*
- *anxiety*
- *autoimmunity*

Visit our website

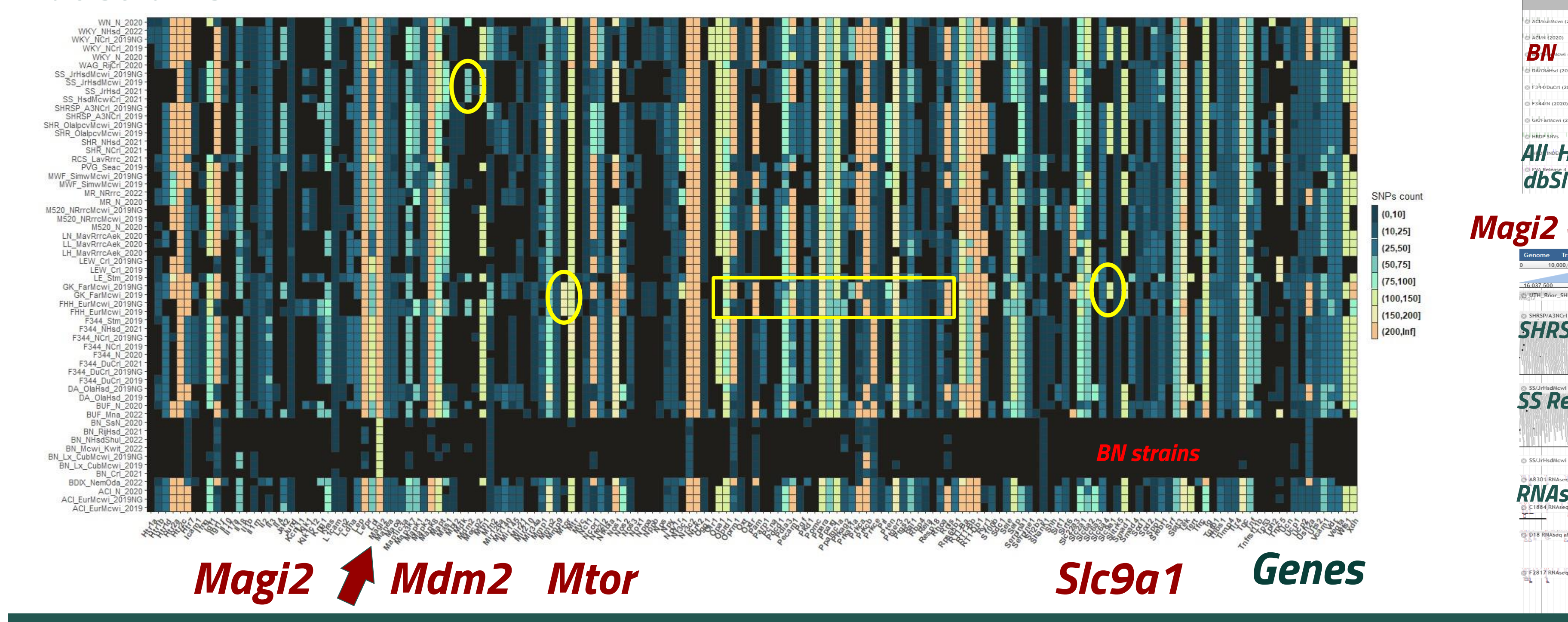
HRDP portal

Mtor
Slc9a1
Mdm2

Gene Annotator

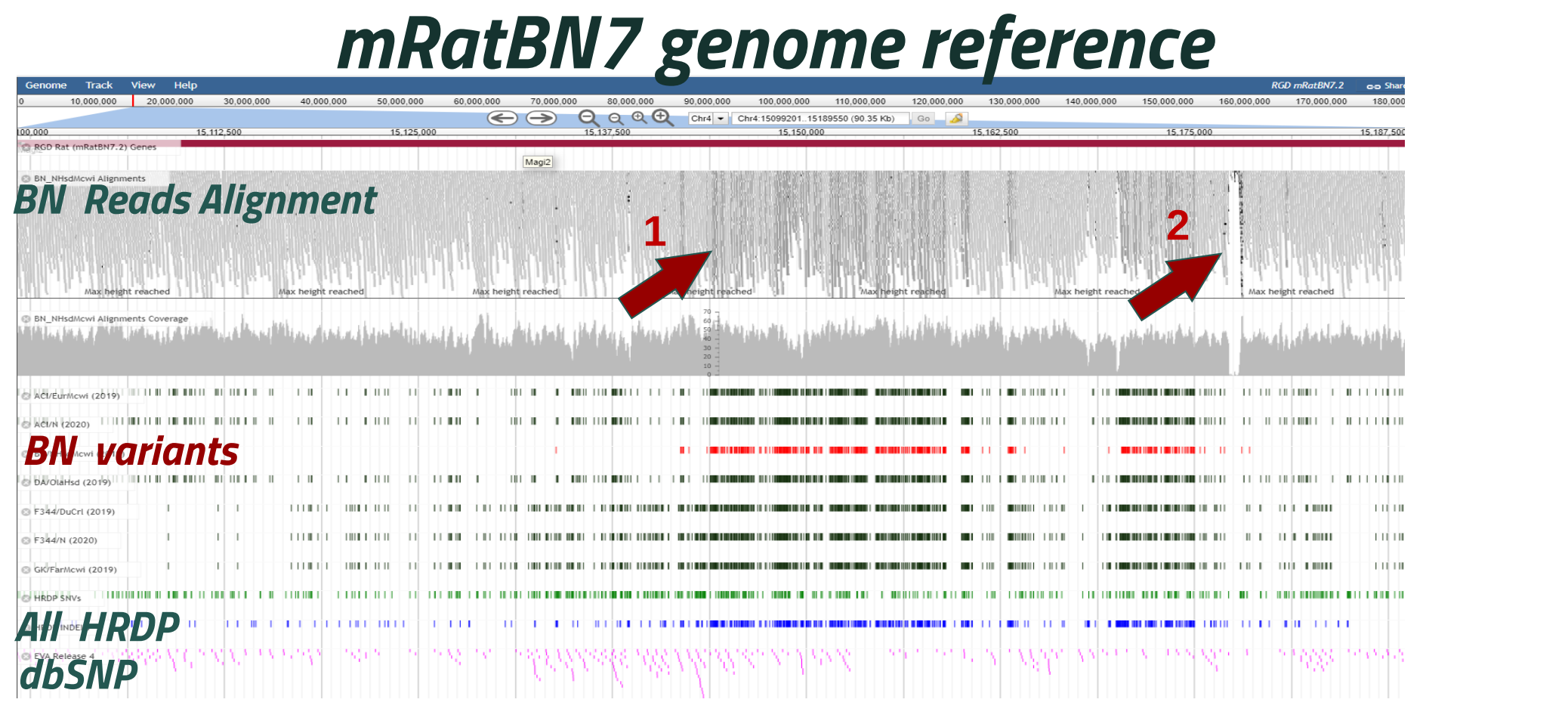

Analysis: Annotations Annotation Distribution Comparison Heat Map All Analysis Tools
Download: All Gen
Update X Axis: GO: Biological Process
Update Y Axis: Disease
Gene List
Mtor (1)
Slc9a1 (2)
Mdm2 (3)
biological_process vs disease

Rat strains



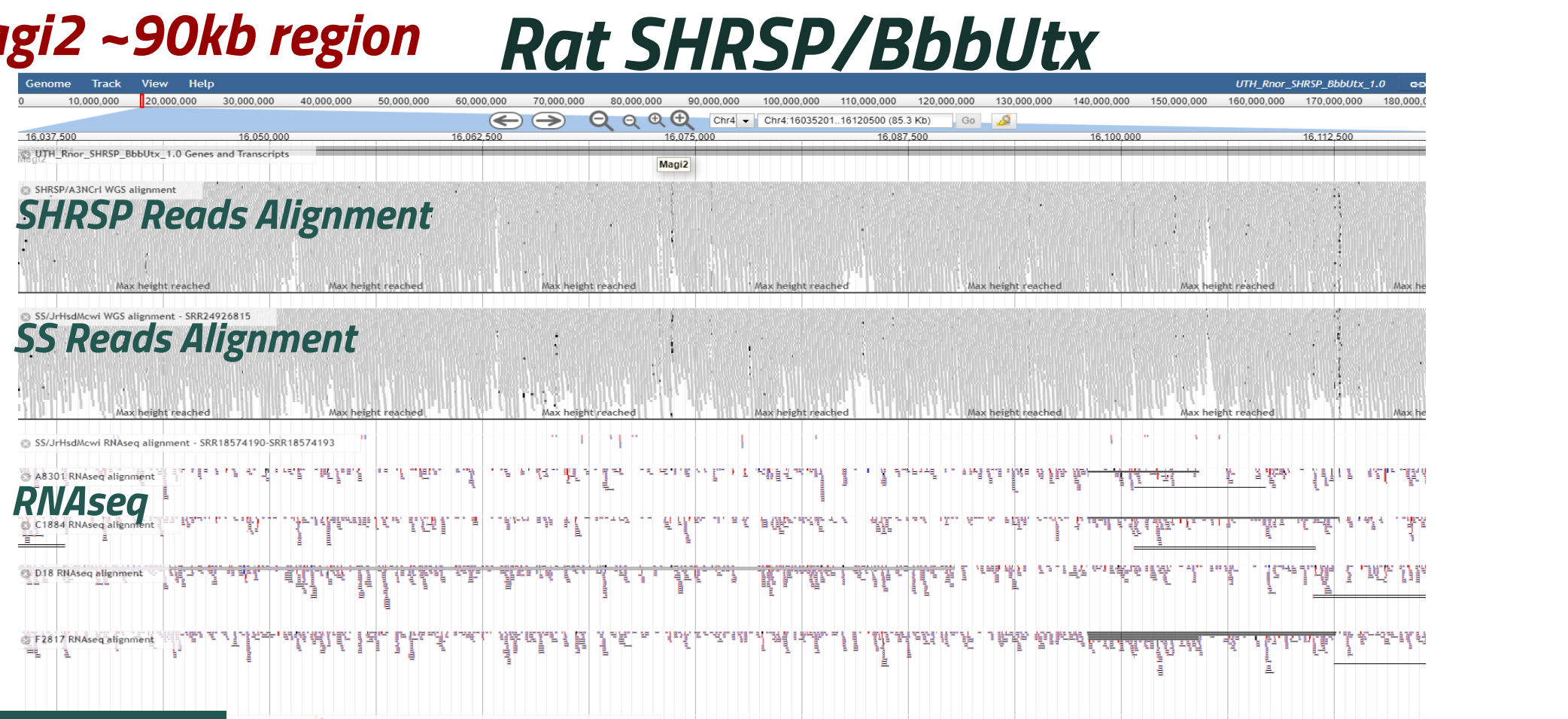
mRatBN7 genome reference

1. *Low mapping quality region, but variants are still called; single bp problems*
2. *Major assembly issue - gap*



Magi2 ~90kb region

Rat SHRSP/BbbUtx



Summary

- ❑ *Variants discovery consists of multiple filtering, annotation and comparison steps involving both careful inspection of variant data and sequence alignments.*
- ❑ *Integrative analysis of variant data with functional transcriptomics, proteomics, and metabolomics data collection could help to validate the impact of identified variants, to guide diagnosis or the disease model design.*
- ❑ *Whole genome sequencing of rat diversity panel provides high confidence variants that represent the sequence cohort heterogeneity but also variants that require additional quality testing. Analyzing variants in non-reference rat genomes can help discover true positive variants..*
- ❑ *Variants collection from the HRDP rat strains helps to generate high resolution association mapping and investigate complex traits.*
- ❑ *HRDP Portal available on RGD site provides data mining and visualization functions to retrieve, correlate and interpret genomic and phenotypic data across species.*

Funding provided by:

The Hybrid Rat Diversity Program (grant 5R24OD024617 from the NIH Office of the Director),
National Heart, Lung, and Blood Institute (NHLBI; R01HL064541)