Pursuing the discovery of impactful genetic variation in selected rat population Hybrid Rat Diversity Panel

Monika Tutaj[1], Kent Brodie[3], Jeffrey L De Pons[1], Wendy M Demos[1], Adam Gibson[1], G Thomas Hayman[1], Morgan L Hill[1], Mary L Kaldunski[1], Logan Lamers[1], Lynn Lazcares[2], Stanley JF Laulederkind[1], Harika S Nalabolu[1], Rebecca Shilling[2], Jennifer R Smith[1], Akiko Takizawa[2], Ketaki Thorat[1], Jyothi Thota[1], Marek A Tutaj[1], Mahima Vedi[1], Shur-Jen Wang[1],Stacy Zacher[4], Anne E Kwitek[2], Melinda R Dwinell[1,2]

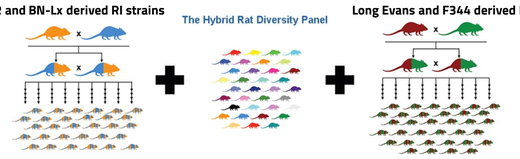
[1]Rat Genome Database, Department of Biomedical Engineering, [2]Department of Physiology, [3]Clinical & Translational Science Institute, [4]Finanse and Administration, Medical College of Wisconsin, Milwaukee, WI 53226

Abstract

The Hybrid Rat Diversity Panel consists of 33 classic inbred strains and two recombinant inbred panels: 33 FXLE/LEXF strains from Japan and 30 HXB/BXH strains from the Czech Republic. These rats were carefully selected for genetic diversity in order to study genetic loci associated with complex traits. Common human diseases are complex traits and are shaped by the additive effect of many genetic variants. Here we provide insight into variant diversity in a population of 72 rats, the distribution of SNVs and indels in QTLs and genes associated with disease phenotypes for which the strains were selected, as well as functional consequences of variants shared between strains representing the same disease model, like hypertension. We aligned whole genome sequences (Illumina short-reads) to the high-quality rat reference mRatBN7, performed variant calls (GATK4) and analyzed their impact on specific genes (SnpEff, RGD OLGA ontology tool). As disease-associated loci contain many genes it is difficult to identify the compromised ones and even more difficult to distinguish causal variants and their phenotypic effects. Researchers increasingly utilize multiple organism models exploiting their advantages in pursuance of comprehensive knowledge. Toward that end, the Rat Genome Database integrates multi-species data, develops tools to improve multilevel navigation and discovery of valuable information.

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**



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

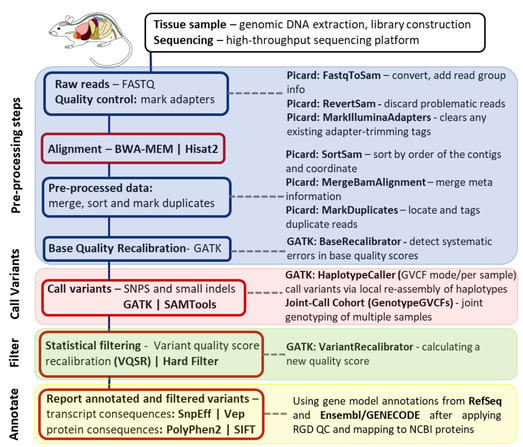
- cardiovascular diseases
- renal diseases
- obesity
- diabetes
- pulmonary diseases
- behavioral variation
- anxiety
- autoimmunity

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

Hybrid Rat Diversity Panel was selected to:

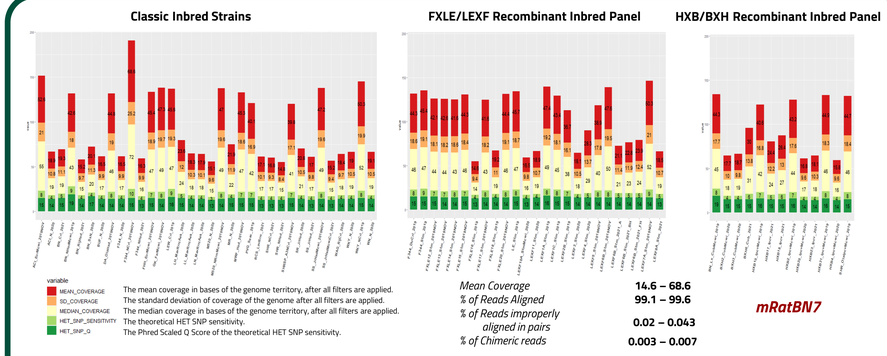
1. Provide stable genetic and phenotypic strains to allow researchers to conduct reproducible experiments
2. Maximize the genetic diversity among strains and to maximize power to detect specific genetic loci associated with a complex trait (QTL mapping resolution)
3. Extend the whole genome sequencing 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

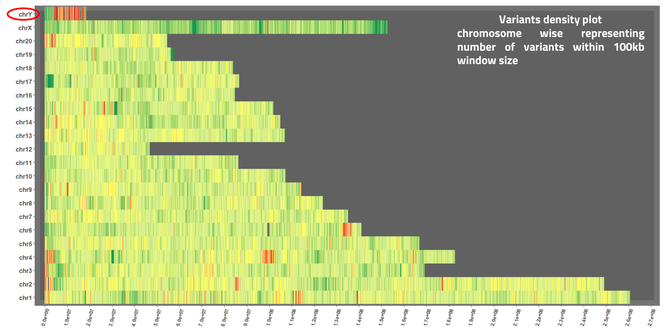


1. Pre-processing steps involve marking adapter sequences, alignment to the new rat genome reference mRatBN7 and marking duplicates.
2. Base Quality Scores Recalibration corrects biases introduced by sequencing platforms and assigns scores empirically determined from the read data using validated variants.
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.

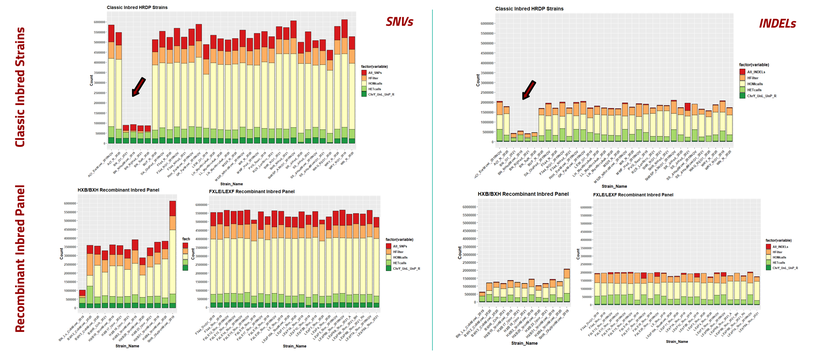
Variants Analysis Results



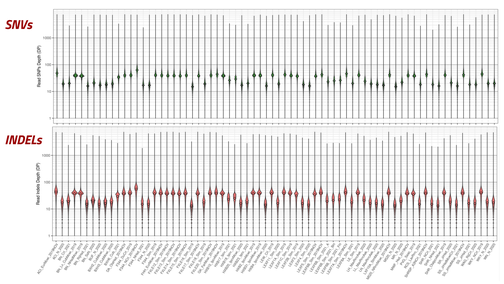
Mapping analysis shows the coverage and performance of the whole genome sequencing. Mean base coverage ranges from 15x to 68x and the estimated sensitivity to detect heterozygous sites (as a function of coverage and base quality distribution ranges) drops quickly when mean coverage is below 20X.



Variants density plot shows genome location with high number of variants accumulation that require further analysis.



Number of SNVs and indels subsets per sample - total number of SNVs per sample was up to 6mln and around 2mln of indels, except Brown Norway - classic and recombinant - strains that have ~5-6x lower number of variants.

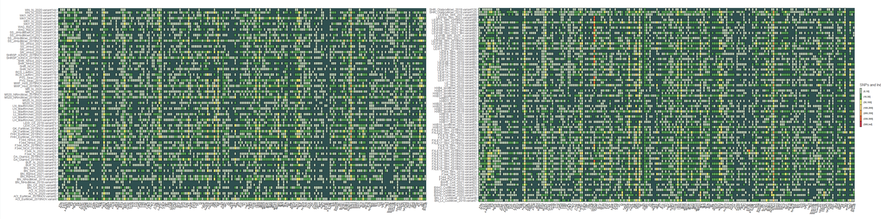


Hard Filter - 6 parameters PASS

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

All INDELs	5069666
HF INDELs	4790898
All SNPs	16805098
HF SNPs	12718917

High confidence variants - Number of variants decrease after applying hard filtering parameters and only variants supported by more than 10 reads were selected as high confidence set.

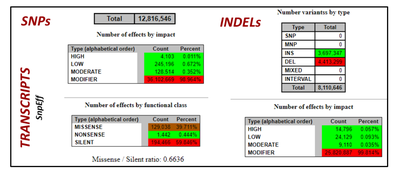


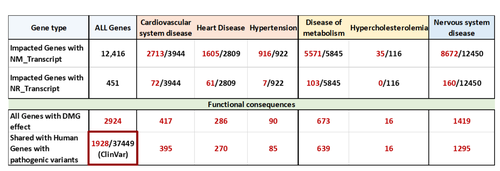
Genes per rat strain with number of variants after exclusion of shared variants with reference Brown Norway strain.

HRDP variants in RGD

	Atanur et al. 2013	Hermsen et al. 2015	RGD 2018	HRDP 2022	
Number of analyzed rat strains	27	40	25	47	72
Number of identified high confidence SNVs	9,665,340	9,183,702	8,953,897	11,585,641	12,354,591 12,718,917
Reference Genome	RGSC 3.4 - chromosomes	RGSC 5.0 - whole	RGSC 6.0 - whole	RGSC 6.0 - whole	mRatBN7v2- whole
Alignment Software	BWA-0.5.8c	BWA mem –M 0.7.5a	BWA mem 0.7.15	BWA mem 0.7.17	
Genomic variants call Variant quality recalibration (VQSR) - true training set	GATK v. 1.0.6001	GATK HaplotypeCaller v2.8-1	GATK HaplotypeCaller v3.6	GATK HaplotypeCaller v.4.1.3.0	
	Top 30% of high quality SNVs	Not defined	273,568 selected SNVs	In preparation	
dbSNPs	dbSNP125 - 41,658 (1,291) (35,186)	dbSNP138 - 5,076,239 (5,043,831)	dbSNP149 - 5,075,461 (5,042,280) 4,721,043	dbSNP149 - 5,075,461 (5,042,280) 4,721,043	dbSNP_EVAv3 9653928
Tranche sensitivity threshold	99.0	99.5	95.0	In preparation	

Comparison of previously reported analysis of single nucleotide variants identified in different rat populations.

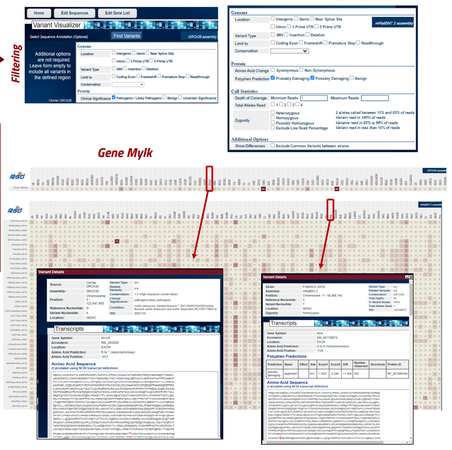




Gene type	ALL Genes	Cardiovascular system disease	Heart Disease	Hypertension	Disease of metabolism	Hypercholesterolemia	Nervous system disease
Impacted Genes with NM_Transcript	12,416	2713/3944	1605/2809	916/922	5571/5845	35/116	8672/12450
Impacted Genes with NR_Transcript	451	72/3944	61/2809	7/922	103/5845	0/116	160/12450
				Functional consequences			
All Genes with DMG effect	2924	417	286	90	673	16	1419
Shared with Human Genes with pathogenic variants	1928/37449 (ClinVar)	395	270	85	639	16	1295

Gene counts with damaging variants by Disease Ontology terms

OLGA Cross Analysis: tool allows to select terms from Disease, Pathway, Mammalian Phenotype, Biological Process, Cellular Component, Molecular Function and ChEBI ontologies for a given list of genes and further analyze it with other RGD tools.



***Variant Visualizer** shows variants distribution. First select **organism/breeds/strains** of interest, input **gene list** or define **genomic position**, set parameters/filters for the type(s) of desired variants and the tool will return all of the SNVs and/or indels which match the input criteria, including information on read depth, zygosity, conservation score and more.*

Grant support: R24OD022617, R01HL064541

Summary

- ***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.***
- ***Genetic variation analysis in rat strains selected for HRDP helps to generate high resolution association mapping and complete systems genetics on complex traits***
- ***We are building the HRDP Portal within RGD that will provide data mining and visualization functions for genomic and phenotypic data***