Pangenome reconstruction in rats enhances genotype-phenotype mapping and novel variant discovery

Flavia Villani[1], Andrea Guarracino[1,2], Rachel R Ward[3], Tomomi Green[3], Madeleine Emms[4], Michal Pravenec[5], Pjotr Prins[1], Erik Garrison[1], Robert W. Williams[1], Hao Chen[3] and Vincenza Colonna[1,4*]

1. Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA
2. Human Technopole, Viale Rita Levi-Montalcini 1, 20157, Milan, Italy
3. Department of Pharmacology, Addiction Science, and Toxicology, University of Tennessee Health Science Center
4. Institute of Genetics and Biophysics, National Research Council, Naples, 80111, Italy
5. Institute of Physiology, Czech Academy of Sciences, 14200 Prague, Czech Republic

*Corresponding author. E-mail: vicenza.colonna@uthsc.edu

Abstract

The HXB/BXH family of recombinant inbred rat strains is a unique genetic resource that has been extensively phenotyped over 25 years, resulting in a vast dataset of quantitative molecular and physiological phenotypes. We built a pangenome graph from 10x Genomics linked-read data for 31 recombinant inbred rats to study genetic variation and association mapping. The pangenome length was on average 2.4 times greater than the corresponding length of the reference mRatBN7.2, confirming the capture of substantial additional variation. We validated variants in challenging regions, including complex structural variants resolving into multiple haplotypes. Phenome-wide association analysis of validated SNPs uncovered variants associated with glucose/insulin levels and hippocampal gene expression. We propose an interaction between Pirl1l1, Cromogranine expression, TNF-α levels, and insulin regulation. This study demonstrates the utility of linked-read pangenomes for comprehensive variant detection and mapping phenotypic diversity in a widely used rat genetic reference panel.

Keywords

Rat, Pangenome, Genotype-Phenotype, Recombinant Inbred, Glucose, Insulin, Chromogranin expression

Introduction

Rattus norvegicus has a long history of scientific research and has been used as a model organism to study various human diseases.[1] One notable example is the HXB/BXH (referred to as HXB in general hereafter) family of recombinant inbred (RI) strains of rats. The HXB family started with reciprocal crosses between the SHR/OlaIpcv (H) and BN-Lx/Cub (B) strains, with the goal of creating distinct genetic lines. Currently, there are a total of 30 RI strains, all having undergone more than 50 generations of inbreeding. Among them, 21 belong to the HXB/Ipcv group, and 11 to the BXH/Cub group, with DNA samples preserved from almost all the original strains for the purpose of constructing genetic maps.[2]

The HXB is currently the most widely used RI family of laboratory rats and therefore represents therefore a great resource for the identification and mapping of quantitative trait loci (QTL) associated with complex phenotypes, motivating efforts to produce a comprehensive catalog of genetic variants. HXB have been utilized for studying the genetics of a wide range of phenotypes related to human diseases, such as blood pressure,[3] cardiac conduction,[4] Pavlovian conditioning,[5] ethanol metabolism[6] and hippocampal neurogenesis.[4]

Pangenomic methods directly compare all genomes to each other, enabling a comprehensive analysis of genomic diversity and relationships.[7] These methods have previously proven successful in representing human genetic variation[8] and addressing long-standing evolutionary questions.[9] The design, implementation, and application of pangenomic models constitute an active area of research. Ideally, pangenomes are constructed from long read data, but this technology is not always available, and the accuracy and utility of pangenome construction from suboptimal sequence data have been insufficiently explored.

In this study, we build a pangenome graph from 10x Chromium Linked-Reads sequence data from 31 HXB rats plus BN/NHsdMcwi (mRatBN7.2 reference genome), and use newly identified/discovered genetic variation discovered from it to perform phenome-wide association analysis and explore complex genetic variability within structural variants. We demonstrate the impact of using pangenomes in the mapping of complex traits through examples of genetic association of novel variants discovered from the pangenome and traits related to insulin and glucose metabolism.

Results

Pangenome building from linked-reads

We performed whole genome sequencing using the 10x Chromium Linked-Reads technology and de novo genome assembly of 31 inbred strains from the RI rat panel[10], using the sSupernova tool[11] **(Figure S1)**. Evaluation of genome assembly within genic regions[12] shows that the majority of the genes are complete and present in a single-copy configuration **(Figure S1)**, which is indicative of a high quality.

In our classical reference-based approach as implemented in joint calling (JC) of variants using DdeepVvariant/GLNnexus,[13] we identified 7,520,223 variant sites from the HXB/BXH panel, approximately 24.1 – 53.5 % of the alleles at these sites of each RI strain were derived from SHR/OlaIpcv. We found that the majority of the strains were highly inbred, with close to 98% of the variants being homozygous, one exception was BXH2, in which 7.7% of variants were heterozygous, likely due to a recent breeding error.[6] Based on these results, we considered the assembly as haploid and applied PGGB[8] to build pangenome graphs for all chromosomes from *de novo* genome assemblies of the full HXB family (generated from 10x linked reads) and the mRatBN7.2 (rn7) reference genome. The lengths of such chromosome pangenome graphs were calculated as the sum of their node lengths and resulted, on average, 2.4 times greater than the corresponding chromosome length of the reference mRatBN7.2 suggesting that the pangenome captures genetic variation beyond that contained in the reference. This is expected because the pangenome graphs we built represent the genetic diversity of all the strains they represent, including SNPs, INDELs, and SVs. In particular, each allele would be represented with a dedicated node, therefore increasing the total graph length. For this reason, pangenome graphs can be longer than the individual genomes they represent. However, the high difference in length we observed (chromosome graphs are 2.4 times longer than the reference chromosomes, on average) is partially due to the uncalled nucleotides in our assemblies (runs of Ns in the sequences) **(Table S1)**. To demonstrate the utility of a pangenomic approach, we focused on chromosome 12 (chr12), one of the shortest chromosomes. The pangenome graph of scaffolds that map to chr12 consists of 1M nodes, 1.6M edges, and 28.5k paths, with a total length of 78M bp, 1.7 times greater than the length of mRatBN7.2 chr12 reported in GenBank **(Figure 1A)**.

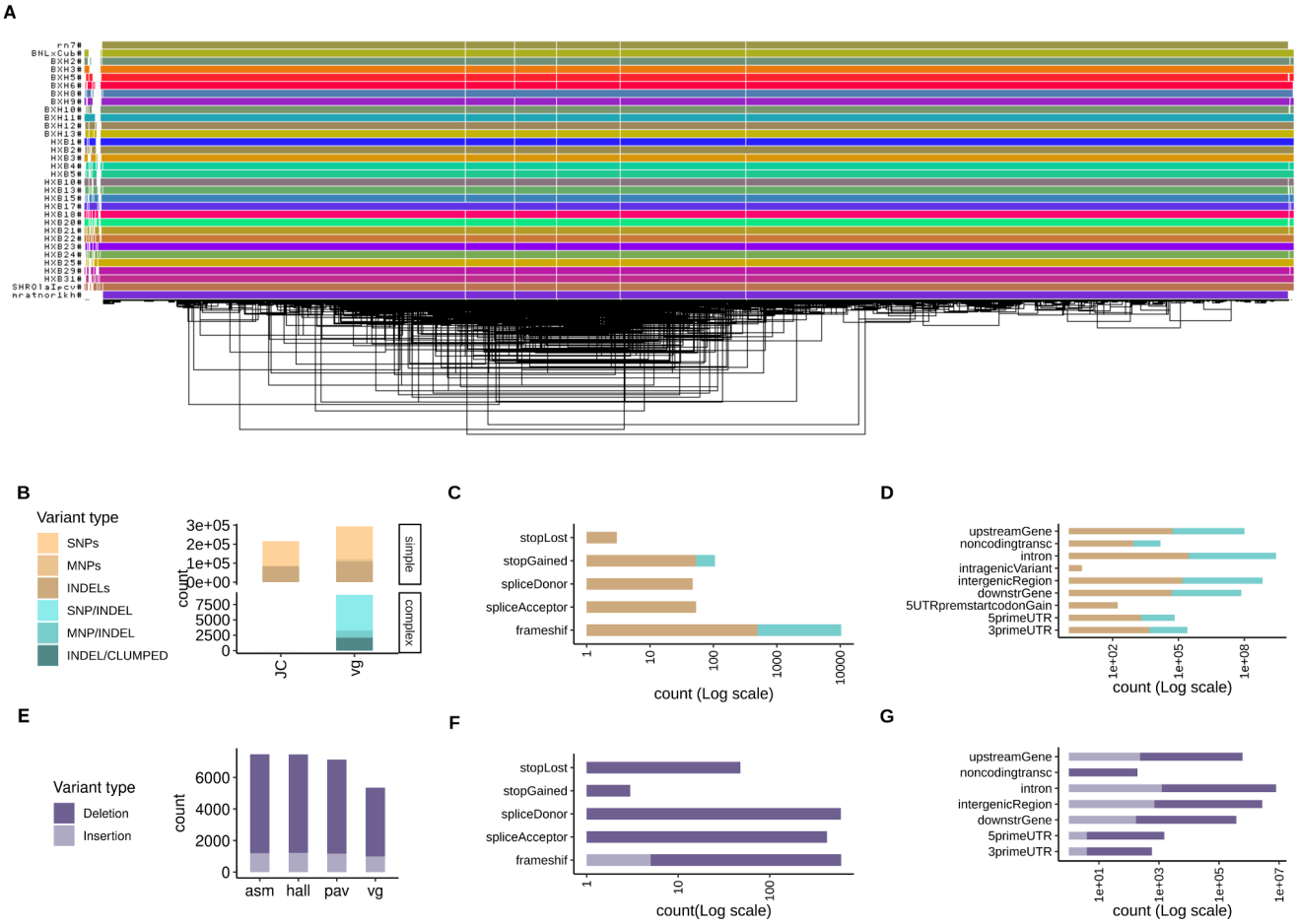

Figure 1. Pangenome graph of chromosome 12 from 31 HXB/BXH rats and genetic variants derived from it. (A) odgi-viz visualization of the pangenome graph for chromosome 12 from 31 rats. Horizontal colored lines denote haplotypes, with vertical white stripes representing insertions/deletions. Includes the mRatBN7.2 reference genome (top line). **(B)** Comparison between small variants (<50bp) called from the pangenome graph using vg (vg) with variants joint-called directly from 10x data using GLNnexus (JC), stratified by genome complexity. vg called 36% and 99.95% more simple and complex variants, respectively, compared to JC. **(C-D)** Predicted functional annotations for small variants called from the pangenome graph. **(E)** Comparison of three assembly-based methods (asm, hall, pav) and one graph-based method (vg) to call structural variants (SVs, >50bp) **(F-G)** Predicted functional annotations for SVs called from the pangenome graph.

Small variants analysis and validation in the pangenome

Small variants (<50 bp in length) were called from the pangenome using graph decomposition to identify bubbles corresponding to non-overlapping variant sites, as implemented in vg tools.[14] We also performed a classical reference-based approach as implemented in joint calling (JC) of variants using dDeepVvariant/GLNnexus.[13] Variant calling from the pangenome of chr12 identified 302,048 variants (single nucleotide polymorphisms, multi-nucleotide polymorphisms, insertions, deletions, and complex, i.e. nested mixed variants) *versus* 215,431 variants called by JC. Of these, vg called 36% and 99.95% more simple and complex variants respectively **(Figure 1B)**. With a focus on size up to 50 bp (small variants), we predicted functional consequences of variants on genes and proteins **(Table S2)**. The majority of variants are classified as modifiers (99.0%). A smaller proportion of variants have a high impact (0.1%) and are located within 364 protein-coding genes. For each class of variant impact (high, moderate, low, and modifier) annotations are more abundant for simple variants **(Figure 1C-D)**. Finally, with respect to gene structure, the majority of variants were located in intergenic regions and introns **(Table S2)**.

Small variants called from pangenome using vg were validated over JC call set (gold standard) in the SHR/OlaIpcv sample, which represents one of the best assemblies among the strains in our pangenome in terms of quality **(Table S3, Figure S1)**. Genome-wide accuracy of vg calls is ~90% **(Figure S2)**, with the exceptions of chromosome Y and 16, possibly due to the complexities in assembling chromosome Y and the abundance of complex variants for chromosome 16 **(Figure S2C)**. Further validation by resequencing was restricted to twenty SNPs within easy regions[15] of chr12. Selection criteria for these small variants were: support by original reads determined by mapping reads against the pangenome with vg giraffe,[16] and location in regions supposed to be particularly challenging (14 were within repetitive regions and 6 within 1500 bp from repetitive regions). Twelve out of the 20 variants were confirmed by Sanger sequencing **(Table S4)**, which translates into a 0.6 rate of validation. Approximately 90% of variants (7 out of 8) that failed the validation were within repeats.

Novel pangenome variants associated with glucose, insulin and chromogranin expression

We leveraged phenotypic data present in GeneNetwork[17] to perform phenome-wide association analysis (PheWAS) for the twelve SNPs validated through Sanger sequencing. PheWAS was restricted to 30 rat strains with phenotypic data, and 60 phenotype classes with at least two phenotypes per class. Two SNPs were significantly associated with six phenotypes within five phenotype classes (LRS>16, **Figure 2A-B, Table S5**). We further considered three of the six significant associations that were validated using a linear mixed model corrected for kinship implemented in GeneNetwork **(Figure 2C-E)**.

The first association is between blood glucose concentration and a SNP (chr12_4347739) within the long non-coding RNAs gene *Zfp958l1*. Rats carrying the H allele have significantly higher blood glucose concentrations **(Figure 2C)**. chr12_4347739 variant is located within a mapped QTL that controls insulin/glucose ratio, mapped in an independent cohort of 185 F2 rats (Insglur6; LOD score 18.97, p-value: 0.001).[18]

The second relevant association is between a variant and two phenotypes, insulin concentration in blood and gene expression of chromogranin (CGA) in the hippocampus, and an intronic SNP (chr12_18797475) within an annotated locus (*LOC685157*). Rats with the H allele at chr12_18797475 have higher expression of CGA in the hippocampus and lower blood insulin concentration **(Figure 2D-E)**. CGA expression is ubiquitous throughout the central nervous system,[19] especially in the hypothalamus and amygdala regions, and stimulates the production of tumor necrosis factor alpha (TNF-α)[20] a key inflammatory molecule that promotes insulin resistance and glucose accumulation. We hypothesize that in rats with the H allele, increased CGA expression and subsequent TNF-α production leads to insulin resistance, impairing cellular glucose uptake and resulting in reduced blood insulin levels **(Figure 2F)**. This proposed mechanism is consistent with the observed reduction in insulin concentration associated with the other significant PheWAS variant **(Figure 2C)**.

Remarkably, LOC685157 is similar to the paired immunoglobin-like type 2 receptor beta 1 like 1 gene (Pilrb1l1, Ensembl: LOC681182) predicted to code for a protein putatively located on the membrane. The associated intronic variant chr12_18797475 is located just 0.7kb upstream of Pilrb1l1. Pilrb1l1 orthologs to human PILRa code for a lectin that recognizes both O-glycan and aglycon[20] and a missense variant (rs1859788, p.Arg78Gly) in the functional

immunoglobulin like domain seems to be protective for Alzheimer's disease.[21] In mice, Pilrb1l1 is abundantly expressed on myeloid cells, and required for cell movement out of the circulatory system and towards the site of tissue damage or infection.[22] Pilrb−/− mice have reduced serum or bronchoalveolar lavage fluid levels of tumor necrosis factor alpha (TNF-α)[23] suggesting that increased CGA expression and consequent TNF-α production might compensate the lack of function of Pilrb1l1. The relationship between increased CGA expression, and insulin concentration aligns with previous findings, chromogranin A regulates catecholamine storage and release, variation affecting CGA levels could impact catecholamines that can suppress insulin secretion and affect glucose metabolism.[21,22] Finally, chr12_18797475 is located at 0.8 Mb from a mapped QTL that controls hyperglycemia related to diabetes and increased susceptibility to autoimmune disorder (Iddm2; LOD 18.97).[18]

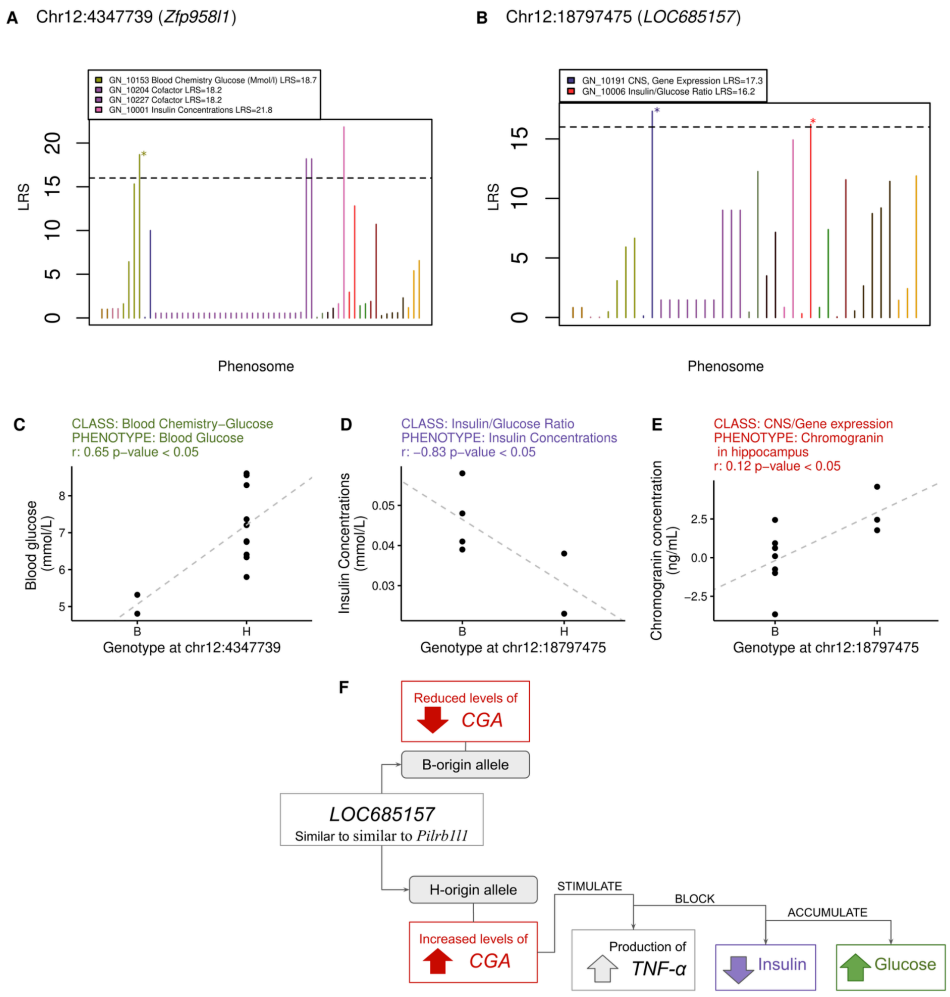

Figure 2. Mapping of validated pangenomic variants against phenotypes in GeneNetwork. (A,B) PheWAS and results for two variants in the long non-coding RNAs gene *Zfp958l1* and the annotated gene *LOC685157*, which is similar to *Pilrb1l1*. Colors refer to phenotype classes, the dashed horizontal line indicates threshold for significance in the PheWAS. Stars mark genotype-phenotype associations that were validated also through Spearman correlation analysis (adjusted p-value < 0.05) reported in **(C-E).** In **F** we propose a hypothesis that reconciles our genotype-specific findings with what is known in literature about these genes and their relationship.

Novel complex structural variants discovered from the pangenome

Structural variants (SV, >50bp in length) of chr12 from 31 HXB were called using four assembly-based methods (PAV, SVIM, hall, vg)[8], one of which is a graph-based method (vg). We retained 7,811 SVs supported by at least two tools out of 9,405 identified by at least one tool **Figure S3**). Contrary to small variants, here the graph-based method calls less SVs than other methods **(Figure 1E)**. Out of the 7,811 SVs, 1,282 are insertions and 6,526 deletions. We predicted the functional consequences of variants on genes and proteins **(Table S6, Figure 1F-G)**. The majority of variants are classified as modifiers (56.0%). A big proportion of variants have a high impact (44%) and are located within 918 protein-coding genes. Notably, the majority of variants were located in exon regions and introns **(Table S6)**.

We further validated SVs of the SHR/OlaIpcv sample called from the pangenome. We considered 2,481 variants seen by at least two methods out of the 4,853 identified by at least one tool **(Figure S3)**. We selected ten variants with high impact and size between 50-500bp that were examined using Oxford Nanopore long-read data from SHR/OlaIpcv rat. Two insertions and four deletions were confirmed, suggesting a validation rate of 0.6 **(Figure S4)**.

All six validated SVs are detected in additional non SHR strains, with variant frequency varying from 0.35 to 1 **(Table S7)**. A deletion of 300bp in the Lmtk2 gene is found in SHR/OlaIpcv and all other samples **(Figure 3A)**. A second deletion of 357bp in the *Mcemp1* gene found in SHR/OlaIpcv is also found in 17 rats **(Figure 3B)**. This G-rich SV contains a non-LTR retrotransposon that belongs to the Long Interspersed Elements (LINEs). It also contains two B1F repetitive elements, or *Alu,* that belong to the class of retroelements called short interspersed elements (SINEs, **Table S7**).

We also describe one deletion and one insertion found in SHR/OlaIpcv which represent examples of complex variation that is fully resolved by the pangenome graph that captures all the realized haplotypes. The first complex deletion SV (117 bp, supported by two methods) is found in the *LOC679924* gene, predicted to enable ATP binding activity, and overlaps for 80bp with a SINE element **(Figure 3C)**. The second complex insertion SV (82 bp, supported by all four methods) falls within the *Cd209c* gene, and, despite its shorter length, is made of seven variable blocks, which combine into six haplotypes, four of which are seen less than two times **(Figure 3D)**. It contains a Lx8b_3end, a non-LTR retrotransposon in rodents (LINE). The Human ortholog of *Cd209c* is implicated in Coronavirus infectious disease,[23] aspergillosis,[24] dengue disease,[25] and tuberculosis.[26]

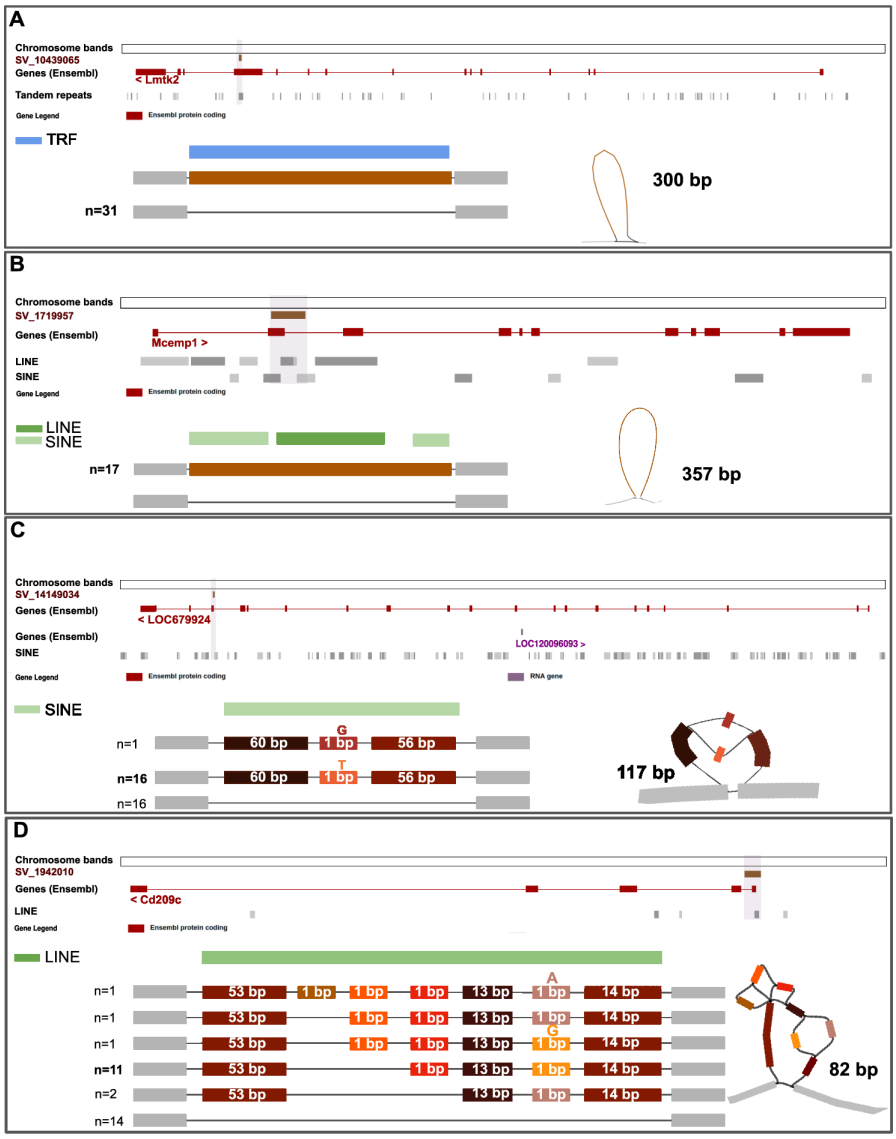

Figure 3. Resolution of complex haplotypes in validated structural variants (SVs) discovered using pangenome graphs. Visualization using ODGI and Bandage of four SVs predicted to have high impact and validated with Oxford Nanopore sequencing technology. **(A)** A deletion of 300bp in the *Lmtk2* gene supported by all four methods of identification, is found in all 31 samples and contains a Tandem Repeats element (TRF). **(B)** A deletion of 357bp in the *Mcemp1* gene is supported by all four methods and is found in 17 rats. The gene contains SINE and LINE elements. **(C)** A deletion of 117 bp in the *LOC679924* gene is supported by two methods, and includes a single nucleotide variant. The SV overlaps for 80 bp with a SINE element. **(D)** A complex insertion of 82 bp in the *Cd209c* gene, supported by all four methods. The insertion resolves in six haplotypes of which two are most abundant, and falls completely in a non-LTR retrotransposon (LINE).

Discussion

In this study, we built the first rat pangenome to explore the genetic variation landscape of the rat HXB/BXH family and conduct association mapping with emphasis on novel variants identified by our pangenomic approaches on chr12. The pangenome graph we built enabled the discovery of novel small variants and SVs. Although the sequencing data we used (i.e., 10x linked reads) were not ideal, our research highlights the potential for adequate pangenomic reconstruction.

One key finding of this study lies in the significant increase in the number and complexity of variants identified by the pangenome.

The pangenome lengths per chromosome are, on average, 2.4 times greater than the typical chromosome length. Variants in the same genomic regions across multiple samples lead to multiple representations of that region, and the graph accounts for all the different alleles. For this reason, pangenome graphs can be longer than the individual genomes they represent. Unresolved regions are not aligned between the genomes and then are represented in the graphs as separated nodes, leading to an increase in graph lengths. This also explains why we observed 4,86 Gb of non-reference sequence (obtained with respect to mRatBN7.2). Such high values overestimate the real non-reference bases in our pangenome and reflect the technological limitation of applying Linked-Reads in *de novo* assembly. Most of the novel

pangenomic variation is complex, confirming that the pangenome captures genetic variation beyond that contained in the reference genome.

Validation in SHR/OlaIpcv shows high rates of concordance and, except for complex repeated regions, we validated through resequencing novel variants in challenging regions. We provide validated examples of how the pangenome fully resolves complex nested variants within repetitive regions. We identified and validated novel complex SVs falling within genes implicated in immunity and enriched for transposable elements, a major source of genetic polymorphism, known to act as regulatory elements within immune regulatory networks, and facilitate rapid adaptive evolution of immune responses.[27] Mutations in the *Lmtk2* gene have been linked to multiple neurological disorders, including Alzheimer's disease, Parkinson's disease, and schizophrenia in humans. *Mcemp1* seems to be involved in the regulation of mast cell differentiation and immune responses. Its human orthologous *MCEMP1* is one of the top inducible genes in many inflammatory diseases, such as asthma,[28] idiopathic pulmonary fibrosis,[29] cancer,[30] sepsis,[31] and stroke.[32] *MCEMP1* is a critical factor in allergic and inflammatory lung diseases,[33] with the level of *MCEMP1* expression positively correlating with disease severity. These examples provide crucial insights into regions of the genome that play a key role in adaptation, but have been poorly investigated so far due to their challenging resolution.

We demonstrate that novel variants identified by the pangenome are valuable for mapping quantitative traits, demonstrating that access to this fullest spectrum of genetic diversity is key for studying genotype-phenotype connections. We used validated SNPs to conduct phenome-wide association analysis, to discover two significant associations: blood glucose concentration linked to a variant (chr12_4347739) within the *Zfp958l1* gene, and insulin levels in the blood and the expression of chromogranin in the hippocampus, mediated by an intronic variant (chr12_18797475) near a gene (*LOC685157*) similar to *Pilrb1l1*.

Overall, this pangenomic study enhances our understanding of the rat HXB/BXH family and sets a new standard for genomic research in model organisms, underscoring the importance of comprehensive genetic analysis for revealing the complexity of genotype-phenotype connections.

Limitations of the study

Although this pioneering rat pangenome analysis demonstrates the capability to capture complex variation beyond the reference genome and enable more complete genotype-phenotype mapping, we note a few limitations. The linked-read technology is not ideal for *de novo* assembly and leads to an overestimation of non-reference sequences due to unresolved regions. Additionally, complex regions remain challenging to fully resolve despite validation efforts. The combination of linked-read and long-read sequencing technologies offers the promise of constructing higher-quality assemblies and then pangenomes. As sequencing technologies and algorithms continue advancing, we expect to achieve complete pangenome assembly and whole genome structural variation characterization that will reveal the fullest extent of genome variation across rat strains. Despite current limitations, our study demonstrates the capability of pangenomes to enable more accurate, even more complete genotype-phenotype mapping.

Acknowledgments

The work of H.C. E.G. and R.W.W. is supported by NIH NIDA U01DA047638.
M.P. was supported by a program from the National Institute for Research of Metabolic and Cardiovascular Diseases (Program EXCELES, ID Project No. LX22NPO5104) funded by the European Union - Next Generation EU and by grant LUAUS23095 within the INTER-EXCELLENCE program of the Ministry of Education, Youth, and Sports of the Czech Republic.

Author contributions

Conceptualization, F.V., A.G., M.E., E.G., R.W.W., H.C., V.C.; Formal Analysis, F.V., A.G., M.E., H.E., V.C.; Visualization, E.G., R.W.W., H.C., V.C.; Investigation, R.W., T.G., M.P.; Resources, P.P., R.W.W., H.C.; Data Curation, F.V., A.G., R.W., T.G., E.G., R.W.W., H.C., V.C.; Writing - Review & Editing, F.V., A.G., R.W., T.G., E.G., P.P., R.W.W., H.C., V.C.; Supervision, E.G., R.W.W., H.C., V.C.

Declaration of interests

The authors declare no competing interests.

Methods

Supernova assembly

Fastq files were used as the input for the Supernova (version 2.1.1) program provided by 10x genomics[11] to generate *de novo* assembly with default settings. To evaluate genome assembly completeness, we used the Compleasm[34] tool to assess representation of universal single-copy orthologs expected to be present across mammals (BUSCO Mammalia gene set). Compleasm searched each Supernova genome assembly for the presence and completeness of these conserved orthologs. The reference mRatBN7.2/rn7 genome was analyzed comparably as a benchmark.[10]

Pangenome generation and evaluation

Supernova haploid assemblies obtained by 10x chromium linked reads technology were mapped against the mRatBN7.2/rn7.fa reference genome using wfmash[35] to partition the assemblies' contigs by chromosome. Mapped assemblies were used to build the pangenome with PGGB (v.c1c3a1565604fc41f880bccd9f46d0a709f3e774)[36] using this combination of parameters: -p 98 -s 2000 -n 32 -F 0.001 -k 79 -P asm5 -O 0.03 -G 4001,4507 -V rn7:# -t 48 -T 48. Validation of the call set was performed on SHR/OlaIpcv sample, using one true positive set obtained from joint-calling conducted using GLNnexus.[13] Prior to comparison, the pangenome-derived and the validated call set were processed to remove missing data, sites where alleles are stretches of Ns, homozygous reference genotypes and variants greater than 50bp before normalization and decomposition using bcftools[37] under default parameters. While the pangenome-derived VCF was based on haploid assemblies, for comparison purposes the calls were considered as homozygous diploid in the assumption that SHR/OlaIpcv is isogenic. Comparison of the two call sets was performed with RTG tools (v.3.12.1)[38] using the --squash-ploidy option. The RTG tool used for the variants' evaluation gives us some information about the variants in common, called by vg and JC. From these results we considered only the vcf file which contains variants called by vg-only. RepeatMasker[38] was used to mask complex regions, easy regions that do not contain low complexity regions and repeats. Hard regions contain the whole genome regions. RepeatMasker, Ensembl Genome Browser[39] and Dfam[40] databases are used to check the

abundance of repeats across the genome. We mapped raw reads of SHR/OlaIpcv against the pangenome using vg giraffe (v. 1.41.0).[16] For the pangenome graph statistics we used ODGI (v.0.7.3).[41] For the pangenome graph visualization we used ODGI and Bandage.[42] To check the length of the rat genome assemblies (mRatBN7.2/rn7 reference) we used the National Center for Biotechnology Information (NCBI) online resource.[43] For the vcf statistics we used bcftools[37] stats and vt peek.[44]

Variant calling

The variant calling of variants on the pangenome was obtained using vg (v. 1.41.0)[14] and in particular -V rn7:# into the pggb command line. Small variants are variants with length <50 bp, simple variants are SNPs/MNPs/indels, complex variants are allelic variants that overlap but do not cover the same range. For the variant classification in simple and complex variants we used vt resource (https://genome.sph.umich.edu/wiki/Variant_classification). For variant calling based on the reference genome, fastq files of 10x chromium linked data were mapped against mRatBN7.2 using LongRanger (version 2.2.2) (https://github.com/10XGenomics/longrangere Long Ranger code is available at https://github.com/10XGenomics/longranger). DeepvVariant (ver 1.0.0)[38] was used to call variants for each sample. Joint calling of variants for all the samples was conducted using GLNnexus.[13]

To optimize a pangenomic perspective and capture SVs that might be missed with traditional reference-based methods, we applied several SV calling pipelines. As a graph-based method, we used PGGB and vg to generate a comprehensive VCF file. Variants were normalized using vcfwawe[30] and then filtered by keeping only variants >50 bp. As assembly-based methods, we applied 3 pipelines as described in.[8] To evaluate the concordance between all the approaches, we conducted an overlap analysis using truvari (v.4.0),[31] a tool for assessing SV calls following some criteria 1) we merged SVs using a maximum allowed distance of 1000 bp, as measured pairwise between breakpoints (begin1 vs begin2, end1 vs end2) 2) we reported calls supported by at least 2 callers and they have to agree on the type and on the strand of the SV 3) we removed variants with more than 10% of Ns in REF/ALT. For the SV's pangenome graph extraction we used odgi (v.0.7.3) and for the visualization we used Bandage and odgi. Subsequently, we investigated the impact of small variants and SVs using

SnpEff (v5.1),[30] a well established tool for annotating and predicting variant effects. Through SnpEff analysis, we gained insights into the potential functional consequences of the detected genetic variants on the identified genes.

Sanger resequencing for variant validation

Variants only found with the pangenome approach were randomly selected from Chr 12 for the SHR/OlaIpcv sample. Primers were designed using NCBI primerblast and synthesized by IDT at 25 nm scale with standard desalt purification (Coralville, Iowa). DNA from a male adult SHR/OlaIpcv rat was extracted from brain tissue using DNEasy Blood and Tissue Kit (Qiagen, Cat no. 69504) and amplified using Phusion High-Fidelity PCR Kit (ThermoFisher, Cat no. F553L) in a Thermocycler for 30 cycles. PCR products were inspected on 1% agarose gel. PCR products with one distinctive band were submitted to GeneWiz (South Plainfield, NJ) for sequencing. Sequencing results were mapped against the target sequence using the blast2seq application and mutations were manually examined for confirmation.

PheWAS analyses and functional annotation

A Phenotype Wide Association Study (PheWAS) was conducted to identify potential traits related to the validated SNPs. The Phenome of the HXB/BXH family was downloaded from GeneNetwork.org (version 2) using an API. Out of 324 phenotypes we included 60 phenotypes, with data from at least 2 observations per phenotype within phenotypic class. Phenotypes were correlated via the 'hxb.phenotypes' object, using an adaptation of BXDtools[48], an R package containing various genetic tools to work in particular with model organisms. BXDtools employs a simpler linear model. Spearman correlations between genotype and phenotype were carried out using R, with p-values adjusted (Bonferroni) according to the number of samples each phenotype contains. The statistical significance level was set to $p\text{-adj} < 0.05$. This relationship was visualized in a genotype-phenotype plot. Ensembl Genome Browser[35] and Rat Genome Database (RGD)[30] were used to identify genes. Genome assembly (mRatBN7.2/rn7) for *Rattus norvegicus* was used.[34,4] To enhance the visualization and understand the distribution of these genes and their variant impacts on chr12, we employed a specialized R package called karyotypeR.[49]

SVs validation with nanopore adaptive sequencing

To validate the SVs identified, we employed Adaptive Sampling, a nanopore sequencing-specific software-controlled enrichment method. For Oxford Nanopore Adaptive Sequencing, 4.4 µg of DNA was extracted from 22 mg of microdissected rat brain tissue using the Qiagen DNeasy Blood & Tissue Kit and quantified using Nanodrop. The extracted DNA was fragmented using Covaris' g-TUBE for 30 seconds at 9100 x g at room temperature, for a median fragment size of ~3kb . Library prep was performed with ONT's Ligation Kit and libraries were quantified with Qubit. Following ONT's guidelines for the use of their Ligation Kit with Adaptive Sampling and ONT Community input, 100 fmol of library was loaded onto a MinION flow cell and run for 72 hours with Adaptive Sampling selected in run setup. Nanopore validated reads are mapped against the mRatBN7.2/rn7.fa reference genome with minimap2,[50] and the bam files are visualized using Integrative Genomics Viewer (IGV).[51]

Data and code availability

- Linked-read WGS data for the HXB/BXH family is available from NIH SRA. The SRA identifications for each sample is provided in (S8 Table).
- All original code has been deposited at GitHub (https://github.com/Flavia95/HXB_rat_pangenome_manuscript/).

Supplementary Information

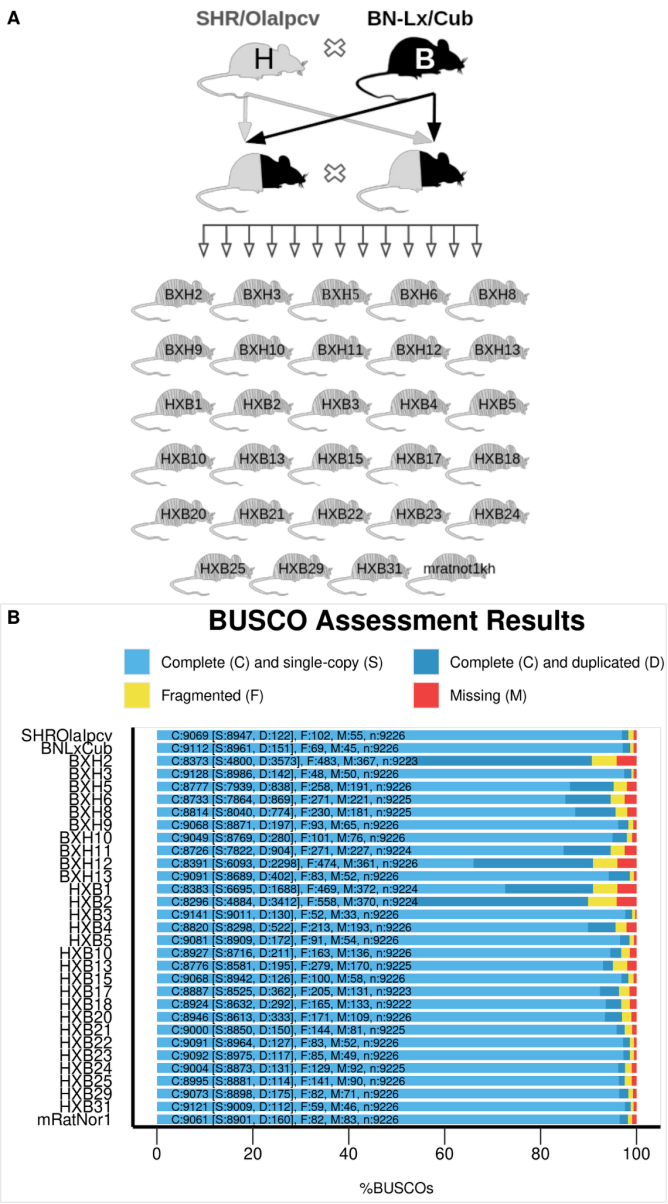

Figure S1. Description of rats used in this study and quality of the genome assembly. (A) Overview of the HXB/BXH recombinant inbred rat family describing the origins and breeding history of the HXB/BXH rats. **(B)** BUSCO completeness of the genome assembly used to build the pangenome for genomics data quality control. Bar charts show proportions classified as complete (C, blues), complete single-copy (S, light blue), complete duplicated (D, dark blue), fragmented (F, yellow), and missing (M, red).

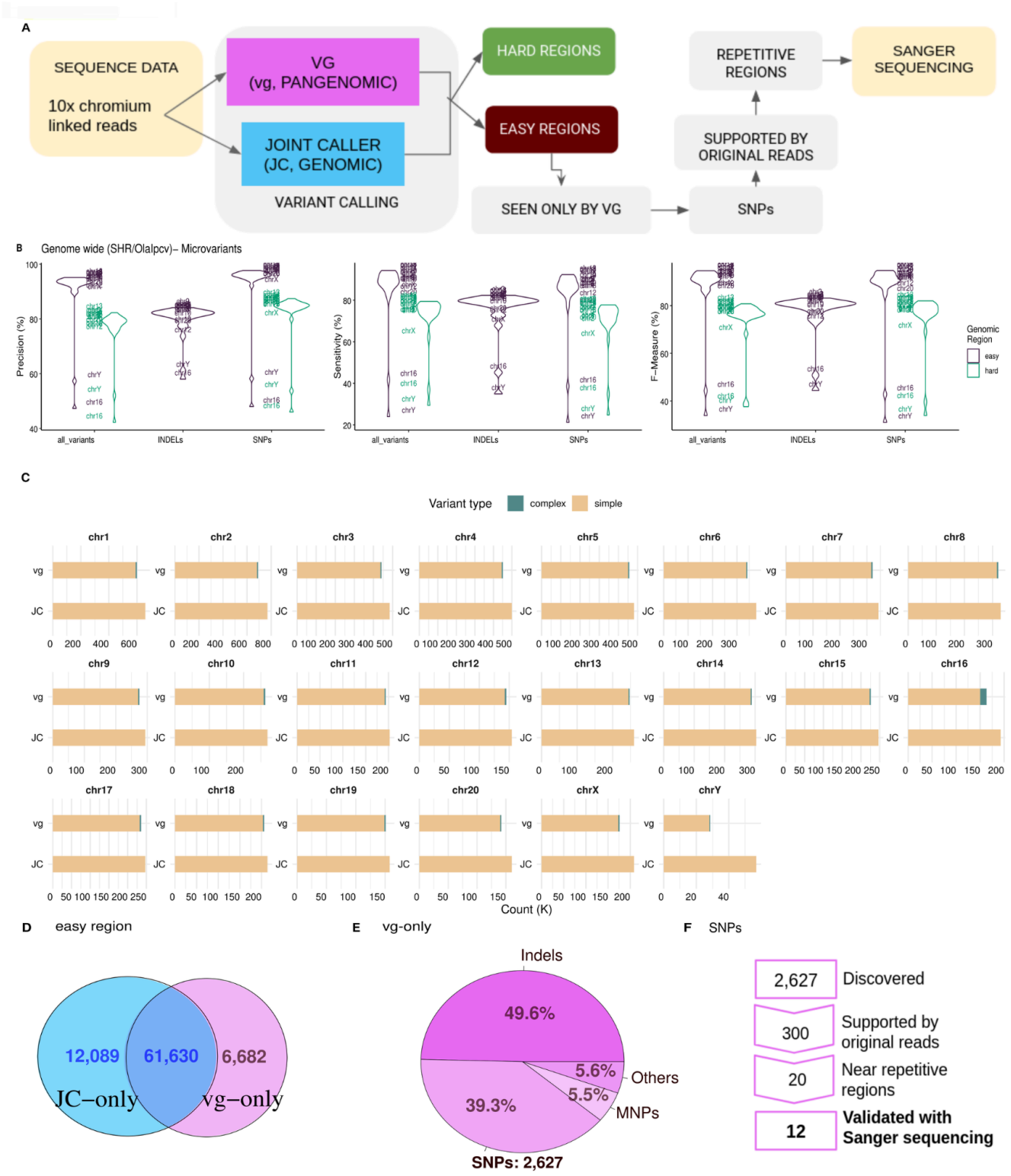
A
SEQUENCE DATA
10x chromium linked reads
VG (vg, PANGENOMIC)
JOINT CALLER (JC, GENOMIC)
VARIANT CALLING
HARD REGIONS
EASY REGIONS
SEEN ONLY BY VG
SNPs
SUPPORTED BY ORIGINAL READS
REPETITIVE REGIONS
SANGER SEQUENCING
B
Genome wide (SHR/OlaIpcv)– Microvariants
Precision (%)
Sensitivity (%)
F–Measure (%)
all_variants
INDELs
SNPs
Genomic Region
easy
hard
C
Variant type
complex
simple
Count (K)
D
easy region
12,089
61,630
6,682
JC–only
vg–only
E
vg-only
Indels
49.6%
5.6%
Others
5.5%
MNPs
39.3%
SNPs: 2,627
F
SNPs
2,627
Discovered
300
Supported by original reads
20
Near repetitive regions
12
Validated with Sanger sequencing

Figure S2. Validation in the SHR/Olapavic sample of the small variants called from the genome graph using vg. Small variants were validated over the JC call set (gold standard) in the SHR/Olalpcv sample according to the scheme in **(A). (B)** Genome-wide accuracy of vg calls is ~100% in the easy regions for SNPs, ~90% in the easy regions and ~80% in the easy regions for Indels and hard regions of the genome. Exceptions are seen for chromosomes 16 and Y, which are enriched for complex variation **(C)**. Validation through Sanger sequencing was restricted to easy regions **(D)**, to SNPs **(E)** supported by original reads and in challenging, repetitive regions **(F).**

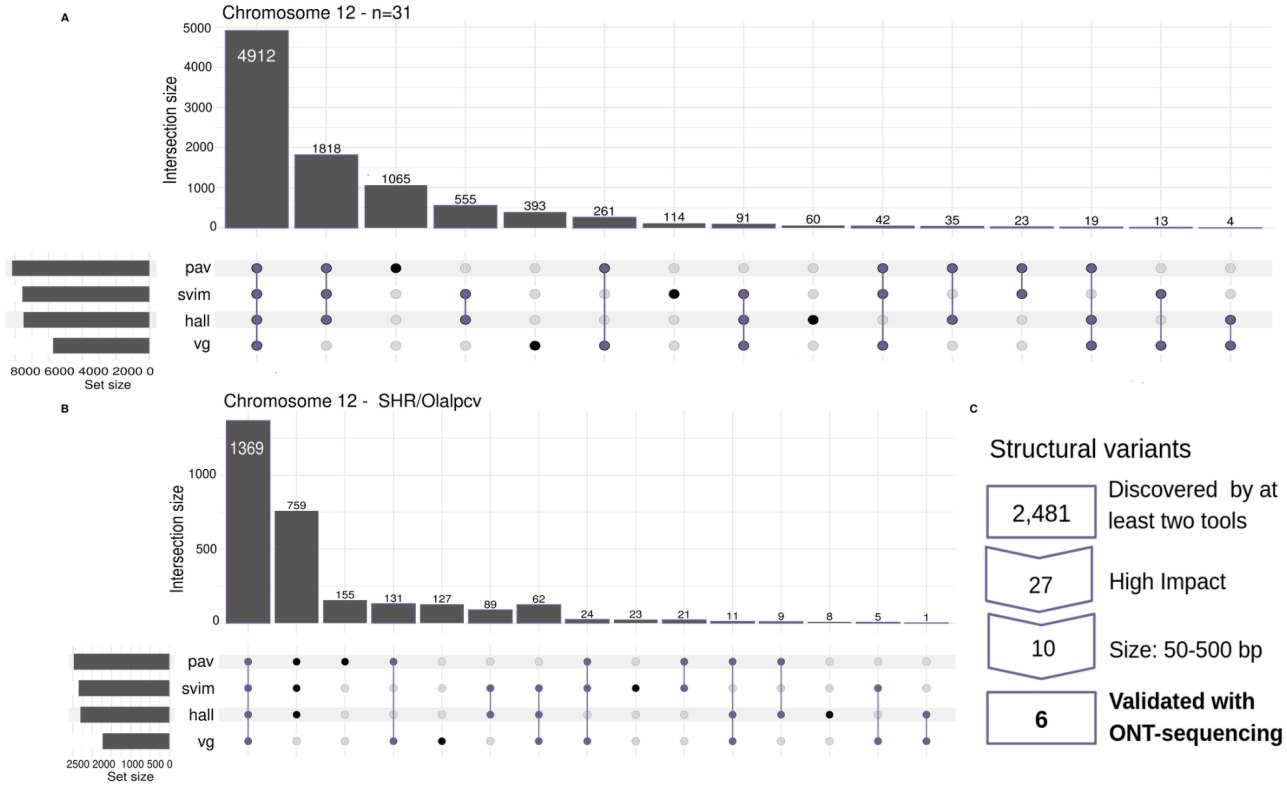

Figure S3. Validation in the SHR/Olapavic sample of the Structural Variants (SVs) called from the genome graph using vg. (A) Overlap of the call sets obtained by the three assembly-based methods (pav, svim, hall), and the graph-based method (vg) for chromosome 12 using the data from all rats. **(B)** Same as (A) for SHR/Olalpcv only. **(C)** Scheme of the validation for SVs

Figure S4. Integrated Genomic View of the validated SVs. For each SV gray bars are validated reads mapped against the mRatBN7.2/rn7.fa reference genome, red bars define the boundaries of the SV.

Supplementary tables: Supplementary Tables

References

1. Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature *428*, 493–521. 10.1038/nature02426.
2. Printz, M.P., Jirout, M., Jaworski, R., Alemayehu, A., and Kren, V. (2003). Invited Review: HXB/BXH rat recombinant inbred strain platform: a newly enhanced tool for cardiovascular, behavioral, and developmental genetics and genomics. J. Appl. Physiol. *94*, 2510–2522. 10.1152/japplphysiol.00064.2003.

3. Dmitrieva, R.I., Hinojos, C.A., Grove, M.L., Bell, R.J., Boerwinkle, E., Fornage, M., and Doris, P.A. (2009). Genome-Wide Identification of Allelic Expression in Hypertensive Rats. Circ. Cardiovasc. Genet. *2*, 106–115. 10.1161/CIRCGENETICS.108.809509.
4. Adriaens, M.E., Lodder, E.M., Moreno-Moral, A., Šilhavý, J., Heinig, M., Glinge, C., Belterman, C., Wolswinkel, R., Petretto, E., Pravenec, M., et al. (2018). Systems Genetics Approaches in Rat Identify Novel Genes and Gene Networks Associated With Cardiac Conduction. J. Am. Heart Assoc. *7*, e009243. 10.1161/JAHA.118.009243.
5. Bielavská, E., Kr, V., and Pravenec, M. Genome Scanning of the HXB/BXH Sets of Recombinant Inbred Strains of the Rat for Quantitative Trait Loci Associated with Conditioned Taste Aversion.
6. Lusk, R., Saba, L.M., Vanderlinden, L.A., Zidek, V., Silhavy, J., Pravenec, M., Hoffman, P.L., and Tabakoff, B. (2018). Unsupervised, Statistically Based Systems Biology Approach for Unraveling the Genetics of Complex Traits: A Demonstration with Ethanol Metabolism. Alcohol. Clin. Exp. Res. *42*, 1177–1191. 10.1111/acer.13763.
7. Eizenga, J.M., Novak, A.M., Sibbesen, J.A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J.D., Rounthwaite, R., Ebler, J., et al. (2020). Pangenome Graphs. Annu. Rev. Genomics Hum. Genet. *21*, 139–162. 10.1146/annurev-genom-120219-080406.
8. Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., et al. (2023). A draft human pangenome reference. Nature *617*, 312–324. 10.1038/s41586-023-05896-x.
9. Guarracino, A., Buonaiuto, S., de Lima, L.G., Potapova, T., Rhie, A., Koren, S., Rubinstein, B., Fischer, C., Human Pangenome Reference Consortium, Abel, H.J., et al. (2023). Recombination between heterologous human acrocentric chromosomes. Nature *617*, 335–343. 10.1038/s41586-023-05976-y.
10. de Jong, T.V., Pan, Y., Rastas, P., Munro, D., Tutaj, M., Akil, H., Benner, C., Chitre, A.S., Chow, W., Colonna, V., et al. (2023). A revamped rat reference genome improves the discovery of genetic diversity in laboratory rats (Genomics) 10.1101/2023.04.13.536694.
11. Marks, P., Garcia, S., Barrio, A.M., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A., et al. (2019). Resolving the full spectrum of human genome variation using Linked-Reads. Genome Res. *29*, 635–645. 10.1101/gr.234443.118.
12. Huang, N., and Li, H. (2023). miniBUSCO: a faster and more accurate reimplementation of BUSCO (Genomics) 10.1101/2023.06.03.543588.
13. Yun, T., Li, H., Chang, P.-C., Lin, M.F., Carroll, A., and McLean, C.Y. (2021). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. Bioinformatics *36*, 5582–5589. 10.1093/bioinformatics/btaa1081.
14. Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T., Jones, W., Garg, S., Markello, C., Lin, M.F., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat. Biotechnol. *36*, 875–879. 10.1038/nbt.4227.
15. Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinforma. *Chapter 4*, 4.10.1-4.10.14. 10.1002/0471250953.bi0410s25.
16. Sirén, J., Monlong, J., Chang, X., Novak, A.M., Eizenga, J.M., Markello, C., Sibbesen, J.A., Hickey, G., Chang, P.-C., Carroll, A., et al. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science *374*, abg8871. 10.1126/science.abg8871.
17. Mulligan, M.K., Mozhui, K., Prins, P., and Williams, R.W. (2017). GeneNetwork: A Toolbox for Systems Genetics. In Systems Genetics Methods in Molecular Biology., K. Schughart and R. W. Williams, eds. (Springer New York), pp. 75–120. 10.1007/978-1-4939-6427-7_4.
18. Marissal-Arvy, N., Heliès, J.-M., Tridon, C., Moisan, M.-P., and Mormède, P. (2014). Quantitative Trait Loci Influencing Abdominal Fat Deposition and Functional Variability of the HPA Axis in the Rat. Horm. Metab. Res. *46*, 635–643. 10.1055/s-0034-1383574.
19. Schafer, M.K.-H., Mahata, S.K., Stroth, N., Eiden, L.E., and Weihe, E. (2010). Cellular distribution

of chromogranin A in excitatory, inhibitory, aminergic and peptidergic neurons of the rodent central nervous system. Regul. Pept. *165*, 36–44. 10.1016/j.regpep.2009.11.021.
20. Ciesielski-Treska, J., Ulrich, G., Taupenot, L., Chasserot-Golaz, S., Corti, A., Aunis, D., and Bader, M.-F. (1998). Chromogranin A Induces a Neurotoxic Phenotype in Brain Microglial Cells. J. Biol. Chem. *273*, 14339–14346. 10.1074/jbc.273.23.14339.
21. Jirout, M.L., Friese, R.S., Mahapatra, N.R., Mahata, M., Taupenot, L., Mahata, S.K., Křen, V., Zídek, V., Fischer, J., Maatz, H., et al. (2010). Genetic regulation of catecholamine synthesis, storage and secretion in the spontaneously hypertensive rat. Hum. Mol. Genet. *19*, 2567–2580. 10.1093/hmg/ddq135.
22. Zhang, K., Mir, S.A., Hightower, C.M., Miramontes-Gonzalez, J.P., Maihofer, A.X., Chen, Y., Mahata, S.K., Nievergelt, C.M., Schork, N.J., Freedman, B.I., et al. (2015). Molecular Mechanism for Hypertensive Renal Disease: Differential Regulation of Chromogranin A Expression at 3′-Untranslated Region Polymorphism C+87T by MicroRNA-107. J. Am. Soc. Nephrol. *26*, 1816–1825. 10.1681/ASN.2014060537.
23. Nagozir, S., Shakouri Khomartash, M., Parsania, M., Vahidi, M., and Ghorbani, M. (2023). Association between genetic variants in the CD209 gene and susceptibility to COVID-19 in Iranian population. Hum. Gene *38*, 201215. 10.1016/j.humgen.2023.201215.
24. Morton, C.O., Fliesser, M., Dittrich, M., Mueller, T., Bauer, R., Kneitz, S., Hope, W., Rogers, T.R., Einsele, H., and Loeffler, J. (2014). Gene Expression Profiles of Human Dendritic Cells Interacting with Aspergillus fumigatus in a Bilayer Model of the Alveolar Epithelium/Endothelium Interface. PLoS ONE *9*, e98279. 10.1371/journal.pone.0098279.
25. Sakuntabhai, A., Turbpaiboon, C., Casadémont, I., Chuansumrit, A., Lowhnoo, T., Kajaste-Rudnitski, A., Kalayanarooj, S.M., Tangnararatchakit, K., Tangthawornchaikul, N., Vasanawathana, S., et al. (2005). A variant in the CD209 promoter is associated with severity of dengue disease. Nat. Genet. *37*, 507–513. 10.1038/ng1550.
26. Vannberg, F.O., Chapman, S.J., Khor, C.C., Tosh, K., Floyd, S., Jackson-Sillah, D., Crampin, A., Sichali, L., Bah, B., Gustafson, P., et al. (2008). CD209 Genetic Polymorphism and Tuberculosis Disease. PLoS ONE *3*, e1388. 10.1371/journal.pone.0001388.
27. Ivancevic, A., and Chuong, E.B. (2020). Transposable elements teach T cells new tricks. Proc. Natl. Acad. Sci. *117*, 9145–9147. 10.1073/pnas.2004493117.
28. Bosco, A., McKenna, K.L., Firth, M.J., Sly, P.D., and Holt, P.G. (2009). A Network Modeling Approach to Analysis of the Th2 Memory Responses Underlying Human Atopic Disease. J. Immunol. *182*, 6011–6021. 10.4049/jimmunol.0804125.
29. Herazo-Maya, J.D., Noth, I., Duncan, S.R., Kim, S., Ma, S.-F., Tseng, G.C., Feingold, E., Juan-Guardela, B.M., Richards, T.J., Lussier, Y., et al. (2013). Peripheral Blood Mononuclear Cell Gene Expression Profiles Predict Poor Outcome in Idiopathic Pulmonary Fibrosis. Sci. Transl. Med. *5*. 10.1126/scitranslmed.3005964.
30. Zeng, D., Wu, J., Luo, H., Li, Y., Xiao, J., Peng, J., Ye, Z., Zhou, R., Yu, Y., Wang, G., et al. (2021). Tumor microenvironment evaluation promotes precise checkpoint immunotherapy of advanced gastric cancer. J. Immunother. Cancer *9*, e002467. 10.1136/jitc-2021-002467.
31. Chen, J., Xu, X., and Zhang, S. (2019). Silence of long noncoding RNA NEAT1 exerts suppressive effects on immunity during sepsis by promoting microRNA-125-dependent MCEMP1 downregulation. IUBMB Life *71*, 956–968. 10.1002/iub.2033.
32. Wood, H. (2016). MCEMP1 — a new prognostic and diagnostic biomarker for stroke? Nat. Rev. Neurol. *12*, 127–127. 10.1038/nrneurol.2016.17.
33. Choi, Y.J., Yoo, J.-S., Jung, K., Rice, L., Kim, D., Zlojutro, V., Frimel, M., Madden, E., Choi, U.Y., Foo, S.-S., et al. (2023). Lung-specific MCEMP1 functions as an adaptor for KIT to promote SCF-mediated mast cell proliferation. Nat. Commun. *14*, 2045. 10.1038/s41467-023-37873-3.
34. Huang, N., and Li, H. (2023). compleasm: a faster and more accurate reimplementation of BUSCO. Bioinformatics *39*, btad595. 10.1093/bioinformatics/btad595.

35. Guarracino,Garrison 2021 wfmash: a pangenome-scale aligner.
36. Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Marco-Sola, S., Kubica, C., et al. (2023). Building pangenome graphs (Bioinformatics) 10.1101/2023.04.05.535718.
37. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. 10.1093/gigascience/giab008.
38. Cleary, J.G., Braithwaite, R., Gaastra, K., Hilbush, B.S., Inglis, S., Irvine, S.A., Jackson, A., Littin, R., Rathod, M., Ware, D., et al. (2015). Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines (Bioinformatics) 10.1101/023754.
39. Martin, F.J., Amode, M.R., Aneja, A., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., et al. (2023). Ensembl 2023. Nucleic Acids Res. *51*, D933–D941. 10.1093/nar/gkac958.
40. Wheeler, T.J., Clements, J., Eddy, S.R., Hubley, R., Jones, T.A., Jurka, J., Smit, A.F.A., and Finn, R.D. (2012). Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res. *41*, D70–D82. 10.1093/nar/gks1265.
41. Guarracino, A., Heumos, S., Nahnsen, S., Prins, P., and Garrison, E. (2022). ODGI: understanding pangenome graphs. Bioinforma. Oxf. Engl. *38*, 3319–3326. 10.1093/bioinformatics/btac308.
42. Wick, R.R., Schultz, M.B., Zobel, J., and Holt, K.E. (2015). Bandage: interactive visualization of *de novo* genome assemblies. Bioinformatics *31*, 3350–3352. 10.1093/bioinformatics/btv383.
43. Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., et al. (2022). Database resources of the national center for biotechnology information. Nucleic Acids Res. *50*, D20–D26. 10.1093/nar/gkab1112.
44. Tan, A., Abecasis, G.R., and Kang, H.M. (2015). Unified representation of genetic variants. Bioinformatics *31*, 2202–2204. 10.1093/bioinformatics/btv112.
45. Garrison, E., Kronenberg, Z.N., Dawson, E.T., Pedersen, B.S., and Prins, P. (2022). A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. PLOS Comput. Biol. *18*, e1009123. 10.1371/journal.pcbi.1009123.
46. English, A.C., Menon, V.K., Gibbs, R.A., Metcalf, G.A., and Sedlazeck, F.J. (2022). Truvari: refined structural variant comparison preserves allelic diversity. Genome Biol. *23*, 271. 10.1186/s13059-022-02840-6.
47. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w ^{1118} ; iso-2; iso-3. Fly (Austin) *6*, 80–92. 10.4161/fly.19695.
48. Arends,D (2017)) BXDtools.
49. Gel, B., and Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics *33*, 3088–3090. 10.1093/bioinformatics/btx346.
50. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. 10.1093/bioinformatics/bty191.
51. Robinson, J.T. (2011). Integrative genomics viewer. C O Rresp O N N Ce *29*.