bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Extended regions of suspected mis-assembly in the rat reference genome

Shweta Ramdas[1,*], Ayse Bilge Ozel[2,*], Katie Holl[3], Myrna Mandell[4], Leah Solberg Woods[5,#], Jun Z. Li[2,6,#]

[1] Program in Bioinformatics, University of Michigan Ann Arbor

[2] Department of Human Genetics, University of Michigan Ann Arbor

[3] Department of Pediatrics, Human and Molecular Genetics Center and Children's Research Institute, Medical College of Wisconsin, Milwaukee, Wisconsin

[4] National Institutes of Health, Bethesda

[5] Department of Internal Medicine, Wake Forest School of Medicine

[6] Department of Computational Medicine and Bioinformatics, University of Michigan Ann Arbor

* Co-first authors

Co-corresponding authors

Correspondence should be addressed to:

junzli@med.umich.edu or

lsolberg@wakehealth.edu

bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

We performed whole-genome sequencing for eight inbred rat strains commonly used in genetic mapping studies, and they are the founders of the NIH heterogeneous stock (HS) outbred colony. We provide their sequences and variant calls to the rat genomics community. When analyzing the variant calls we identified regions with unusually high heterozygosity. We show that these regions are consistent across the eight inbred strains, including the BN strain, which was the basis of the rat reference genome. These regions show significantly higher read depth than other regions in the genome. The evidence suggests that these regions may contain segmental duplications that are incorrectly overlaid in the reference genome. We provide masks for these suspected regions of mis-assembly as a resource for the community to flag potentially false interpretations of mapping results or functional data.

bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Introduction

The laboratory rat (*Rattus norvegicus*) is an important model organism for studying the genetic and functional basis of physiological traits. Compared to the mouse, the rat shows a greater similarity to humans[1] and has been widely used in physiological, behavioral and pharmacological research. With the arrival of high-throughput genotyping and sequencing technologies, the rat has also been used in genetic studies to map causal loci, or identify genes that affect disease-related traits.

A fundamental resource in such genetic studies is the rat reference genome[2], which provides the coordinate system to organize our rapidly increasing knowledge of rat genes, their regulatory elements and functional profiles of diverse tissues, as well as gene variants and dysregulation in disease models. The reference genome is also the basic map in comparative analyses that focus on the evolutionary relationship among rat strains or between the rat and other organisms.

Gene discovery studies can be roughly classified by the type of mapping populations adopted. Currently, popular genetic systems include naturally occurring outbred populations, laboratory-maintained diversity outbred populations, inbred line-based crosses (e.g., F2-crosses, or advance inbred lines), recombinant inbred lines, and many others. Regardless of the system, a comprehensive knowledge of DNA variation in the mapping population is essential for both the study design and biological interpretation. In this study, we sought to use whole-genome sequencing (WGS) to ascertain DNA variants in eight inbred strains: ACI, BN, BUF, F344, M520, MR, WKY and WN, which are founders of the NIH Heterogeneous Stock (HS) population. The HS rat has been used for genetic studies of metabolic and behavioral traits [3-8]. WGS data for these eight strains have been previously described[9,10], using the SOLiD technology. Here we present WGS results from the Illumina technology, containing genotypes at 16.4 million single-nucleotide variant (SNV) sites. We expect that the sequences of the eight HS founders and fully-ascertained DNA variations can aid the imputation, haplotyping, and fine mapping efforts by the rat genomics community.

When analyzing the SNV data we noted that, while eight founders are inbred, all contain an unusually high amount of heterozygous nucleotide positions. Remarkably, these sites tend to concentrate in hundreds of discrete genomic regions, which collectively span 6-9% of the genome. We show that the heterozygous genotypes tend to recur in multiple, if not all, of the eight strains, and that the suspected regions tend to have higher-than-average read depths. We propose that these regions can be explained by mis-assembly of the rat reference genome, where many of the highly repetitive segments may exist in tandem or dispersed in distant regions, but have been erroneously "folded" in the current coordinate system, causing the reads of high homology that originate in homozygotic regions to falsely converge to produce apparent heterozygous calls. This interpretation is not unexpected when one compares the genome assembly statistics between mouse and rat: the latest release of the mouse reference genome (GRCm38.p6[11]) contains 885 contigs and a median contig length of 32.3 megabases; whereas the rat reference genome (rn6[2]) has 75,687 contigs and an median length of 100.5 kilobases. With this report we release mask files for the suspected regions, so that they can be used as flag questionably results in current genomic studies until the time when a revised, more accurate reference assembly becomes available.

Materials and Methods

Animals, DNA samples, whole-genome sequencing

bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Eight animals, one for each of the eight founders of the HS population, were used in the study. Genomic DNA of the original founder strains was obtained from NIH. DNAwas extracted at the University of Michigan (Ann Arbor, MI) from the liver of seven animals (ACI/N, BUF/N, F344/N, M520/N, MR/N, WKY/N, WN/N), and from the tail of a BN/N animal. All animals were female. The DNeasy Blood and Tissue Kit from Qiagen (Hilden, Germany) was used for DNA extraction. Samples were further QC'd and sequenced at Novogene (Beijing, China) following the standard Illumina protocols. Library preparation produced fragment libraries of <350 bp insert length. Sequencing was done on Illumina HiSeqX-Ten to collect 150 bp paired-end data, aiming for an average depth of 25X per strain.

In this report we abbreviate the strain names by the first two letters: AC, BN, BU, F3, M5, MR, WK and WN.

Sequence alignment and variant calling

We aligned the raw sequence reads to the rat reference genome (rn6) using *BWA* version 0.5.9[12] , removed duplicates using *Picard* v1.76[13] , and performed realignment, recalibration and joint variant calling across eight strains with the UnifiedGenotyper with *GATK* v3.4 [14]. We removed variant sites with fewer than 10 reads in eight samples, and variant site quality score (QUAL) <= 30. We chose not to use the HaplotypeCaller as we have only eight inbred lines, which are not the population-based samples suitable for building haplotypes.

ChrX data showed the same pattern of heterozygosity as the autosomes in all eight animals, thus confirmed that they are female. We excluded the Y chromosome calls in downstream analysis. We did not call indels in this data release.

For comparison purposes we also ran the analysis with (1) two earlier versions of the reference genome, rn4 and rn5; (2) two other aligners. The first is by feeding the BWA aligned files into *Stampy*[15] version 1.0.32. Stampy alignment shows higher sensitivity than BWA, especially when reads include sequence variation[15]. (The use of BWA-alignment as input for Stampy is to increases alignment speed without reducing sensitivity). The second is *Bowtie2* v2.1.0[16]. All the post-alignment processes followed the same *Picard* and *GATK* steps.

Comparison with the previously published variant calls using SOLiD

We compared our variant call set (for BWA alignment and rn6) with that by Hermsen et al.[10], which was based on the SOLiD sequencing data. As that call set was aligned to rn5, we lifted over the variants to the rn6.

Defining regions of unusually high rates of heterozygosity

The final call set contains >16.4M SNV sites. We divided the genome into 1000-site windows, with a median window length of 221,100 bases. For each of the eight samples and in each window we computed the fraction of heterozygous sites in that window (only using the number of non-missing sites in that window as the denominator). Based on the inflection point in the empirical distribution of this per-window heterozygosity fraction (**Figure 2**) we chose a cutoff of 25% to designate a windows as of high-heterozygosity. We concatenated adjacent windows of high-heterozygosity into the same segment, and in a second step, merged adjacent high-het segments if they are separated by a single “low-heterozygosity” window, if that window had more than 0.175 heterozygote rate. After merging, there is no evidence of many very short low-het segments separating high-het segments (**Figure S4**).

bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Data records released

We share data files to cover multiple levels of processing.

1. Raw FASTQ/BAM files will be deposited in to the Rat Genome Database (RGD) [17].

2. Variant calls from the UnifiedGenotyper (using BWA and rn6), as VCF files for the eight strains.

3. Mask files on our GitHub page[18], documenting the regions of high heterozygosity, for each of the eight founder strains, as well as regions that are highly-heterozygous in all of the eight.

Results

Description of the variant calls

Median read depth over the genome ranges 24X - 28X across the eight samples. Joint variant calling revealed 16,405,184 post-filter single-nucleotide variant sites on the autosomes and chromosome X. The number of heterozygous sites per strain varies from 1,560,708 (BN) to 2,114,990 (WKY) (**Supplementary Table 1**). BN represents the reference genome, and had more Ref/Ref than Alt/Alt genotypes. In contrast, the other seven strains had a comparable number of Ref/Ref calls as Alt/Alt calls.

We compared our genotype calls with those from Hermsen *et al*.[10] by calculating between-study concordance rates at sites reported in both, and using genotypes that do not include the missing calls. **Supplementary Table 2** shows that each of eight lines can be correctly matched between the two datasets, confirming the sample identity even when the two studies were based on different animals for a given line. BN had the highest between-study concordance: 0.95. Six other lines has concordance > 0.86. However, MR had the lowest concordance, 0.69. To determine if an animal from another line was mislabeled as MR, we compared our MR data with a larger panel of 42 strains previously published[10] and found that our MR had the highest match with MR and WAG-Rij in that study. This suggests that the MR lines in different laboratories may have diverged to an unusually large degree.

Regions of unexpected high-heterozygosity

The eight founder strains were kept as inbred lines over many generations, thus were expected to show low heterozygosity across the genome[19]. However, when we calculated the fraction of heterozygous genotypes in consecutive 1000-SNV windows, we observed highly varied distribution of this metric along the genome. Not only were there many windows of high heterozygosity (>0.25), they also tended to recur in multiple lines (**Figure 1**). Some of these windows were found in all 8 strains, including BN, the strain of the reference genome (**Supplementary Figure 1**).

We chose a het cutoff value of 25% according to the distribution of per-window heterozygosity (**Figure 2**). After merging neighboring high-het windows if they are separated by a single low-het window with het rate > 0.175, we obtained 304-482 contiguous high-het regions across the eight lines (median 452), covering 176.4- 254.8 Mb, equivalent to 6.3%-9.2% % of the genome (average 8.4%). The distribution of segment lengths is shown in **Supplementary Figure 2**. Of the heterozygous calls in each line, 28-31% fall in the high-het regions (mean 29%).

The high-heterozygous regions found in all eight lines contain 1,756 Ensembl genes. These genes were enriched for G-protein coupled receptors and olfactory receptors (fold enrichment of 3.2 and 2.0 respectively, BenjaminiHochberg P-value < 0.01), which are known to have many paralogous copies in

bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

mammalian genomes[20]. There are 4,963 missense variants, 123 stop-gain variants, and 154 splice donor/acceptor variants in 372 genes in these regions.

Heterozygous calls tend to be recurrent and show higher read depths

While **Supplementary Figure 1** shows that 1000-variant windows of high heterozygosity tend to appear in multiple lines, we also analyzed individual heterozygous genotypes to see how much they tend to recur in multiple lines. We divided the ~16.4M variant sites according to the number of heterozygous genotypes observed in the eight lines, thus defining nine variant categories, from 0 to 8 heterozygotes (**Figure 3**). If the heterozygous genotypes appear independently in the eight lines with a probability of p, the expected chance of seeing a site with two heterozygotes (that is, in two of the eight lines) would be proportional to p^2, and in three lines: p^3. We estimated the upper limit of p by counting the fraction of heterogeneous genotypes over all 8 lines in all sites, knowing that this fraction is already biased upward due to the highly recurrent heterozygous sites. The expected probability of seeing k heterozygous sites, $a(k)p^k$, where a(k) is the sampling coefficient of the binomial distribution, drops much faster than the observed k-het counts (**Figure 3**). For instance, the observed number of sites that are heterozygous in all eight lines is more than five orders of magnitude higher than expectation.

The tendency for an individual site to appear heterozygous in multiple lines is related to higher read depth at these sites (**Figure 4A**). **Figure 4B** shows an example where low-het regions tend to have lower, and more stable, read depth than nearby regions of high heterozygosity. Notably, the read depth is higher in these regions for both homozygous and heterozygous genotypes.

Concordance with results from a different platform

We analyzed the previously released variant calls from Hermsen *et al.* 2015 (after lifting over these calls to the rn6 genome from the original rn5) to see if our high-heterozygosity segments were also found in their calls. The Hermsen dataset revealed 5-18% (median 14%) of the genome in high-het windows. We found that these high-het regions tend to match in the two call sets (an example in **Supplementary Figure 3**), with an overlap of 24%-36% as defined by the length of the intersect of segments called high-heterozygosity in both calls, divided by the union of regions defined as high-heterozygosity in either call.

We have compiled a list of these high-het regions of the genome and provided them as bed files on our Github page: http://github.com/shwetaramdas/rataccessibleregions/.

Discussion

In this study we created WGS-based variant call sets for eight inbred lines which represent the genomic source of the multi-parental HS population. Our data suggest that the current rat reference genome contain hundreds of problematic regions where inbred lines show increased apparent heterogeneity and higher-than-usual read depth. These regions make up ~8.4% of the genome, and likely represent regions of mis-assembly. In other words, a properly unfolded genome may be ~8% longer.

Several lines of evidence led us to the interpretation that regions of high-heterozygosity likely represent an artifact when two or more segments of highly similarity being incorrectly "collapsed", or "folded", into the same segment in the rat reference genome. First, the rate of heterozygous calls is higher than what one would expect for inbred lines; and these heterozygous calls cluster in discrete regions. Second, these high-heterozygosity regions tend to recur, sometimes in all eight lines, indicating that they are unlikely to have arisen from genetic drift or strain-specific selection. Finally, these regions tend to have higher read depth than surrounding regions. The rat reference genome, unlike the human and the

bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

mouse reference genomes, was assembled using a hybrid of shotgun-sequencing and clone-based approaches. Our results suggest that this mixed approach has not been sufficient to resolve all the repetitive regions. As a result, Illumina sequence reads that originated in distinct but highly similar regions – some of these may fit the strict definition of paralogous regions – are incorrectly aligned to the same collapsed region in the reference genome, producing false heterozygous calls.

Guryev *et al*[21] in 2008 studied the rat reference genome (rn4 at the time) and used the read-depth distribution to identify 73 regions of suspected mis-assembly, which make up ~1% of the genome. Only 2 of these 73 regions lift over to the rn6 genome and both were called high-het in our data. However, they didn't analyze the distribution of heterozygous genotypes. Another study performed WGS on the inbred strain SHR/Olalpcv and observed higher-than-expected levels of heterozygosity. The authors suggested that this could have resulted from collapsing of reads from segmental duplications[22], an interpretation similar to ours. Because they analyzed a single strain, the data did not show the consistency of suspected regions across strains.

In an early draft of the mouse genome segmental duplications were inadvertently misassembled[23]. Our estimates of misassembled regions underscore a similar problem with the current rat reference genome. These regions harbor more than 1,500 genes, with 5,000 apparent missense variants that may be an artifact. Genomic studies that fail to flag these regions and variants are at risk of reporting incorrect candidates for downstream studies. We propose that the rat community apply the mask files provided here, until a more refined reference genome becomes available.

bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure legends

Figure 1. Consistent heterozygosity patterns across the eight lines. Shown are the fraction of heterozygous genotypes in non-overlapping 1000-SNV windows in chromosome 1, displayed for the eight inbred lines.

Figure 2. Distribution of the heterozygosity level in 1000-SNV windows for each of the eight lines, showing that 0.25 is a reasonable cutoff for defining high-het windows.

Figure 3. Recurrence pattern of heterozygous genotypes across 8 founders. All variant sites were categorized as having 0, 1, …, or up to 8 heterozygous genotypes. The bar graph displays the histogram of the observed sites, while the dots show the expected number of sites if recurrence is random, estimated under a simple binomial model. The expected values for 0 and 1 founders exceed the limits of the plot.

Figure 4. Higher read depths in high-heterozygosity windows. **A**. Boxplot of per-window average read depth stratified by the number of lines for which a given window is classified as high-het. It shows that windows with high-het in all eight lines tend to show higher average read depth. **B**. An example of a 5Mb region in Chr 1 and line AC, with a high-het region in the middle. Y axis shows the read depth for individual sites, showing that in h=the high-het window, both the het calls (upper panel) and hom calls (lower panel) show higher read depth and often higher variance of read depth.

bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

1 Iannaccone, P. M. & Jacob, H. J. Rats! *Disease models & mechanisms* **2**, 206-210, doi:10.1242/dmm.002733 (2009).

2 Gibbs, R. A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493-521, doi:10.1038/nature02426 (2004).

3 Solberg Woods, L. C., Holl, K., Tschannen, M. & Valdar, W. Fine-mapping a locus for glucose tolerance using heterogeneous stock rats. *Physiological genomics* **41**, 102-108, doi:10.1152/physiolgenomics.00178.2009 (2010).

4 Solberg Woods, L. C. *et al.* Fine-mapping diabetes-related traits, including insulin resistance, in heterogeneous stock rats. *Physiological genomics* **44**, 1013-1026, doi:10.1152/physiolgenomics.00040.2012 (2012).

5 Keele, G. R. *et al.* Genetic Fine-Mapping and Identification of Candidate Genes and Variants for Adiposity Traits in Outbred Rats. *Obesity* **26**, 213-222, doi:10.1002/oby.22075 (2018).

6 Woods, L. C. & Mott, R. Heterogeneous Stock Populations for Analysis of Complex Traits. *Methods in molecular biology* **1488**, 31-44, doi:10.1007/978-1-4939-6427-7_2 (2017).

7 Holl, K. *et al.* Heterogeneous stock rats: a model to study the genetics of despair-like behavior in adolescence. *Genes, brain, and behavior* **17**, 139-148, doi:10.1111/gbb.12410 (2018).

8 Solberg Woods, L. C. *et al.* Heterogeneous stock rats: a new model to study the genetics of renal phenotypes. *American journal of physiology. Renal physiology* **298**, F1484-1491, doi:10.1152/ajprenal.00002.2010 (2010).

9 Baud, A. *et al.* Genomes and phenomes of a population of outbred rats and its progenitors. *Scientific data* **1**, 140011, doi:10.1038/sdata.2014.11 (2014).

10 Hermsen, R. *et al.* Genomic landscape of rat strain and substrain variation. *BMC genomics* **16**, 357, doi:10.1186/s12864-015-1594-1 (2015).

11 Church, D. M. *et al.* Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS biology* **7**, e1000112, doi:10.1371/journal.pbio.1000112 (2009).

12 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

13 *Picard*, <http://broadinstitute.github.io/picard/> (

14 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

15 Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research* **21**, 936-939, doi:10.1101/gr.111120.110 (2011).

16 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).

17 Shimoyama, M. *et al.* The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res* **43**, D743-750, doi:10.1093/nar/gku1026 (2015).

18 Ramdas, S. <www.github.com/shwetaramdas/rataccessibleregions/> (2018).

19 Brenner, S., Miller, J. H., Broughton, W. & ScienceDirect (Online service). *Encyclopedia of genetics*. (Academic Press, 2002).

20 Hughes, G. M. *et al.* The birth and death of olfactory receptor gene families in mammalian niche adaptation. *Molecular biology and evolution*, doi:10.1093/molbev/msy028 (2018).

21 Guryev, V. *et al.* Distribution and functional impact of DNA copy number variation in the rat. *Nature genetics* **40**, 538-545, doi:10.1038/ng.141 (2008).

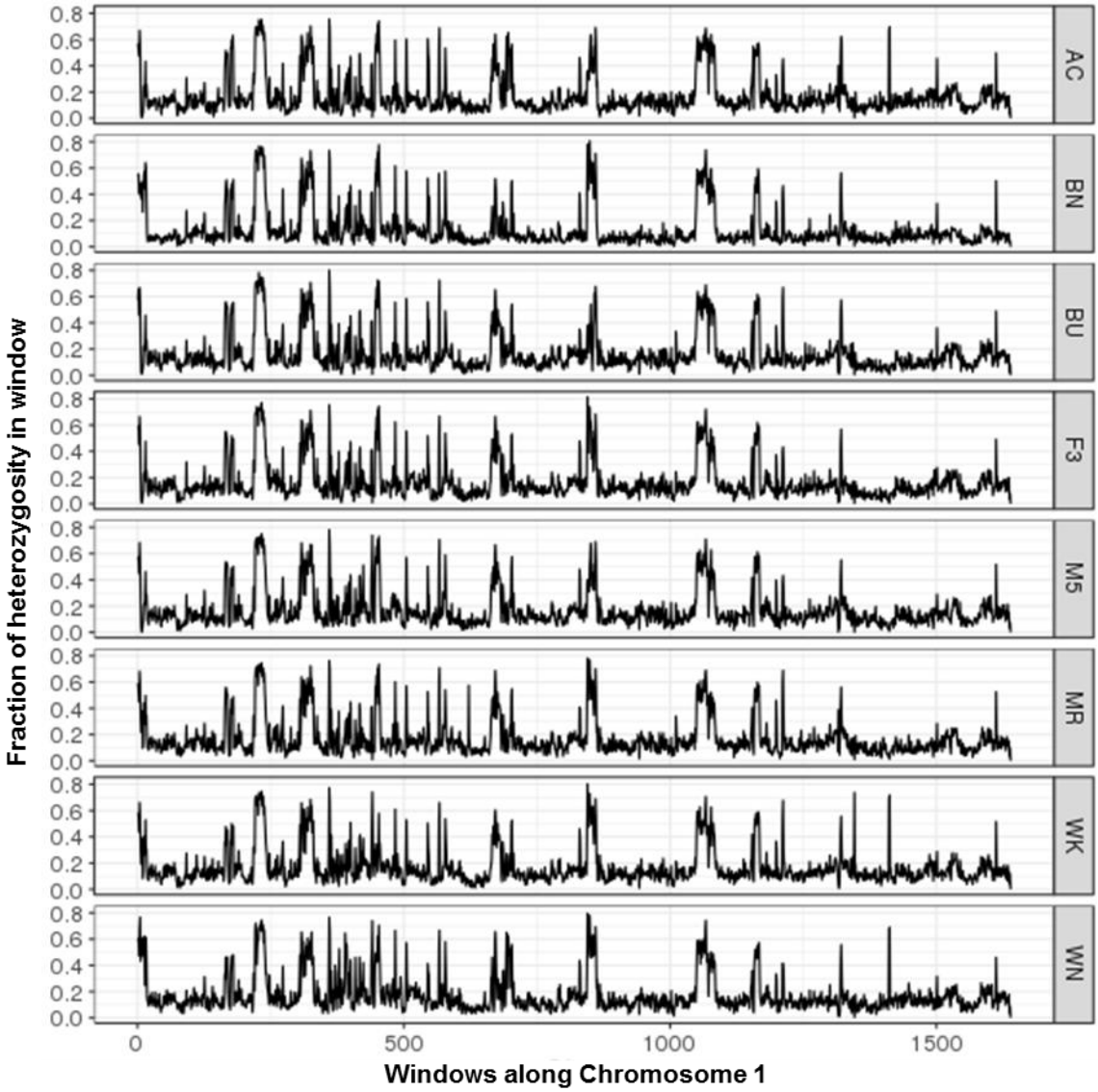
bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

22 Atanur, S. S. *et al.* The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. *Genome research* **20**, 791-803, doi:10.1101/gr.103499.109 (2010).

23 Bailey, J. A., Church, D. M., Ventura, M., Rocchi, M. & Eichler, E. E. Analysis of segmental duplications and genome assembly in the mouse. *Genome research* **14**, 789-801, doi:10.1101/gr.2238404 (2004).

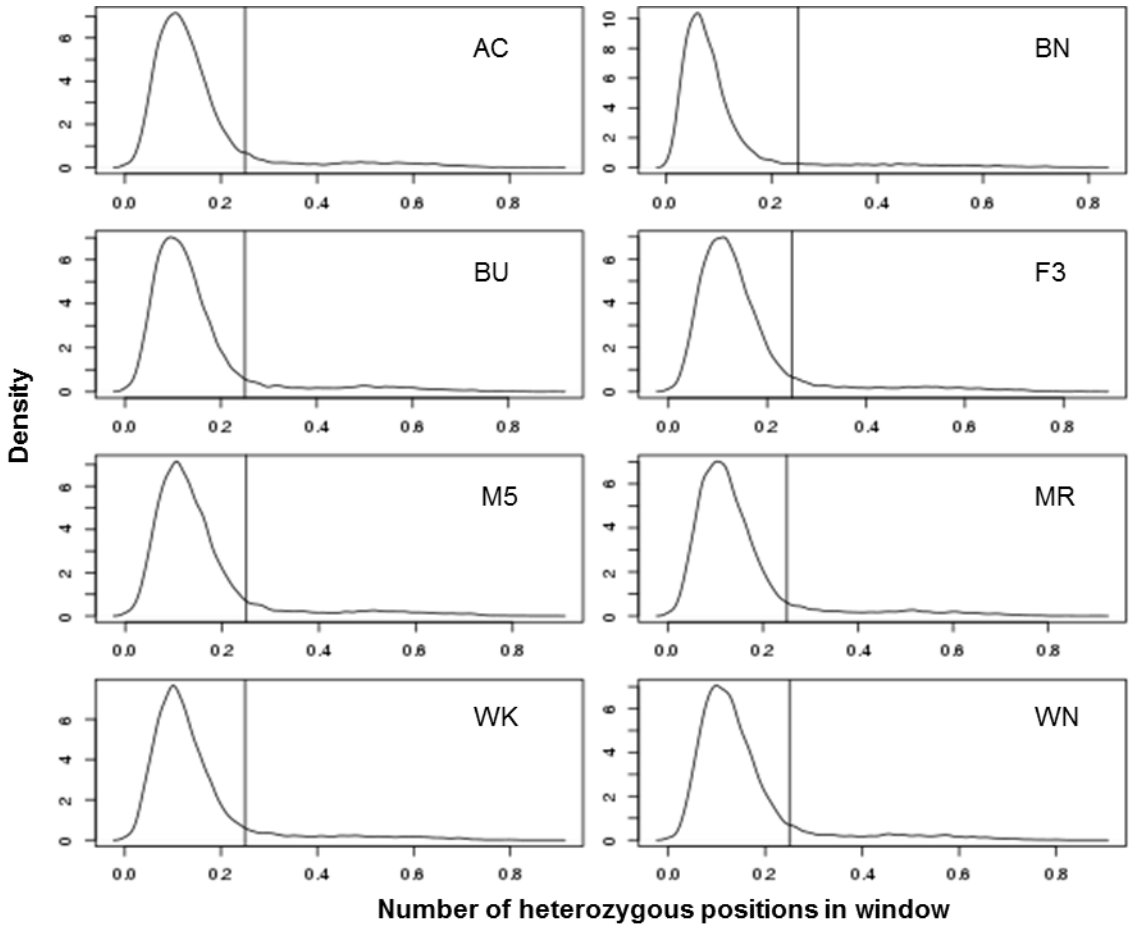
bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 1



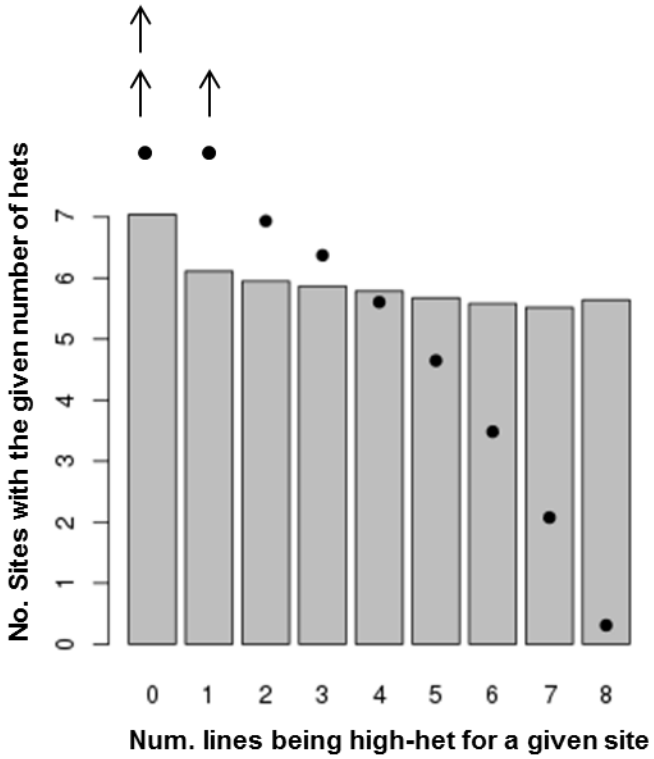
bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 2



bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

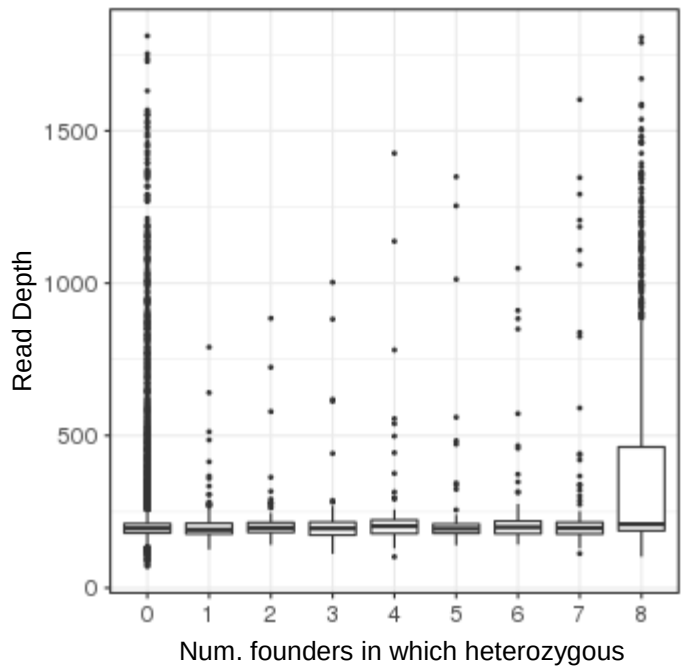
Figure 3



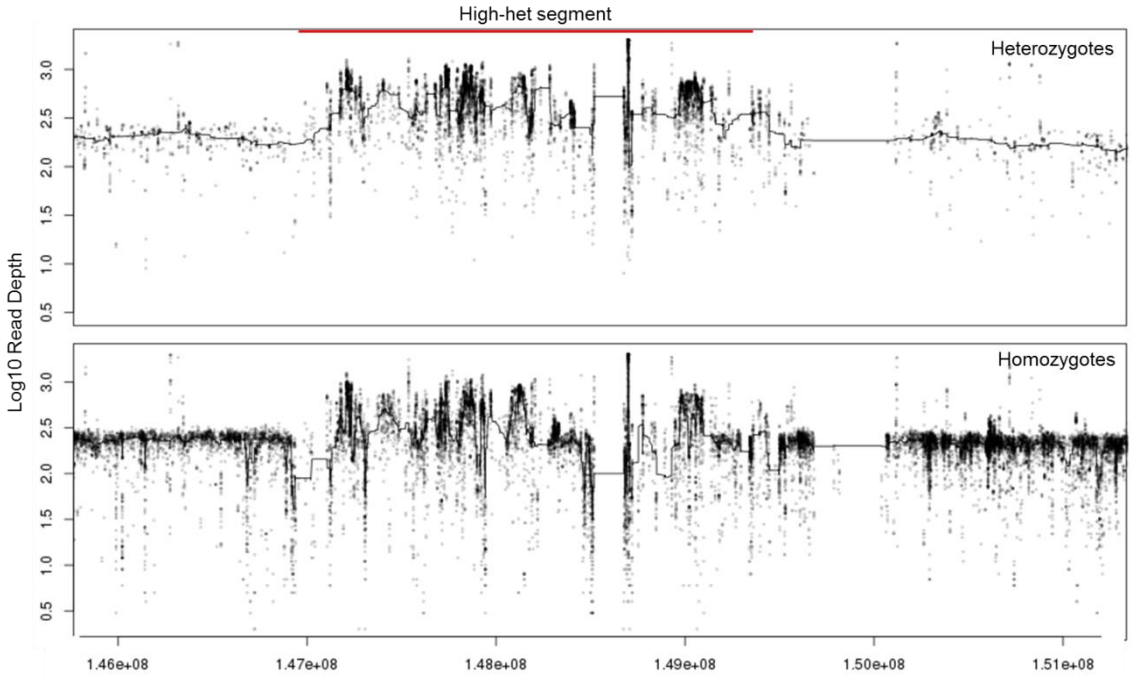
bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 4

A



B



bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Table 1. Genotype counts and heterozygote frequencies in the eight lines. BN represents the reference genome and shows an outlier pattern for most metrics.

Strain	0/0	0/1	1/1	./.	Sites with > 2 alleles	Het frequencies
AC	6,249,928	2,096,993	5,421,143	2,385,991	251,129	0.15
BN	12,324,354	1,560,708	803,800	1,623,744	92,578	0.106
BU	6,688,095	2,058,214	5,192,072	2,218,374	248,429	0.145
F3	6,587,382	2,101,058	5,282,082	2,181,114	253,548	0.148
M5	6,628,579	2,160,032	5,329,841	2,031,084	255,648	0.15
MR	6,736,380	2,111,216	5,243,452	2,063,655	250,481	0.147
WK	6,599,091	1,976,801	5,130,247	2,456,145	242,900	0.142
WN	5,993,693	2,114,990	5,734,351	2,304,637	257,513	0.15

bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Table 2. Concordance rate between our variant calls and the previous variant calls from Hermsen et al, 2015. The 8-by-8 table contains concordance values for all pair of samples Here, we define concordance as the number of variant sites with the same genotype calls in both call sets divided by the number of variant sites with non-missing calls in both call sets.

		Hermsen et al. 2015 calls							
		AC	**BN**	**BU**	**F3**	**M5**	**MR**	**WK**	**WN**
Our variant calls	**AC**	0.874	0.478	0.577	0.588	0.584	0.599	0.514	0.547
	BN	0.52	0.952	0.548	0.541	0.55	0.559	0.497	0.551
	BU	0.586	0.511	0.871	0.628	0.631	0.609	0.52	0.599
	F3	0.594	0.501	0.625	0.873	0.674	0.593	0.52	0.594
	M5	0.591	0.511	0.63	0.676	0.873	0.601	0.514	0.592
	MR	0.585	0.521	0.605	0.599	0.598	0.685	0.526	0.603
	WK	0.518	0.453	0.516	0.518	0.51	0.536	0.859	0.508
	WN	0.55	0.511	0.594	0.591	0.588	0.599	0.508	0.88

bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Table 3. Consistency of high-heterozygosity regions among three alignment methods. Shown are the total length of high-het regions for the 8 lines and with the use of the three aligners (*BWA*, *Bowtie2* and *BWA-Stampy*).

As the background level of heterozygosity differ among the three call sets, we applied different het fraction thresholds to define high-het windows: 0.25 for *BWA*, 0.175 for *Bowtie2*, and 0.2 for *Stampy*. The following table lists concordances between heterozygous segments called by different aligners.

	BWA	*Bowtie*	*Stampy*	*BWA:Bowtie*	*BWA:Stampy*	*Bowtie:Stampy*
AC	238,523,232	260,711,962	302,820,837	0.57	0.54	0.51
BN	171,689,865	189,496,386	213,984,578	0.61	0.6	0.58
BU	223,339,767	247,060,856	289,737,173	0.55	0.55	0.49
F3	236,379,041	252,011,886	293,848,725	0.56	0.54	0.51
M5	243,992,401	243,643,168	294,212,048	0.58	0.54	0.5
MR	225,952,435	263,629,202	276,523,184	0.55	0.55	0.5
WK	242,849,594	239,016,453	288,897,012	0.57	0.54	0.52
WN	228,435,487	249,395,735	298,824,484	0.53	0.52	0.47

bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Table 4. Consistence across the three versions of the reference genome. The tables show the cross-tabulation of the number of genotype calls (in unit of 1,000) between rn4 and rn6 (upper table), and between rn5 versus rn6 (lower). The overall concordance is 0.98 between rn5 and rn6, and 0.97 between rn4 and rn6.

		rn6			
		0/0	**0/1**	**1/1**	**./.**
rn4	**0/0**	37,196.7	594	140.1	152
	0/1	48	839	133.3	10.6
	1/1	5.9	303.2	24,958.4	155.1
	./.	537.8	498.6	1,112.4	1,920.8

	rn6				
		0/0	**0/1**	**1/1**	**./.**
rn5	**0/0**	13,908.1	374.7	59.8	240.6
	0/1	48.2	1,410.5	74.6	19.7
	1/1	2.5	170.9	9,606.8	351.7
	./.	72.3	65.6	103.3	364

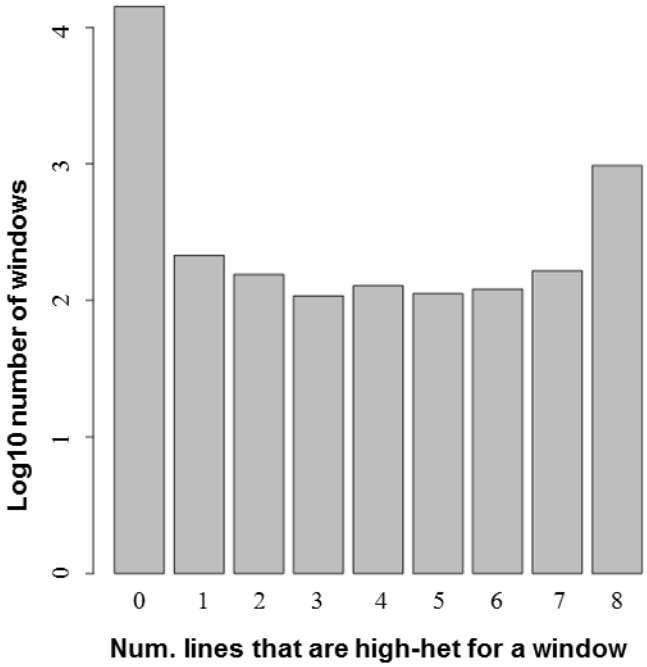
bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Table 5. High levels of genotype concordance among variant calls obtained from different aligners (*BWA*, *Bowtie*, *BWA-Stampy*) followed by calling with UnifiedGenotyper. As before, we define concordance as the number of variant sites with the same genotype calls in both call sets divided by the number of variant sites with non-missing calls in both call sets.

	BWA	*Bowtie*	*Stampy*
BWA	1	0.972	0.972
Bowtie		1	0.974
Stampy			1

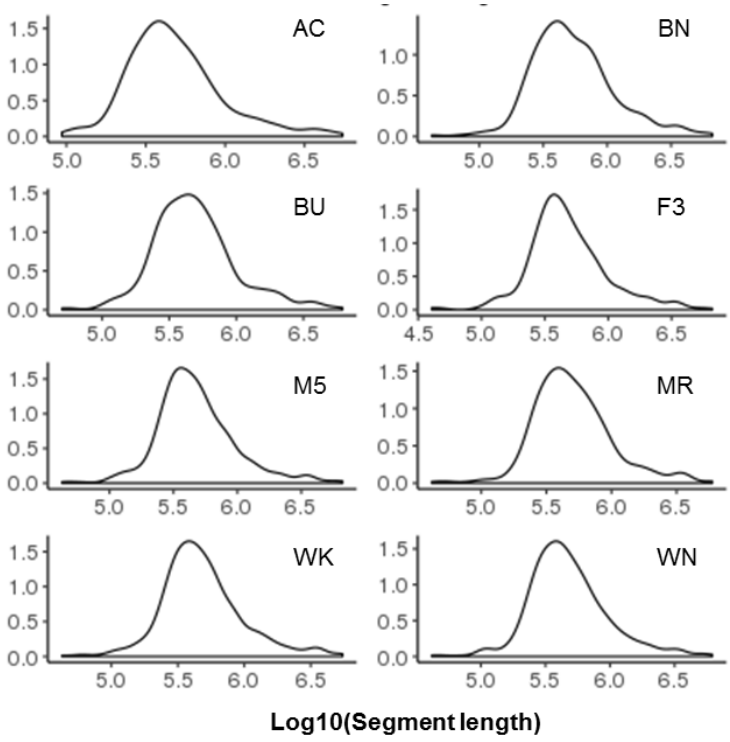
bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Figure 1. Histogram of the number of windows that are high-het (heterozygosity > 0.25) in 0, 1, …, 8 founders.



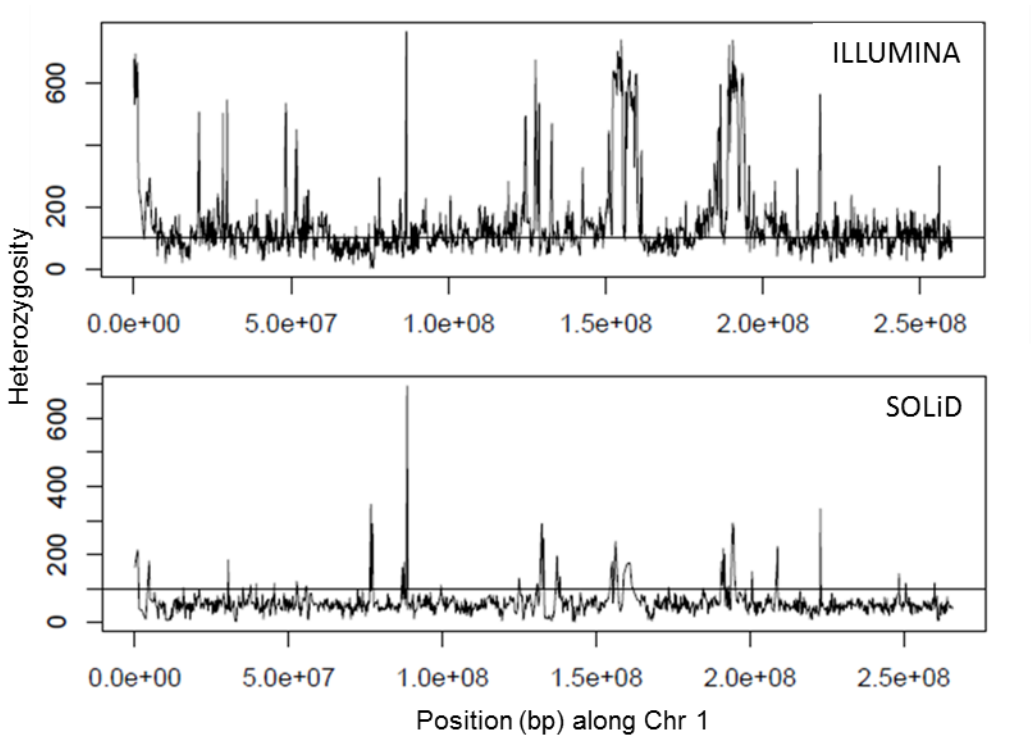
bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

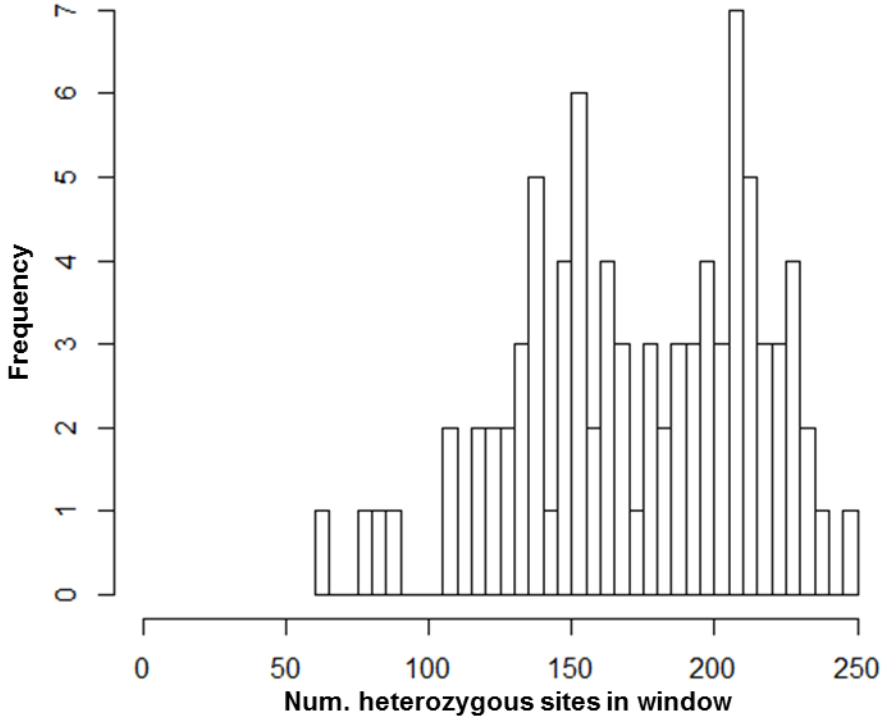
Supplementary Figure 2. Distribution of high-het segment lengths in the 8 lines.



bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Figure 3. Similar regional patterns of high-het segments in Chr 2 between this study and the previously reported SOLiD calls.



bioRxiv preprint doi: https://doi.org/10.1101/332932; this version posted May 29, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Figure 4. Distribution of heterozygosity rates in windows that are initially classified as low-het, but are immediately flanked by high-het windows on both sides. The bimodal distribution suggests those with >175 hets may be merged with adjacent high-het segments.