Whole genome sequencing of nearly isogeneic WMI and WLI inbred rats identifies genes potentially involved in depression

Tristan de Jong, Panjun Kim, Victor Guryev, Megan Mulligan, Robert W Williams, Eva E Redei, Hao Chen

Abstract

Background: The WMI and WLI inbred rat substrains were generated from the stress-prone, and not yet fully inbred, Wistar Kyoto (WKY) strain using bi-directional selection for immobility in the forced swim test followed by over 38 generations of inbreeding. Despite the low level of genetic diversity among WKY progenitors, the WMI substrain is more vulnerable to stress relative to its WLI control substrain. Here we quantify numbers and classes of sequence variants distinguishing these substrains and test the hypothesis that they are nearly isogenic.

Results: The WLI and WMI genomic DNA were sequenced using Illumina xTen, IonTorrent and 10X Chromium technologies to obtain a combined coverage of over 100X. We identified 4,296 high quality homozygous SNPs and indels that differ between the WMI and WLI substrains. Gene ontology analysis of these variants showed an enrichment for neurogenesis related pathways. In addition, high impact variations were detected in genes previously implicated in depression (e.g. *Gnat2*), depression-like behavior (e.g. *Prlr*, *Nlrp1a*), other psychiatric disease (e.g. *Pou6f2*, *Kdm5a*, *Reep3*, *Wdfy3*) or stress response (e.g. *Pigr*).

Conclusions: The high coverage sequencing data confirms the near isogenic nature of the two substrains, which combined with the variants detected can lead to the identification of genetic factors underlying greater susceptibility for depression, stress reactivity, and addiction.

Keywords

WMI, WLI, WKY, substrains, Whole-Genome-Sequencing, depression, rats, WKY isogenic strains, genetic variants, stress-reactivity

Background

Major depressive disorder (MDD) is a common, debilitating disease that is the leading cause of "years lived with disability" worldwide [1]. Genetic factors play important roles in the etiology of MDD. Heritability of MDD is estimated to be between 28 and 44% [2,3], although recent estimates are over 50% [4]. Genomic variants contributing to depression have been difficult to identify, but large genome-wide association studies (GWAS) [5] are starting to identify candidates, including variants near *SIRT1*, *LHPP* [6], *OLFM4*, *MEF2C*, and *TMEM161B* [7]. Meta-analysis of GWAS based on self-reported depression also identified a larger number of independent and significant loci [8,9], although relying on self diagnosis may have reduced the reproducibility of the findings [6]. Even when the MDD diagnosis is not based on self-report, the current diagnostic methods are still comparatively subjective and cannot truly characterize subgroups of this complex disease, which are likely affected by differences in genetics. Thus, identification of sequence variants associated with the disease, and the genetic etiology of MDD, remains largely unsolved.

Compared to the high levels of genetic variations among humans (6 million between any two individuals), well defined animal models can tightly constrain both genomic and environmental variables. Many genetic mapping strategies have been developed for model organisms. For example, the reduced complexity cross (RCC) uses offspring from two genetically similar parents that have divergent phenotypes. The number of segregating variants in an RCC is orders of magnitude smaller than in conventional cross. It therefore often permits immediate identification of causal variants [10].

In this report we analyze a genetic model of depression and its nearly isogenic control strain, both bred from Wistar Kyoto (WKY) rats. The WKY strain had been developed as the normotensive control for the spontaneously hypertensive rat strain and was distributed to vendors and universities between the 4th and 11th generation of inbreeding [11]. At this early stage, the stock varied widely in behaviors [12]. The Redei lab obtained WKY rats from Harlan Laboratories (Madison, WI), where they had been bred for 65 generations. However, it is not known whether the sublines (breeding pairs) Harlan obtained at the beginning of the breeding were maintained as sublines or interbred. The WKY strain has become a well-established model of adult and adolescent depression and comorbid anxiety [13–15]. Its behavior mirrors several symptoms of human MDD and anxiety, including anhedonia, disturbed sleep, a reduced appetite and reduced energy, and the attenuation of depression-like behaviors after treatment with antidepressants [16–19].

A large variability in behavioral and psychological measurements were noted within the WKY strain [20,21]. The variability of behavior in the forced swim test (FST)—one of the most widely utilized tests for depressive behavior in rodents—motivated the bi-directional selection of the animals based on their level of immobility in the FST [22]. Males and females with the least mobility and lowest climbing scores in the FST were mated, producing the WKY *More Immobile* (WMI) line. Males and females with the highest mobility and highest climbing scores were mated, producing the WKY *Less Immobile* (WLI) line. Those animals showing the most extreme FST behavior within each line were selected for subsequent breeding, specifically avoiding sibling mating until the fifth G generation, when filial F matings were initiated.

Throughout the generations, the WMIs consistently have shown significantly greater immobility behavior in the FST than the WLIs [23]. The sex differences observed in the developmental pattern of MDD and its comorbidity with anxiety parallel differences observed in humans [24]. Maternal characteristics of the WMI after birth show similarities to that of women with postpartum depression [25]. Antidepressant treatments, specifically the tricyclic desipramine and the MAO inhibitor phenelzine, but not fluoxetine, alleviate depression-like

behavior of WMIs [22], and enriched environment in adulthood does the same [26]. Resting state functional connectivity differences between WMIs and WLIs, measured by fMRI, are similar to those found in depressed patients [27,28]. Behavioral and hormonal responsiveness to acute and chronic stress also differ between the strains [26,29]. In humans, posttraumatic stress disorder (PTSD) and alcohol use disorder have high comorbidity with major depression. As hypothesized, the stress-reactive WMI strain showed increased fear memory in a model of PTSD, the stress-enhanced fear learning behavior compared to the isogenic WLI strain [29]. Additionally, WMIs consume more alcohol than WLIs when tested using an operant licking procedure [29]. In human studies depression has been noted as a risk factor for dementia in females. Similarly, middle-aged WMI females show cognitive decline compared to middle-aged WLI females [30]. Together, these data establish WMI as a suitable model to study human depression.

The WMI and WLI strains also differ in their brain and blood gene expression profiles [23]. A panel of blood transcriptomic markers, developed using the WMI strain, can diagnose major depression in humans. These blood transcriptomic markers are able to distinguish adolescent and adult subjects with major depression from those with no disorder with a high level of reliability [31–33]. Additionally, the expression of these markers correlated with depression symptoms in pregnant women [34]. These data provide tantalizing evidence that genetically determined gene expression differences between the WMI and WLI substrains can potentially lead to the discoveries of molecular mechanisms of depression in humans.

Full genome sequencing provides an abundance of genetic information (single nucleotide polymorphisms, inserts and deletions, and large structural variants) and can allow for comparative genomics between the rat model and humans. Comparing the genome of WMI and WLI to each other as well as the reference genome could provide insights to the underpinnings of their distinctive behavioral phenotypes. Because the WMI and WLI strains were both derived from WKY founders, we hypothesized that a small number of genetic variants between these strains contribute to behavioral and physiological differences in

depression-associated traits between WMI and WLI. Here we describe the whole genome sequencing of these two strains using data obtained from three different platforms (Illumina xTen, Ion Proton, and 10X Chromium linked-read) and the identification of genetic variants between these strains.

Results

To discover SNPs associated with the depression phenotype in WLI and WMI rats, whole genome sequencing data was obtained using three different platforms: Ion Torrent Proton, 10X Chromium and Illumina xTen from male WLI and WMI rats. Each technique covered an average depth of 41, 27 and 43 for both strains, respectively (Figure 1). X-chromosomal coverage was expected to be half of autosomal coverage but was found to be much higher on IonProton and Illumina X-ten sequencing results (Figure 1).

Sequencing data were mapped to the rat reference genome rn6 using bwa (Illumina and IonProton data) or LongRanger (10X Chromium data). The resulting bam files were used as the input to DeepVariant to report genomic variants (i.e. SNPs and small indels) for each sample. GLNexus was then used to conduct a joint analysis of variants across all six samples. Over 12 million unique variants were identified before filtering. The analysis workflow was designed to take full advantage of the data provided from three sequencing technologies. We are interested in variants that have a Phred quality score above 30, have a clear call for either reference, homozygous or heterozygous, have no matching calls between WLI and WMI and must not have both a high quality reference and alternative allele called on different sequencing platforms within the same strain (Figure 2).

A large portion of the variants had a Phred quality score below 10 and were excluded from subsequent analysis. In total, 99,465, 25,937, and 6,454 homozygous variants had a Phred quality score greater than 10, 20, and 30 in at least one sample, respectively. For heterozygous calls the number of variants were ~3 million, ~1 million and ~200 thousand, for

quality scores of 10, 20 and 30 respectively. The number of high-quality calls for homozygous variants varied per sequencing technology (Figure 3).

The majority of high confidence heterozygous calls came from a single technique, Ion proton (Figure 3). Closer inspection revealed that the majority (>90%) of these calls was detected on homopolymeric nucleotide sequences. In addition, approximately 95% of these were deletions rather than SNPs, further confirming that these calls are due to errors in base calling homopolymeric sequences. To filter out this common sequencing error, all deletions on homopolymeric regions which were not supported by at least one other sequencing technique were removed (Supplementary figure 1).

As a final result, 2,232 and 2,064 homozygous high confidence variants were discovered on WLI and WMI respectively (Table 1). The majority were insertions (45.3%) followed by SNPs (36.9%), and finally deletions (17.8%) (Table 1). Of these SNPs, approximately 57% were transitions, meaning a purine nucleotide was mutated to another purine or a pyrimidine nucleotide to another pyrimidine. The other 43% were transversion SNPs, in which a purine was replaced by a pyrimidine or vice versa (Table 1). A total of 655 and 894 heterozygous variants were identified for WLI and WMI. It should be noted that the heterozygous variants contained higher coverage than average as compared to homozygous variants (supplemental Figure 2). This implies a large portion of these could be homozygous SNPs aligned to collapsed regions on the reference genome.

Cross comparison of the genomic positions of variants discovered on WLI and WMI with variants discovered in a panel of 44 inbred rat strains (our unpublished data) allowed us to gain an indication into whether the variants were likely de novo. Out of the 2232 variants on WLI and the 2,064 variants on WMI, 1215 and 856 were unique among all strains, respectively.

Table 1. Overview of the number of variants, insertions and deletions in the final selection per strain.

	WLI	WMI
Transition SNP	478	428
Transversion SNP	325	356
Insertions	1090	855
Deletions	339	425

In addition, 79 and 119 homozygous variants were identified for WLI and WMI, respectively, with a Phred quality score of at least 10 in all three sequencing technologies. Though verified across technologies, quality scores cannot be simply summed. For certitude these were not included in the final selection.

We used SnpEff [35] to identify the impact, location (table 2), and the nearest gene in proximity of these variants. About half of the variants (52%) are located within intergenic regions, whilst some (a total of 62) variants fall within exons, 2,432 are within introns, and 450 are located within 5 KB upstream of a gene (Supplementary table 1).

Table 2. Position of selected variants in regions of interest.

Type (alphabetical order)	WLI	WMI
3_prime_UTR_variant	12	11
5_prime_UTR_variant	3	4
downstream_gene_variant	177	117
frameshift_variant	6	5

intergenic_region	1,440	1,334
intragenic_variant	1	5
intron_variant	874	810
missense_variant	9	3
non_coding_transcript_exon_variant	5	2
splice_acceptor_variant	0	0
splice_donor_variant	1	4
splice_region_variant	7	7
stop_lost	1	0
synonymous_variant	5	0
upstream_gene_variant	177	128
Total:	2718	2430

In total, 1491 unique genes were in closest proximity to the final selection of homozygous variants across both strains. Of these, 744 genes were found in WMI and 866 in WLI (119 were found in both strains).

In total, 9 WMI variants and 11 WLI variants were estimated to have a large impact on the final protein product. These included changes to splice sites, missense mutations, loss of stop codons or frameshifts (Table 3). These genes included *Asxl1*, *Zfp292*, *Wrap73*, *Col5a3*, *Abcc5*, *Fscn1*, *Wdfy3*, *Pou6f2*, *Svil*, *Prlr*, *Gnat2*, *Slc30a7*, *Kdm5a*, *Slco1a2*, *Nlrp1a*, *Crlf3*, *Tpcn1*, *Pigr*, *Pou6f2* and *Reep3*. Of these, five variants are likely de novo, whilst 15 are also found on other strains. (Table 3). Among these genes, *GNAT*2 has a variant in

human (rs6537837) that was reported to be associated with unipolar depression with a genome wide significance of p=1e-6 [36], while *Prlr* and *Nlrp1a* were implicated in depression-like behavior in animal models [37,38]. Further, *Pou6f2*, *Kdm5a*, *Reep3*, *Wdfy3* have been implicated in psychiatric diseases such as autism [39–41] and *Pigr* was found to be involved in stress response [42].

Table 3. Overview of variants of high and moderate impact, likely de novo status, their impact and the gene affected.

Strain	Chr	Position	REF	ALT	De novo	Gene name	Ensembl ID	Modification
WMI	3	148895880	T	TA	T	*Asxl1*	ENSRNOG00000061603	splice donor variant & splice region variant & intron variant
WMI	5	50287827	C	A	F	*Zfp292*	ENSRNOG00000031031	missense variant
WMI	5	171455130	CG	C	F	*Wrap73*	ENSRNOG00000014805	splice donor variant & splice region variant & intron variant
WMI	8	21788512	G	A	F	*Col5a3*	ENSRNOG00000020525	missense variant & splice region variant
WMI	11	84399496	CG	C	F	*Abcc5*	ENSRNOG00000029178	frameshift variant
WMI	12	13660098	CG	C	F	*Fscn1*	ENSRNOG00000056585	frameshift variant
WMI	14	9266419	G	GA	F	*Wdfy3*	ENSRNOG00000061121	frameshift variant
WMI	17	49440318	T	G	F	*Pou6f2*	ENSRNOG00000013237	splice donor variant & intron variant
WMI	17	55289842	C	T	F	*Svil*	ENSRNOG00000018110	missense variant
WLI	2	60302395	C	CCT	T	*Prlr*	ENSRNOG00000057557	frameshift variant & splice region variant
WLI	2	210884044	G	A	T	*Gnat2*	ENSRNOG00000019296	missense variant
WLI	2	218889177	T	C	F	*Slc30a7*	ENSRNOG00000013912	stop lost
WLI	4	152938803	C	T	T	*Kdm5a*	ENSRNOG00000010591	missense variant
WLI	4	176505968	C	T	F	*Slco1a2*	ENSRNOG00000031249	splice donor variant & intron variant
WLI	10	57738003	T	A	F	*Nlrp1a*	ENSRNOG00000023143	missense variant
WLI	10	67392591	C	CA	F	*Crlf3*	ENSRNOG00000050657	frameshift variant & splice region variant
WLI	12	41544356	T	C	F	*Tpcn1*	ENSRNOG00000059344	missense variant
WLI	13	47589399	A	G	F	*Pigr*	ENSRNOG00000004405	missense variant
WLI	17	49440316	A	AG	F	*Pou6f2*	ENSRNOG00000013237	frameshift variant
WLI	20	22913769	G	A	T	*Reep3*	ENSRNOG00000000645	missense variant

We also leveraged Gene Ontology-term (GO-term) and KEGG-term enrichment analysis using G-profiler (https://biit.cs.ut.ee/gprofiler/gost) to explore the biological functions of genes in close proximity to sequence variants. We found an over-representation of several neurogenesis, behavioral and locomotion related pathways. Over-represented terms for WLI

included locomotory behavior, behavior, nervous system development, neuron projection and neurogenesis (supplementary table 2a). Over-represented terms for WMI included Par-3-KIF3A-PKC-zeta complex, actin-mediated cell contraction, neuronal related and cellular stress related pathways (supplementary table 2b).

Of the genes found in close proximity to high impact variants (based on SNPeff annotations) in WLI, 30 were annotated with the GO-term neuron to neuron synapse (GO:0098984). We further examined these genes using RatsPub [43], an online tool that allows us to conduct automated searches for genes associated with depression, addiction, stress, or other psychological afflictions from PubMed. We found 23 genes that were associated with psychiatric disease in previous research. These genes included; *Syt1*, *Stxbp5*, *Sorcs2*, *Rs1*, *Ptprd*, *Prkcz*, *Pdlim5*, *Lyn*, *Itga8*, *Igsf11*, *Grm3*, *Erbb4*, *Epha7*, *Epha4*, *Dlgap1*, *Dgki*, *Cdkl5*, *Cacna1c*, *Atp2b2*, *Ank2*, *Als2*, *Add3*, *Adcy8* (Supplementary table 3).

Lastly, we validated our genome sequencing results by selecting 224 SNPs, half unique to each strain, and using multiplex PCR to amplify each target region (150 bp flanking each variant) in genomic DNA collected from 8 rats, including four WMI and four WLI with equal number of males and females. We then constructed sequencing libraries using these PCR products and sequenced them on an Illumina instrument. We were able to obtain PCR products from 89 and 87 primers sets targeting WLI and WMI specific variants, respectively. Among them, 75 WLI and 76 WMI targets met the following two criteria: 1. homozygous alternative in at least three rats of the target strain; 2. homozygous alternative in none of the rats of the opposite strain. Therefore, the positive rate of our stringent empirical validation using 8 rats was 85.8%.

Discussion

The goal of this research is to give us genetic markers for WLI and WMI in context of other strains in reduced complexity crosses and to give us candidate variants for immediate scrutiny

of linkage to depression. We used three leading next-generation sequencing technologies to obtain a combined coverage of approximately 100X for each genome of two closely related inbred rat strains, the WMI and WLI. We identified 4,296 homozygous variants with high fidelity that are located in close proximity to 1,491 unique genes that differ between these two strains. The SNPs and indels identified in this dataset offer new opportunities for the identification of genes related to the phenotypic differences between the WLI and WMI strains.

Each of the three sequencing methods we used has its own merits and flaws. For example, compared to the widely used Illumina platform, the Ion Torrent platform provides high quality data at a lower cost. However, it suffers at homopolymer regions. The 10x Chromium linked reads technology attaches barcodes to high molecular weight DNA before library preparation and can detect large structural variants. But obtaining good quality HMW DNA is technically challenging and is associated with increased cost. Further, when utilizing sequencing data from a single technique, technical biases are likely to make their way into the final result. By removing variants called differently by sequencing platforms, the technical bias is mitigated across the final selection of variants.

We used DeepVariant to identify SNP and small indels across all sequencing techniques [44]. Deepvariant has been shown to outperform GATK in different tests, especially in calling indels [45]. It also handles data from diverse sequencing platforms without additional calibration. We also used LongRanger to map the 10X linked reads sequencing data to the reference genome because it incorporates the molecular barcodes into the mapping algorithm. Following DeepVariant analysis, we used GLNexus to conduct a joint analysis to obtain a raw list of genomic variants. Joint analysis empowers variant discovery by leveraging population-wide information from a cohort of multiple samples, allowing us to detect variants with great sensitivity and genotype samples as accurately as possible [46].

With the combination of different sequencing methods, a higher certainty of variant calling between WLI and WMI has been made possible, though throughout this experiment strict filtering was performed. We first removed any variants detected with any certainty in both WLI and WMI, because we are interested in the differences between these two strains, rather than the common differences to the reference genome. One caveat of this approach is that variants incorrectly called by DeepVariant (e.g. due to low coverage in a single method) can lead to the exclusion of potentially interesting targets. Similarly, a strict quality score cutoff of 30 was used whilst un-opposed by any other sample with a Phred quality score of 10 or higher. Based on the abundance of data we have collected, including thousands of identified variants, we decided that 1 in 1000 variants was a sensible cutoff to avoid hundreds of false positives in the final set of variants.

The current reference genome (rn6) consists of 75,697 contigs and 1,395 scaffolds with N50 lengths of 100.5 KB and 14.99 Mb respectively. These sequences combine into a golden path of approximately 2.8 billion bases. Due to the fragmented nature of the reference genome, the identification of structural variants has proven to be difficult. One example of this is that it is often not possible to establish whether sequence variation is strain specific or related to a problem with the reference genome. In addition to the 4,296 high quality homozygous variants discovered in this research, an additional 15,268 variants were discovered in either WLI or WMI with no significant coverage or significant phred score on the opposing strain (called as ./.). Without a high quality read on both strains we cannot verify newly discovered variants. These low quality calls could be caused by heterozygosity, low coverage or overlapping variants. Initial studies using a new rat reference genome (mRatBN7.1, yet to be annotated) has shown that low quality variant calls become much more sporadic with the updated reference genome (de Jong et al. Unpublished results). For this reason, we have opted to exclude these high quality variant calls without a quality call on the opposing strain.

In this investigation 655 and 894 heterozygous variants were discovered on WLI and WMI respectively. Despite both strains being fully inbred, there is a chance that de-novo mutations could propagate as heterozygous variants within each substrain. A look at the coverage of these positions reveals an average two-fold coverage, implying these variants are homozygous on collapsed regions on the reference genome, or duplicated and mutated within either strain. With an updated reference genome these regions could be resolved and can contribute in a meaningful way to the identification of variants that contribute to phenotypic differences between WLI and WMI substrains.

Ongoing research has identified over 40,000 variants in multiple BN/NHsdMcwi samples (the strain used to generate data for rn6). This means some caution is required when identifying variants for WLI or WMI strains based on comparison to the current rat reference genome. There is a chance that variants found in both strains could potentially be due to base level errors in the reference, i.e., there is no variant present at all. Similarly, when variant is only reported in strain A, there exists a small chance that the variant actually is located on strain B (i.e. the base level error in the reference happens to be the same as the sequence in strain B). Thus, a small percentage of the reported mutations in WLI strain could potentially be present in WMI. This might contribute, to some degree, to the enrichment of neuronal GO-term annotation for genes located within the vicinity of WLI sequence variants.

GO-term annotation enrichment for genes in the nearest proximity of variants detected in WLI included locomotor behavior and neuron projection. This provides some evidence that these variants could be capable of producing an impact on behavior, however this will require further investigation. As locomotory behavior is a complex trait, a combination of variants can be causal. For WMI the terms: neuron to neuron synapse (GO:0098984), nervous system development (GO:0007399), generation of neurons (GO:0048699), and finally, the Par-3-KIF3A-PKC-zeta complex (CORUM:899) was significantly over-represented. The Par-3-

KIF3A-PKC-zeta complex is interesting as both parts are in proximity of variants detected on WMI and it is involved in the establishment of neuronal polarity [47].

The ancestral WKY strain was noted for its highly variable behavior [19,20]. The WLI and WMI have been selected for both depressive and non-depressive behavior. With the discovery of variants associated with psychiatric phenotypes in both strains it should be kept in mind that variants could have been both selected for and against. In addition, as discussed above, there is a small chance some variants are located on the opposite strain due to potential errors in the reference genome. For this reason, we have only included variants which are different between WLI and WMI and not those that are different relative to the reference genome.

The WMI strain has been assigned to be a genetic model of depressive behavior, but the functional selection using the forced swim test could illuminate the potential connectedness of multiple phenotypic differences between the strains. The forced swim test is arguably thought of as a measure of stress-coping strategy [48], therefore, many behavioral phenotypes that employ stress coping could differ between the strains.

The small number of variants between the WMI and WLI strains indicates that they are close to isogenic. And yet, there exist numerous phenotypical differences between these two strains. This provides an opportunity to use genetic mapping strategies such as reduced complexity cross [10] to discover causal variants mediating behavioral phenotypes such as susceptibility to depression, stress reactivity, learning, memory , aging and drug abuse.

Methods

Animals. Liver tissue from 4 adult WLI (2 males and 2 females) and 4 adult WMI (2 males and 2 females) rats were collected. Equal amounts of tissue from males and females were pooled for each strain (total weight = 20 mg). DNA were extracted using the Qiagen DNeasy blood and tissue kit (Cat# 69506).

Whole Genome Sequencing. For sequencing using the HiSeq X Ten instrument, DNA whole genome shotgun sequencing libraries were generated using 200 ng of genomic DNA as input

for the TruSeq Nano DNA Library Prep Kit (Illumina). Indexed libraries were sequenced as pools of eight samples on a full slide (8 lanes) on an Illumina HiSeq X Ten sequencer using HiSeq X Ten v2.5 reagents. For sequencing using the Ion Torrent instrument, 1 µg of genomic DNA was sheared to an average size of 200 bp using a Covaris S2 Sonicator. Then 500 ng of the sheared DNA was used to prepare libraries for sequencing using the AB Library Builder™ Fragment library Kit on a Library Builder system. Libraries were used without amplification and size selected on a 2% Pippin Prep gel. After quantification using qPCR, the libraries (190 pg) were then used to prepare beads for sequencing using an Ion Torrent One Touch instrument. DNA on these beads then sequenced on an Ion Torrent Proton sequencer using Hi-Q chemistry and a P1 chip. For 10X Chromium sequencing, the Qiagen MagAttract HMW DNA kit was used for DNA isolation. Sequencing library was then constructed from 1 ng of high molecular weight (~ 50kb) genomic DNA using the Chromium Genome Library kit and sequenced on Illumina Hi-Seq (150 bp PE).

Mapping. Illumina and Ion proton data were mapped to the rat reference genome (rn6) using bwa (reference). 10x Chromium data were mapped to rn6 using LongRanger (ver 2.2.2). DeepVariant (ver 1.0.0) was used to call SNPs and small indels from the bam files and GLnexus was used for joint calling of variants.

Analysis. Variant identification was performed separately for each strain and sequencing method. A total of 6 samples spread over 2 strains and 3 sequencing technologies were analyzed. Variants with less than 10 reads across all samples or more than 300 on a single sample for a variant were removed. Variants with the same highest quality call for WLI and WMI were removed. Variants with an identical call for all three sequencing technologies within either WLI or WMI were stored for further analysis. Variants with 5 out of 6 uncertain calls (./.) were removed. Variants with the same highest quality call for WMI and WLI were removed. Variants with 5 out of 6 identical calls of which the last had a quality score less than 10 were removed. If the majority (>90%) of reads were of the same variant call across all reads and both strains shared at least 25% of all reads, the variant was removed.

Variants were selected based on the highest quality call per method and removed if disputed by variants called on another sequencing method with call quality of at least 30 within the same strain. Only variants were included in which the call for WLI differed from WMI and one of two strains was called as 0/0 (reference allele). Finally, all deletions on a position consisting of two identical nucleotides (homopolymeric) which were not supported by multiple sequencing techniques were removed (Figure 2). The final selection was exported to VCF per strain and type of call (homozygous or heterozygous).

Variants were identified in a panel of 44 inbred rats samples (5 BN samples, 32 HXB recombinant inbred strains including parental strains and 7 other inbred strains, our unpublished data). The genomic positions of variants were cross-compared with the variants of WMI and WLI without specific filtering or pre-selection to identify the number of likely de novo variants on WMI and WLI.

SnpEff (v4_3t_core) was used for nearest gene identification, impact estimation and annotation of the VCF for selected variants [35]. Impact and nearest genes were estimated separately per strain, as well as heterozygous and homozygous variants. Variants marked as high or moderate impact were separated and placed in table 3. The annotated VCF is available for reference. g:Profiler version e101_eg48_p14_baf17f0 was used for GO-term enrichment analysis, standard settings were used, no background dataset was utilized [49]. RatsPub [43]) was used to explore a small set of genes nearest to variants enriched with the GO-term: neuron to neuron synapse (GO:0098984).

Validation of variants and small indels by targeted re-sequencing. Ear punches from four WLI, four WMI (equal number of males and females) were used to extract genomic DNA. A total of 112 variants unique to WMI and 112 variants unique to WLI were selected from the final list of variants, with approximately equal distribution across the genome. Individual primer pairs were designed using Batch Primer 3 (http://probes.pw.usda.gov/batchprimer3/) at default settings for generic primers with total amplicon size set as an optimum of 100bp with

the amplified region containing the target SNP (or region of interest). The primer sequences and genomic DNA were submitted to Floodlight Genomics (FG, Knoxville, TN) for processing using a Hi-Plex targeted sequencing approach [50]. The Hi-Plex approach pools primers to PCR amplify targets and adds a barcode sequence during the amplification process. The resulting target library is then sequenced on an Illumina instrument. Data were then aligned to the fasta file containing the targeting target variants using bwa. Genotypes for each sample were called using DeepVariant.

Figure legends

Figure 1. Mean depth of coverage per chromosome per sample. Deepvariant called a total of 12,764,518 unique variants across 20 chromosomes plus X and Y with varying quality scores on either WLI or WMI samples. Depth of coverage is shown per A) WLI, IonProton, B) WLI 10X Chromium, C) WLI Illumina xTen, D) WMI IonProton, E) WMI 10X Chromium, F) WMI Illumina xTen.

Figure 2. Flowchart of each filtering step and the number of variants removed per step. The initial 8 steps were performed in Python, the last 2 were performed in R.

Figure 3. A) Number of ALT calls by deepVariant. Samples are separated based on quality score: quality 10 = ($p < 0.1$), quality 20 = ($p < 0.01$), quality 30 = ($p < 0.001$). B) Number of HET calls by deepVariant in different samples separated by quality.

Figure 4. From outside to inside: 1) Smoothed summed coverage of variant calls per technique for WMI samples (blue) and WLI samples (red). 2) Hotspots of homozygous SNPs on each chromosome found only in WMI (Blue) or WLI (Red). 3) Hotspots of heterozygous variants on each chromosome found only in WMI (light blue) or WLI (light red).

4) Smoothed summed quality of variant calls per technique for WMI samples (blue) and WLI samples (red).

Supplementary figure 1. Supplementary figure 1. Total number of homozygous (ALT) and heterozygous (HET) variants after final selection before and after homopolymer removal per strain.

Supplementary figure 2. Coverage of reference called variants to alternative on opposing strains (REF), homozygous variants (ALT) and heterozygous variants per sequencing technology. The coverage of heterozygous variants is overal twice as high as reference calls to homozygous variants on the opposing strain and homozygous variants.

Supplementary table 1. Positions of variants relative to genes on the genome for both WLI and WMI. Some regions overlap in classification.

Supplementary table 2 A). GO-term enrichment analysis for WLI and **B)** WMI of selected homozygous variants.

Supplementary table 3. Overview of 23 out of 30 genes associated with the enriched GO-term: neuron to neuron synapse (GO:0098984) that have been associated with psychiatric disease in previous studies.

References

1. WHO | Disease burden and mortality estimates. World Health Organization; 2018 [cited 2018 Mar 7]; Available from: http://www.who.int/healthinfo/global_burden_disease/estimates/en/index2.html

2. Sullivan PF, Neale MC, Kendler KS. Genetic epidemiology of major depression: review and meta-analysis. Am J Psychiatry. 2000;157:1552–62.

3. Fernandez-Pujals AM, Adams MJ, Thomson P, McKechanie AG, Blackwood DHR, Smith BH, et al. Epidemiology and Heritability of Major Depressive Disorder, Stratified by Age of Onset, Sex, and Illness Course in Generation Scotland: Scottish Family Health Study (GS:SFHS). PLoS One. 2015;10:e0142197.

4. Wang K, Gaitsch H, Poon H, Cox NJ, Rzhetsky A. Classification of common human diseases derived from shared genetic and environmental determinants. Nat Genet. 2017;49:1319–25.

5. Flint J, Kendler KS. The genetics of major depression. Neuron. 2014;81:484–503.

6. CONVERGE consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature. 2015;523:588–91.

7. Hyde CL, Nagle MW, Tian C, Chen X, Paciga SA, Wendland JR, et al. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. Nat Genet. 2016;48:1031–6.

8. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet. 2018;50:668–81.

9. Howard DM, Adams MJ, Clarke T-K, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci. 2019;22:343–52.

10. Bryant CD, Smith DJ, Kantak KM, Nowak TS Jr, Williams RW, Damaj MI, et al. Facilitating Complex Trait Analysis via Reduced Complexity Crosses. Trends Genet. 2020;36:549–62.

11. Louis WJ, Howes LG. Genealogy of the spontaneously hypertensive rat and Wistar-Kyoto rat strains: implications for studies of inherited hypertension. J Cardiovasc Pharmacol. 1990;16 Suppl 7:S1–5.

12. Kurtz TW, Montano M, Chan L, Kabra P. Molecular evidence of genetic heterogeneity in Wistar-Kyoto rats: implications for research with the spontaneously hypertensive rat. Hypertension. 1989;13:188–92.

13. Paré WP, Redei E. Sex differences and stress response of WKY rats. Physiol Behav. 1993;54:1179–85.

14. Solberg LC, Baum AE, Ahmadiyeh N, Shimomura K, Li R, Turek FW, et al. Sex- and lineage-specific inheritance of depression-like behavior in the rat. Mamm Genome. 2004;15:648–62.

15. Malkesman O, Braw Y, Maayan R, Weizman A, Overstreet DH, Shabat-Simon M, et al. Two different putative genetic animal models of childhood depression. Biol Psychiatry. 2006;59:17–23.

16. Dugovic C, Solberg LC, Redei E, Van Reeth O, Turek FW. Sleep in the Wistar-Kyoto rat, a putative genetic animal model for depression. Neuroreport. 2000;11:627–31.

17. Baum AE, Solberg LC, Churchill GA, Ahmadiyeh N, Takahashi JS, Redei EE. Test- and behavior-specific genetic factors affect WKY hypoactivity in tests of emotionality. Behav Brain Res. 2006;169:220–30.

18. Solberg LC, Olson SL, Turek FW, Redei E. Altered hormone levels and circadian rhythm of activity in the WKY rat, a putative animal model of depression. Am J Physiol Regul Integr Comp Physiol. 2001;281:R786–94.

19. Schaffer DJ, Tunc-Ozcan E, Shukla PK, Volenec A, Redei EE. Nuclear orphan receptor Nor-1 contributes to depressive behavior in the Wistar-Kyoto rat model of depression. Brain Res. 2010;1362:32–9.

20. Kurtz TW, Morris RC Jr. Biological variability in Wistar-Kyoto rats. Implications for research with the spontaneously hypertensive rat. Hypertension. 1987;10:127–31.

21. Paré WP, Kluczynski J. Differences in the stress response of Wistar-Kyoto (WKY) rats from different vendors. Physiol Behav. 1997;62:643–8.

22. Will CC, Aird F, Redei EE. Selectively bred Wistar–Kyoto rats: an animal model of depression and hyper-responsiveness to antidepressants. Mol Psychiatry. Nature Publishing Group; 2003;8:925–32.

23. Andrus BM, Blizinsky K, Vedell PT, Dennis K, Shukla PK, Schaffer DJ, et al. Gene expression patterns in the hippocampus and amygdala of endogenous depression and chronic stress models. Mol Psychiatry. 2012;17:49–61.

24. Mehta NS, Wang L, Redei EE. Sex differences in depressive, anxious behaviors and hippocampal transcript levels in a genetic rat model. Genes Brain Behav. 2013;12:695–704.

25. Luo W, Lim PH, Wert SL, Gacek SA, Chen H, Redei EE. Hypothalamic Gene Expression and Postpartum Behavior in a Genetic Rat Model of Depression. Front Behav Neurosci. 2020;14:190.

26. Mehta-Raghavan NS, Wert SL, Morley C, Graf EN, Redei EE. Nature and nurture: environmental influences on a genetic rat model of depression. Transl Psychiatry. Nature Publishing Group; 2016;6:e770.

27. Williams KA, Mehta NS, Redei EE, Wang L, Procissi D. Aberrant resting-state functional connectivity in a genetic rat model of depression. Psychiatry Res. 2014;222:111–3.

28. Mulders PC, van Eijndhoven PF, Schene AH, Beckmann CF, Tendolkar I. Resting-state functional connectivity in major depressive disorder: A review. Neurosci Biobehav Rev. 2015;56:330–44.

29. Lim PH, Shi G, Wang T, Jenz ST, Mulligan MK, Redei EE, et al. Genetic Model to Study the Co-Morbid Phenotypes of Increased Alcohol Intake and Prior Stress-Induced Enhanced Fear Memory. Front Genet. 2018;9:566.

30. Lim PH, Wert SL, Tunc-Ozcan E, Marr R, Ferreira A, Redei EE. Premature hippocampus-dependent memory decline in middle-aged females of a genetic rat model of depression. Behav Brain Res [Internet]. 2018; Available from: http://dx.doi.org/10.1016/j.bbr.2018.02.030

31. Pajer K, Andrus BM, Gardner W, Lourie A, Strange B, Campo J, et al. Discovery of blood transcriptomic markers for depression in animal models and pilot validation in subjects with early-onset major depression. Transl Psychiatry. Nature Publishing Group; 2012;2:e101.

32. Redei EE, Andrus BM, Kwasny MJ, Seok J, Cai X, Ho J, et al. Blood transcriptomic biomarkers in adult primary care patients with major depressive disorder undergoing cognitive behavioral therapy. Transl Psychiatry. 2014;4:e442.

33. Yu JS, Xue AY, Redei EE, Bagheri N. A support vector machine model provides an accurate transcript-level-based diagnostic for major depressive disorder. Transl Psychiatry. 2016;6:e931.

34. Redei EE, Ciolino JD, Wert SL, Yang A, Kim S, Clark C, Zumpf KB, Wisner, KL. Pilot validation of blood-based biomarkers during pregnancy and postpartum in women with prior or current depression. Transl Psychiatry. 2020;

35. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly . 2012;6:80–92.

36. Shi J, Potash JB, Knowles JA, Weissman MM, Coryell W, Scheftner WA, et al. Genome-wide association study of recurrent early-onset major depressive disorder. Mol Psychiatry. 2011;16:193–201.

37. Tian R-H, Bai Y, Li J-Y, Guo K-M. Reducing PRLR expression and JAK2 activity results in an increase in BDNF expression and inhibits the apoptosis of CA3 hippocampal neurons in a chronic mild stress model of depression. Brain Res. 2019;1725:146472.

38. Song A-Q, Gao B, Fan J-J, Zhu Y-J, Zhou J, Wang Y-L, et al. NLRP1 inflammasome contributes to chronic stress-induced depressive-like behaviors in mice. J Neuroinflammation. 2020;17:178.

39. Napoli E, Song G, Panoutsopoulos A, Riyadh MA, Kaushik G, Halmai J, et al. Beyond autophagy: a novel role for autism-linked Wdfy3 in brain mitophagy. Sci Rep. 2018;8:11348.

40. Chen K, Luan X, Liu Q, Wang J, Chang X, Snijders AM, et al. Drosophila Histone Demethylase KDM5 Regulates Social Behavior through Immune Control and Gut Microbiota Maintenance. Cell Host Microbe. 2019;25:537–52.e8.

41. Castermans D, Vermeesch JR, Fryns J-P, Steyaert JG, Van de Ven WJM, Creemers JWM, et al. Identification and characterization of the TRIP8 and REEP3 genes on chromosome 10q21.3 as novel candidate genes for autism. Eur J Hum Genet. 2007;15:422–31.

42. Campos-Rodríguez R, Godínez-Victoria M, Abarca-Rojano E, Pacheco-Yépez J, Reyna-Garfias H, Barbosa-Cabrera RE, et al. Stress modulates intestinal secretory immunoglobulin A. Front Integr Neurosci. 2013;7:86.

43. Gunturkun MH, Flashner E, Wang T, Mulligan MK, Williams RW, Prins P, et al. RatsPub: a webservice aided by deep learning to mine PubMed for addiction-related genes [Internet]. Cold Spring Harbor Laboratory. 2020 [cited 2020 Oct 25]. p. 2020.09.17.297358. Available from: https://www.biorxiv.org/content/10.1101/2020.09.17.297358v1

44. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol [Internet]. 2018; Available from: http://dx.doi.org/10.1038/nbt.4235

45. Supernat A, Vidarsson OV, Steen VM, Stokowy T. Comparison of three variant callers for human whole genome sequencing. Sci Rep. 2018;8:17851.

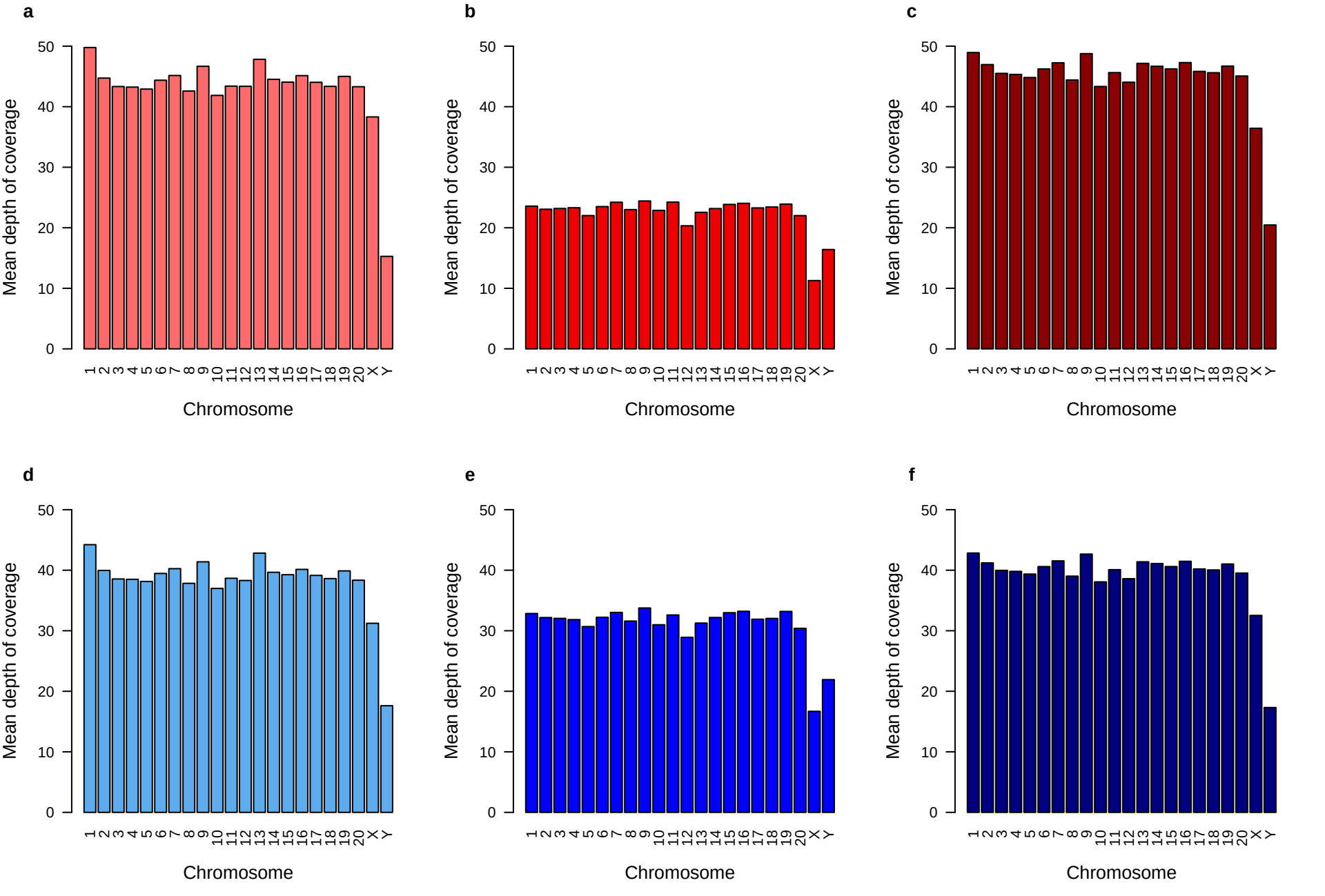
46. Brouard J-S, Schenkel F, Marete A, Bissonnette N. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. J Anim Sci Biotechnol. 2019;10:44.

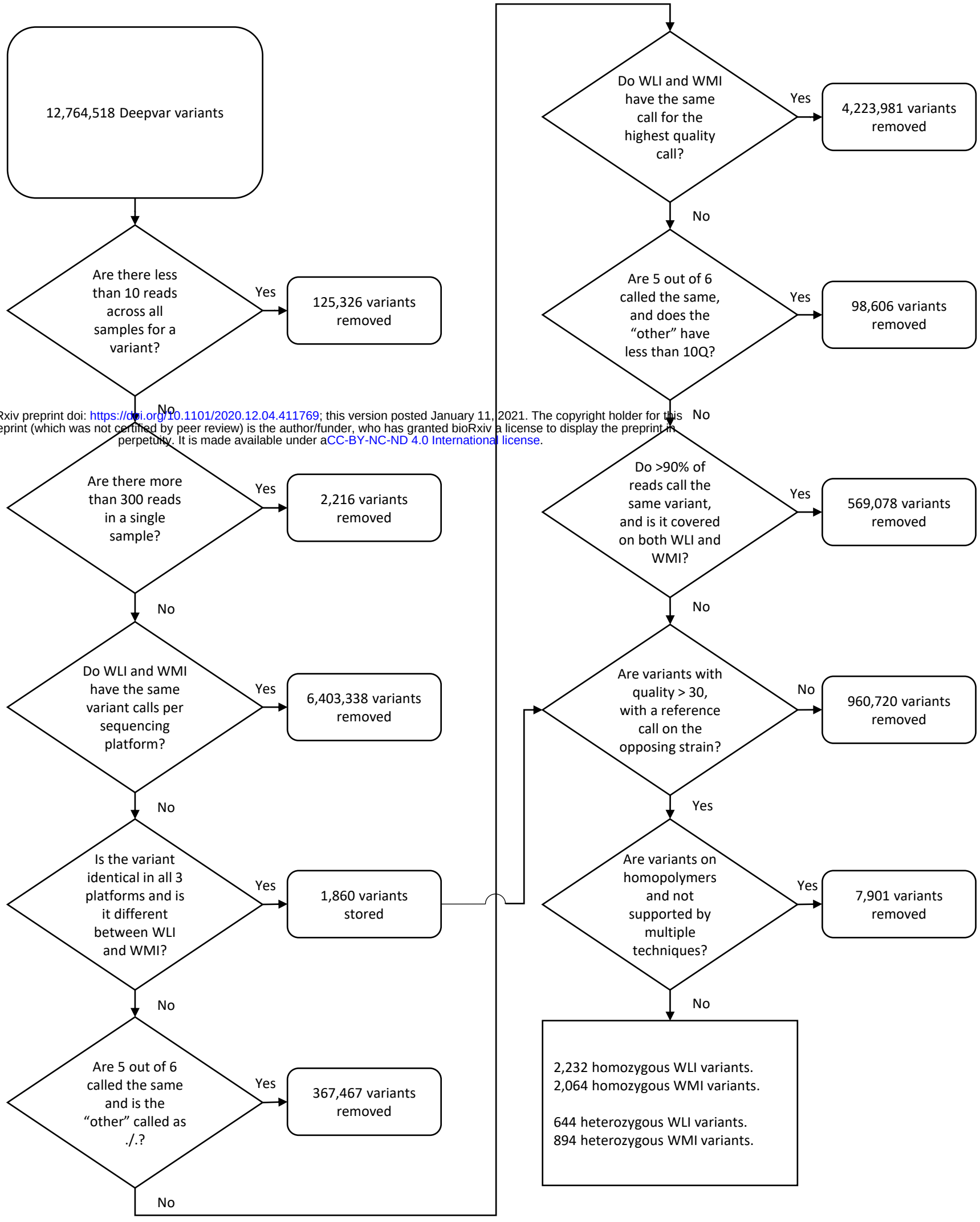
47. Nishimura T, Kato K, Yamaguchi T, Fukata Y, Ohno S, Kaibuchi K. Role of the PAR-3-KIF3 complex in the establishment of neuronal polarity. Nat Cell Biol. 2004;6:328–34.

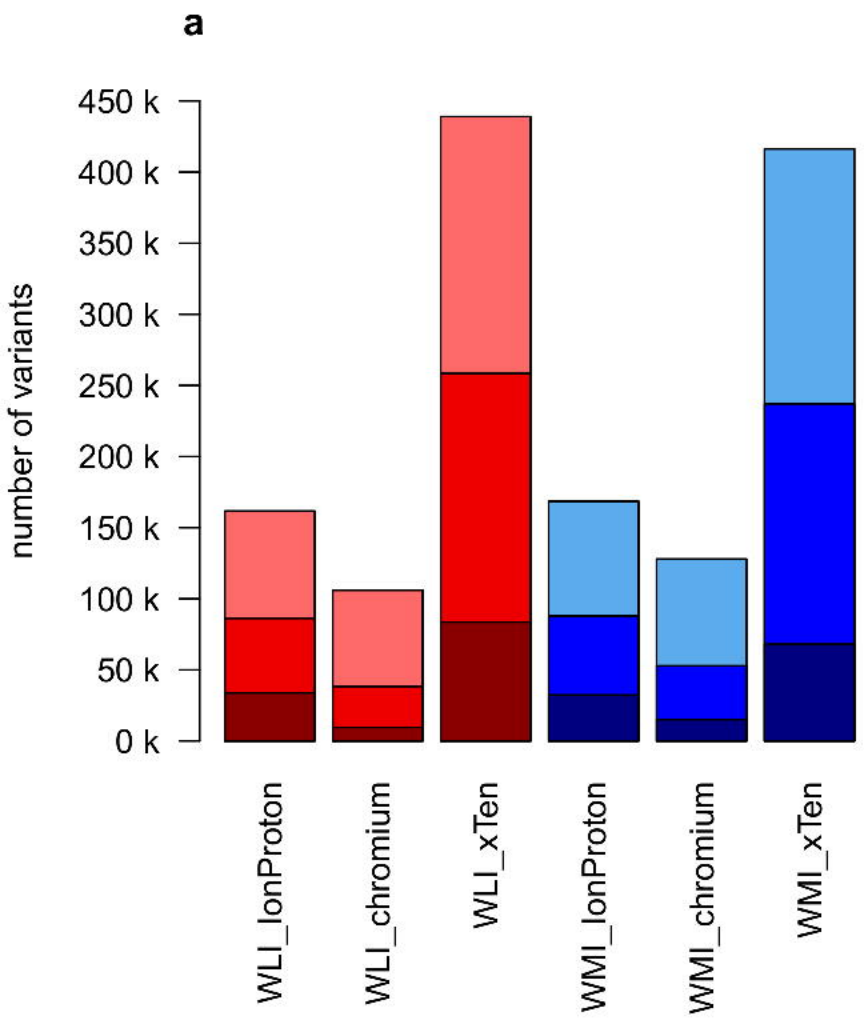
48. Molendijk ML, de Kloet ER. Coping with the forced swim stressor: Current state-of-the-art. Behav Brain Res. 2019;364:1–10.

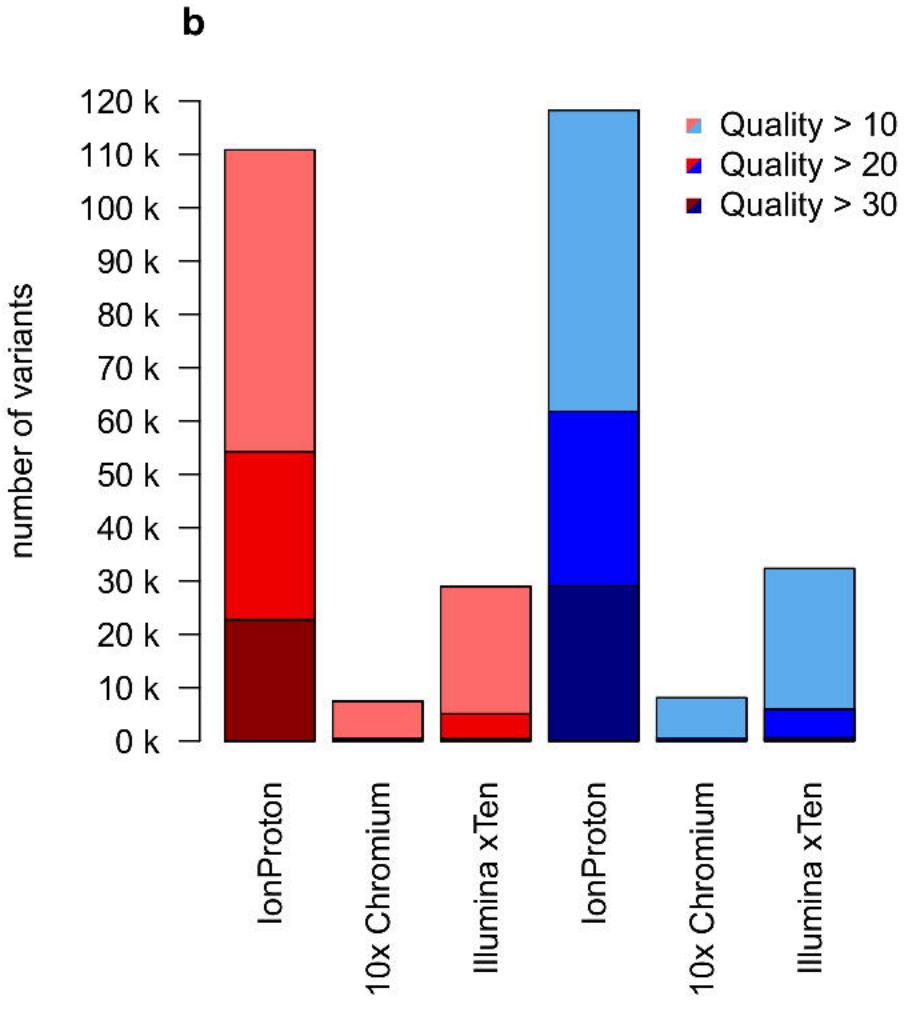
49. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47:W191–8.

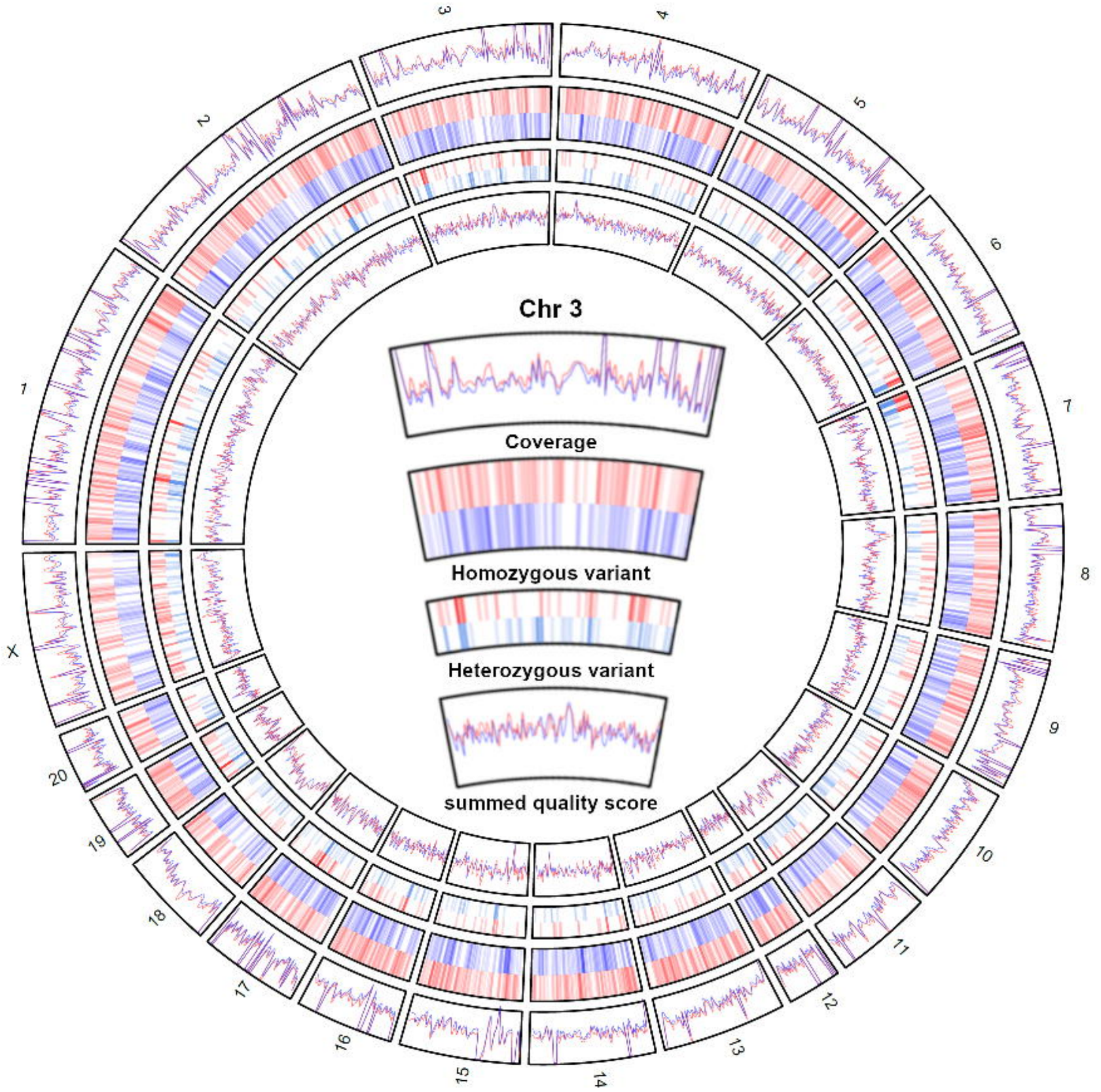
50. Nguyen-Dumont T, Pope BJ, Hammet F, Southey MC, Park DJ. A high-plex PCR approach for massively parallel sequencing. Biotechniques. 2013;55:69–74.


a
b
c
d
e
f
Mean depth of coverage
Chromosome
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
X
Y




a
number of variants
450 k
400 k
350 k
300 k
250 k
200 k
150 k
100 k
50 k
0 k
WLI_IonProton
WLI_chromium
WLI_xTen
WMI_IonProton
WMI_chromium
WMI_xTen


b
number of variants
120 k
110 k
100 k
90 k
80 k
70 k
60 k
50 k
40 k
30 k
20 k
10 k
0 k
Quality > 10
Quality > 20
Quality > 30
IonProton
10x Chromium
Illumina xTen
IonProton
10x Chromium
Illumina xTen


Chr 3
Coverage
Homozygous variant
Heterozygous variant
summed quality score
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
X