bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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-NC-ND 4.0 International license.

Haplotype-resolved and integrated genome analysis of the cancer cell line HepG2

Bo Zhou[1,2,#], Steve S. Ho[1,2,#], Stephanie U. Greer[3], Noah Spies[2,4,5], John M. Bell[6], Xianglong Zhang[1,2], Xiaowei Zhu[1,2], Joseph G. Arthur[7], Seunggyu Byeon[8], Reenal Pattni[1,2], Ishan Saha[2], Yiling Huang[1,2], Giltae Song[8], Dimitri Perrin[9], Wing H. Wong[7,10], Hanlee P. Ji[3,6], Alexej Abyzov[11], Alexander E. Urban[1,2,*]

[1]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, California 94305, USA
[2]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA
[3]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, California 94305, USA
[4]Department of Pathology, Stanford University School of Medicine, Stanford, California 94305, USA
[5]Genome-scale Measurements Group, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA
[6]Stanford Genome Technology Center, Stanford University, Palo Alto, California 94304, USA
[7]Department of Statistics, Stanford University, Stanford, California 94305, USA
[8]School of Computer Science and Engineering, College of Engineering, Pusan National University, Busan 46241, South Korea
[9]Science and Engineering Faculty, Queensland University of Technology, Brisbane, QLD 4001, Australia
[10]Department of Biomedical Data Science, Bio-X Program, Stanford University, Stanford, California 94305, USA
[11]Department of Health Sciences Research, Center for Individualized Medicine, Mayo Clinic, Rochester, Minnesota 55905, USA

# These authors contributed equally.

*****Correspondence**:** aeurban@stanford.edu

Running title: Haplotype-resolved and integrated global analysis of the HepG2 genome

Keywords: HepG2, cancer genomics, ENCODE, CRISPR/Cas9, linked-reads, haplotype phasing, structural variation (SV), retrotransposons, allele-specific expression, allele-specific DNA methylation

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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-NC-ND 4.0 International license.

SUMMARY

The HepG2 cancer cell line is one of the most widely-used biomedical research and one of the main cell lines of ENCODE. Vast numbers of functional genomics and epigenomics datasets have been produced to characterize its biology. However, the correct interpretation such data requires an understanding of the cell line's genome sequence and genome structure. Using a variety of sequencing and analysis methods, we identified a wide spectrum of HepG2 genome characteristics: copy numbers of chromosomal segments, SNVs and Indels (corrected for aneuploidy), phased haplotypes extending to entire chromosome arms, loss of heterozygosity, retrotransposon insertions, structural variants (SVs) including complex and somatic genomic rearrangements. We also identified allele-specific expression and DNA methylation genome-wide and assembled an allele-specific CRISPR/Cas9 targeting map.

SIGNIFICANCE

Haplotype-resolved and comprehensive whole-genome analysis of a widely-used cell line for cancer research and ENCODE, HepG2, serves as an essential resource for unlocking complex cancer gene regulation using a genome-integrated framework and also provides genomic context for the analysis of ~1,000 functional datasets to date on ENCODE for biological discovery. We also demonstrate how deeper insights into genomic regulatory complexity are gained by adopting a genome-integrated framework.

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

INTRODUCTION

Genomic instability is a hallmark of cancer where critical genomic changes create gene fusions, the disruption of tumor-suppressor, and the amplification of oncogenes (Adey et al., 2013; Hanahan and Weinberg, 2011; Negrini et al., 2010). A comprehensive knowledge of the mutations and larger structural changes that underlie a cancer genome is not only critical for a deeper understanding of the biological processes that drive tumor progression and evolution but also for the development of targeted cancer therapies. The HepG2 cell line is one of the most widely used cancer cell lines used in many areas biomedical research due to its extreme versatility, contributing to over 23,000 publications to date, even more than K562. It is a hepatoblastoma cell line derived from a 15-year-old Caucasian male (Aden et al., 1979; López-Terrada et al., 2009a). Representing the human endodermal lineage, HepG2 cells are widely used as models for human toxicology studies (Dias da Silva et al., 2013; Kamalian et al., 2015; Menezes et al., 2013; Sahu et al., 2012; Schoonen et al., 2005), including toxicogenomic screens using CRISPR-Cas9 (Xia et al., 2016), in addition to studies on drug metabolism (Alzeer and Ellis, 2014), cancer (Xu et al., 2013), liver disease (Hao et al., 2014), gene regulatory mechanisms (Huan et al., 2014), and biomarker discovery (Mangrum et al., 2015). As one of the main cell lines of the ENCyclopedia Of DNA Elements Project (ENCODE), HepG2 has been used to generate close to 1,000 datasets for ENCODE (Sloan et al., 2016).

The functional genomic and epigenomics aspects of HepG2 cells have been extensively studied with approximately 325 ChIP-Seq, 300 RNA-Seq, and 180 eCLIP datasets available through ENCODE in addition to recent single-cell methylome and transcriptome datasets (Hou et al., 2016). However, the genome sequence and higher-order genomic structural features of HepG2 have never been characterized in a comprehensive manner, even though the HepG2 cell line has been known to contain multiple chromosomal abnormalities (Chen et al., 1993; Simon et al., 1982). As a result, the extensive HepG2 functional genomics and epigenomics studies conducted to date were done without reliable genomic contexts for accurate

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

interpretation.

Here, we report the first global, integrated, and haplotype-resolved whole-genome characterization of the HepG2 cancer genome that includes copy numbers (CN) of large chromosomal regions at high-resolution, single-nucleotide variants (SNVs, also including single-nucleotide polymorphisms, i.e. SNPs) and small insertions and deletions (indels) with allele-frequencies corrected by CN in aneuploid regions, loss of heterozygosity, mega-base-scale phased haplotypes, and structural variants (SVs), many of which are haplotype-resolved (Figure 1, Figure S1). The datasets generated in this study form an integrated, phased, high-fidelity genomic resource that can provide the proper contexts for future experiments that rely on HepG2's unique characteristics. We show how knowledge about HepG2's genomic sequence and structural variations can enhance the interpretation of functional genomics and epigenomics data. For example, we integrated HepG2 RNA-Seq data and whole-genome bisulfite sequencing data with ploidy and phasing information and identified many cases of allele-specific gene expression and allele-specific DNA methylation. We also compiled a phased CRISPR map of loci suitable for allele specific-targeted genome editing or screening. Finally, we demonstrate the power of this resource by providing compelling insights into the mutational history of HepG2 and oncogene regulatory complexity derived from our datasets. The technical framework demonstrated in this study is also suitable for the study of other cancer cell lines and primary tumor samples.

RESULTS

Karyotyping

We obtained HepG2 cells from the Stanford ENCODE Production Center. The cells exhibit a hyperdiploid karyotype of 49 to 52 chromosomes (Figure 2A). Of the 20 metaphase HepG2 cells analyzed using the GTW banding method, 15 of cells were complexly and variably abnormal and also characterized by multiple structural and numerical abnormalities. These include translocation between the chromosome 1p and 21p, trisomies of chromosomes 2, 16,

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

and 17, tetrasomy of chromosome 20, uncharacterized arrangements of chromosomes 16 and 17, and a variable number of marker chromosomes. Five cells demonstrated greater than 100 chromosomes and represent a tetraploid expansion of the stemline described. This tetraploid expansion is consistent with previously published results (Simon et al. 1982) but also absent from other published cytogenetic analyses of HepG2 (Chen et al. 1993), suggesting the clonal evolution arose during tumor-genesis or early in the establishment of the HepG2 cell line. Although the ploidies of all chromosomes in our HepG2 cell line were supported by previous published karyotypes (Chen et al., 1993; Simon et al., 1982; Wong et al., 2000), variations do exist and also among the various published analyses especially for chromosomes 16 and 17, suggesting that karyotypic differences exist between different HepG2 cell lines (Table S1).

High-Resolution Ploidy Changes in HepG2

To obtain a high-resolution aneuploid map i.e. large CN changes by chromosomal region in HepG2, WGS coverage across the genome was first calculated in 10 kb bins and plotted against percent GC content where four distinct clusters were clearly observed (Abyzov et al., 2011) (Figure S2). CNs were assigned to each cluster based on the ratio between its mean coverage and that of the lowest cluster (CN=1). These assigned large CN changes by chromosomal region confirm the hyperdiploid state of the HepG2 genome as identified by karyotyping (Figure 2A, Figure S2, Table S2). We see that 74.1% of the HepG2 genome has a baseline copy number of two (consistent with karyotype), 15.5% copy number of three, 2.7% copy number of four, 0.7% has a copy number of five, and 6.9% in a haploid state (Figure 2B, Table S2). Furthermore, these high-resolution CN changes across the HepG2 genome were also confirmed by two independent replicates of Illumina Infinium Multi-Ethnic Global-8 arrays (MEGA) array data (Figure S3A, Supplementary Data). We found increased CN (CN=3) over the oncogene *VEGFA* (6p21.1), which was found to be recurrently duplicated in cases of hepatocellular carcinoma (Cancer Genome Atlas Research Network, 2017).

SNVs and indels

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

We identified SNVs and indels in HepG2 by taking into account the CN of the chromosomal regions in which they reside so that heterozygous allele frequencies can be assigned accordingly (e.g. 0.33 and 0.67 in triploid regions; 0.25, 0.50, and 0.75 in tetraploid regions). Using GATK Haplotypecaller (McKenna et al., 2010), we identified a total of ~3.34M SNVs (1.90M heterozygous, 1.44M homozygous) and 0.90M indels (0.60M heterozygous, 0.29M homozygous) (Table 1, Dataset 1). Interestingly, there are 12,375 heterozygous SNVs and indels that have more than two haplotypes in chromosomal regions with CN>2 (Dataset 1). In addition, chromosome 22 and large continuous stretches of chromosomes 6, 11, and 14 show striking loss of heterozygosity (LOH) (Figure 1, Table S3). Since genomic data from healthy tissue that correspond to HepG2 cells is not available, we intersected these SNVs and indels with dbsnp138 (Sherry et al., 2001) and found the overlapping proportion to be ~98% and ~78% respectively (Figure 2C). This suggests that HepG2 has accumulated a significant number of SNVs and indels relative to inherited. We found that 377 SNVs and 255 indels are private protein-altering (PPA) after filtering out those that overlapped with The 1000 Genomes Project (The 1000 Genomes Project Consortium et al., 2015) or the Exome Sequencing Project (Fu et al., 2012) (Table 1, Table S4). Moreover, the intersection between the filtered PPA variants and the Catalogue of Somatic Mutations in Cancer (COSMIC) is 39% and 16% for SNVs and indels respectively (Table S5). The gene overlap between HepG2 PPA and the Sanger Cancer Gene Census is 19 (Table S6). HepG2 PPA variants include oncogenes and tumor suppressors such as *NRAS* (Pylayeva-Gupta et al., 2011), *STK11/LKB1* (Zhou et al., 2014), and *PREX2* (Berger et al., 2012; Yang et al., 2016) as well as other genes recently found to play critical roles in driving cancer such as *CDK12* (Paculová and Kohoutek, 2017) and *IKBKB* (Kai et al., 2014; Xia et al., 2012). *RP1L1*, which was recently found to be significantly mutated in hepatocellular carcinoma (Cancer Genome Atlas Research Network 2017) is also present among the PPA variants.

Resolving Haplotypes

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

We phased the heterozygous SNVs and indels in the HepG2 genome by performing 10X Genomics Chromium linked-read library preparation and sequencing (Marks et al., 2018; Zheng et al., 2016). Post sequencing quality control analysis show that 1.49 ng or approximately 447 genomic equivalents of high molecular weight (HMW) genomic DNA fragments (mean=68 kb, 96.1% > 20 kb, 22.0% >100 kb) were partitioned into 1.53 million oil droplets and uniquely barcoded (16 bp). This library was sequenced (2x151 bp) to 67x genome coverage with half of all reads coming from HMW DNA molecules with at least 61 linked reads (N50 Linked-Reads per Molecule) (Table 1). We estimate the actual physical coverage (C_F) to be 247x. Coverage of the mean insert by sequencing (C_R) is 18,176 bp (284 bp X 64) or 30.8%, thus the overall sequencing coverage $C = C_R \times C_F = 67x$. Distributed over 1628 haplotype blocks (Table 1, Dataset 2), 1.87M (98.7%) of heterozygous SNVs and 0.67M (77.9%) of indels in HepG2 were successfully phased. The longest phased haplotype block is 31.1Mbp (N50=6.80Mbp) (Figure 2D, Table 1, Dataset 2); however, haplotype block lengths vary widely across different chromosomes (Figure S4, Figure 1). Poorly phased regions correspond to regions exhibiting LOH (Table S3, Figure 1, Dataset 2).

Construction of Mega-Haplotypes of Entire Chromosome Arms

We constructed mega-haplotypes of entire chromosome arms by leveraging the haplotype imbalance in aneuploid regions in the HepG2 genome where phased haplotype blocks derived from linked-reads were “stitched” together (Table 1, Figure 2E, Supplementary Data). Briefly, by using a recently developed method (Bell et al., 2017) specifically for cancer genomes, we counted linked-read barcodes for each phased heterozygous SNVs in haplotype blocks with ≥100 phased SNVs (Dataset 2). Because each barcode is specific for a HMW DNA molecule, the total number of unique barcodes is directly correlated with the number of individual HMW DNA molecules that were sequenced. The fractional representation of a particular genomic sequence (or locus) can be obtained by counting the total number of unique barcodes associated with that particular genomic sequence. Consequently, for each phased

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

haplotype with CN>2, major and minor haplotypes can be assigned according to the number of barcodes associated with each haplotype (Figure 2E), where the major haplotype is the haplotype with more associated unique barcodes. In genomic regions where CN=2, the two haplotypes are expected to have similar numbers of unique barcodes. In this method (Bell et al., 2017), a matched control for comparison is required to confidently discriminate between the major and minor haplotypes. Here, we used NA12878 as normal control because no matching normal tissue sample is available for HepG2 (Figure 2E). After performing the normalization procedures and statistical tests described in (Bell et al., 2017) to verify haplotype imbalance or aneuploidy genomic regions in HepG2, we then "stitched" together contiguous blocks of phased major and minor haplotypes respectively. Using this approach, a total of 6 autosomal mega-haplotypes were constructed (Table 1, Supplementary Data); 4 of which encompass entire (or >96%) chromosome arms: 2p, 2q, 6p, and 16q (Fig. 3). The largest mega-haplotype is approximately 144 Mb long (2q).

Using Linked-Reads to Identify and Reconstruct Large and Complex SVs

From the linked reads, breakpoints of large-scale SVs can be identified by searching for distant genomic regions with linked-reads that have large numbers of overlapping barcodes. SVs can also be assigned to specific haplotypes if the breakpoint-supporting reads contain phased SNVs or indels (Marks et al., 2018; Zheng et al., 2016). Using this approach (implemented by the Long Ranger software from 10X Genomics), we identified 97 large SVs >30 kb (99% phased) (Dataset 3) and 3,473 deletions between 50 bp and 30 kb (78% phased) (Dataset 4). The large SVs include inter- and intra-chromosomal rearrangements (Figure 3A, B), duplications (Figure 3C, D), deletions (Figure 3E, F), and inversions (Figure 3G, H). A remarkable example is the haplotype-resolved translocation between chromosomes 16 and 6 (Figure 3A) resulting in the disruption of the non-receptor Fyn-related tyrosine kinase gene *FRK,* which has been identified as a tumor suppressor (Brauer and Tyner, 2009; Yim et al., 2009). Another example is the 127 kb tandem duplication on chromosome 7 (Figure 3C) that results in

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

the partial duplication of genes *PMS2*, encoding a mismatch repair endonuclease, and *USP42*, encoding the ubiquitin-specific protease 42. An interesting large SV is the 395 kb duplication within *PRKG1* (Figure 3D), which encodes the soluble I-alpha and I-beta isoforms of cyclic GMP-dependent protein kinase. We also identified a 193 kb homozygous deletion in *PDE4D* for HepG2 using linked-read sequencing where six internal exons within the gene are deleted (Figure 3D).

Furthermore, we also used the long-range information from the deep linked-reads sequencing dataset to identify, assemble, and reconstruct the breakpoints of SVs in the HepG2 genome using a recently developed method called Genome-wide Reconstruction of Complex Structural Variants (GROC-SVs) (Spies et al., 2017). Here, HMW DNA fragments that span breakpoints are statistically inferred and refined by quantifying the barcode similarity of linked-reads between pairs of genomic regions similar to Long Ranger (Zheng et al., 2016). Sequence reconstruction is then achieved by assembling the relevant linked reads around the identified breakpoints from which SVs are automatically reconstructed. Breakpoints that also have supporting evidence from the 3kb-mate pair dataset (see Methods) are indicated as high-confidence events. GROC-SVs called a total of 140 high-confidence breakpoints including 4 inter-chromosomal events (Figure 1, Dataset 5, Figure 4A-D); 138 of the breakpoints were successfully sequence-assembled with nucleotide-level resolution of breakpoints as well the exact sequence in the cases where nucleotides have been added or deleted. We identified striking examples of inter-chromosomal rearrangements or translocations in HepG2 between chromosomes 1 and 4 (Figure 4A) and between chromosomes 6 and 17 (Figure 4B) as well as breakpoint-assembled large genomic deletions (Figure 4C, Dataset 5). This break-point assembled 335 kb heterozygous deletion is within the *NEDD4L* on chromosome 18. Lastly, we identified a large (1.3 mb) intra-chromosomal rearrangement that deletes large portions of *RBFOX1* and *RP11420N32* in one haplotype on chromosome 16 using deep linked-read sequencing (Figure 4D, Dataset 3, Dataset 5).

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

We then employed "gemtools" (Greer et al., 2017) to resolve and phase large and complex SVs in the HepG2 genome. We identified a complex SV on chromosome 8 that involves a small deletion downstream of *ADAM2* which is also within a larger tandem duplication leading to the amplification of the oncogene *IDO1* (Platten et al., 2015) and the first half of *IDO2* (Figure 5). Two allele-specific deletions 700 kb and 200 kb respectively were identified in the *PDE4D* on chromosome 5 (Figure 5). Since chromosome 5 is triploid in HepG2 (Figure 1, Figure 2A, Supplementary Information), we see approximately twice as much linked-reads barcode representation for the allele harboring the 200 kb deletion, suggesting that this allele of *PDE4D* has two copies and the allele harboring the 700 kb deletion has one copy (Figure 5). Similarly, we also identified two allele-specific deletions, 290 kb and 160 kb respectively within *AUTS2* on chromosome 7 (Figure 5). Interestingly, for the allele harboring the 160 kb deletion, the non-deleted reference allele is also present at much larger frequency as indicted by the total number of linked-read barcodes, suggesting that the allele harboring the 160 kb deletion within *AUTS2* occurs in a fraction of HepG2 cells or sub-clonally (Figure 5). Estimating from the total number of linked-read barcodes that are associated with this 160 kb allele-specific deletion in *AUTS2*, we estimate that this deletion occur at a frequency of ~10%. In other words, 10% of HepG2 cells have both deletions rendering a large portion of *AUTS2* in HepG2 cells deleted completely in ~10% of HepG2 cells whereas the other 90% of HepG2 cells carry only the 290 kb deletion within *AUTS2*. All breakpoints identified using "gemtools" were individual PCR and Sanger sequencing verified (Table S7).

SVs from Mate-Pair Sequencing

To obtain increase sensitivity in the detection medium-sized (1 kb-100 kb) SVs in HepG2, we prepared a 3kb-mate pair library and sequenced (2x151 bp) it to a genome coverage of 7.9x after duplicate removal. The sequencing coverage of each 3 kb insert (C_R) is 302 bp (or 10% of the insert size) which translates to a physical coverage (C_F) of 79x. Deletions, inversions, and tandem duplications from the mate-pair library were identified from

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

analysis of discordant read pairs and split reads using LUMPY (Layer et al., 2014). Only SVs that are supported by both discordant read-pair and split-read supports were retained. Using this approach, we identified 122 deletions, 41 inversions, and 133 tandem duplications (Dataset 6). Approximately 76% of these SVs are between 1 kb-10 kb, 86% are between 1kb-100kb (9% between 10 kb-100 kb), and 3% are greater than 100 kb **(**Dataset 6). Twenty SVs (16 deletions and 4 duplications) were randomly selected for experimental validation using PCR and Sanger sequencing in which 15/16 were successfully validated (93.8%) (Table S7).

SVs identified from deep short-insert WGS

Deletions, inversions, insertions, and tandem duplications, were identified from the HepG2 WGS dataset using Pindel (Ye et al., 2009), BreakDancer (Chen et al., 2009), and BreakSeq (Lam et al., 2010). Since similar categories of SVs were also identified using mate-pair and linked-read sequencing, these SVs were combined with the SVs identified previously using Long Ranger and LUMPY where variations with support from multiple methods and with greater than 50% reciprocal overlap were merged. In total, 6,405 SVs were obtained from all methods that include 5,226 deletions, 245 duplications, 428 inversions, and 494 insertions (only BreakDancer (Chen et al., 2009) was designed to call insertions) (Supplementary Data). A set of deletion (n=27) and tandem duplication calls (n=4) were randomly selected to confirm by PCR and Sanger sequencing, and 30/32 (94%) events were successfully validated (Table S7). Consistent with previous analysis (Lam et al., 2012), deletions show the highest concordance among the various methods of detection compared to duplication and inversion calls (Figure S5). As expected, we detected a 520 bp deletion in exon 3 of the β-catenin (*CTNNB1*) gene (Dataset 4, Supplementary Data), which was previously documented to exist in HepG2 (López-Terrada et al., 2009b). Interestingly, we found no SVs or PPA mutations in the Wnt-pathway gene *CAPRIN2* (Ding et al., 2008), which had been previously reported for hepatoblastoma (Jia et al., 2014).

Identification of Non-Reference Alu and LINE1 Insertions

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

From our deep-coverage short-insert WGS data, we also analyzed the HepG2 genome for non-reference LINE1 and Alu retrotransposon insertions using RetroSeq (Keane et al., 2013) with some modifications. These insertions were identified from paired-end reads that have one of the pair mapping to hg19 uniquely and other mapping to an Alu or LINE1 consensus sequence in full or split fashion (see Methods). Retrotransposon insertion events with greater than five supporting reads were categorized as high confidence and retained (Table S8). We identified 1,899 and 351 non-reference Alu and LINE1 insertions in the HepG2 genome respectively (Figure 1). We randomly chose 8 Alu and 10 LINE1 insertions with split-read support for confirmation using PCR and Sanger sequencing where 87.5% and 100% respectively were successfully validated (Table S8).

Allele-Specific Gene Expression

Due to the abundance of aneuploidy in the HepG2 genome, CN changes of genomic regions should be taken into account when analyzing for allele-specific gene expression in order to reduce false positives and false negatives. Using the heterozygous SNV allele frequencies in HepG2 (Dataset 1), we re-analyzed two replicates of HepG2 ENCODE RNA-Seq data**.** We identified 3,189 and 3,022 genes that show allele-specific expression ($p < 0.05$) in replicates one and two, respectively (Figure 1, Table S9). Furthermore, we also identified 862 and 911 genes that would have been falsely identified to have allele-specific expression (false positives) if the copy numbers of SNV allele frequencies were not taken into consideration as well as 446 and 407 genes that would not have been identified (false negatives) in replicates one and two, respectively (Table S10).

Allele-Specific DNA Methylation

Using the phasing information for HepG2 SNVs (Dataset 2), we also identified 384 CpG islands (CGIs) that exhibit allele-specific DNA methylation (Figure 1, Table S11). We obtained two independent replicates of HepG2 whole-genome bisulfite sequencing data (2x125 bp, experiment ENCSR881XOU) from the ENCODE Portal (Sloan et al., 2016). Read alignment to

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

hg19 was performed using Bismark (Krueger and Andrews, 2011); 70.0% of reads were uniquely aligned and a striking 44.7% of cytosines were methylated in a CpG context. We then phased methylated and unmethylated CpGs to their respective haplotypes by identifying reads that overlap both CpGs and phased heterozygous SNVs (Dataset 2). We grouped the phased individual CpGs into CGIs and totaled the number of reads that contain methylated and unmethylated cytosines for each CGI allele, normalizing by CN in cases of aneuploidy. Fisher's exact test was used to evaluate allele-specific methylation, and significant results were selected using a target false discovery rate of 10% (Storey and Tibshirani, 2003) (see Methods). In total, 98 CGIs reside within promoter regions (defined as 1 kb upstream of a gene); 277 are intragenic, and 96 lie within 1 kb downstream of 348 different genes. The following 11 genes are within 1 kb of a differentially methylated CGI and also overlap with the Sanger Cancer Gene Census: *FOXA1, GNAS, HOXD13, PDE4DIP, PRDM16, PRRX1, SALL4, STIL, TAL1, and ZNF331.*

Allele-Specific CRISPR Targets

We identified 38,551 targets in the HepG2 genome suitable for allele-specific CRSIPR targeting (Figure 1, Table S12). Phased variant sequences (including reverse complement) that differ by >1 bp between the alleles were extracted to identify all possible CRISPR targets by pattern searching for [G, C, or A]N_{20}GG (see Methods). Only conserved high-quality targets were retained by using a selection method previously described and validated (Sunagawa et al., 2016). We took the high-quality target filtering process further by taking the gRNA function and structure into account. Targets with multiple exact matches, extreme GC fractions, and those with TTTT sequence (which might disrupt the secondary structure of gRNA) were removed. Furthermore, we used the Vienna RNA-fold package (Lorenz et al., 2011) to identify gRNA secondary structure and eliminated targets for which the stem loop structure for Cas9 recognition is not able to form (Nishimasu et al., 2014). Finally, we calculated the off-target risk score using the tool as described for this purpose (Ran et al., 2013). A very strict off-target

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

threshold score was chosen in which candidates with a score below 75 were rejected to ensure that all targets are as reliable and as specific as possible.

Genomic Sequence and Structural Context Provides Insight into Regulatory Complexity

We show examples of how deeper insights into gene regulation and regulatory complexity can be obtained by integrating genomic sequence and structural contexts with functional genomics and epigenomics data (Figure 5A-D). One example is the allele-specific RNA expression and allele-specific DNA methylation in HepG2 at the *PLK2* locus on chromosome 5 (Figure 5A). By incorporating the genomic context in which *PLK2* is expressed in HepG2 cells, we see that *PLK2* RNA is only expressed from Haplotype 1 (p = 4.66E-10) in which the CGI within the gene is completely unmethylated (p = 1.51E-66) in the expressed allele and completely methylated in the non-expressed allele (Figure 5A, C, D). The second example is allele-specific RNA expression and allele-specific DNA methylation of the *TBX2* gene in HepG2 (Figure 5B). The *TBX2* locus on chromosome 17 is triploid, and we see that *TBX2* is preferentially expressed from Haplotype 1 which has one copy and lower expression is observed from the two copies of Haplotype 2 (p = 0.0179) (Figure 5B, C). We also observed highly preferential DNA methylation of the CGI in Haplotype 1 (p = 1.55E-32) (Figure 5B, D). In addition, there is also an allele-specific CRISPR targeting site for both haplotypes in the promoter region of *TBX2* and inside CGI 22251 (1,937 bp upstream of *TBX2* gene and 2,259 bp downstream of the 5' end of CGI 22251) (Figure 5B).

DISCUSSION

When the HepG2 cell line was first established in 1979, it was mistakenly reported as of hepatocellular carcinoma origin (Aden et al., 1979) and also curated such in the ATCC repository (Rockville, MD, USA). This misclassification has generated much confusion among investigators in the past decades and in the published literature. Review of the original tumor specimen by the original investigators firmly established HepG2 to be of epithelial hepatoblastoma origin rather than hepatocellular carcinoma (López-Terrada et al., 2009a). As

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

one of the most widely used cell lines in biomedical research and one of the main cell lines of ENCODE, HepG2's genomic sequence and structural features have never been characterized in a comprehensive manner beyond its karyotype (Chen et al., 1993; Simon et al., 1982) and SNVs identified from ChIP-Seq data and 10x coverage WGS that do not take aneuploidy or CN into consideration (Cavalli et al., 2016; Huang and Ovcharenko, 2015). Studies conducted using the extensive collection of functional genomics and epigenomics datasets for HepG2 have previously relied on the human reference genome. Here, we present the first global and phased characterization of the HepG2 genome. By using deep short-insert WGS, 3 kb-insert mate-pair sequencing, array analysis, karyotyping, deep linked-reads sequencing, and integrating a collection of novel and established computational methods, we performed a comprehensive analysis of genomic structural features (Figure 1) for the HepG2 cell line that includes SNVs (Dataset 1), Indels (Dataset 1), large CN or ploidy changes across chromosomal regions at 10 kb resolution (Table S2), phased haplotype blocks (Dataset 2), phased CRISPR targets (Table S12), novel retrotransposon insertions (Table S8), and SVs (Dataset 3, Dataset 4, Dataset 5, Dataset 6) including deletions, duplications, inversions, translocations, and those that are the result of complex genomic rearrangements. Many of the HepG2 SVs are also phased, assembled, and experimentally verified (Dataset 5, Table S7, and Table S8).

We also illustrate, using *PLK2* and *TBX2* (Figure 5A, B), examples where knowing the genomic sequence and structural context can enhance the interpretation of function genomics and epigenomics data to derive novel insights into the complexity of oncogene regulation. The Polo-like kinase gene *PLK2* (*SNK*) is a transcriptional target of p53 and also a cancer biomarker (Burns et al., 2003; Coley et al., 2012). It has been studied in the contexts of many human cancers (Burns et al., 2003; Ou et al., 2016; Pellegrino et al., 2010; Syed et al., 2011). Disruption of PLK2 has also been proposed to have therapeutic value in sensitizing chemo-resistant tumors. Its roles in Burkitt's lymphoma (Syed et al., 2006), hepatocellular carcinoma (Pellegrino et al., 2010), and epithelial ovarian cancer (Syed et al., 2011) are consistent with

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

that of tumor suppressors while its role in colorectal cancer is consistent with that of an oncogene (Ou et al., 2016). Interestingly, promoter methylation and/or LOH were linked to the down-regulation of PLK2 in human hepatocellular carcinoma (Pellegrino et al., 2010). Chemotherapy resistance of epithelial ovarian cancer can be conferred by the down-regulation of PLK2 at the transcriptional level via DNA methylation of the CpG island in the *PLK2* promoter (Syed et al., 2011). Here we show that the down-regulation of PLK2 in HepG2 cancer cells could be achieved through what appears to be allele-specific transcriptional silencing via allele-specific DNA methylation of a large CGI within the gene body (Figure 5A).

The T-box transcription factor TBX2 is a critical regulator of cell fate decisions, cell migration, and morphogenesis in the development of many organs (Cho et al., 2011; Harrelson et al., 2004; Manning et al., 2006; Suzuki et al., 2005). It regulates cell cycle progression (Bilican and Goding, 2006), and its overexpression has been demonstrated in promoting or maintaining the proliferation of many cancers including melanomas (Vance et al., 2005), nasopharyngeal cancer (Lv et al., 2017), breast cancer (D'Costa et al., 2014; Wang et al., 2012), prostate cancer (Du et al., 2017), and gastric cancer (Yu et al., 2015). Here, we show that three copies of the *TBX2* gene exist in HepG2 cancer cells as a result of duplication in Haplotype 2. However, it is preferentially expressed in Haplotype 1 possibly due to the highly allele-specific DNA methylation in the CGI that span its promoter region and most of the gene body (Figure 5B). It is plausible that overexpression of TBX2 in other cancer types are caused by similar genomic rearrangements and/or epigenetic mechanisms where duplication of *TBX2* may result in the overexpression and DNA methylation (possibly allele-specific) may contribute an additive effect to TBX2 overexpression or act as the sole contributor where *TBX2* is not duplicated.

The duplication of *TBX2* is a direct consequence of aneuploidy. Previous studies have also revealed the aneuploid karyotype of the HepG2 cell line (Chen et al., 1993; Simon et al., 1982). The karyotype of HepG2 cells in our analysis was largely supported by previously

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

published karyotypes for all chromosomes (Table S1). While karyotypes provide a general guide of the degree of aneuploidy in HepG2 cells, we point researchers to the results of our high-resolution read depth analysis of large CN changes across the HepG2 genome for a clearer picture (Table S2). See ***Supplementary Discussion*** for detailed discussion of other oncogenes, tumor-suppressors, and other genes associated with cancer that are disrupted as a consequence of genomic variation in HepG2.

The identification of SNVs and Indels with sensitivity and accuracy requires deep WGS coverage (>33x for SNVs and >60x for Indels) (Bentley et al., 2008; Fang et al., 2014). From our WGS dataset (>70.3x coverage) we identified large numbers of SNVs and Indels that we subsequently corrected for their allele frequencies according to chromosomal CN. In addition to being essential for haplotype discovery, correct allele frequencies of variants are also needed for functional genomics or epigenomics studies such as the identification of allele-specific gene expression or of allele-specific transcription factor binding in HepG2. A statistically significant increase in transcription factor binding signal for ChIP-Seq or transcription signal in RNA-Seq for one allele compared to the other at a heterozygous locus may be seen as a case of allele-specific expression or allele-specific transcription factor binding which usually indicates allele-specific gene regulation at this locus. However, if aneuploidy can be taken into account and the RNA-Seq or ChIP-Seq signals are normalized by CN, the case observed might simply due to increased CN at that particular locus rather than the preferential activation of one allele over the other. This was often the case in our re-analysis of two replicates of HepG2 RNA-Seq data where we identified 862 and 911 genes that would have been falsely identified to have allele-specific expression in addition to 446 and 407 genes that would not have been identified to have allele-specific expression in replicates one and two, respectively, if chromosomal CN or haplotype allele frequency was not taken into consideration (Table S10).

The haplotype phase of the genomic variants is an essential aspect of human genetics but entirely ignored by current standard WGS approaches. To obtain haplotype phasing

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

information for HepG2 (Dataset 2), we performed deep coverage linked-read sequencing (Zheng et al., 2016). However, due to high LOH (Figure 1 and Table S3), chromosomes 22 and large portions of 6, 11, and 14 were difficult to phase resulting in much shorter haplotype blocks compared to other chromosomes (Dataset 2, Figure 1, Figure S4). We further extended on the phasing capabilities of Long Ranger and constructed mega-haplotypes that encompass entire HepG2 chromosome arms (Figure 2E, Table 1, Supplementary Data). This was achieved by leveraging the inherent haplotype imbalance in aneuploid genomic regions in cancer genomes using a recently published method developed specifically for this purpose (Bell et al., 2017). Heterozygous loci in aneuploidy regions that contain more than two haplotypes were excluded from phasing due to software limitations (Zheng et al., 2016). The phase information of these loci could be resolved, in principle, from our linked-read data should new algorithms become available.

Combining orthogonal methods and signals greatly improves SV-calling sensitivity and accuracy (Layer et al., 2014; Mohiyuddin et al., 2015). Here, we combined deep short-insert WGS, mate-pair sequencing, and linked-read sequencing with a combination of several computational SV calling methods to identify a spectrum of structural variants that includes deletions, duplications, and inversions as well as CN-corrected complex rearrangements. We compared SVs identified from various methods. For deletions, we see significant overlap as well as variant calls that are specific to each method (Figure S5A), but overlap is less for duplications (Figure S5B) and inversions (Figure S5C). This is consistent with what has been shown previously (Lam et al., 2012) as inversions and duplications are more difficult in principle to accurately resolve. Since each SV detection method is designed to use different types of signals and also optimized to identify different classes of SVs, such overlaps are also expected. Experimental validation of individual SVs of interest should be conducted prior to functional follow up studies.

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

All data and results generated from this global whole-genome analysis of HepG2 is publicly available on the ENCODE portal (Sloan et al., 2016). This analysis serves as a valuable reference for further understanding the vast amount of existing HepG2 ENCODE data, such as determining whether a known or potential sequence regulatory element has been altered by SNVs, Indels, Alu or LINE1 insertions, CN changes of that given element, or subjected to allele-specific regulation.

Our results also guide future study designs that utilize HepG2 cells including CRISPR experiments where knowledge of the phased genomic variants can extend or modify the number of editing targets including those that are allele-specific (Table S12) while knowledge of aberrant chromosomal CN changes will allow for more accurate interpretation of functional data in non-diploid regions. This study on the HepG2 genome may serve as a technical archetype for advanced, integrated, and global analysis of genomic sequence and structural variants for other widely cell lines with complex genomes.

Researchers should consider that HepG2 and other widely utilized cell lines have been passaged for long periods of time across many different laboratories and encounter opportunities for additional genome variation to occur, especially if they are HepG2 cells that are long separated from the ENCODE HepG2 production line we used in our analysis. Many of the analyses we discuss here are supported by previous study, for instance, our karyotyping and reported mutation in *CTNNB1*, but there are minor differences such as the lack of a mutation in *CAPRIN2* which has been previously reported. We expect that the vast majority of genomic variants that we describe here can be found across the different versions of HepG2 cells but it is always possible that distinct lines of HepG2 cells may harbor slight variations. This also applies when analyzing the various functional genomic datasets available for HepG2 on the ENCODE Portal (Sloan et al., 2016) that have accumulated over the years, especially if follow up work for individual loci is conducted on a different HepG2 line. A first step should always be to experimentally confirm the presence of the particular genomic variant of interest in that

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

particular working line of HepG2. For global analyses that integrate multi-omics datasets, we expect the majority of the genomic variants described and catalogued here to exist: datasets presented here should be well-suited for global perspectives and substantial insights can be expected to be gained. While the aneuploidy in HepG2 renders the design and interpretation of HepG2 genomic and epigenomic studies more challenging, the results of this study enables researchers to continue to use HepG2 to investigate the effects of different types of genomic variations on the multiple layers of functionality and regulation for which ENCODE data is already available and continues to be produced. Although our studies reveal that the genome of HepG2 highly complex, taking our results into account for future analyses of HepG2 ENCODE data should not be considered to make the process more challenging but rather potentially much more insightful and rewarding. This study is primarily focused on the utilization of Illumina sequencing to resolve the genome structure of HepG2. In the future, we plan to utilize other long-read technologies such as Pacific Biosciences and Oxford Nanopore.

METHODS

To characterize the aneuploid genome of HepG2 in a comprehensive manner, we integrated karyotyping (Figure S1), high-density oligonucleotide arrays, deep (70.3x non-duplicate coverage) short-insert whole-genome sequencing (WGS), 3kb-mate-pair sequencing, and 10X Genomics linked-reads sequencing (Zheng et al., 2016). Using read-depth analysis of WGS data (Abyzov et al., 2011), we first obtained a high-resolution aneuploid map or large copy number (CN) changes by chromosomal region. CN changes were also validated by karyotyping and two independent replicates of arrays. This high-resolution aneuploid map was then integrated into the identification of SNVs and indels and in determining their heterozygous allele-frequencies. By leveraging the inherent haplotype imbalance in aneuploid genomic regions, SNV haplotype blocks were “stitched” into mega-haplotypes that encompass entire chromosome arms (Bell et al., 2017). In addition, SVs, such as deletions, duplications, inversions, and non-reference retrotransposon (LINE1 and Alu) insertions, were also identified

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

from the deep-coverage WGS dataset. The 10X Genomics linked-reads dataset was used to phase haplotypes as well as to identify, assemble, and phase primarily large (>30kb) SVs that include translocations (Marks et al., 2018; Spies et al., 2017; Zheng et al., 2016). The 3kb-mate pair was used to validate the large SVs resolved from using the linked-read sequencing data and also to identify additional SVs, mostly in the medium size-range (1kb-100kb). Finally, we used the phased haplotype information to produce an allele-specific CRISPR targeting map for HepG2 and identified cases of allele-specific expression and allele-specific DNA methylation. See Supplementary Methods for detailed descriptions.

DATA AVAILABILITY

Raw and processed data files are publicly released on the ENCODE portal (encodeproject.org) via experimental accession numbers ENCSR350RUO, ENCSR276ECO, and ENCSR319QHO. Datasets 1-6 can be accessed via ENCODE accession numbers ENCFF336CFC, ENCFF853HHD, ENCFF467ETN, ENCFF717TPE, ENCFF330UFT, ENCFF241CEK respectively. For immediate review of Datasets and Supplementary Data, files are also available under

https://stanfordmedicine.box.com/s/w1fa3ncuhje8zabgugvsw6wmeu7ojv3k

DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST

The authors of this manuscript declare no conflicts of interest.

AUTHORS' CONTRIBUTIONS

B.Z. and A.E.U conceived and designed the study. B.Z., R.P., I.S., and Y.H. performed experiments. B.Z., S.S.H., S.U.G., N.S., J.M.B., XL.Z., XW.Z., J.G.A., S.B., G.S., D.P., and A.A. performed analysis. H.P.J, W.H.W., and A.E.U contributed resources and supervised the study. B.Z., S.S.H., and A.E.U. wrote the manuscript.

ACKNOWLEDGEMENTS

We thank Arineh Khechaduri for performing genomic DNA preparation. Aditi Narayanan, Dr. Idan Gabdank, Nathaniel Watson, Zachary A. Myers, and Dr. Cricket Sloan for data

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

organization and upload to ENCODE. Dr. Athena Cherry and the Stanford Cytogenetics Laboratory for karyotype analysis. A.E.U. was supported by NIH grant HG007735 and the Stanford Medicine Faculty Innovation Program. W.H.W. received support from NIH grants HG007834 and HG007735. B.Z. was additionally funded by NIH Grant T32 HL110952. J.G.A was funded by the NSF Graduate Research Fellowship and NIH T32-GM096982.

REFERENCES

Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. *21*, 974–984.

Aden, D.P., Fogel, A., Plotkin, S., Damjanov, I., and Knowles, B.B. (1979). Controlled synthesis of HBsAg in a differentiated human liver carcinoma-derived cell line. Nature *282*, 615–616.

Adey, A., Burton, J.N., Kitzman, J.O., Hiatt, J.B., Lewis, A.P., Martin, B.K., Qiu, R., Lee, C., and Shendure, J. (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. Nature *500*, 207–211.

Alzeer, S., and Ellis, E.M. (2014). Metabolism of gamma hydroxybutyrate in human hepatoma HepG2 cells by the aldo-keto reductase AKR1A1. Biochem. Pharmacol. *92*, 499–505.

Bell, J.M., Lau, B.T., Greer, S.U., Wood-Bouwens, C., Xia, L.C., Connolly, I.D., Gephart, M.H., and Ji, H.P. (2017). Chromosome-scale mega-haplotypes enable digital karyotyping of cancer aneuploidy. Nucleic Acids Res. *45*, e162.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.

Berger, M.F., Hodis, E., Heffernan, T.P., Deribe, Y.L., Lawrence, M.S., Protopopov, A., Ivanova, E., Watson, I.R., Nickerson, E., Ghosh, P., et al. (2012). Melanoma genome sequencing reveals frequent PREX2 mutations. Nature *485*, 502–506.

Bilican, B., and Goding, C.R. (2006). Cell cycle regulation of the T-box transcription factor tbx2.

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

Exp. Cell Res. *312*, 2358–2366.

Brauer, P.M., and Tyner, A. (2009). RAKing in AKT: A tumor suppressor function for the intracellular tyrosine kinase FRK. Cell Cycle *8*, 2728–2732.

Burns, T.F., Fei, P., Scata, K.A., Dicker, D.T., and El-Deiry, W.S. (2003). Silencing of the Novel p53 Target Gene Snk/Plk2 Leads to Mitotic Catastrophe in Paclitaxel (Taxol)-Exposed Cells. Mol. Cell. Biol. *23*, 5556–5571.

Cancer Genome Atlas Research Network (2017). Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell *169*, 1327–1341.e23.

Cavalli, M., Pan, G., Nord, H., Wallén Arzt, E., Wallerman, O., and Wadelius, C. (2016). Allele-specific transcription factor binding in liver and cervix cells unveils many likely drivers of GWAS signals. Genomics *107*, 248–254.

Chen, H.-L., Chiu, T.-S., Chen, P.-J., and Chen, D.-S. (1993). Cytogenetic studies on human liver cancer cell lines. Cancer Genet. Cytogenet. *65*, 161–166.

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods *6*, 677–681.

Cho, G.-S., Choi, S.-C., Park, E.C., and Han, J.-K. (2011). Role of Tbx2 in defining the territory of the pronephric nephron. Development *138*, 465–474.

Coley, H.M., Hatzmichael, E., Blagden, S.P., McNeish, I.A., Thompson, A., Crook, T., and Syed, N. (2012). Polo Like Kinase 2 Tumour Suppressor and cancer biomarker: new perspectives on drug sensitivity/resistance in ovarian cancer. Oncotarget *3*.

D'Costa, Z.C., Higgins, C., Ong, C.W., Irwin, G.W., Boyle, D., McArt, D.G., McCloskey, K., Buckley, N.E., Crawford, N.T., Thiagarajan, L., et al. (2014). TBX2 represses CST6 resulting in uncontrolled legumain activity to sustain breast cancer proliferation: a novel cancer-selective target pathway with therapeutic opportunities. Oncotarget *5*, 1609–1620.

Dias da Silva, D., Carmo, H., Lynch, A., and Silva, E. (2013). An insight into the hepatocellular

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

death induced by amphetamines, individually and in combination: the involvement of necrosis and apoptosis. Arch. Toxicol. *87*, 2165–2185.

Ding, Y., Xi, Y., Chen, T., Wang, J., Tao, D., Wu, Z.-L., Li, Y., Li, C., Zeng, R., and Li, L. (2008). Caprin-2 enhances canonical Wnt signaling through regulating LRP5/6 phosphorylation. J. Cell Biol. *182*, 865–872.

Du, W.-L., Fang, Q., Chen, Y., Teng, J.-W., Xiao, Y.-S., Xie, P., Jin, B., and Wang, J.-Q. (2017). Effect of silencing the T-Box transcription factor TBX2 in prostate cancer PC3 and LNCaP cells. Mol. Med. Rep. *16*, 6050–6058.

Fang, H., Wu, Y., Narzisi, G., O'Rawe, J.A., Barrón, L.T.J., Rosenbaum, J., Ronemus, M., Iossifov, I., Schatz, M.C., and Lyon, G.J. (2014). Reducing INDEL calling errors in whole genome and exome sequencing data. Genome Med. *6*, 89.

Fernandez-Banet, J., Lee, N.P., Chan, K.T., Gao, H., Liu, X., Sung, W.-K., Tan, W., Fan, S.T., Poon, R.T., Li, S., et al. (2014). Decoding complex patterns of genomic rearrangement in hepatocellular carcinoma. Genomics *103*, 189–203.

Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Altshuler, D., Shendure, J., Nickerson, D.A., et al. (2012). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216–220.

Greer, S.U., Nadauld, L.D., Lau, B.T., Chen, J., Wood-Bouwens, C., Ford, J.M., Kuo, C.J., and Ji, H.P. (2017). Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. Genome Med. *9*, 57.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*, 646–674.

Hao, L., Ito, K., Huang, K.-H., Sae-tan, S., Lambert, J.D., and Ross, A.C. (2014). Shifts in dietary carbohydrate-lipid exposure regulate expression of the non-alcoholic fatty liver disease-associated gene PNPLA3/adiponutrin in mouse liver and HepG2 human liver cells. Metabolism. *63*, 1352–1362.

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

Harrelson, Z., Kelly, R.G., Goldin, S.N., Gibson-Brown, J.J., Bollag, R.J., Silver, L.M., and Papaioannou, V.E. (2004). Tbx2 is essential for patterning the atrioventricular canal and for morphogenesis of the outflow tract during heart development. Development *131*, 5041–5052.

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., Huang, Y., et al. (2016). Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res. *26*, 304–319.

Huan, L.C., Wu, J.-C., Chiou, B.-H., Chen, C.-H., Ma, N., Chang, C.Y., Tsen, Y.-K., and Chen, S.C. (2014). MicroRNA regulation of DNA repair gene expression in 4-aminobiphenyl-treated HepG2 cells. Toxicology *322*, 69–77.

Huang, D., and Ovcharenko, I. (2015). Identifying causal regulatory SNPs in ChIP-seq enhancers. Nucleic Acids Res. *43*, 225–236.

Jia, D., Dong, R., Jing, Y., Xu, D., Wang, Q., Chen, L., Li, Q., Huang, Y., Zhang, Y., Zhang, Z., et al. (2014). Exome sequencing of hepatoblastoma reveals novel mutations and cancer genes in the Wnt pathway and ubiquitin ligase complex. Hepatology *60*, 1686–1696.

Kai, X., Chellappa, V., Donado, C., Reyon, D., Sekigami, Y., Ataca, D., Louissaint, A., Mattoo, H., Joung, J.K., and Pillai, S. (2014). IκB Kinase β (IKBKB) Mutations in Lymphomas That Constitutively Activate Canonical Nuclear Factor κB (NFκB) Signaling. J. Biol. Chem. *289*, 26960–26972.

Kamalian, L., Chadwick, A.E., Bayliss, M., French, N.S., Monshouwer, M., Snoeys, J., and Park, B.K. (2015). The utility of HepG2 cells to identify direct mitochondrial dysfunction in the absence of cell death. Toxicol. Vitr. *29*, 732–740.

Keane, T.M., Wong, K., and Adams, D.J. (2013). RetroSeq: Transposable element discovery from next-generation sequencing data. Bioinformatics *29*, 389–390.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science *318*, 420–426.

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

Krueger, F., and Andrews, S.R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.

Lam, H.Y.K., Mu, X.J., Stütz, A.M., Tanzer, A., Cayting, P.D., Snyder, M., Kim, P.M., Korbel, J.O., and Gerstein, M.B. (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat. Biotechnol. *28*, 47–55.

Lam, H.Y.K., Pan, C., Clark, M.J., Lacroute, P., Chen, R., Haraksingh, R., O'Huallachain, M., Gerstein, M.B., Kidd, J.M., Bustamante, C.D., et al. (2012). Detecting and annotating genetic variations using the HugeSeq pipeline. Nat. Biotechnol. *30*, 226–229.

Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. *15*, R84.

López-Terrada, D., Cheung, S.W., Finegold, M.J., and Knowles, B.B. (2009a). Hep G2 is a hepatoblastoma-derived cell line. Hum. Pathol. *40*, 1512–1515.

López-Terrada, D., Gunaratne, P.H., Adesina, A.M., Pulliam, J., Hoang, D.M., Nguyen, Y., Mistretta, T.-A., Margolin, J., and Finegold, M.J. (2009b). Histologic subtypes of hepatoblastoma are characterized by differential canonical Wnt and Notch pathway activation in DLK+ precursors. Hum. Pathol. *40*, 783–794.

Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.

Lv, Y., Si, M., Chen, N., Li, Y., Ma, X., Yang, H., Zhang, L., Zhu, H., Xu, G.-Y., Wu, G.-P., et al. (2017). TBX2 over-expression promotes nasopharyngeal cancer cell proliferation and invasion. Oncotarget *8*, 52699–52707.

Mangrum, J.B., Martin, E.J., Brophy, D.F., and Hawkridge, A.M. (2015). Intact stable isotope labeled plasma proteins from the SILAC-labeled HepG2 secretome. Proteomics *15*, 3104–3115.

Manning, L., Ohyama, K., Saeger, B., Hatano, O., Wilson, S.A., Logan, M., and Placzek, M. (2006). Regional Morphogenesis in the Hypothalamus: A BMP-Tbx2 Pathway Coordinates Fate and Proliferation through Shh Downregulation. Dev. Cell *11*, 873–885.

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

Marks, P., Garcia, S., Barrio, A.M., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A., et al. (2018). Resolving the Full Spectrum of Human Genome Variation using Linked-Reads. Preprint at.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Menezes, C., Alverca, E., Dias, E., Sam-Bento, F., and Pereira, P. (2013). Involvement of endoplasmic reticulum and autophagy in microcystin-LR toxicity in Vero-E6 and HepG2 cell lines. Toxicol. In Vitro *27*, 138–148.

Mohiyuddin, M., Mu, J.C., Li, J., Bani Asadi, N., Gerstein, M.B., Abyzov, A., Wong, W.H., and Lam, H.Y.K. (2015). MetaSV: An accurate and integrative structural-variant caller for next generation sequencing. Bioinformatics *31*, 2741–2744.

Negrini, S., Gorgoulis, V.G., and Halazonetis, T.D. (2010). Genomic instability--an evolving hallmark of cancer. Nat. Rev. Mol. Cell Biol. *11*, 220–228.

Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell *156*, 935–949.

Ou, B., Zhao, J., Guan, S., Wangpu, X., Zhu, C., Zong, Y., Ma, J., Sun, J., Zheng, M., Feng, H., et al. (2016). Plk2 promotes tumor growth and inhibits apoptosis by targeting Fbxw7/Cyclin E in colorectal cancer. Cancer Lett. *380*, 457–466.

Paculová, H., and Kohoutek, J. (2017). The emerging roles of CDK12 in tumorigenesis. Cell Div. *12*, 7.

Pellegrino, R., Calvisi, D.F., Ladu, S., Ehemann, V., Staniscia, T., Evert, M., Dombrowski, F., Schirmacher, P., and Longerich, T. (2010). Oncogenic and tumor suppressive roles of polo-like kinases in human hepatocellular carcinoma. Hepatology NA-NA.

Platten, M., von Knebel Doeberitz, N., Oezen, I., Wick, W., and Ochs, K. (2015). Cancer

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

Immunotherapy by Targeting IDO1/TDO and Their Downstream Effectors. Front. Immunol. *5*.

Pylayeva-Gupta, Y., Grabocka, E., and Bar-Sagi, D. (2011). RAS oncogenes: weaving a tumorigenic web. Nat. Rev. Cancer *11*, 761–774.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. *8*, 2281–2308.

Sahu, S.C., O'Donnell, M.W., and Sprando, R.L. (2012). Interactive toxicity of usnic acid and lipopolysaccharides in human liver HepG2 cells. J. Appl. Toxicol. *32*, 739–749.

Schoonen, W.G.E.J., Westerink, W.M.A., de Roos, J.A.D.M., and Débiton, E. (2005). Cytotoxic effects of 100 reference compounds on Hep G2 and HeLa cells and of 60 compounds on ECC-1 and CHO cells. I mechanistic assays on ROS, glutathione depletion and calcein uptake. Toxicol. In Vitro *19*, 505–516.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Simon, D., Aden, D.P., and Knowles, B.B. (1982). Chromosomes of human hepatoma cell lines. Int. J. Cancer *30*, 27–33.

Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T., et al. (2016). ENCODE data at the ENCODE portal. Nucleic Acids Res. *44*, D726–D732.

Spies, N., Weng, Z., Bishara, A., McDaniel, J., Catoe, D., Zook, J.M., Salit, M., West, R.B., Batzoglou, S., and Sidow, A. (2017). Genome-wide reconstruction of complex structural variants using read clouds. Nat. Methods.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U. S. A. *100*, 9440–9445.

Sunagawa, G.A., Sumiyama, K., Ukai-Tadenuma, M., Perrin, D., Fujishima, H., Ukai, H., Nishimura, O., Shi, S., Ohno, R.-I., Narumi, R., et al. (2016). Mammalian Reverse Genetics without Crossing Reveals Nr3a as a Short-Sleeper Gene. Cell Rep. *14*, 662–677.

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

Suzuki, T., Takeuchi, J., Koshiba-Takeuchi, K., and Ogura, T. (2005). Tbx Genes Specify Posterior Digit Identity through Shh and BMP Signaling. Dev. Cell *8*, 971–972.

Syed, N., Smith, P., Sullivan, A., Spender, L.C., Dyer, M., Karran, L., O'Nions, J., Allday, M., Hoffmann, I., Crawford, D., et al. (2006). Transcriptional silencing of Polo-like kinase 2 (SNK/PLK2) is a frequent event in B-cell malignancies. Blood *107*, 250–256.

Syed, N., Coley, H.M., Sehouli, J., Koensgen, D., Mustea, A., Szlosarek, P., McNeish, I., Blagden, S.P., Schmid, P., Lovell, D.P., et al. (2011). Polo-like Kinase Plk2 Is an Epigenetic Determinant of Chemosensitivity and Clinical Outcomes in Ovarian Cancer. Cancer Res. *71*, 3317–3327.

The 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Vance, K.W., Carreira, S., Brosch, G., and Goding, C.R. (2005). Tbx2 is overexpressed and plays an important role in maintaining proliferation and suppression of senescence in melanomas. Cancer Res. *65*, 2260–2268.

Wang, B., Lindley, L.E., Fernandez-Vega, V., Rieger, M.E., Sims, A.H., and Briegel, K.J. (2012). The T Box Transcription Factor TBX2 Promotes Epithelial-Mesenchymal Transition and Invasion of Normal and Malignant Breast Epithelial Cells. PLoS One *7*, e41355.

Wong, N., Lai, P., Pang, E., Leung, T.W., Lau, J.W., and Johnson, P.J. (2000). A comprehensive karyotypic study on human hepatocellular carcinoma by spectral karyotyping. Hepatology *32*, 1060–1068.

Xia, P., Zhang, X., Xie, Y., Guan, M., Villeneuve, D.L., and Yu, H. (2016). Functional Toxicogenomic Assessment of Triclosan in Human HepG2 Cells Using Genome-Wide CRISPR-Cas9 Screening. Environ. Sci. Technol. *50*, 10682–10692.

Xia, Y., Yeddula, N., Leblanc, M., Ke, E., Zhang, Y., Oldfield, E., Shaw, R.J., and Verma, I.M. (2012). Reduced cell proliferation by IKK2 depletion in a mouse lung-cancer model. Nat. Cell

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

Biol. *14*, 257–265.

Xu, D., He, X., Chang, Y., Xu, C., Jiang, X., Sun, S., and Lin, J. (2013). Inhibition of miR-96 expression reduces cell proliferation and clonogenicity of HepG2 hepatoma cells. Oncol. Rep. *29*, 653–661.

Yang, J., Gong, X., Ouyang, L., He, W., Xiao, R., and Tan, L. (2016). PREX2 promotes the proliferation, invasion and migration of pancreatic cancer cells by modulating the PI3K signaling pathway. Oncol. Lett. *12*, 1139–1143.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865–2871.

Yim, E.-K., Siwko, S., and Lin, S.-Y. (2009). Exploring Rak tyrosine kinase function in breast cancer. Cell Cycle *8*, 2360–2364.

Yu, H., Liu, B.O., Liu, A., Li, K., and Zhao, H. (2015). T-box 2 expression predicts poor prognosis in gastric cancer. Oncol. Lett. *10*, 1689–1693.

Zheng, G.X.Y., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., Terry, J.M., et al. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat. Biotechnol. *34*, 303–311.

Zhou, W., Zhang, J., and Marcus, A.I. (2014). LKB1 tumor suppressor: Therapeutic opportunities knock when LKB1 is inactivated. Genes Dis. *1*, 64–74.

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

FIGURE LEGENDS

Figure 1. Comprehensive Overview of the HepG2 Genome

Circos visualization (Krzywinski et al. 2009) of HepG2 genome variants with the following tracks in concentric order starting with outermost "ring": human genome reference track (hg19); large CN changes (colors correspond to different CN, see legend panel); in 1.5 Mb windows, merged SV density (deletions, duplications, inversions) called using BreakDancer, BreakSeq, PINDEL, LUMPY, and Long Ranger; phased haplotype blocks (demarcated with 4 colors for clearer visualization); SNV density in 1 Mb windows; Indel density in 1 Mb windows; dominant zygosity (heterozygous or homozygous > 50%) in 1 Mb windows; regions with loss of heterozygosity; allele-specific expression; CpG islands exhibiting allele-specific DNA methylation; non-reference LINE1 and Alu insertions; allele-specific CRISPR target sites; large-scale SVs resolved by using Long Ranger (peach: intrachromosomal: dark maroon: interchromosomal); by using GROC-SVs (light-purple: intrachromsomal; dark-purple: interchromosomal).

Figure 2. HepG2 Karyogram and Callset Overview

(A) Representative karyogram of HepG2 cells by GTW banding which shows multiple numerical and structural abnormalities including a translocation between the short arms of chromosomes 1 and 21, trisomies of chromosomes 12, 16, and 17, tetrasomy of chromosome 20, uncharacterized rearrangements of chromosomes 16 and 17 and a two marker chromosomes. ISCN 2013 description:
49~52,XY,t(1:21)(p22;p11),+2,+16,add(16)(p13),?+17,?add(17)(p11.2),+20,+20,+1~3mar[cp15]/101~106,idemx2[cp5]. (B) CNs (by percentage) across the HepG2 genome. (C) Percentage of HepG2 SNVs and Indels that are novel and known (in dbSNP). (D) Violin plot with overlaid boxplot of phased haplotype block sizes, with N50 represented as a dashed line (N50 = 6,792,324 bp) with log-scaled Y-axis. (E) X-axis: chromosome coordinate (Mb). Y-axis: difference in unique linked-read barcode counts between major and minor haplotypes, normalized by SNV density. Haplotype blocks from of normal control sample (NA12878) in blue

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

and from HepG2 in dark gray. Density plots on the right reflects the distribution of the differences in haplotype-specific barcode counts for control sample HepG2. Significant difference (one-sided t-test, $p<0.001$) in haplotype-specific barcode counts indicate aneuploidy and haplotype imbalance. Haplotype blocks (with ≥100 phased SNVs) generated from Long Ranger (Dataset 2) for the major and minor haplotypes were then "stitched" to mega-haplotypes encompassing the entire triploid chromosome arms of 2p and 2q.

Figure 3. Large SVs in HepG2 Resolved from Linked-Read Sequencing using Long Ranger

HepG2 SVs resolved by identifying identical linked-read barcodes in distant genomic regions with non-expected barcode overlap for identified using Long Ranger (Marks et al., 2018; Zheng et al., 2016). (A) Disruption of *FRK* by translocation between chromosomes 6 and 16. (B) 2.47 Mb intra-chromosomal rearrangement between *MALRD1* and *MLLT10* on chromosome 10. (C) 127 kb duplication on chromosome 7 resulting in partial duplications of *USP42* and *PMS2.* (D) 395 kb duplication within *PRKG1* on chromosome 10. ~~(D) 194 kb deletion within *PDE4D* on chromosome 5 (E) 286 kb deletion within *AUTS2* on chromosome 7~~ (E) 31.3 kb inversion within *GUSBP1* on chromosome 5. (F) 60.4 kb inversion that disrupts *PPL* and *SEC14L5*.

Figure 4. HepG2 SVs Reconstructed and Assembled Using GROC-SVs in HepG2

(A-D) Each line depicts a fragment inferred from 10X-Genomics data based on clustering of reads with identical barcodes (Y-axis) identified from GROC-SVs (Spies et al., 2017). Abrupt ending (dashed vertical line) of fragments indicates location of SV breakpoint. All breakpoints depicted are validated by 3 kb-mate-pair sequencing data. Fragments are phased locally with respect to surrounding SNVs (haplotype-specific) are in orange, and black when no informative SNVs are found nearby. Gray lines indicate portions of fragments that do not support the current breakpoint. (A) Translocation between chromosomes 1 and 4. Linked-read fragments containing overlapping barcodes that map to chromosome 1 end abruptly near 248.60 mb

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

indicating a breakpoint, and then continues simultaneously near 168.75 mb on chromosome 4. (B) Translocation between chromosomes 6 and 17. Linked-read fragments containing overlapping barcodes that map to chromosome 17 end abruptly near 36.17 mb indicating a breakpoint and then continues simultaneously near 113.52 mb on chromosome 6. (C) Large (335 kb) heterozygous deletion within *NEDD4L* on chromosome 18. (D) Large (1.3 mb) intra-chromosomal rearrangement that deletes large portions of *RBFOX1* and *RP11420N32* on chromosome 16.

Figure 5. Large and complex haplotype-resolved SVs using gemtools (Greer et al., 2017). Each SV is identified from linked-reads clustered by identical barcodes (i.e. SV-specific barcodes, Y-axis) indicative of single HMW DNA molecules (depicted by each row) that span SV breakpoints. Haplotype-specific SVs are represented in blue and red. X-axis: hg19 genomic coordinate. (Top) Complex SV on chromosome 8 involving a 4585 bp deletion downstream of *ADAM2*. This deletion is within a tandem duplication leading to the amplification of the *IDO1* and the first half of *IDO2*. The presence of HMW molecules sharing the same linked-read barcodes spanning both breakpoints indicates a *cis* orientation and occurrence on only one allele of this locus. Schematic diagram of the rearranged structures drawn above the plot. (Middle) Two haplotype-resolved deletions 700 kb (blue) and 200 kb (red) respectively occurring on two separate alleles within of *PDE4D* on chromosome 5 – the spanning HMW molecules for each deletion do not share SV-specific barcodes, indicating that these deletions are in *trans*. (C) Two haplotype-resolved deletions, 290 kb (red) and 160 kb (blue) respectively, within *AUTS2* on chromosome 7. The reference allele of *AUTS2* without the deletion (Haplotype 2) is also detected and resolved by linked-reads (blue, bottom panel). The 160 kb deletion on Haplotype 2 occurs sub-clonally.

Figure 6. Genomic Sequence and Structural Context Provides Insight into Regulatory Complexity in HepG2

(A) Chr5:57,755,334-57,756,803 locus containing the serine/threonine-protein kinase gene

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

PLK2 and CGI 6693 (1,463 bp) where phased Haplotype 1 and Haplotype 2. Allele-specific transcription of *PLK2* from Haplotype 2 only. CpGs in CGI 6693 are mostly unmethylated in Haplotype 2 (expressed) and highly methylated in Haplotype 1 (repressed). (B) Chr17:59,473,060-59,483,266 locus (triploid in HepG2) containing T-box transcription factor gene *TBX2* and CpG Island (CGI) 22251 (10,206 bp) where phased Haplotype 2 has two copies and Haplotype 1 has one copy. Allele-specific transcription of *TBX2* from Haplotype 2 only. CpGs in CGI 22251 are unmethylated in Haplotype 1 (repressed) and methylated in Haplotype 2 (expressed). Allele-specific CRISPR targeting site 1937 bp inside the 5' region of *TBX2* for both Haplotypes. (C) Number of allele-specific RNA-Seq reads in Haplotypes 1 and 2 for *PLK2* and *TBX2* where both genes exhibit allele-specific RNA expression (p=0.4.66E-10 and p=0.0179 respectively). (D) Number of methylated and unmethylated phased whole-genome bisulfite-sequencing reads for Haplotypes 1 and 2 in CGI 6693 and CGI 22251 where both CGIs exhibit allele-specific DNA methylation (p=1.51E-66 and p=1.55E-32 respectively).

Figure S1. Illustration of the Sequencing-Based Methodologies Used

Deep short-insert WGS (70x non-duplicate coverage), 3 kb-mate-pair sequencing (Korbel et al., 2007), and 10X-Genomics linked-reads sequencing (Zheng et al., 2016) were used to comprehensively characterize the genome of HepG2. The WGS dataset was used to obtain CN i.e. ploidy by chromosome segments using read-depth analysis (Abyzov et al., 2011), SNVs and Indels using GATK Haplotypecaller with CN taken into account (McKenna et al., 2010), non-reference LINE-1 and Alu insertions (Keane et al., 2013), and SVs, such as deletions, duplications, inversions, and insertions using BreakDancer (Chen et al., 2009), BreakSeq (Lam et al., 2010), and Pindel (Ye et al., 2009). The linked-read sequencing data was used to phase heterozygous SNVs and Indels as well as to identify, assemble, and phase large (>30 kb) and SVs using Long Ranger (Zheng et al., 2016) and GROC-SVs (Spies et al., 2017). Deletions <30 kb were also identified by Long Ranger. The SV assembly file from GROC-SVs is in BAM format (Dataset 5). The 3 kb-mate-pair sequencing data was used to call additional structural

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

variants, mostly in the medium size-range (1 kb-100 kb) using LUMPY (Layer et al., 2014) and also used to validate large and complex SVs. The union of non-complex SVs identified merged into a single VCF file (Supplementary Data).

Figure S2. HepG2 WGS coverage vs. % GC content

Y-axis: HepG2 short-insert WGS coverage in 10 kb bins across the genome; X-axis: % GC content of bins. coverage). Clusters correspond to CN (i.e. ploidy).

Figure S3. Orthogonal Validation of CN as Determined by Read-Depth Analysis in HepG2 by Illumina MEGA Array (2 Replicates) (Chromosomes 16 & 17)

(A) Upper panel: WGS coverage plot. X-axis genomic coordinate in kb. Y-axis: WGS coverage. Purple: CN4, Blue: CN3, Green: CN2. Lower panel: Y-axis: array probe signal intensity. X-axis: Genomic coordinate. For complete chromosomes (1-22, X), see Supplementary Data. (B) Copy numbers of genomic regions on chromosome 16 of HepG2. X-axis genomic coordinate in kb. Y-axis: WGS coverage. Purple: CN4, Blue: CN3, Green: CN2.

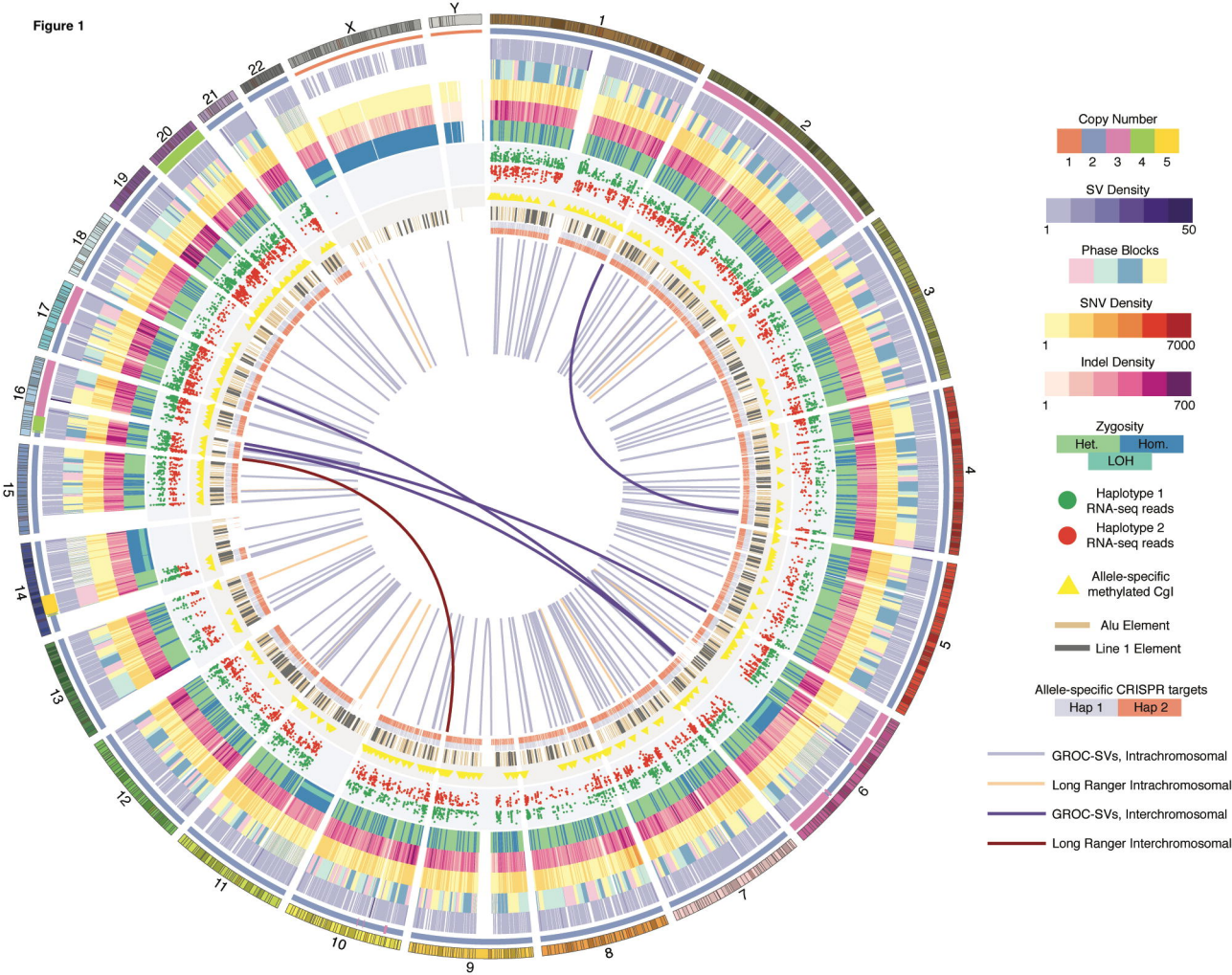
Figure S4. Size Distribution of Phased Haplotype Blocks by Chromosome

Violin plots (with overlaid boxplot) of HepG2 phased haplotype blocks (Dataset 2) by chromosomes. Y-axis: size in log-scale.

Figure S5. Overlap Between SV Callers

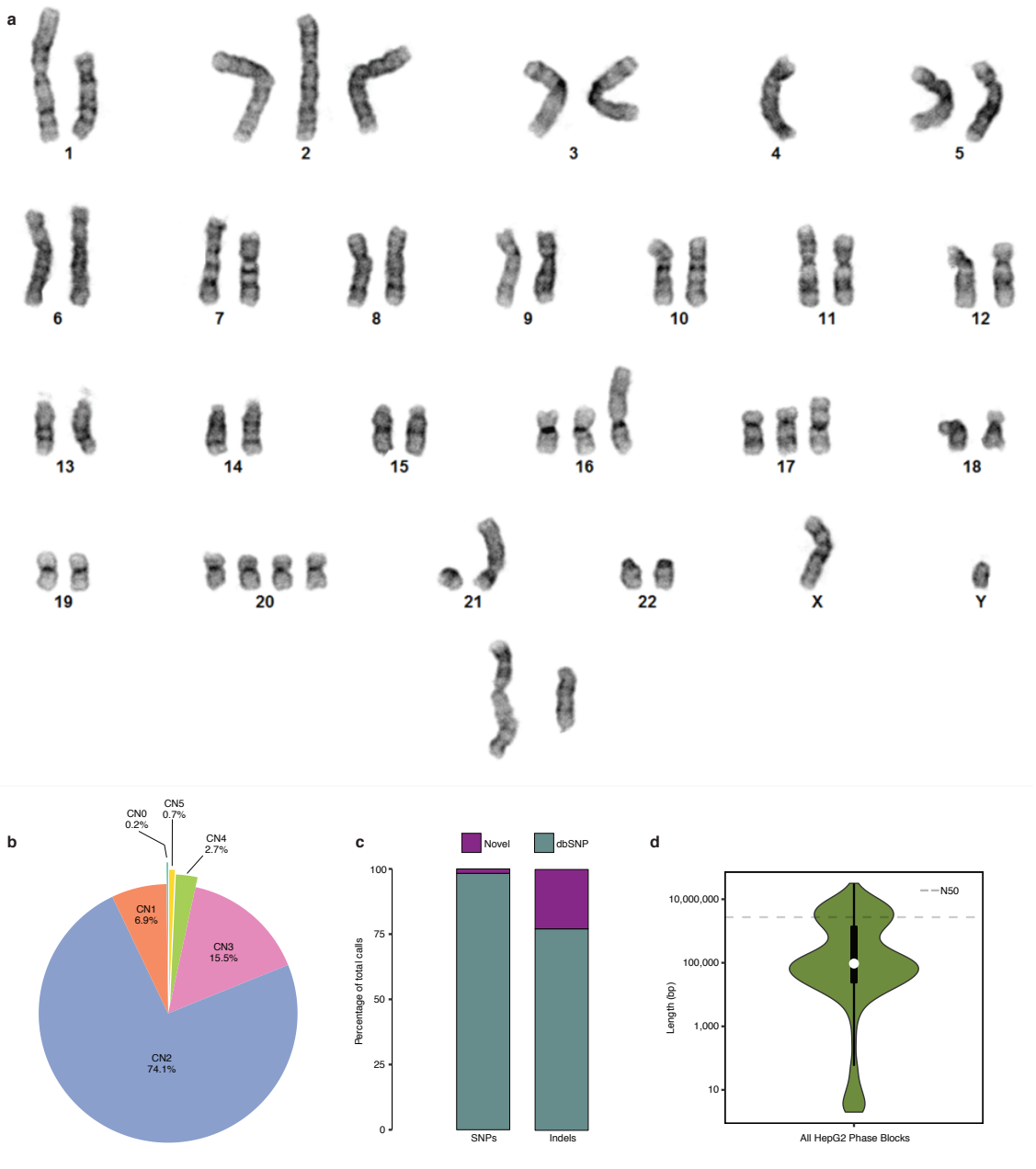
Venn diagram of overlaps (>50% reciprocal) between HepG2 SVs identified in WGS using BreakDancer (Chen et al., 2009), Pindel (Ye et al., 2009), and BreakSeq (Lam et al., 2010), in 3kb-mate-pair sequencing using LUMPY (Layer et al., 2014), and in linked-read sequencing using Long Ranger (Marks et al., 2018; Zheng et al., 2016) for (A) deletions (>50 bp), (B) tandem duplications, and (C) inversions.

Figure 1


X
Y
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Copy Number
1 2 3 4 5
SV Density
1
50
Phase Blocks
SNV Density
1
7000
Indel Density
1
700
Zygosity
Het.
Hom.
LOH
Haplotype 1 RNA-seq reads
Haplotype 2 RNA-seq reads
Allele-specific methylated CgI
Alu Element
Line 1 Element
Allele-specific CRISPR targets
Hap 1
Hap 2
GROC-SVs, Intrachromosomal
Long Ranger Intrachromosomal
GROC-SVs, Interchromosomal
Long Ranger Interchromosomal

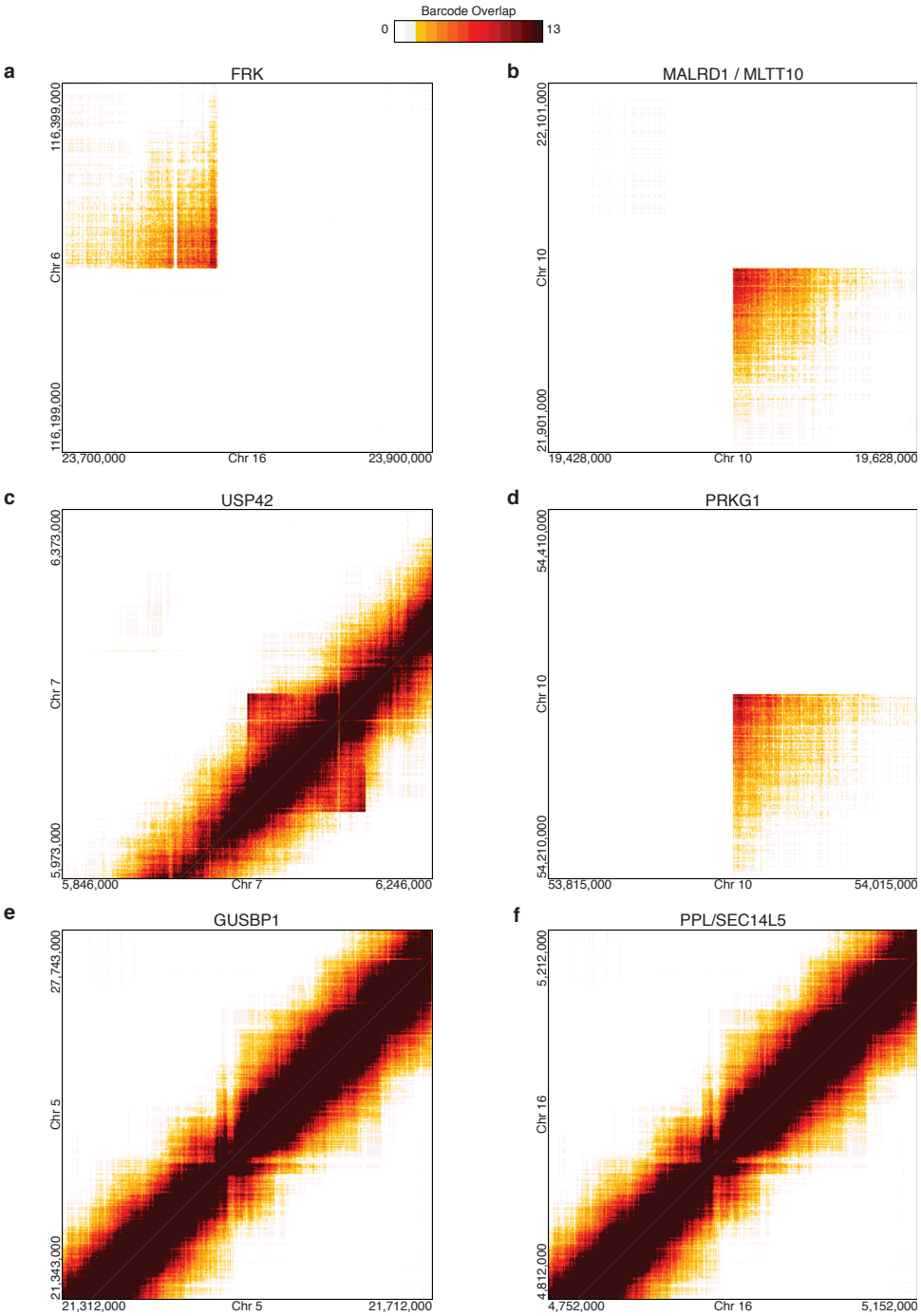
bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2



bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

Figure 3



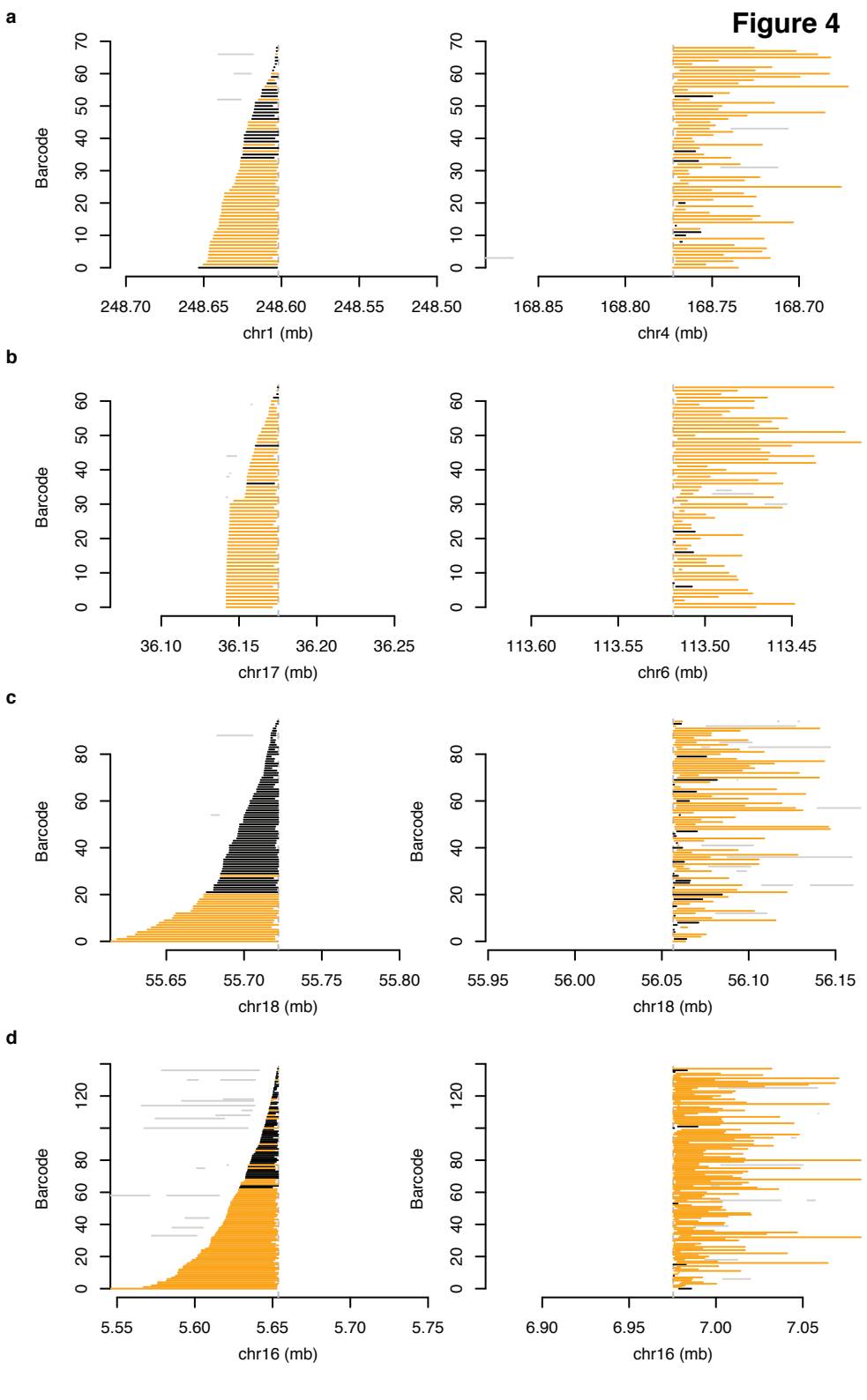


Figure 4
a
Barcode
0
10
20
30
40
50
60
70
248.70
248.65
248.60
248.55
248.50
chr1 (mb)
168.85
168.80
168.75
168.70
chr4 (mb)
b
Barcode
0
10
20
30
40
50
60
36.10
36.15
36.20
36.25
chr17 (mb)
113.60
113.55
113.50
113.45
chr6 (mb)
c
Barcode
0
20
40
60
80
55.65
55.70
55.75
55.80
chr18 (mb)
55.95
56.00
56.05
56.10
56.15
chr18 (mb)
d
Barcode
0
20
40
60
80
120
5.55
5.60
5.65
5.70
5.75
chr16 (mb)
6.90
6.95
7.00
7.05
chr16 (mb)

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

Figure 5

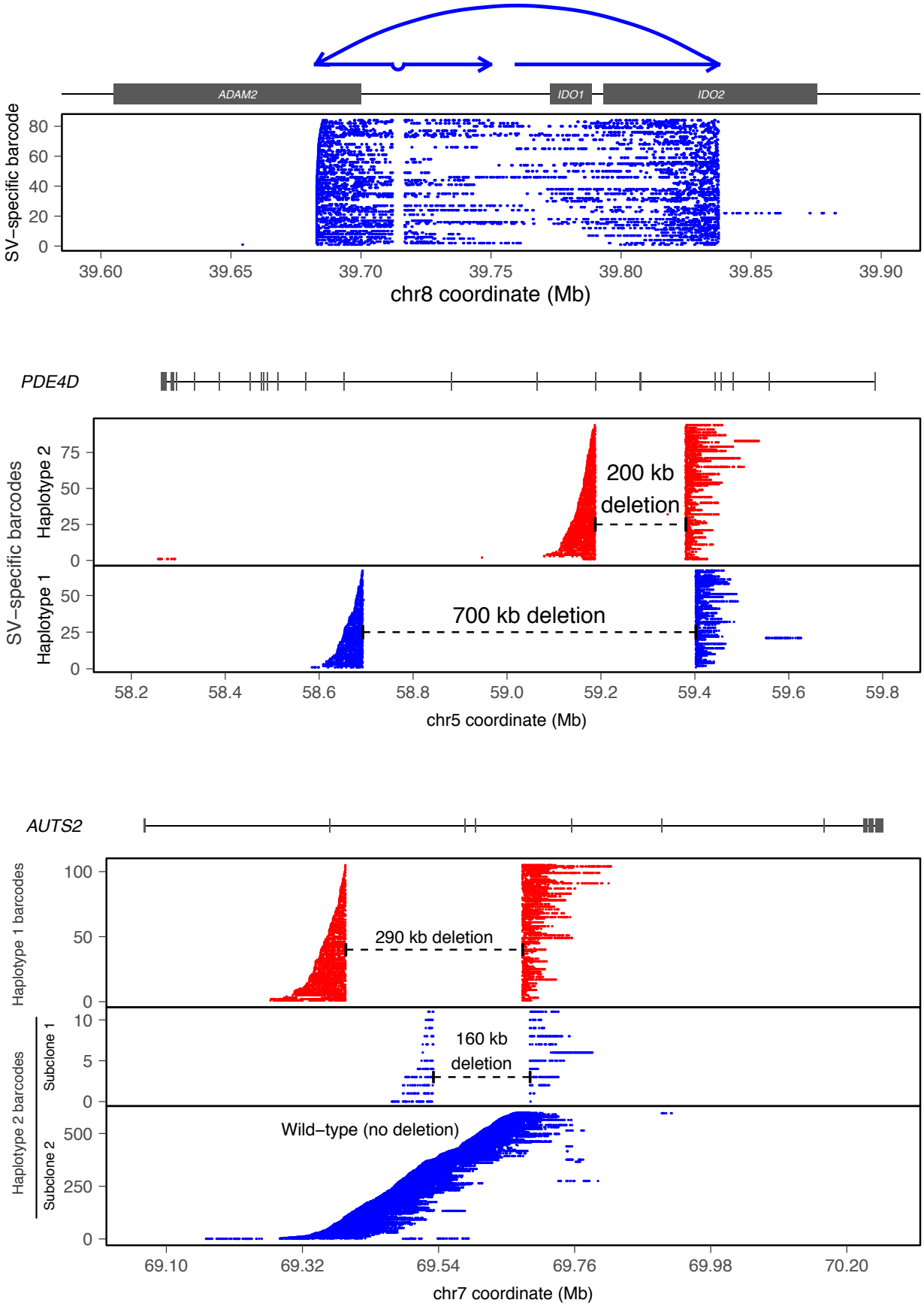
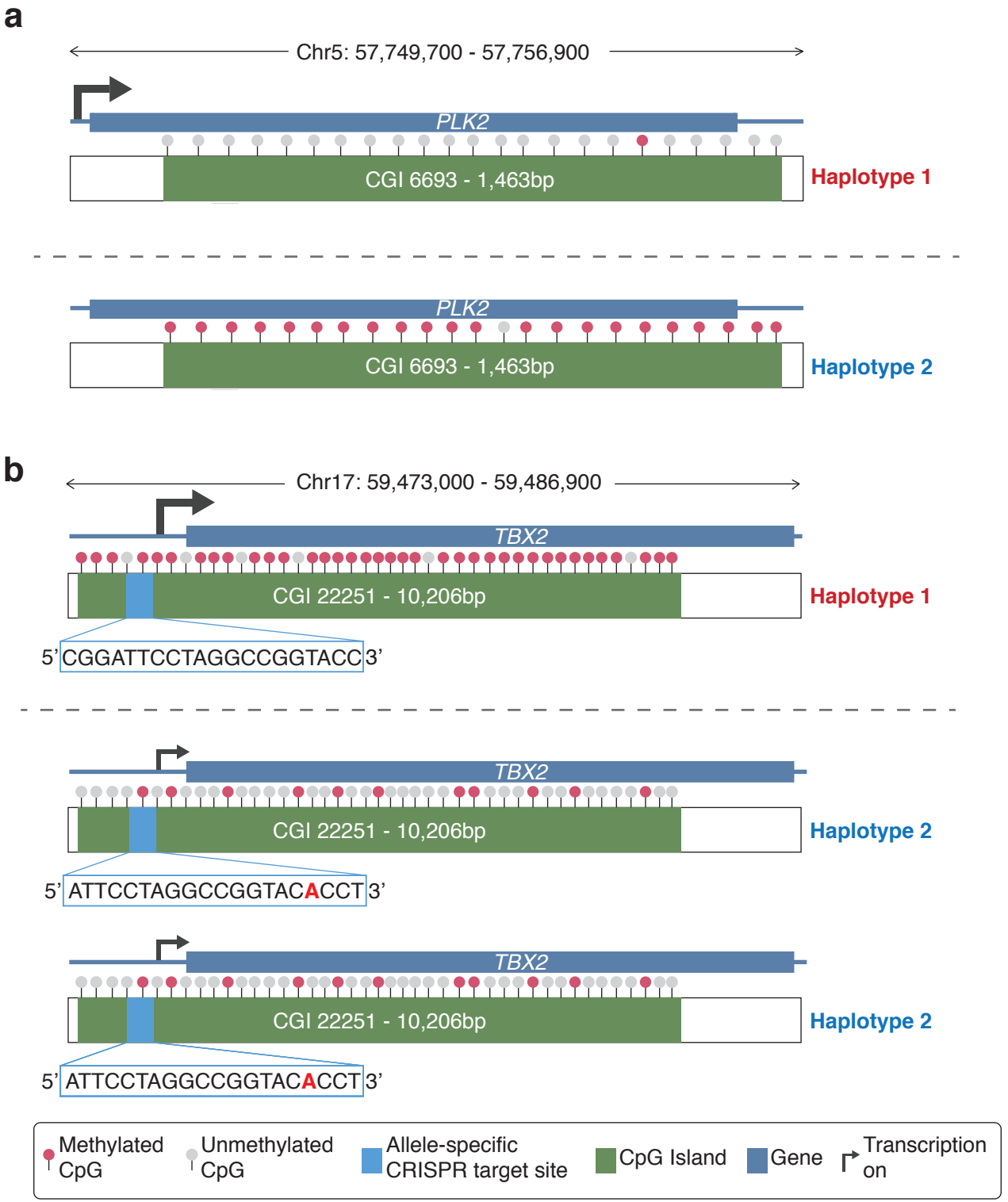
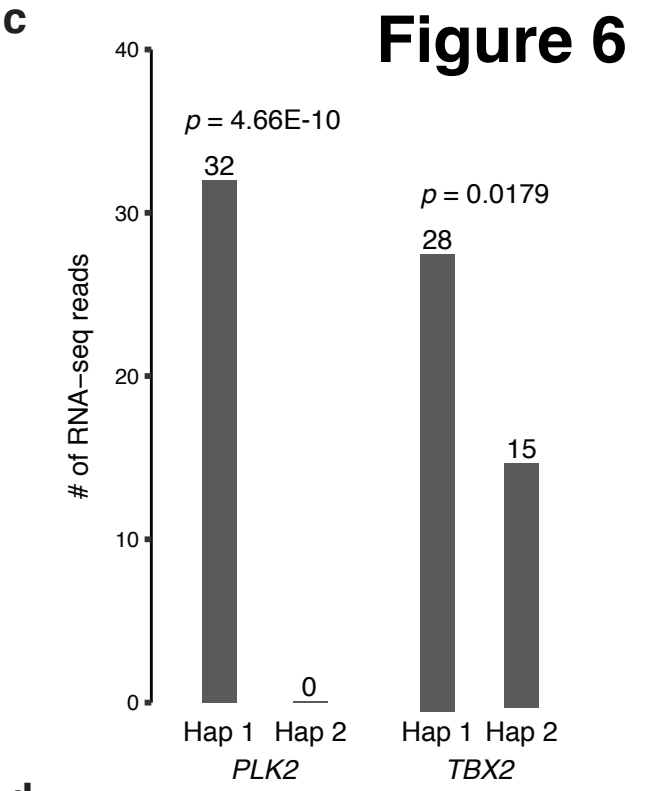
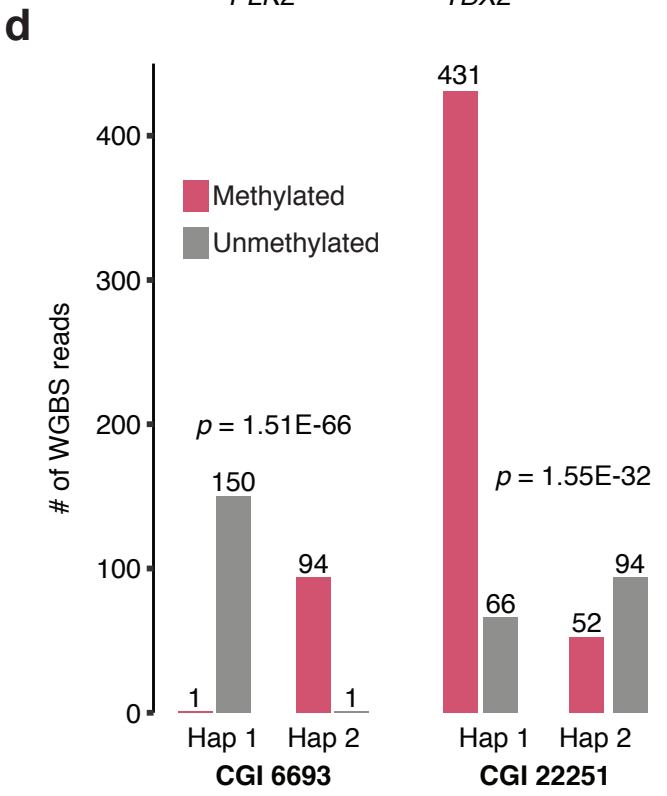


Figure 6


a
Chr5: 57,749,700 - 57,756,900
PLK2
CGI 6693 - 1,463bp
Haplotype 1
PLK2
CGI 6693 - 1,463bp
Haplotype 2
b
Chr17: 59,473,000 - 59,486,900
TBX2
CGI 22251 - 10,206bp
Haplotype 1
5' CGGATTCCTAGGCCGGTACC 3'
TBX2
CGI 22251 - 10,206bp
Haplotype 2
5' ATTCCTAGGCCGGTACACCT 3'
TBX2
CGI 22251 - 10,206bp
Haplotype 2
5' ATTCCTAGGCCGGTACACCT 3'
Methylated CpG
Unmethylated CpG
Allele-specific CRISPR target site
CpG Island
Gene
Transcription on


c
# of RNA−seq reads
40
30
20
10
0
p = 4.66E-10
32
0
p = 0.0179
28
15
Hap 1
Hap 2
PLK2
Hap 1
Hap 2
TBX2


d
# of WGBS reads
400
300
200
100
0
Methylated
Unmethylated
p = 1.51E-66
1
150
94
1
p = 1.55E-32
431
66
52
94
Hap 1
Hap 2
CGI 6693
Hap 1
Hap 2
CGI 22251

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

Table 1 - Summary of HepG2 Small Variant Calls and Phasing Results

	SNPs	INDELs
All	3337361	892019
Heterozygous/homozygous	1898493/1438868	598882/293137
Protein altering	11460 (0.3%)	1347 (0.2%)
dbSNP138	3279135 (98%)	693348 (78%)
Heterozygous/homozygous	1845345/1433790	439143/254205
Novel	58226 (2%)	198671 (22%)
Heterozygous/homozygous	53148/5078	159739/38932
1000 Genomes Project & Exome Sequencing Project Overlap (with protein altering variants)	11083 (97%)	1092 (81%)
Novel Protein Altering	377	255
COSMIC Overlap	148 (39%)	42 (16%)

Phased WGS	
% phased heterozygous SNPs	99
% phased INDELs	78
Longest phase block	31106135
Number of phase blocks	1628
N50 phase block	6792324
N50 Linked-reads per molecule	61
Barcodes detected	1532287
Mean DNA per barcode (bp)	633889

Mega-haplotypes

Chromosome	Start	End	Chromosome Arm	% of Arm Covered	*p*-Value
2	21,888	89,128,628	2p	98%	2.20E-16
2	98,803,025	243,046,591	2q	98%	2.20E-16
6	269,211	56,501,036	6p	96%	8.70E-07
6	62,383,957	170,631,019	6q	99%	3.92E-13
16	46,511,762	90,230,343	16q	99%	3.87E-05
17	34,819,191	80,982,386	17q	83%	4.25E-06

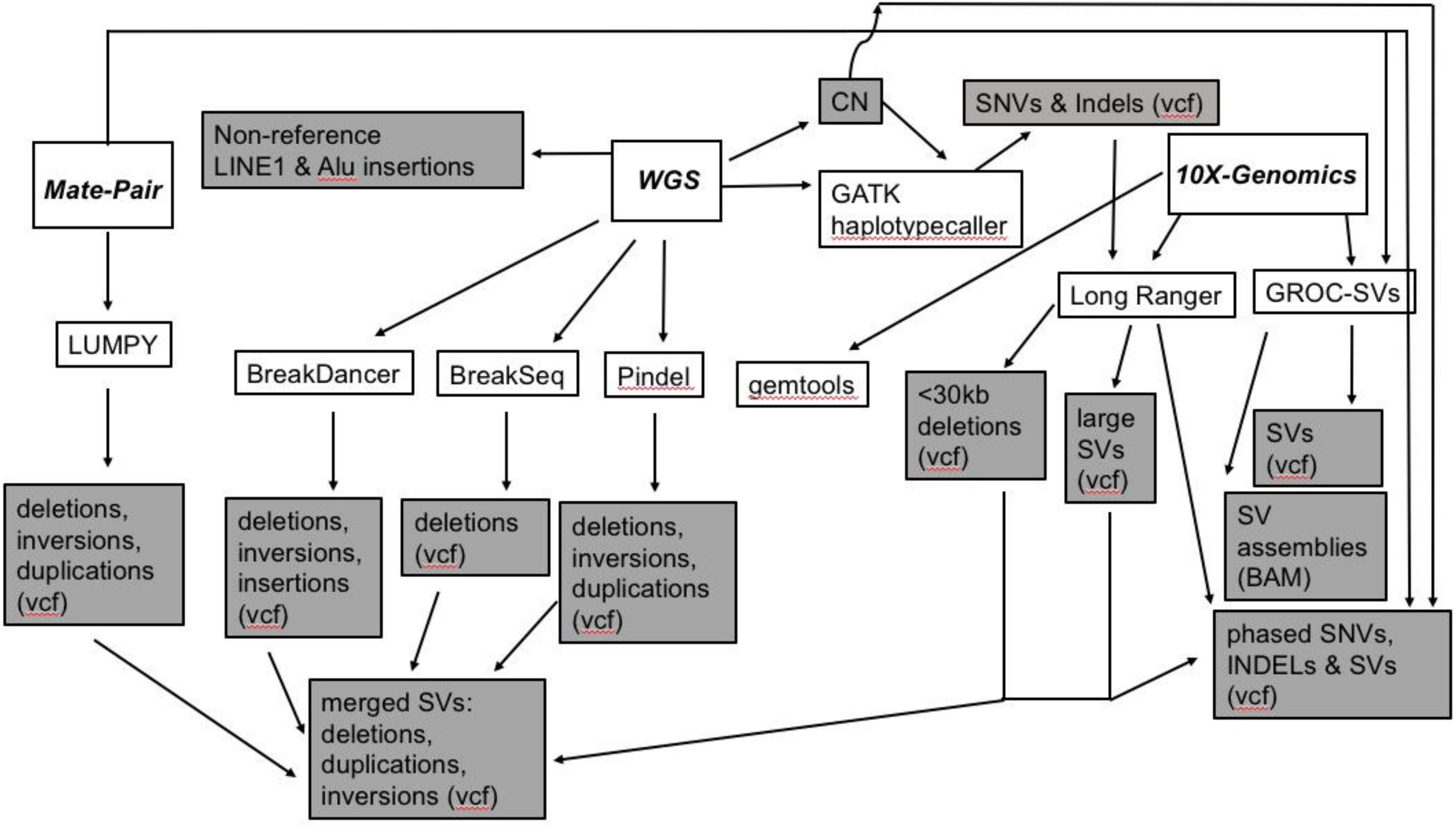
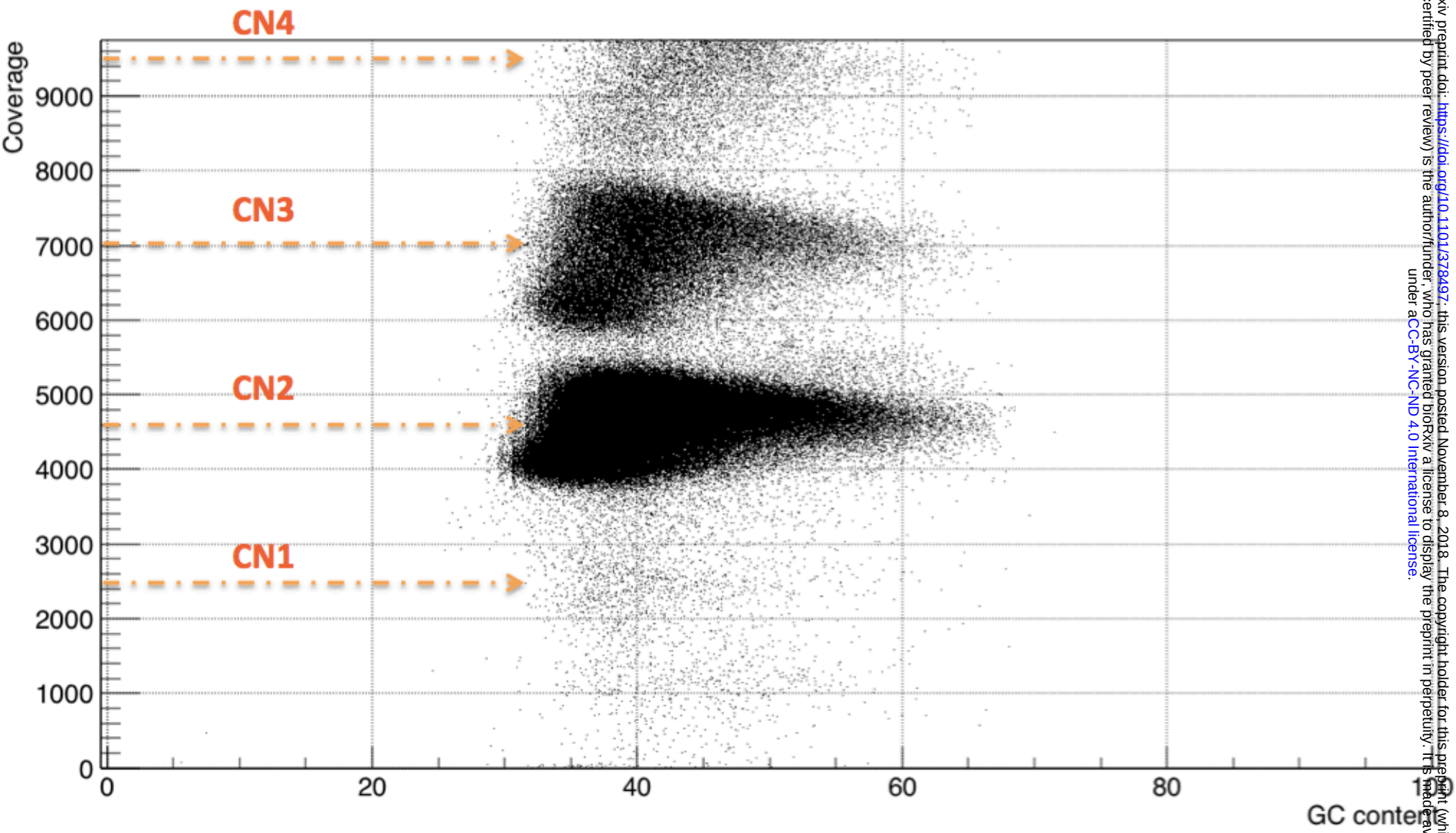

Mate-Pair
Non-reference LINE1 & Alu insertions
WGS
CN
SNVs & Indels (vcf)
GATK haplotypecaller
10X-Genomics
LUMPY
BreakDancer
BreakSeq
Pindel
gemtools
Long Ranger
GROC-SVs
<30kb deletions (vcf)
large SVs (vcf)
SVs (vcf)
SV assemblies (BAM)
deletions, inversions, duplications (vcf)
deletions, inversions, insertions (vcf)
deletions (vcf)
deletions, inversions, duplications (vcf)
phased SNVs, INDELs & SVs (vcf)
merged SVs: deletions, duplications, inversions (vcf)

Figure S2

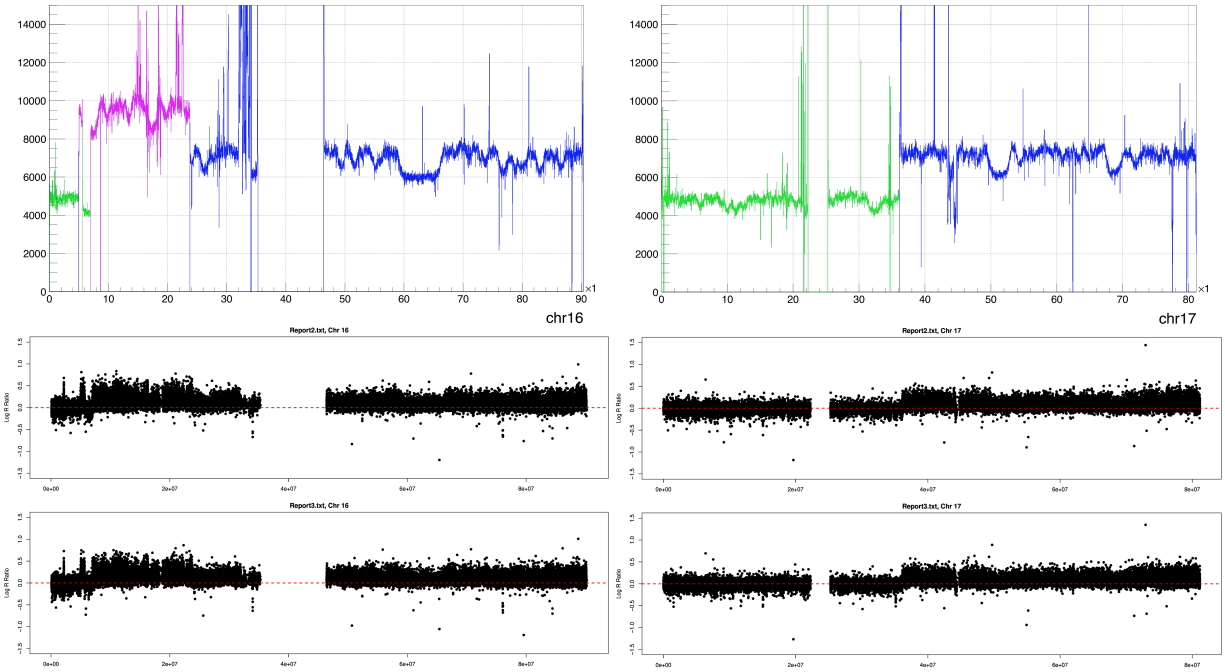


bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 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-NC-ND 4.0 International license.

Figure S3

A



B

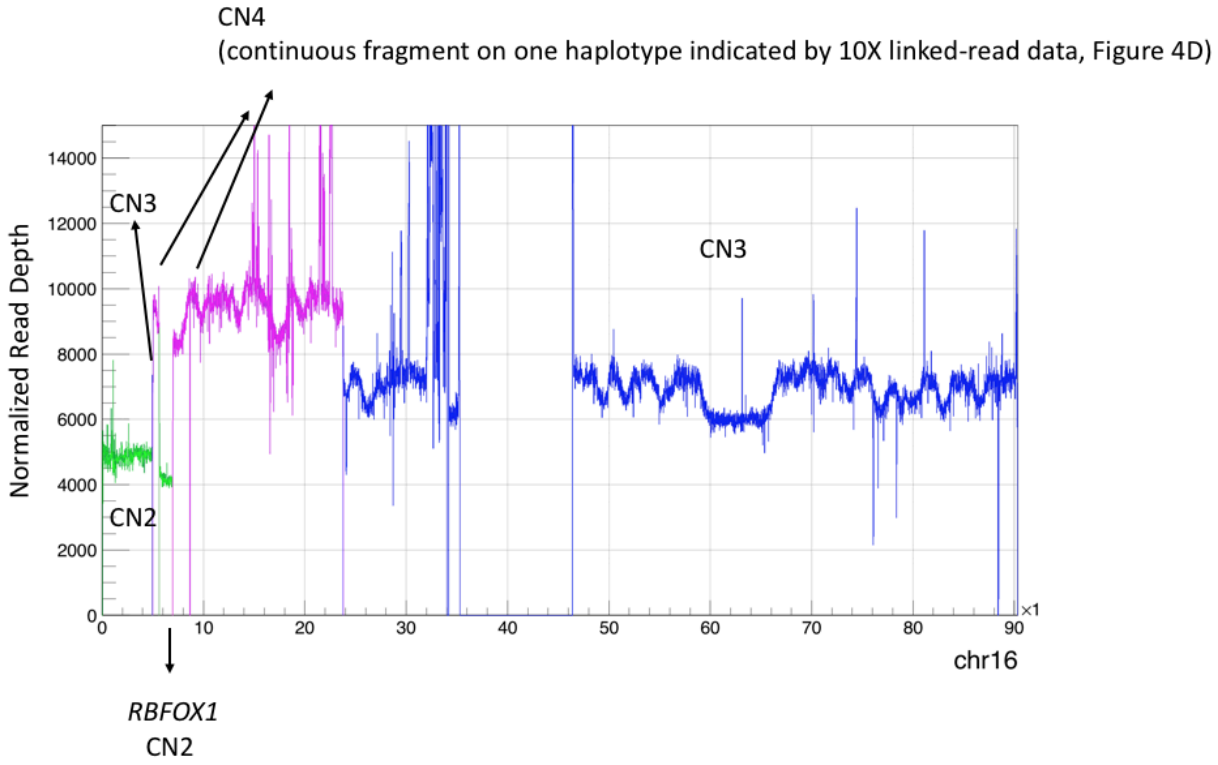
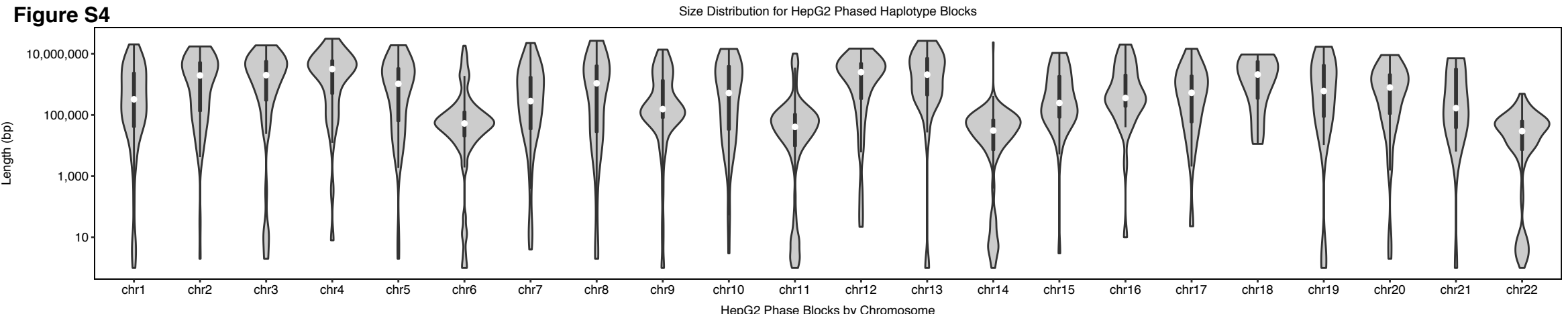


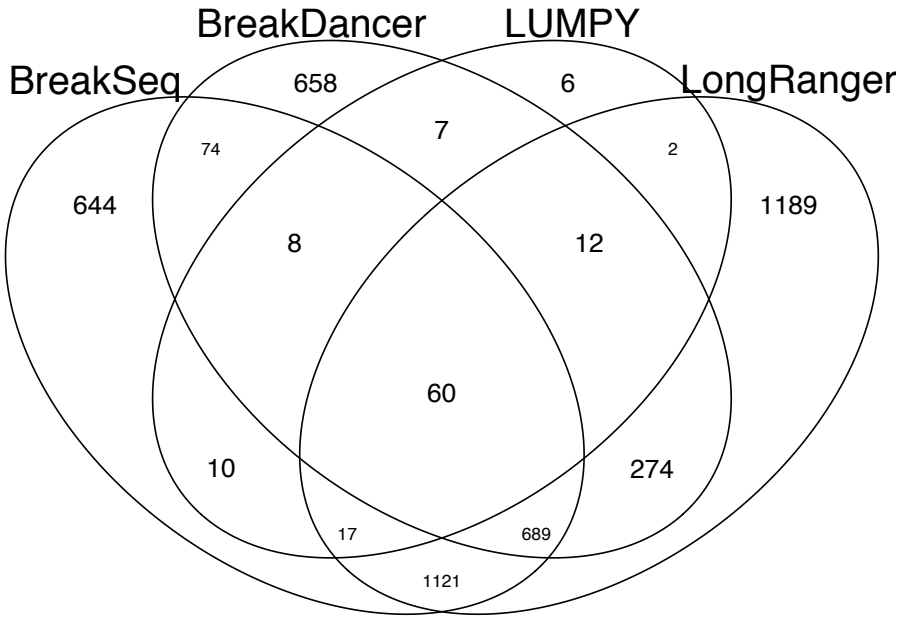
Figure S4


Size Distribution for HepG2 Phased Haplotype Blocks
Length (bp)
10,000,000
100,000
1,000
10
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
HepG2 Phase Blocks by Chromosome

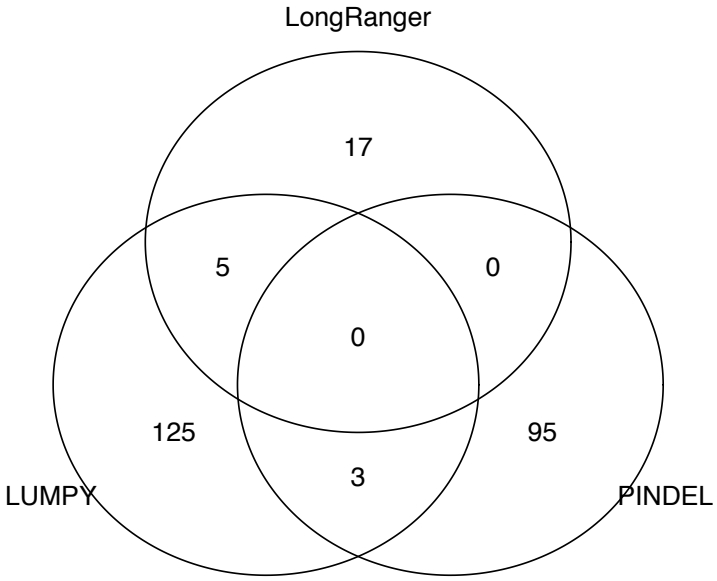
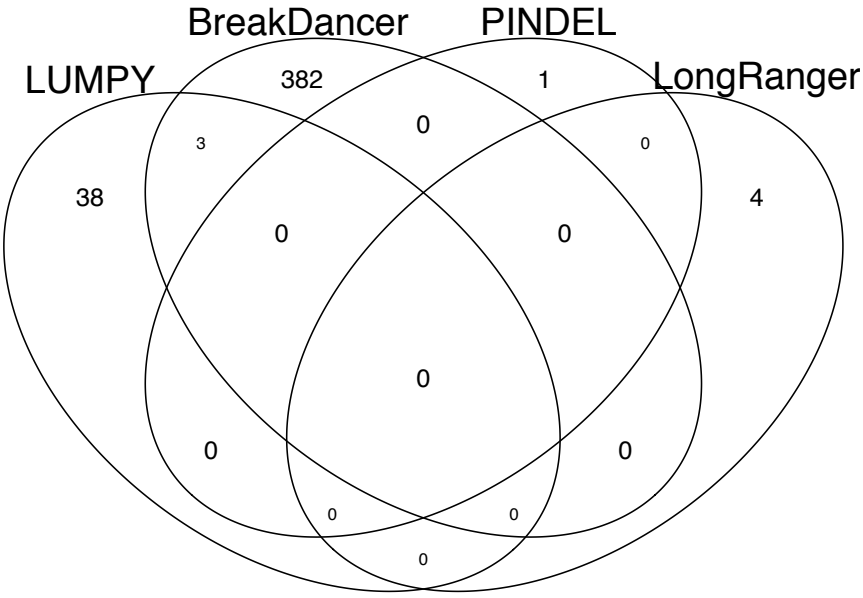
bioRxiv preprint doi: https://doi.org/10.1101/378497; this version posted November 8, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S5

A



B



C