Haplotype-resolved and integrated genome analysis of ENCODE 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], Giltae Song[8], Hanlee P. Ji[3,6], Dimitri Perrin[9], Wing H. Wong[7,10], 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.

Corresponding author:

Alexander E. Urban
Program on Genetics of Brain Function
Stanford Center for Genomics and Personalized Medicine
Tasha and John Morgridge Faculty Scholar
Stanford Child Health Research Institute
3165 Porter Drive, Room 2180
Palo Alto, CA 94304-1213
USA
aeurban@stanford.edu

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

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

ABSTRACT

HepG2 is one of the most widely used human cell lines in biomedical research and one of the main cell lines of ENCODE. Although the functional genomic and epigenomic characteristics of HepG2 are extensively studied, its genome sequence has never been comprehensively analyzed and higher-order structural features of its genome beyond its karyotype were only cursorily known. The high degree of aneuploidy in HepG2 renders traditional genome variant analysis methods challenging and partially ineffective. Correct and complete interpretation of the extensive functional genomics data from HepG2 requires an understanding of the cell line's genome sequence and genome structure. We performed deep whole-genome sequencing, mate-pair sequencing and linked-read sequencing to identify a wide spectrum of genome characteristics in HepG2: copy numbers of chromosomal segments, SNVs and Indels (both corrected for copy-number), phased haplotype blocks, structural variants (SVs) including complex genomic rearrangements, and novel mobile element insertions. A large number of SVs were phased, sequence assembled and experimentally validated. Several chromosomes show striking loss of heterozygosity. We re-analyzed HepG2 RNA-Seq and whole-genome bisulfite sequencing data for allele-specific expression and phased DNA methylation. We show examples where deeper insights into genomic regulatory complexity could be gained by taking knowledge of genomic structural contexts into account. Furthermore, we used the haplotype information to produce an allele-specific CRISPR targeting map. This comprehensive whole-genome analysis serves as a resource for future studies that utilize HepG2.

INTRODUCTION

HepG2 is a hepatoblastoma cell line derived from a 15-year-old Caucasian male (Aden et al. 1979; López-Terrada et al. 2009a). 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 one of the main cell lines of the ENCyclopedia Of DNA Elements Project (ENCODE), HepG2 has contributed to over 800 datasets on the ENCODE portal (Sloan et al. 2016) and to over 23,000 publications to date, even more than K562. In addition to being a main ENCODE cell line, the HepG2 cell line is commonly used in other areas of biomedical research due to its extreme versatility. Representing the human endodermal lineage, HepG2 cells are widely used as models for human toxicology studies (Schoonen et al. 2005; Sahu et al. 2012; Dias da Silva et al. 2013; Menezes et al. 2013; Kamalian et al. 2015), 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).

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. 2016b). However, the sequence and higher-order structural features of the HepG2 cancer genome have never been analyzed and documented in a comprehensive manner, even though the HepG2 cell line has been known to contain multiple chromosomal abnormalities (Simon et al. 1982; Chen et al. 1993). As a result, the extensive HepG2 functional genomics and epigenomics studies conducted to date were done without reliable genomic contexts for accurate 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.

RESULTS

Method Overview

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. In addition, SVs, such as deletions, duplications, inversions, and non-reference retrotransposon (LINE1 and Alu) insertions, were also identified 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 (Spies et al. 2017; Zheng et al. 2016; Marks et al. 2018). 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.

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, 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 (Simon et al. 1982; Chen et al. 1993; 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 S3, 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

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* (Xia et al. 2012; Kai et al. 2014). *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

We phased the heterozygous SNVs and indels in the HepG2 genome by performing 10X Genomics Chromium linked-read library preparation and sequencing (Zheng et al. 2016; Marks et al. 2018). 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).

Using Linked-Reads to Identify and Reconstruct 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 (Zheng et al. 2016; Marks et al. 2018). 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 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). This deletion is not only extremely interesting in and of itself, but also provides profound insight into the mutational history of HepG2 genome (see Discussion).

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 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) (Supplemental 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, Supplemental 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

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 S1), 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 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 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

HepG2 is one of the most widely used laboratory cell lines in biomedical research and one of the main cell lines of ENCODE. However, despite its wide usage, its genomic sequence and structural features have never been characterized in a comprehensive manner beyond its karyotype (Simon et al. 1982; Chen et al. 1993) and SNVs identified from ChIP-Seq data and 10x coverage WGS that do not take aneuploidy or CN into consideration (Huang and Ovcharenko 2015; Cavalli et al. 2016). 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 S1), Indels (Dataset S1), large CN or ploidy changes across chromosomal regions at 10 kb resolution (Table S2), phased haplotype blocks (Dataset S2), phased CRISPR targets (Table S12), novel retrotransposon insertions (Table S8), and SVs (Dataset S3, Dataset S4, Dataset S5, Dataset S6) 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 S5, Table S7, and Table S8).

Information regarding the various regulatory pathways that might contribute to HepG2 cancer progression is already evident from our very general and first-pass analysis of these datasets. For example, we identified CN changes (CN=3) over the oncogene *VEGFA* (6p21.1) which is widely known as a key mediator of angiogenesis in cancer (Carmeliet 2005). We see mutations in oncogenes and tumor-suppressor genes in HepG2, 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 in the Wnt-pathway gene *CTNNB1* due to the presence of either HepG2-specific protein-altering changes or small-scale structural changes (Table S5, Table S6, Dataset 1, Dataset 4, Supplemental Data). Moreover, we also identified mutations in genes recently found to play critical roles in driving cancer such as *CDK12* (Paculová and Kohoutek 2017) and *IKBKB* (Xia et al. 2012; Kai et al. 2014). These mutations also implicate the dis-regulation of their associated molecular pathways in HepG2. For *CTNNB1*, dis-regulation and mutations have been extensively documented for hepatoblastoma in general (Bläker et al. 1999; Taniguchi et al. 2002; Curia et al. 2008; López-Terrada et al. 2009b; Jia et al. 2014; Crippa et al. 2017).

From the large-scale SVs resolved using linked-read sequencing, we identified structural disruptions of *FRK* (Figure 3A), *PMS2*, *USP42* (Figure 3C), *PRKG1* (Figure 3D), *PDE4D* (Figure 3E), *NEDD4L* (Figure 4C), and *RBFOX1* (Figure 4D). The non-receptor Fyn-related tyrosine kinase FRK is distantly related to the Src family kinases.

FRK resides in a region on 6q22.31 that is known to be destabilized by LOH in malignant melanomas (Becher et al. 1983), lung cancers (Girard et al. 2000), and nearly 50% of breast tumors (Sheng et al. 1996). In HepG2, we observed a partial LOH for the 5' portion *FRK* (Table S3). Since its discovery in 1993, *FRK* was first attributed as a tumor suppressor (Craven et al. 1995; Oberg-Welsh et al. 1998; Brauer and Tyner 2009; Kim et al. 2015); however, further investigations show that *FRK* appears to play a tissue-specific oncogenic role in liver, pancreatic, breast and brain cancers and has been proposed as a therapeutic target (Goel and Lukong 2016). Interestingly, it was shown ectopic overexpression of *FRK* in HepG2 had no effect on cell proliferation, while this overexpression results in a significant growth reduction (p-value < 0.05) was observed in HepG2 cells with *FRK*-knockdown (Chen et al. 2013). The truncation of *FRK* by the translocation of chromosomes 6 and 16 in HepG2 (Figure 3A) suggests that a truncated form of *FRK* might also be expressed in HepG2 cells, which influences the interpretation of results from previous knockdown and overexpression assays of *FRK* carried out in HepG2 cells (Chen et al. 2012, 2013). It is possible that this truncated form of *FRK* confers new functions to the protein, leading to oncogenic properties. It is also plausible that the translocation, while "knocking-out" one normal allele of *FRK* which might play a tumor-suppressive role, concomitantly allows the truncated form to be expressed from another location at a higher level. This would explain why ectopic overexpression of a non-truncated allele of *FRK* reduces HepG2 proliferation only in *FRK*-knockdown cells as demonstrated in (Chen et al. 2013). This scenario might very well be the case in other types of cancers, hence confounding the interpretations of results that suggest *FRK* acts as an oncogene in certain tissue types while acting as a

tumor-suppressor gene in others. There may be other forms of *FRK* that play different roles in different cancers. Only further studies or new interpretations of existing data taking this knowledge which we have presented here into account will resolve this complicated issue.

USP42 is a deubiquitinating enzyme that is widely expressed across human tissues (Quesada et al. 2004) and has also been implicated in the pathogenesis of many types of tumors (Fraile et al. 2012). USP42 overexpression has recently been shown to be a potential prognostic maker and a possible therapeutic target for gastric cancer (Hou et al. 2016a). It is plausible that one of the mechanisms behind USP42 overexpression in various cancers is partly due to gene duplication (Figure 3C). PRKG1 activity was found to be important for promoting DNA synthesis and cell proliferation in human ovarian cancer (Leung et al. 2010). It is also plausible that the normal function of these pathways in HepG2 are dis-regulated due to the disruption of *PRKG1* by its internal rearrangement (Figure 3D). Homozygous microdeletions within the phosphodiesterase subtype 4D gene *PDE4D* frequently occur in primary solid tumors where, unexpectedly, PDE4D protein expression is up-regulated rather than lost (Lin et al. 2013). Ectopic expression of the shorter isoforms of truncated *PDE4D* can enhance the *in vitro* and *in vivo* proliferation of cancer cells, and endogenous shRNA knockdown of PDE4D expression can inhibit the growth of breast, lung, ovary, endometrium, gastric and melanoma cancer cells. This growth inhibition can also be reverted by allowing PDE4D expression to restore. It has been further proposed that this enzyme could be a valuable target for cancer therapeutics. HepG2 also harbor a homozygous deletion (193 kb) within *PDE4D*, deleting six internal exons (Figure 3D),

which might lead to up-regulation of the truncated form of *PDE4D* and enhance its proliferative capabilities.

NEDD4L is an ubiquitin ligase of the *NEDD4* family and is the closest homologue to *NEDD4*, which play a critical regulatory role in the development and progression of many human cancers (Ye et al. 2014). NEDD4L is a regulator of the WNT pathway (Ding et al. 2013; Zhang et al. 2014), and its expression is down-regulated in ovarian and colorectal cancers (Tanksley et al. 2013; Yang et al. 2015). Its down-regulation has also been associated with the aggressive progression and worsening prognosis of malignant glioma and gastric cancer (Gao et al. 2012; He et al. 2012). The large heterozygous deletion within *NEDD4L* very likely also leads to its reduced protein expression in HepG2 cells, providing another example where down-regulation of *NEDD4L* is associated with aggressive cancer cell progression.

Deletions involving *RBFOX1* (also known as *A2BP1*) are very common in human cancers (Dereli-Öz et al. 2011; Rajaram et al. 2013), but remain one of the "unexplained" deletions as to the functionality or selective advantage that they may confer in cancer cells as *RBFOX1* is neither a tumor suppressor gene nor locates in a common fragile site (Bignell et al. 2010; Rajaram et al. 2013). It is most commonly found in colorectal cancers and also occurs at significant frequencies in liver, lung, and ovarian cancers (Rajaram et al. 2013). This phased SV is not only extremely interesting in and of itself but also, perhaps on a more powerful level, serves as a point of entry in piecing together the mutation history of chromosome 16 in HepG2.

This rearrangement involving *RBFOX1* on chromosome 16 is validated by both WGS and 3 kb-mate-pair sequencing. It appears to be a "simple" heterozygous deletion

event. However, the resulting ploidy of the rearranged chromosomal region is 2 and is flanked by two long stretches of CN4 regions on chromosome 16: 5.00 mb to 5.64 mb upstream (640 kb) and 6.96 mb to 23.79 mb downstream (16.8 mb). These two CN4 stretches are then flanked by two CN3 stretches: 4.95 mb to 5.00 mb upstream (50 kb) and 23.79 mb to 90.30 mb downstream (66.51 mb). Since one haplotype is affected by this rearrangement and two haplotypes remain in the resulting CN2 (diploid) region with flanking CN4 and CN3 regions, it is plausible that one original haplotype was first duplicated once resulting in a triploid karyotype (one haplotype with CN2 and one haplotype with CN1) in this arm on chromosome 16 from 4.95 mb to 90.30 mb and then a 640 kb deletion (5.64 mb to 6.975 mb) occurred in the haplotype with CN2 in this triploid state and then followed by a duplication between genomic coordinates 4.95 mb to 23.79 mb resulting in a combination of tetra-ploid and diploid states. It is also plausible that similar mechanisms of duplication and deletion occurred on chromosome 14 between 19.39 mb to 43.80 mb giving rise to the mixture of CN4, CN3, and CN5 regions.

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; Pellegrino et al. 2010; Syed et al. 2011; Ou et al. 2016). 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 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 has been can be conferred by 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 is 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 (Harrelson et al. 2004; Suzuki et al. 2005; Manning et al. 2006; Cho et al. 2011). 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 (Wang et al. 2012; D'Costa et al. 2014), 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 concomitant of aneuploidy, which is highly prevalent in cancer cells. Previous studies have also revealed the abnormal karyotype of the HepG2 cell line (Simon et al. 1982; Chen et al. 1993). The karyotype of HepG2 cells in our analysis was largely supported by previously published karyotypes for all chromosomes (Table S1). While karyotypes provide a general guide of the degree 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).

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 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 S2, Figure 1, Figure S4). 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 (Mohiyuddin et al. 2015; Layer et al. 2014). 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.

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 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.

MATERIALS & METHODS

Genomic DNA extraction

HepG2 cells were obtained from the Stanford ENCODE Product Center (NHGRI Project 1U54HG006996-01). Genomic DNA was extracted using the Qiagen DNeasy Blood & Tissue Kit (Cat No. 69504) and quantified using the Qubit dsDNA HS Assay Kit (Invitrogen, Waltham, MA, USA). DNA was then verified to be pure (OD260/280 > 1.8; OD260/230 > 1.5) using NanoDrop (Thermo Scientific, Waltham, MA, USA) and high molecular weight (mean >30 kb) using field-inversion gel electrophoresis on the Pippin Pulse System (Sage Science, Beverly, MA, USA).

Karyotyping

Karyotyping of HepG2 cells was conducted in the Cytogenetics Laboratory (cytogenetics.stanford.edu) at Stanford University Medical Center (Palo Alto, CA, USA) where twenty metaphase cells were analyzed by GTW banding.

Array analysis

HepG2 genomic DNA was sent to Macrogen (Rockville, MD, USA) for two independent replicates of array hybridization on the Infinium Multi-Ethnic Global-8 arrays (i.e. MEGA array) from Illumina (San Diego, CA, USA). Using the GenomeStudio software suite from Illumina, raw probe intensity data was normalized, and the log R ratio (LRR) and B

allele frequency (BAF) values were exported for all markers. Genomic coordinates are based on hg19.

Short-insert WGS

Aliquots of HepG2 genomic DNA were sent to Macrogen (Rockville, MD, USA) and two independent replicates of short-insert (350 bp) whole-genome sequencing library were prepared and sequenced (2x151 bp) on two lanes of Illumina HiSeq X Ten to achieve >60x coverage. Reads were aligned to hg19 using BWA-MEM (Li and Durbin 2009) followed by marking of PCR duplicates with Picard tools (version 1.129) (http://broadinstitute.github.io/picard/). Local indel realignment and base quality score recalibration was performed using the Genome Analysis Tool Kit (GATK) (McKenna et al. 2010; DePristo et al. 2011).

CN of chromosome segments and allele frequencies of SNVs and Indels

WGS sequencing coverage was calculated in 10 kb bins across the genome and plotted against the % GC content of each bin to verify the existence of discrete clusters corresponding to discrete CNs (Figure S2). CN was assigned to a cluster based on the ratio of its mean coverage to that of the lowest cluster. For an example, the cluster with the lowest mean coverage was assigned CN1, and the cluster with twice as much mean coverage was assigned CN2 and so forth. The ratios for the five discrete clusters observed corresponded almost perfectly to CN1, CN2, CN3, and CN4. WGS coverage across the genome and across each chromosome was examined visually to assign CN for different chromosome segments or entire chromosomes based on the cluster analysis where adjacent chromosomal segments with different CNs could be identified by the clearly visible sharp and steep changes in sequencing coverage (Figure S3,

Supplemental Data). For each chromosome segment, SNVs and Indels were called using by GATK Haplotypecaller (version 3.7) (McKenna et al. 2010) by specifying the CN or ploidy of that chromosome segment (*stand_emit_conf=0.1, variant_index_type=LINEAR, variant_index_parameter=128000, ploidy={CN}*). The resulting Haplotypecaller outputs from all chromosome segments were then concatenated, and variant quality scores were recalibrated using GATK VQSR with training datasets (dbSNP 138, HapMap 3.3, Omni 2.5 genotypes, 1000 Genomes Phase 1) as recommended by GATK Best Practices (Van der Auwera et al. 2013; DePristo et al. 2011) and filtered with the setting *tranche = 99.0*. SNVs and Indels were annotated using dbSNP138 (Sherry et al. 2001) followed by SnpEff (version 4.3; *canonical transcripts*) (Cingolani et al. 2012a) and then filtered for protein altering variants using SnpSift (version 4.3; *'HIGH' and 'MODERATE' putative impact*) (Cingolani et al. 2012b). Protein-altering variants were intersected with the variants from the 1000 Genomes Project (The 1000 Genomes Project Consortium et al. 2015) and the Exome Sequencing Project (http://evs.gs.washington.edu/EVS/) where overlapping variants were removed using Bedtools (version 2.26) (Quinlan and Hall 2010). The resulting PPA variant calls were overlapped against the Catalogue of Somatic Mutations in Cancer (Forbes et al. 2015) and Sanger Cancer Gene Census (Futreal et al. 2004).

Identification of LOH

A Hidden Markov Model (HMM) was used to identify genomic regions exhibiting LOH. The HMM is designed with two states: LOH present, and LOH absent. We used SNVs that were recalibrated and "PASS"-filtered from GATK VQSR as well as overlapped 1000 Genomes Project variants (Sudmant et al. 2015). The genome was split into 40 kb

bins; heterozygous and homozygous SNVs were tallied for each bin, and bins with <12 SNVs were removed. A bin was classified as heterozygous if ≥50% of the SNVs within the bin are heterozygous, otherwise it was classified as homozygous. This classification was used as the HMM emission sequence. The HMM was initialized with the same initiation and transition probabilities (*Prob=10E-8*) (Adey et al., 2013), and the Viterbi algorithm was used to estimate a best path. Adjacent LOH intervals were merged.

10x Genomics linked-read library and sequencing

HepG2 genomic DNA sample (~35 kb-80 kb) was size selected on the BluePippin instrument (Sage Science, Beverly, MA, USA) using the manufacturer's U1 Maker 30 kb High Pass protocol and then diluted to 1 ng/µl and used as input for the Chromium reagent delivery system (Zheng et al. 2016; Marks et al. 2018) from 10x Genomics (Pleasanton, CA, USA) where HMW DNA fragments are partitioned into >1 million droplets, uniquely barcoded (16 bp) within each droplet, and subjected to random priming and isothermal amplification following standard manufacturer's protocol. Afterwards, the emulsion was broken, and the barcoded DNA molecules were released and converted to a Chromium linked-read library in which each library molecule retains its "HMW fragment barcode". Read-pairs generated in this manner, i.e. linked-reads, that come from the same HMW DNA fragment, can be identified by their "HMW fragment barcode" and can be used to construct a virtual long-read that is representative of the sequence of the original HMW genomic DNA fragment. Reads that cover heterozygous SNVs and Indels can be phased by their "HMW fragment barcode". The final library (8 cycles of PCR amplification) was diluted to 5 nM and sent to

Macrogen (Rockville, MD, USA) for sequencing (2x151 bp) on two lanes of the Illumina HiSeq X to achieve ~60x genomic coverage.

Haplotype phasing and variant calling using 10x Genomics linked-reads

Paired-end linked-reads (median insert size 396 bp, duplication rate 7.68%, Q30 Read1 78.7%, Q30 Read2 63.1%) were aligned to hg19 (alignment rate 90.4%, mean coverage 67.0x, zero coverage 0.117%) and analyzed using the Long Ranger Software (version 2.1.5) from 10x Genomics (Zheng et al. 2016; Marks et al. 2018) (Pleasanton, CA, USA). Segmental duplications, reference gaps, unplaced contigs, regions with assembly issues, and highly polymorphic sites (http://cf.10xgenomics.com/supp/genome/hg19/sv_blacklist.bed, http://cf.10xgenomics.com/supp/genome/hg19/segdups.bedpe) were excluded from the analysis. ENSEMBL annotations (http://cf.10xgenomics.com/supp/genome/gene_annotations.gtf.gz) were used for genes and exons. Phasing was performed by specifying the set of pre-called and filtered HepG2 heterozygous SNVs and Indels from GATK (see above) and formatted using *mkvcf* from Long Ranger (version 2.1.5). Heterozygous SNVs and Indels with more than two types of alleles in ploidy>2 regions were excluded from analysis. Large (>30 kb) SVs and large-scale complex rearrangements were identified using both the Long Ranger *wgs* module with the "*--somatic*" option, GROC-SVs (default settings with breakpoint assembly) (Spies et al. 2017), and gemtools (Greer et al. 2017). The "*--somatic*" option increases the sensitivity of the large-scale SV caller for somatic SVs by allowing the detection of sub-haplotype events and does not affect small-scale variant calling. Variants from Long Ranger analysis indicated as "PASS" were retained. SV

breakpoints identified using GROC-SVs were also analyzed for supporting evidence from mate-pair reads (see below).

SV identification from short-insert WGS

From our deep short-insert WGS dataset, we identified structural variants using BreakDancer (version 1.4.5) (Chen et al. 2009), Pindel (version 0.2.4t) (Ye et al. 2009) and BreakSeq (version 2.0) (Lam et al. 2010) with default settings to obtain per-filtered calls. All SV calls were required to be >50 bp. We filtered out BreakDancer calls with less than 2 supporting paired-end reads and with confidence scores less than 90. Pindel calls were filtered for SVs with quality scores greater than 400. No further filtering was performed for BreakSeq calls.

Mate-pair library construction and sequencing

HepG2 mate-pair library (3 kb insert) was constructed using 2.5 μg of high molecular weight genomic DNA (mean >30 kb) as input for the Nextera Mate Pair Library Prep Kit (FC-132-1001) from Illumina (San Diego, CA, USA) following the manufacturer's protocol with 10 cycles of PCR. Insert-size selection (2.7 kb to 3.3 kb) was carried out on the BluePippin instrument from Sage Science (Beverly, MA, USA) using the S1 selection marker. Sequencing (2x151 bp) was performed twice on the Illumina NextSeq 500 using the NextSeq Mid-Output Kit (Illumina catalogue # FC-404-2003). Data from the two sequencing runs were combined for analysis. Illumina external read adapters and the reverse complement sequence of the Nextera circularized single junction adapter sequence (AGATGTGTATAAGAGACAG) were trimmed from the 3' end of reads followed by another trimming of the Nextera circularized single junction adapter (CTGTCTCTTATACACATCT) using the FASTQ Toolkit (version 1.0) application on

Illumina Basespace (basespace.illumina.com). Trimmed reads were aligned to hg19 using BWA-MEM (version 0.7.12-r1039) with the "-*M*" option (Li and Durbin 2009). Duplicates were marked and removed using Samtools (version 1.2) (Li et al. 2009) followed by Picard tools version 1.52 (http://broadinstitute.github.io/picard/). Indels were then locally realigned against the Mills & 1000 Genomes Gold Standard Indels (Mills et al. 2011; The 1000 Genomes Project Consortium et al. 2012) and base scores were recalibrated using GATK following the standard Best Practices workflow (Van der Auwera et al. 2013; DePristo et al. 2011).

SV calling from mate-pair sequencing

SV calls were made using LUMPY (version 0.6.11) (Layer et al. 2014). Split-reads and discordantly-mapped reads were first extracted and sorted from the processed alignment file as described in github.com/arq5x/lumpy-sv (Layer et al. 2014). The *lumpyexpress* command was issued to obtain pre-filtered SV calls. Segmental duplications and reference gaps (hg19) downloaded from the UCSC Genome Browser (Kent et al. 2002; Karolchik et al. 2004) were excluded from the analysis through the "-*x*" option. SV calls <50 bp were filtered out. To select for high-confidence calls, only SVs that have ≥5 supporting reads as well as both discordant and split-read support were retained.

Non-reference LINE1 & Alu insertions

We adapted the RetroSeq package (Keane et al. 2013) to call the non-reference LINE1 and Alu insertions. The calling depends on information from split reads where a part of the read aligns uniquely to hg19 (not adjacent to LINE1 or Alu sequence) and the rest aligns to a library of active LINE1 or Alu consensus sequences from Repbase (Bao et

al. 2015). The calling can also be based on paired-end reads where one of the pair aligns uniquely to hg19 sequence and the other aligns to LINE1 or Alu consensus sequence. The default “alignment-quality” filter for aligning to Alu or LINE1 consensus is >85% identity as set in RetroSeq (Keane et al. 2013). In addition, we require a threshold 6 or more supporting reads (either split or paired-end) to remove low-confidence calls. The boundaries for the Alu or LINE1 insertion junctions (Table S8) are conservative estimates. The “left boundary” is the left-most coordinate of the upstream supporting reads. The “right boundary” is the right-most coordinate of the downstream supporting reads. The minimum distance between “left boundary” and “right boundary” was set to be 1 kb.

Experimental validation of HepG2 SVs

SVs called from WGS data and mate-pair sequencing data were randomly selected from PCR validation. For SVs called from WGS mate-pair sequencing, PCR primers were designed to produce PCR amplicons 200-500 bp in size and span the breakpoint. For retrotransposon insertions, only events with split-read support were chosen for validation. PCR primers were designed to produce products (65-150 bp) in size where one primer anneals to hg19 unique sequence and the other anneals to either the Alu or LINE1 sequence. All PCR amplicons were gel purified and Sanger sequenced.

Allele-specific RNA expression

Two independent replicates of HepG2 polyA mRNA RNA-seq (alignment files ENCFF670LIE & ENCFF074BOV from experiment accession ENCSR000CPE) were downloaded from the ENCODE portal (Sloan et al. 2016). Replicates were analyzed separately. Samtools mpileup (version 0.1.19) (Li et al. 2009) and BCFtools (version

0.1.19) (Narasimhan et al. 2016) were used to count the number of reads mapped to each allele of heterozygous SNVs; SNVs with coverage <10 in RNA-seq data were filtered out. The binomial test was used to determine if the fraction of RNA-seq reads that mapped each allele of a particular SNV significantly ($p < 0.05$) deviated from the expected value of 0.5 for diploid regions or the expected values in aneuploid regions (e.g. 0.33, 0.67 in triploid regions) for each heterozygous SNV.

Allele-specific DNA methylation

Two independent replicates of HepG2 whole-genome shotgun bisulfite sequencing data (under experiment accession ENCSR881XOU) was downloaded from the ENCODE portal (Sloan et al. 2016). Bisulfite reads were aligned to the human reference genome (hg19) using Bismark (version 0.16.3; *--dovetail, --bowtie2, -I 100, -X 500, --unmapped, --ambiguous*) (Krueger and Andrews 2011) resulting in 28.6x non-duplicate coverage. Paired-end reads containing cytosines (with 10-200x sequencing coverage) in a CpG dinucleotide context were selected if they also overlapped with phased heterozygous SNVs (Dataset 2). The haplotypes to which cytosines (methylated or unmethylated) belonged were then assigned based on phased heterozygous SNVs on that same paired-end read. CpGs that overlapped SNVs with disparate haplotypes (< 1%) were filtered out. CpGs were then grouped into CGIs; the phased reads overlapping each CGI were then grouped by haplotype, and the number of reads containing methylated or unmethylated cytosines were then counted. Read counts were normalized by the CN of their corresponding haplotypes, and Fisher's exact test was applied at each CGI to determine if the fraction of methylated to unmethylated reads was significantly different between the alleles. The *p*-values of each CGI from the Fisher's exact tests were used

as input for the q-value package (https://github.com/StoreyLab/qvalue) in R (https://www.R-project.org) to see which Fisher's exact tests were significant given a target false discovery rate (FDR) of 10%. Since each CGI is overlapped by a different set of reads, we expect that each Fisher's exact test performed is independent of the others, thus satisfying the assumptions required for conservative control of the FDR (Storey and Tibshirani 2003).

Allele-specific CRISPR targets

To identify allele-specific CRISPR targets, we started by extracting variants that satisfy the following properties from Dataset S2 (phased variants):

1. They passed quality control (VCF field 'Filter' is equal to "PASS")
2. They are phased (VCF field 'GT' uses "|" as separator rather than "/")
3. The alleles are heterozygous (e.g. "0|1" or "1|0", but not "1|1")
4. The difference between the alleles is > 1 bp (to ensure target specificty)

For the 401,368 variants that satisfy these four properties, we extracted the two haplotype sequences (maximum length: 335). We only worked with the sequences that were present in the phased genotype. Extracted sequences were tagged them according to their haplotype, for instance:

- If the sequence in the 'Ref' field was "GTA", the 'Alt' field sequence was "TA", and phasing was "0|1" i.e. Haplotype_1|Haplotype_2", the sequence containing "GTA" was tagged "1" for Haplotype 1 and the sequence containing "TA" was tagged "2" for Haplotype 2.

- If the ‘Ref’ field was sequence “GTA”, the ‘Alt’ field was “TA,GCTA”, phasing was “1|2”, the sequence containing “TA” is tagged “1”, the sequence containing “GCTA” was tagged “2” (and the sequence containing “GTA” was not used)

A regular expression was used to extract all potential CRISPR targets from these sequences (i.e. all sequences that matched a [G, C, or A]N_{20}GG pattern and those for which the reverse-complement matched this pattern). This yielded 720,239 candidates, which were then filtered to retain only high-quality targets. The process is adapted from a selection method previously described and validated (Sunagawa et al. 2016) and has already been used for more than 20 genes (Tatsuki et al. 2016). A high-quality candidate gRNA needs to target a unique site. All the candidates that have multiple exact matches in hg19 (irrespective of location) were identified using Bowtie2 (Langmead and Salzberg 2012) and removed. We also removed targets with extreme GC content (>80% or <20%), and targets that contain TTTT, which tends to break the gRNA’s secondary structure. We also used the Vienna RNA-fold package (Lorenz et al. 2011) to compute the gRNA’s secondary structure. We eliminated all candidates for which the stem loop structure cannot fold correctly for Cas9 recognition (Nishimasu et al. 2014), except if the folding energy was above -18 (indicating that the “incorrectly-folded” structure is unstable). Finally, we evaluated the off-target risk score using our own implementation of the Zhang tool (Ran et al. 2013). To ensure that all targets are as reliable and specific as possible, we used a very strict threshold and rejected candidates with an off-target risk score less than 75. Candidates that satisfy all these requirements are considered high quality. For each candidate, we report the location of the variant (chromosome and position), the haplotype (“1” or “2” from the “Phased

Genotype“ column i.e. “Haplotype_1 | Haplotype_2”), the gRNA target sequence, its position relative to the start of the variant, its orientation, its off-target score, and the genomic element targeted (gene or enhancer). Note that the position relative to the start of the variant is for the 5'-most end of the target respective to the genome: if the target is 5'-3', it is its 5' end; if a target was extracted on the reverse complement, it is its 3' end. List of protein-coding and RNA genes and enhancers were obtained from GENCODE (“comprehensive gene annotation”, https://www.gencodegenes.org/releases/19.html) and VISTA Enhancer Browser (https://enhancer.lbl.gov) (Visel et al. 2007).

DATA AVAILABILITY

Raw and processed data files are publicly released on the ENCODE portal (Sloan et al. 2016). The Datasets require non-standard formatting on the portal. We are working with the ENCODE portal group to make this data also available. For immediate review of Datasets, files are available under https://stanfordmedicine.box.com/s/w1fa3ncuhje8zabgugvsw6wmeu7ojv3k

ACKNOWLEDGEMENTS

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

DISCLOSURE DECLARATION

The authors of this manuscript declare no conflicts of interest.

REFERENCES

Abyzov A, Urban AE, Snyder M, 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. http://genome.cshlp.org/cgi/doi/10.1101/gr.114876.110.

Aden DP, Fogel A, Plotkin S, Damjanov I, Knowles BB. 1979. Controlled synthesis of HBsAg in a differentiated human liver carcinoma-derived cell line. *Nature* **282**: 615–6. http://www.ncbi.nlm.nih.gov/pubmed/233137.

Alzeer S, Ellis EM. 2014. Metabolism of gamma hydroxybutyrate in human hepatoma HepG2 cells by the aldo-keto reductase AKR1A1. *Biochem Pharmacol* **92**: 499–505. http://www.ncbi.nlm.nih.gov/pubmed/25256836.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**: 11. http://www.ncbi.nlm.nih.gov/pubmed/26045719.

Becher R, Gibas Z, Karakousis C, Sandberg AA. 1983. Nonrandom chromosome changes in malignant melanoma. *Cancer Res* **43**: 5010–6. http://www.ncbi.nlm.nih.gov/pubmed/6883347.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–9. http://www.ncbi.nlm.nih.gov/pubmed/18987734.

Berger MF, Hodis E, Heffernan TP, Deribe YL, Lawrence MS, Protopopov A, Ivanova E, Watson IR, Nickerson E, Ghosh P, et al. 2012. Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485**: 502–506. http://www.nature.com/articles/nature11071.

Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, Buck G, Chen L, Beare D, Latimer C, et al. 2010. Signatures of mutation and selection in the cancer genome. *Nature* **463**: 893–898. http://www.nature.com/doifinder/10.1038/nature08768.

Bilican B, Goding CR. 2006. Cell cycle regulation of the T-box transcription factor tbx2. *Exp Cell Res* **312**: 2358–2366. http://linkinghub.elsevier.com/retrieve/pii/S0014482706001376.

Bläker H, Hofmann WJ, Rieker RJ, Penzel R, Graf M, Otto HF. 1999. Beta-catenin accumulation and mutation of the CTNNB1 gene in hepatoblastoma. *Genes Chromosomes Cancer* **25**: 399–402. http://www.ncbi.nlm.nih.gov/pubmed/10398436.

Brauer PM, Tyner A. 2009. RAKing in AKT: A tumor suppressor function for the intracellular tyrosine kinase FRK. *Cell Cycle* **8**: 2728–2732. http://www.tandfonline.com/doi/abs/10.4161/cc.8.17.9389.

Burns TF, Fei P, Scata KA, Dicker DT, El-Deiry WS. 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. http://mcb.asm.org/cgi/doi/10.1128/MCB.23.16.5556-5571.2003.

Cancer Genome Atlas Research Network. 2017. Comprehensive and Integrative

Genomic Characterization of Hepatocellular Carcinoma. *Cell* **169**: 1327–1341.e23. http://www.ncbi.nlm.nih.gov/pubmed/28622513.

Carmeliet P. 2005. VEGF as a key mediator of angiogenesis in cancer. *Oncology* **69 Suppl 3**: 4–10. http://www.ncbi.nlm.nih.gov/pubmed/16301830.

Cavalli M, Pan G, Nord H, Wallén Arzt E, Wallerman O, Wadelius C. 2016. Allele-specific transcription factor binding in liver and cervix cells unveils many likely drivers of GWAS signals. *Genomics* **107**: 248–254. http://linkinghub.elsevier.com/retrieve/pii/S0888754316300271.

Chen H-L, Chiu T-S, Chen P-J, Chen D-S. 1993. Cytogenetic studies on human liver cancer cell lines. *Cancer Genet Cytogenet* **65**: 161–166. http://linkinghub.elsevier.com/retrieve/pii/016546089390227D.

Chen J-S, Hung W-S, Chan H-H, Tsai S-J, Sun HS. 2013. In silico identification of oncogenic potential of fyn-related kinase in hepatocellular carcinoma. *Bioinformatics* **29**: 420–7. http://www.ncbi.nlm.nih.gov/pubmed/23267173.

Chen J-S, Su I-J, Leu Y-W, Young K-C, Sun HS. 2012. Expression of T-cell lymphoma invasion and metastasis 2 (TIAM2) promotes proliferation and invasion of liver cancer. *Int J cancer* **130**: 1302–13. http://www.ncbi.nlm.nih.gov/pubmed/21469146.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, 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 EC, Han J-K. 2011. Role of Tbx2 in defining the territory of the pronephric nephron. *Development* **138**: 465–74. http://dev.biologists.org/cgi/doi/10.1242/dev.061234.

Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. 2012a. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front Genet* **3**: 1–9.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012b. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**: 80–92.

Coley HM, Hatzmichael E, Blagden SP, McNeish IA, Thompson A, Crook T, Syed N. 2012. Polo Like Kinase 2 Tumour Suppressor and cancer biomarker: new perspectives on drug sensitivity/resistance in ovarian cancer. *Oncotarget* **3**. http://www.oncotarget.com/fulltext/332.

Craven RJ, Cance WG, Liu ET. 1995. The nuclear tyrosine kinase Rak associates with the retinoblastoma protein pRb. *Cancer Res* **55**: 3969–72. http://www.ncbi.nlm.nih.gov/pubmed/7664264.

Crippa S, Ancey P, Vazquez J, Angelino P, Rougemont A, Guettier C, Zoete V, Delorenzi M, Michielin O, Meylan E. 2017. Mutant CTNNB1 and histological heterogeneity define metabolic subtypes of hepatoblastoma. *EMBO Mol Med* **9**: 1589–1604. http://embomolmed.embopress.org/lookup/doi/10.15252/emmm.201707814.

Curia MC, Zuckermann M, De Lellis L, Catalano T, Lattanzio R, Aceto G, Veschi S, Cama A, Otte J-B, Piantelli M, et al. 2008. Sporadic childhood hepatoblastomas show activation of beta-catenin, mismatch repair defects and p53 mutations. *Mod Pathol* **21**: 7–14. http://www.ncbi.nlm.nih.gov/pubmed/17962810.

D'Costa ZC, Higgins C, Ong CW, Irwin GW, Boyle D, McArt DG, McCloskey K, Buckley NE, Crawford NT, 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**: 1609–20. http://www.ncbi.nlm.nih.gov/pubmed/24742492.

DePristo M a, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis A a, del Angel G, Rivas M a, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–8.

Dereli-Öz A, Versini G, Halazonetis TD. 2011. Studies of genomic copy number changes in human cancers reveal signatures of DNA replication stress. *Mol Oncol* **5**: 308–314. http://doi.wiley.com/10.1016/j.molonc.2011.05.002.

Dias da Silva D, Carmo H, Lynch A, Silva E. 2013. An insight into the hepatocellular death induced by amphetamines, individually and in combination: the involvement of necrosis and apoptosis. *Arch Toxicol* **87**: 2165–85. http://www.ncbi.nlm.nih.gov/pubmed/23820845.

Ding Y, Xi Y, Chen T, Wang J, Tao D, Wu Z-L, Li Y, Li C, Zeng R, Li L. 2008. Caprin-2 enhances canonical Wnt signaling through regulating LRP5/6 phosphorylation. *J Cell Biol* **182**: 865–72. http://www.ncbi.nlm.nih.gov/pubmed/18762581.

Ding Y, Zhang Y, Xu C, Tao Q-H, Chen Y-G. 2013. HECT Domain-containing E3 Ubiquitin Ligase NEDD4L Negatively Regulates Wnt Signaling by Targeting Dishevelled for Proteasomal Degradation. *J Biol Chem* **288**: 8289–8298. http://www.jbc.org/lookup/doi/10.1074/jbc.M112.433185.

Du W-L, Fang Q, Chen Y, Teng J-W, Xiao Y-S, Xie P, Jin B, 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. http://www.ncbi.nlm.nih.gov/pubmed/28849151.

Fang H, Wu Y, Narzisi G, O'Rawe JA, Barrón LTJ, Rosenbaum J, Ronemus M, Iossifov I, Schatz MC, Lyon GJ. 2014. Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med* **6**: 89. http://www.ncbi.nlm.nih.gov/pubmed/25426171.

Fernandez-Banet J, Lee NP, Chan KT, Gao H, Liu X, Sung W-K, Tan W, Fan ST, Poon RT, Li S, et al. 2014. Decoding complex patterns of genomic rearrangement in hepatocellular carcinoma. *Genomics* **103**: 189–203. http://www.ncbi.nlm.nih.gov/pubmed/24462510.

Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, et al. 2015. COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43**: D805–D811.

Fraile JM, Quesada V, Rodríguez D, Freije JMP, López-Otín C. 2012. Deubiquitinases in cancer: new functions and therapeutic options. *Oncogene* **31**: 2373–88. http://www.ncbi.nlm.nih.gov/pubmed/21996736.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Altshuler D, Shendure J, Nickerson DA, et al. 2012. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**: 216–220. http://www.nature.com/doifinder/10.1038/nature11690.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A Census of Human Cancer Genes. *Nat Rev Cancer* **4**: 177–183.

Gao C, Pang L, Ren C, Ma T. 2012. Decreased expression of Nedd4L correlates with poor prognosis in gastric cancer patient. *Med Oncol* **29**: 1733–8. http://www.ncbi.nlm.nih.gov/pubmed/21909941.

Girard L, Zöchbauer-Müller S, Virmani AK, Gazdar AF, Minna JD. 2000. Genome-wide allelotyping of lung cancer identifies new regions of allelic loss, differences between small cell lung cancer and non-small cell lung cancer, and loci clustering. *Cancer Res* **60**: 4894–906. http://www.ncbi.nlm.nih.gov/pubmed/10987304.

Goel RK, Lukong KE. 2016. Understanding the cellular roles of Fyn-related kinase (FRK): implications in cancer biology. *Cancer Metastasis Rev* **35**: 179–99. http://www.ncbi.nlm.nih.gov/pubmed/27067725.

Greer SU, Nadauld LD, Lau BT, Chen J, Wood-Bouwens C, Ford JM, Kuo CJ, Ji HP. 2017. Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. *Genome Med* **9**: 57. http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0447-8.

Hao L, Ito K, Huang K-H, Sae-tan S, Lambert JD, Ross AC. 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**: 1352–62. http://www.ncbi.nlm.nih.gov/pubmed/25060692.

Harrelson Z, Kelly RG, Goldin SN, Gibson-Brown JJ, Bollag RJ, Silver LM, Papaioannou VE. 2004. Tbx2 is essential for patterning the atrioventricular canal and for morphogenesis of the outflow tract during heart development. *Development* **131**: 5041–52. http://dev.biologists.org/cgi/doi/10.1242/dev.01378.

He S, Deng J, Li G, Wang B, Cao Y, Tu Y. 2012. Down-regulation of Nedd4L is Associated with the Aggressive Progression and Worse Prognosis of Malignant Glioma. *Jpn J Clin Oncol* **42**: 196–201. https://academic.oup.com/jjco/article-lookup/doi/10.1093/jjco/hyr195.

Hou K, Zhu Z, Wang Y, Zhang C, Yu S, Zhu Q, Yan B. 2016a. Overexpression and Biological Function of Ubiquitin-Specific Protease 42 in Gastric Cancer ed. J.W. Lee. *PLoS One* **11**: e0152997. http://dx.plos.org/10.1371/journal.pone.0152997.

Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, Wu X, Wen L, Tang F, Huang Y, et al. 2016b. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res* **26**: 304–19. http://www.ncbi.nlm.nih.gov/pubmed/26902283.

Huan LC, Wu J-C, Chiou B-H, Chen C-H, Ma N, Chang CY, Tsen Y-K, Chen SC. 2014. MicroRNA regulation of DNA repair gene expression in 4-aminobiphenyl-treated HepG2 cells. *Toxicology* **322**: 69–77. http://www.ncbi.nlm.nih.gov/pubmed/24857880.

Huang D, Ovcharenko I. 2015. Identifying causal regulatory SNPs in ChIP-seq enhancers. *Nucleic Acids Res* **43**: 225–236. http://academic.oup.com/nar/article/43/1/225/2903413/Identifying-causal-regulatory-SNPs-in-ChIPseq.

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. http://doi.wiley.com/10.1002/hep.27243.

Kai X, Chellappa V, Donado C, Reyon D, Sekigami Y, Ataca D, Louissaint A, Mattoo H, Joung JK, 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. http://www.jbc.org/lookup/doi/10.1074/jbc.M114.598763.

Kamalian L, Chadwick AE, Bayliss M, French NS, Monshouwer M, Snoeys J, Park BK. 2015. The utility of HepG2 cells to identify direct mitochondrial dysfunction in the absence of cell death. *Toxicol Vitr* **29**: 732–740. http://linkinghub.elsevier.com/retrieve/pii/S0887233315000284.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**: D493-6. http://www.ncbi.nlm.nih.gov/pubmed/14681465.

Keane TM, Wong K, Adams DJ. 2013. RetroSeq: Transposable element discovery from next-generation sequencing data. *Bioinformatics* **29**: 389–390.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996–1006. http://www.ncbi.nlm.nih.gov/pubmed/12045153.

Kim J-L, Ha G-H, Campo L, Denning MF, Patel TB, Osipo C, Lin S-Y, Breuer E-K. 2015. The role of Rak in the regulation of stability and function of BRCA1. *Oncotarget* **8**: 86799–86815. http://www.ncbi.nlm.nih.gov/pubmed/29156836.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420–6. http://www.ncbi.nlm.nih.gov/pubmed/17901297.

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

Lam HYK, Mu XJ, Stütz AM, Tanzer A, Cayting PD, Snyder M, Kim PM, Korbel JO, Gerstein MB. 2010. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* **28**: 47–55.

Lam HYK, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, O'Huallachain M, Gerstein MB, Kidd JM, Bustamante CD, et al. 2012. Detecting and annotating genetic variations using the HugeSeq pipeline. *Nat Biotechnol* **30**: 226–229. http://www.nature.com/doifinder/10.1038/nbt.2134.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–9. http://www.ncbi.nlm.nih.gov/pubmed/22388286.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**: R84.

Leung EL, Wong JC, Johlfs MG, Tsang BK, Fiscus RR. 2010. Protein Kinase G Type I Activity in Human Ovarian Cancer Cells Significantly Contributes to Enhanced Src Activation and DNA Synthesis/Cell Proliferation. *Mol Cancer Res* **8**: 578–591. http://mcr.aacrjournals.org/cgi/doi/10.1158/1541-7786.MCR-09-0178.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–60. http://www.ncbi.nlm.nih.gov/pubmed/19451168.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–9.

http://www.ncbi.nlm.nih.gov/pubmed/19505943.

Lin D-C, Xu L, Ding L-W, Sharma A, Liu L-Z, Yang H, Tan P, Vadgama J, Karlan BY, Lester J, et al. 2013. Genomic and functional characterizations of phosphodiesterase subtype 4D in human cancers. *Proc Natl Acad Sci* **110**: 6109–6114. http://www.pnas.org/cgi/doi/10.1073/pnas.1218206110.

López-Terrada D, Cheung SW, Finegold MJ, Knowles BB. 2009a. Hep G2 is a hepatoblastoma-derived cell line. *Hum Pathol* **40**: 1512–5. http://www.ncbi.nlm.nih.gov/pubmed/19751877.

López-Terrada D, Gunaratne PH, Adesina AM, Pulliam J, Hoang DM, Nguyen Y, Mistretta T-A, Margolin J, Finegold MJ. 2009b. Histologic subtypes of hepatoblastoma are characterized by differential canonical Wnt and Notch pathway activation in DLK+ precursors. *Hum Pathol* **40**: 783–794. http://linkinghub.elsevier.com/retrieve/pii/S0046817708005625.

Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**: 26. http://www.ncbi.nlm.nih.gov/pubmed/22115189.

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. http://www.ncbi.nlm.nih.gov/pubmed/28881763.

Mangrum JB, Martin EJ, Brophy DF, Hawkridge AM. 2015. Intact stable isotope labeled plasma proteins from the SILAC-labeled HepG2 secretome. *Proteomics* **15**: 3104–15. http://www.ncbi.nlm.nih.gov/pubmed/25504855.

Manning L, Ohyama K, Saeger B, Hatano O, Wilson SA, Logan M, Placzek M. 2006. Regional Morphogenesis in the Hypothalamus: A BMP-Tbx2 Pathway Coordinates Fate and Proliferation through Shh Downregulation. *Dev Cell* **11**: 873–885. http://linkinghub.elsevier.com/retrieve/pii/S1534580706004497.

Marks P, Garcia S, Barrio AM, 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. *bioRxiv* 230946. https://www.biorxiv.org/content/early/2018/01/09/230946.

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**: 1297–303. http://www.ncbi.nlm.nih.gov/pubmed/20644199.

Menezes C, Alverca E, Dias E, Sam-Bento F, Pereira P. 2013. Involvement of endoplasmic reticulum and autophagy in microcystin-LR toxicity in Vero-E6 and HepG2 cell lines. *Toxicol In Vitro* **27**: 138–48. http://www.ncbi.nlm.nih.gov/pubmed/23010415.

Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, Kemeza DM, Strassler DS, Ponting CP, Webber C, et al. 2011. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res* **21**: 830–9. http://www.ncbi.nlm.nih.gov/pubmed/21460062.

Mohiyuddin M, Mu JC, Li J, Bani Asadi N, Gerstein MB, Abyzov A, Wong WH, Lam HYK. 2015. MetaSV: An accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics* **31**: 2741–2744.

Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. 2016. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**: 1749–51. http://www.ncbi.nlm.nih.gov/pubmed/26826718.

Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, Ishitani R, Zhang F, Nureki O. 2014. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**: 935–49. http://www.ncbi.nlm.nih.gov/pubmed/24529477.

Oberg-Welsh C, Annerén C, Welsh M. 1998. Mutation of C-terminal tyrosine residues Y497/Y504 of the Src-family member Bsk/Iyk decreases NIH3T3 cell proliferation. *Growth Factors* **16**: 111–24. http://www.ncbi.nlm.nih.gov/pubmed/9932229.

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**: 457–66. http://www.ncbi.nlm.nih.gov/pubmed/27423313.

Paculová H, Kohoutek J. 2017. The emerging roles of CDK12 in tumorigenesis. *Cell Div* **12**: 7. http://celldiv.biomedcentral.com/articles/10.1186/s13008-017-0033-x.

Pellegrino R, Calvisi DF, Ladu S, Ehemann V, Staniscia T, Evert M, Dombrowski F, Schirmacher P, Longerich T. 2010. Oncogenic and tumor suppressive roles of polo-like kinases in human hepatocellular carcinoma. *Hepatology* NA-NA. http://doi.wiley.com/10.1002/hep.23467.

Pylayeva-Gupta Y, Grabocka E, Bar-Sagi D. 2011. RAS oncogenes: weaving a tumorigenic web. *Nat Rev Cancer* **11**: 761–74. http://www.ncbi.nlm.nih.gov/pubmed/21993244.

Quesada V, Díaz-Perales A, Gutiérrez-Fernández A, Garabaya C, Cal S, López-Otín C. 2004. Cloning and enzymatic analysis of 22 novel human ubiquitin-specific proteases. *Biochem Biophys Res Commun* **314**: 54–62. http://www.ncbi.nlm.nih.gov/pubmed/14715245.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq033.

Rajaram M, Zhang J, Wang T, Li J, Kuscu C, Qi H, Kato M, Grubor V, Weil RJ, Helland A, et al. 2013. Two Distinct Categories of Focal Deletions in Cancer Genomes ed. W.B. Coleman. *PLoS One* **8**: e66264. http://dx.plos.org/10.1371/journal.pone.0066264.

Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. 2013. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**: 2281–2308. http://www.nature.com/doifinder/10.1038/nprot.2013.143.

Sahu SC, O'Donnell MW, Sprando RL. 2012. Interactive toxicity of usnic acid and lipopolysaccharides in human liver HepG2 cells. *J Appl Toxicol* **32**: 739–49. http://www.ncbi.nlm.nih.gov/pubmed/22777745.

Schoonen WGEJ, Westerink WMA, de Roos JADM, 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**: 505–16. http://www.ncbi.nlm.nih.gov/pubmed/15826808.

Sheng ZM, Marchetti A, Buttitta F, Champeme MH, Campani D, Bistocchi M, Lidereau R, Callahan R. 1996. Multiple regions of chromosome 6q affected by loss of heterozygosity in primary human breast carcinomas. *Br J Cancer* **73**: 144–7. http://www.ncbi.nlm.nih.gov/pubmed/8546898.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308–11.

Simon D, Aden DP, Knowles BB. 1982. Chromosomes of human hepatoma cell lines. *Int J Cancer* **30**: 27–33. http://doi.wiley.com/10.1002/ijc.2910300106.

Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT, et al. 2016. ENCODE data at the ENCODE portal. *Nucleic Acids Res* **44**: D726–D732. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1160.

Spies N, Weng Z, Bishara A, McDaniel J, Catoe D, Zook JM, Salit M, West RB, Batzoglou S, Sidow A. 2017. Genome-wide reconstruction of complex structural variants using read clouds. *Nat Methods*. http://www.nature.com/doifinder/10.1038/nmeth.4366.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**: 9440–5. http://www.ncbi.nlm.nih.gov/pubmed/12883005.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81. http://www.nature.com/doifinder/10.1038/nature15394.

Sunagawa GA, 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. http://www.ncbi.nlm.nih.gov/pubmed/26774482.

Suzuki T, Takeuchi J, Koshiba-Takeuchi K, Ogura T. 2005. Tbx Genes Specify Posterior Digit Identity through Shh and BMP Signaling. *Dev Cell* **8**: 971–972. http://linkinghub.elsevier.com/retrieve/pii/S1534580705001802.

Syed N, Coley HM, Sehouli J, Koensgen D, Mustea A, Szlosarek P, McNeish I, Blagden SP, Schmid P, Lovell DP, et al. 2011. Polo-like Kinase Plk2 Is an Epigenetic Determinant of Chemosensitivity and Clinical Outcomes in Ovarian Cancer. *Cancer Res* **71**: 3317–3327. http://cancerres.aacrjournals.org/cgi/doi/10.1158/0008-5472.CAN-10-2048.

Syed N, Smith P, Sullivan A, Spender LC, 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**: 250–6. http://www.bloodjournal.org/cgi/doi/10.1182/blood-2005-03-1194.

Taniguchi K, Roberts LR, Aderca IN, Dong X, Qian C, Murphy LM, Nagorney DM, Burgart LJ, Roche PC, Smith DI, et al. 2002. Mutational spectrum of β-catenin, AXIN1 and AXIN2 in hepatocellular carcinomas and hepatoblastomas. *Oncogene* **21**: 4863–4871. http://www.nature.com/articles/1205591.

Tanksley JP, Chen X, Coffey RJ. 2013. NEDD4L Is Downregulated in Colorectal Cancer and Inhibits Canonical WNT Signaling ed. C. Liu. *PLoS One* **8**: e81514. http://dx.plos.org/10.1371/journal.pone.0081514.

Tatsuki F, Sunagawa GA, Shi S, Susaki EA, Yukinaga H, Perrin D, Sumiyama K, Ukai-

Tadenuma M, Fujishima H, Ohno R, et al. 2016. Involvement of Ca(2+)-Dependent Hyperpolarization in Sleep Duration in Mammals. *Neuron* **90**: 70–85. http://www.ncbi.nlm.nih.gov/pubmed/26996081.

The 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65. http://www.ncbi.nlm.nih.gov/pubmed/23128226.

The 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74. http://www.ncbi.nlm.nih.gov/pubmed/26432245.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma* **43**: 11.10.1-33. http://www.ncbi.nlm.nih.gov/pubmed/25431634.

Vance KW, Carreira S, Brosch G, Goding CR. 2005. Tbx2 is overexpressed and plays an important role in maintaining proliferation and suppression of senescence in melanomas. *Cancer Res* **65**: 2260–8. http://www.ncbi.nlm.nih.gov/pubmed/15781639.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**: D88-92. http://www.ncbi.nlm.nih.gov/pubmed/17130149.

Wang B, Lindley LE, Fernandez-Vega V, Rieger ME, Sims AH, Briegel KJ. 2012. The T

Box Transcription Factor TBX2 Promotes Epithelial-Mesenchymal Transition and Invasion of Normal and Malignant Breast Epithelial Cells ed. P. Derksen. *PLoS One* **7**: e41355. http://dx.plos.org/10.1371/journal.pone.0041355.

Wong N, Lai P, Pang E, Leung TW, Lau JW, Johnson PJ. 2000. A comprehensive karyotypic study on human hepatocellular carcinoma by spectral karyotyping. *Hepatology* **32**: 1060–8. http://doi.wiley.com/10.1053/jhep.2000.19349.

Xia P, Zhang X, Xie Y, Guan M, Villeneuve DL, Yu H. 2016. Functional Toxicogenomic Assessment of Triclosan in Human HepG2 Cells Using Genome-Wide CRISPR-Cas9 Screening. *Environ Sci Technol* **50**: 10682–10692. http://www.ncbi.nlm.nih.gov/pubmed/27459410.

Xia Y, Yeddula N, Leblanc M, Ke E, Zhang Y, Oldfield E, Shaw RJ, Verma IM. 2012. Reduced cell proliferation by IKK2 depletion in a mouse lung-cancer model. *Nat Cell Biol* **14**: 257–265. http://www.nature.com/articles/ncb2428.

Xu D, He X, Chang Y, Xu C, Jiang X, Sun S, Lin J. 2013. Inhibition of miR-96 expression reduces cell proliferation and clonogenicity of HepG2 hepatoma cells. *Oncol Rep* **29**: 653–61. http://www.ncbi.nlm.nih.gov/pubmed/23151657.

Yang J, Gong X, Ouyang L, He W, Xiao R, 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. https://www.spandidos-publications.com/10.3892/ol.2016.4688.

Yang Q, Zhao J, Cui M, Gi S, Wang W, Han X. 2015. Nedd4L expression is decreased in ovarian epithelial cancer tissues compared to ovarian non-cancer tissue. *J Obstet Gynaecol Res* **41**: 1959–1964. http://doi.wiley.com/10.1111/jog.12808.

Ye K, Schulz MH, Long Q, Apweiler R, 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**: 2865–71. http://www.ncbi.nlm.nih.gov/pubmed/19561018 (Accessed June 12, 2017).

Ye X, Wang L, Shang B, Wang Z, Wei W. 2014. NEDD4: a promising target for cancer therapy. *Curr Cancer Drug Targets* **14**: 549–56. http://www.ncbi.nlm.nih.gov/pubmed/25088038.

Yim E-K, Siwko S, Lin S-Y. 2009. Exploring Rak tyrosine kinase function in breast cancer. *Cell Cycle* **8**: 2360–4. http://www.ncbi.nlm.nih.gov/pubmed/19597351.

Yu H, Liu BO, Liu A, Li K, Zhao H. 2015. T-box 2 expression predicts poor prognosis in gastric cancer. *Oncol Lett* **10**: 1689–1693. http://www.ncbi.nlm.nih.gov/pubmed/26622733.

Zhang Y, Ding Y, Chen Y-G, Tao Q. 2014. NEDD4L regulates convergent extension movements in Xenopus embryos via Disheveled-mediated non-canonical Wnt signaling. *Dev Biol* **392**: 15–25. http://www.ncbi.nlm.nih.gov/pubmed/24833518.

Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, et al. 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34**: 303–11. http://www.ncbi.nlm.nih.gov/pubmed/26829319 (Accessed July 14, 2016).

Zhou W, Zhang J, Marcus AI. 2014. LKB1 tumor suppressor: Therapeutic opportunities knock when LKB1 is inactivated. *Genes Dis* **1**: 64–74. http://linkinghub.elsevier.com/retrieve/pii/S2352304214000038.

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.

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 (Zheng et al. 2016; Marks et al. 2018). (A) Disruption of *FRK* by translocation between chromosomes 6 and 16. (B) 127 kb duplication on chromosome 7 resulting in partial duplications of *USP42* and *PMS2.* (C) 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 (F) 2.47 Mb intra-chromosomal rearrangement between *MALRD1* and *MLLT10* on chromosome 10 (G) 60.4 kb inversion that disrupts *PPL* and *SEC14L5*. (H) 31.3 kb inversion within *GUSBP1* on chromosome 5.

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 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. 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 *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 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 (Supplemental 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)

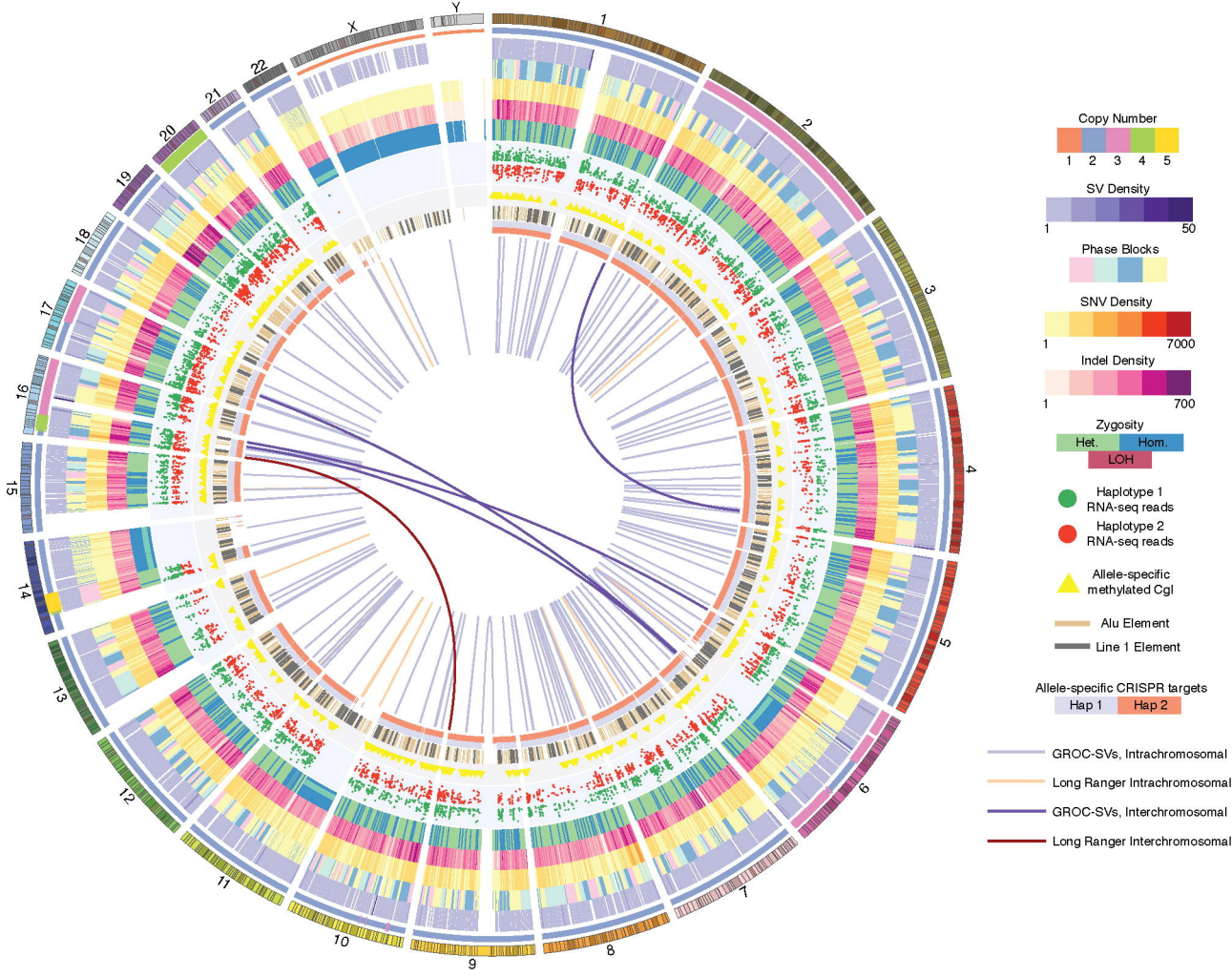
Upper panel: WGS coverage plot. X-axis genomic coordinate in kb. Y-axis: WGS coverage. Purple: CN4, Blue: CN3, Green: CN2. Lower panel: Y-axis: array probe signal intensity. X-axis: Genomic coordinate. For complete chromosomes (1-22, X), see Supplemental Data.

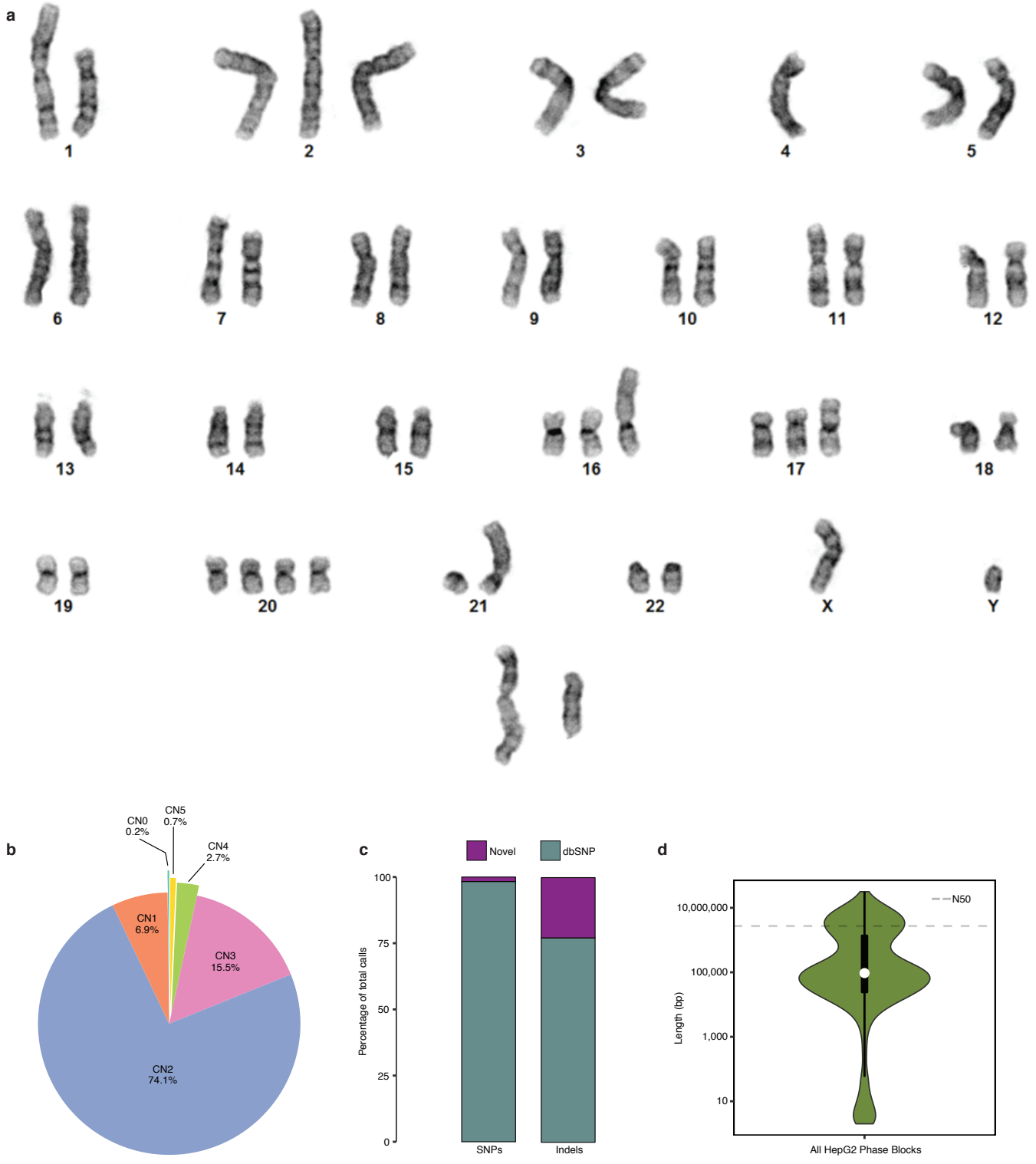
Figure S4. Size Distribution of Phased Haplotype Blocks by Chromosome

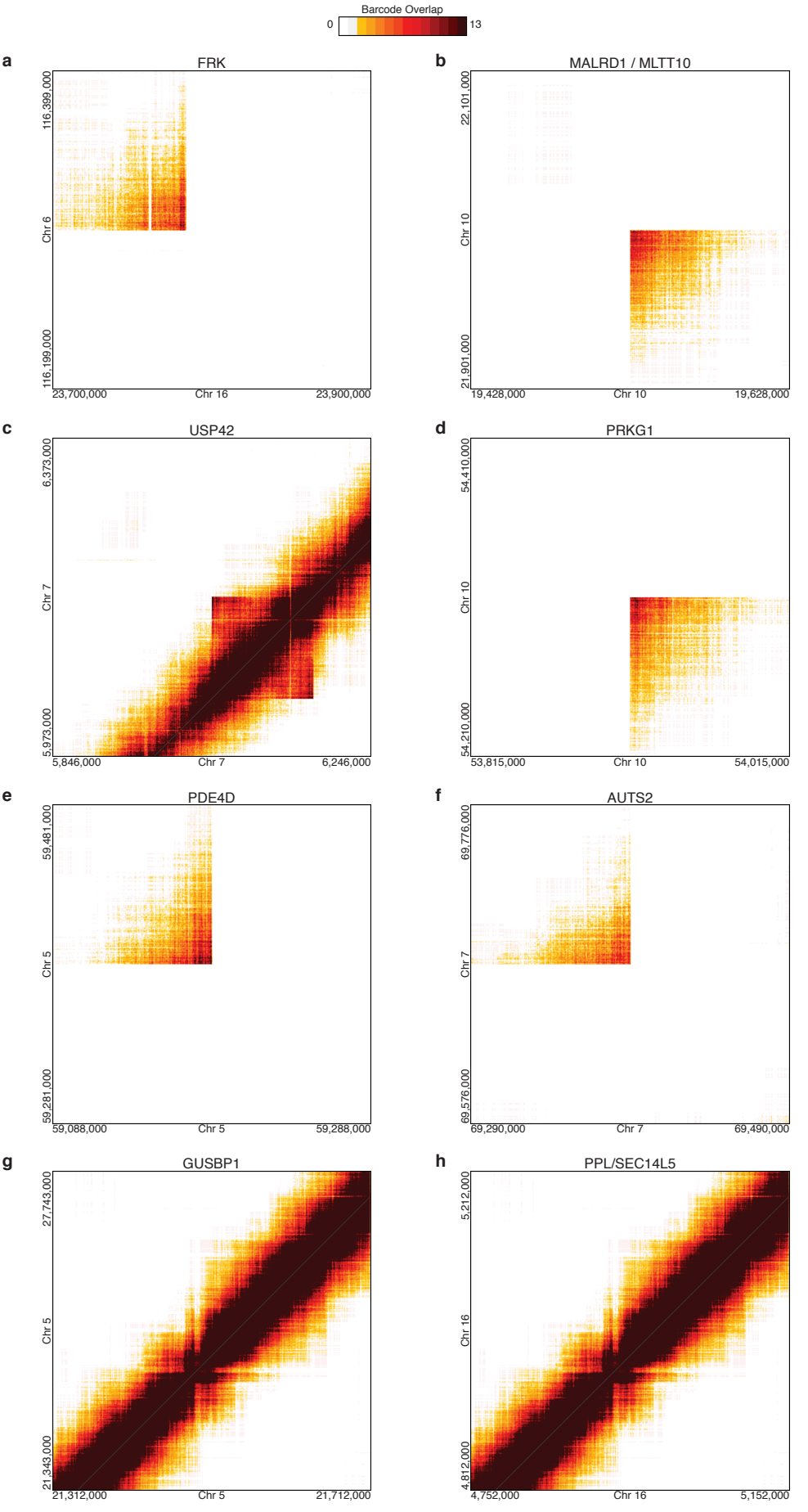
Violin plots (with overlaid boxplot) of HepG2 phased haplotype blocks (Dataset 2) by chromosomes. Y-axis: size in log-scale.

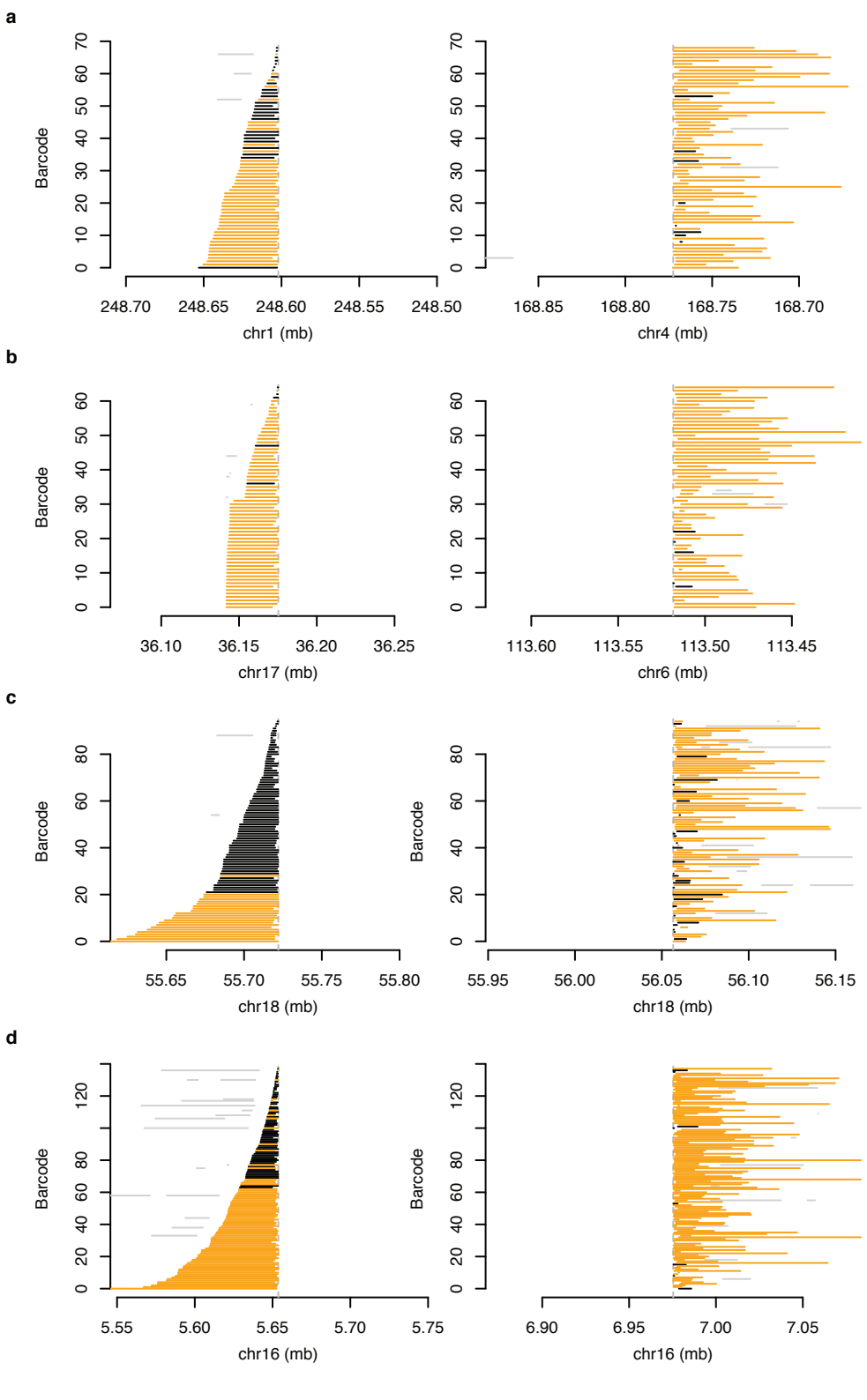
Figure S5. Overlap Between SV Callers

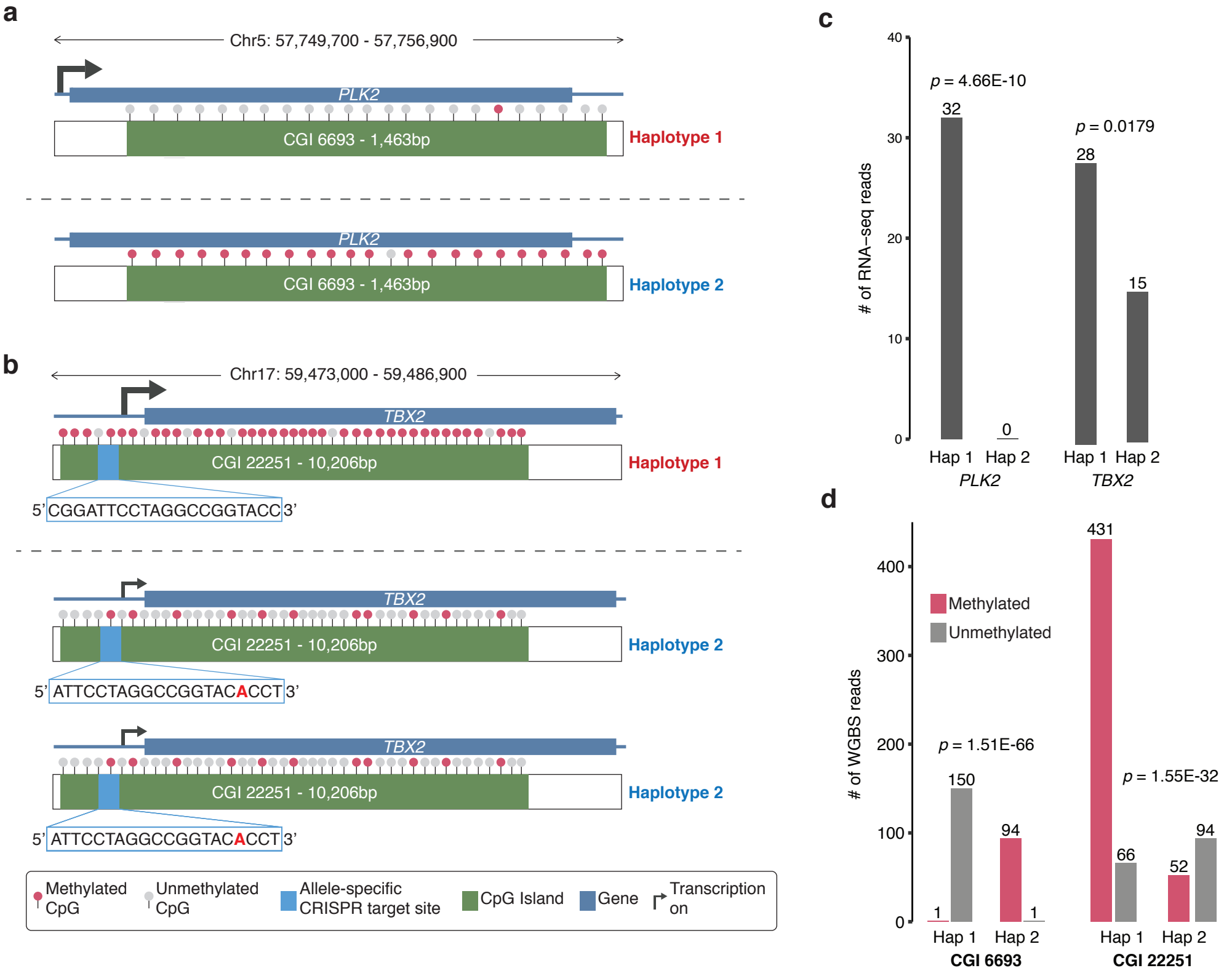
Venn diagram of overlaps (>50% reciprocal) between HepG2 SVs identified in WGS using BreakDancer (Chen et al. 2009), Pindel (Ye et al. 2009), and BreakSeq (Lam et al. 2010), in 3kb-mate-pair sequencing using LUMPY (Layer et al. 2014), and in linked-read sequencing using Long Ranger (Zheng et al. 2016;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

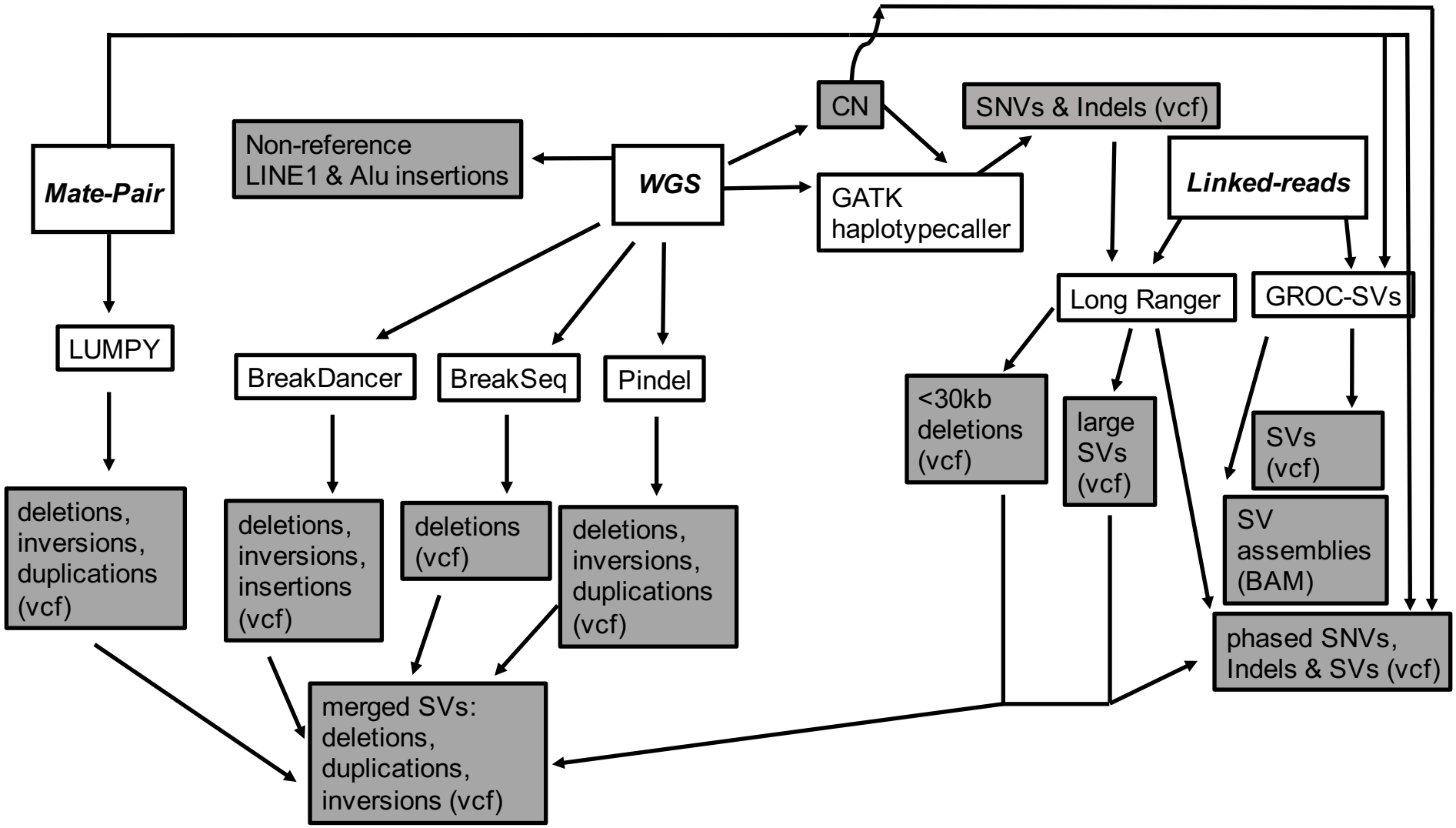

Copy Number
1 2 3 4 5
SV Density
1
50
Phase Blocks
SNV Density
1
7000
Indel Density
1
700
Zygosity
Het.
Hom.
LOH
Haplotype 1 RNA-seq reads
Haplotype 2 RNA-seq reads
Allele-specific methylated CgI
Alu Element
Line 1 Element
Allele-specific CRISPR targets
Hap 1
Hap 2
GROC-SVs, Intrachromosomal
Long Ranger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

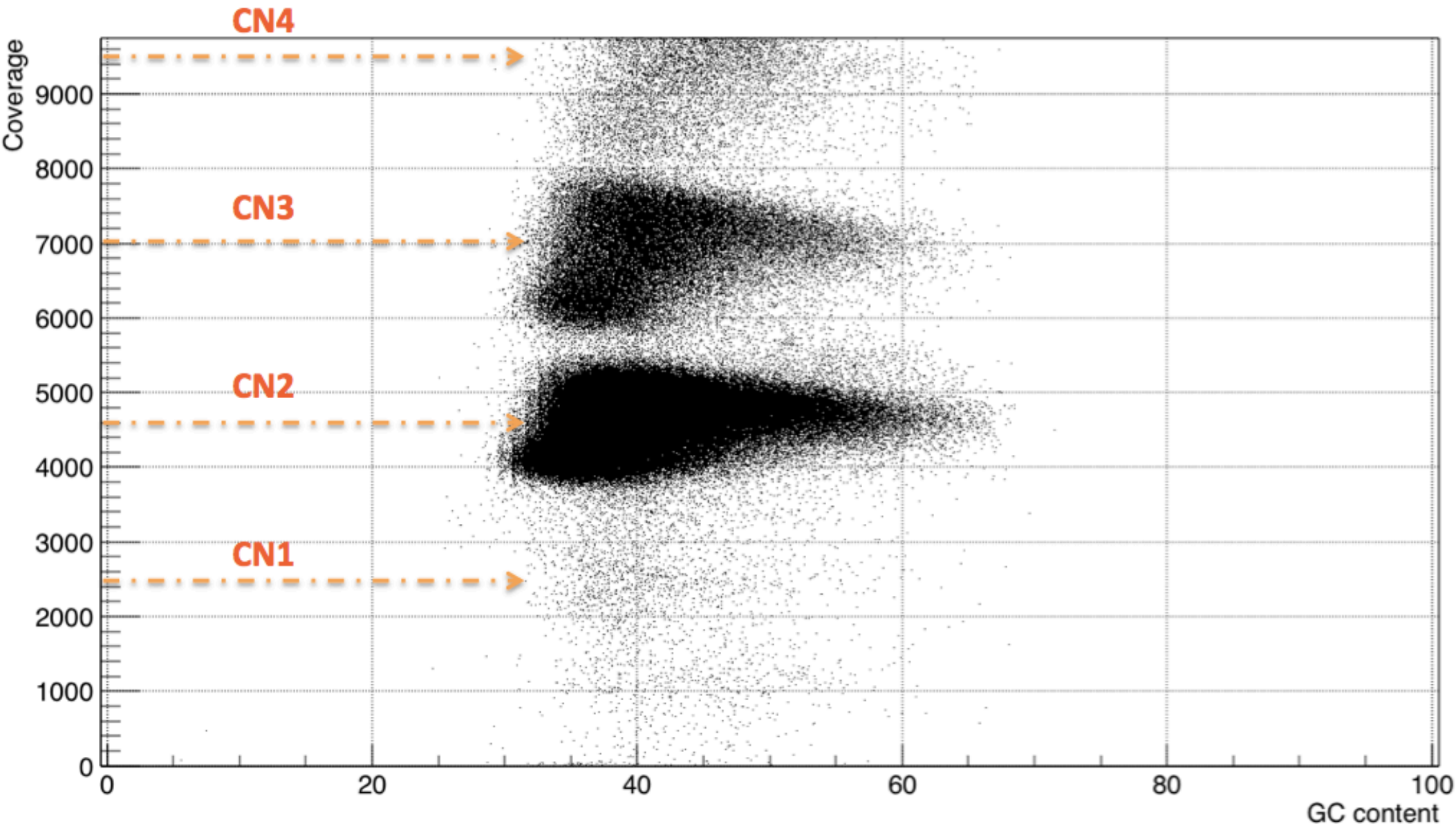

a
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
b
CN0
0.2%
CN5
0.7%
CN4
2.7%
CN1
6.9%
CN3
15.5%
CN2
74.1%
c
Novel
dbSNP
Percentage of total calls
100
75
50
25
0
SNPs
Indels
d
N50
10,000,000
100,000
1,000
10
Length (bp)
All HepG2 Phase Blocks

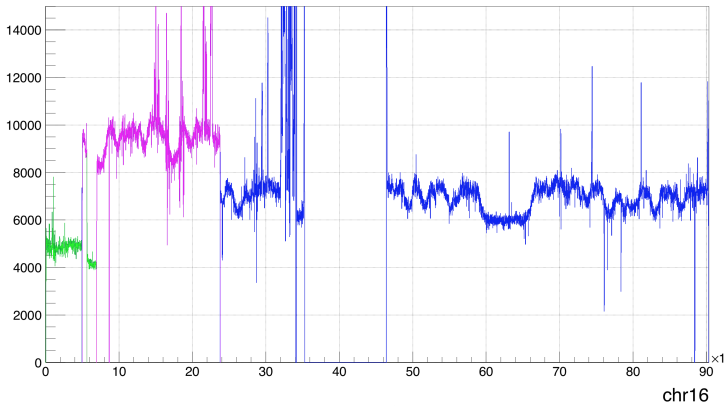

Barcode Overlap
0
13
a
FRK
116,399,000
Chr 6
116,199,000
23,700,000
Chr 16
23,900,000
b
MALRD1 / MLTT10
22,101,000
Chr 10
21,901,000
19,428,000
Chr 10
19,628,000
c
USP42
6,373,000
Chr 7
5,973,000
5,846,000
Chr 7
6,246,000
d
PRKG1
54,410,000
Chr 10
54,210,000
53,815,000
Chr 10
54,015,000
e
PDE4D
59,481,000
Chr 5
59,281,000
59,088,000
Chr 5
59,288,000
f
AUTS2
69,776,000
Chr 7
69,576,000
69,290,000
Chr 7
69,490,000
g
GUSBP1
27,743,000
Chr 5
21,343,000
21,312,000
Chr 5
21,712,000
h
PPL/SEC14L5
5,212,000
Chr 16
4,812,000
4,752,000
Chr 16
5,152,000

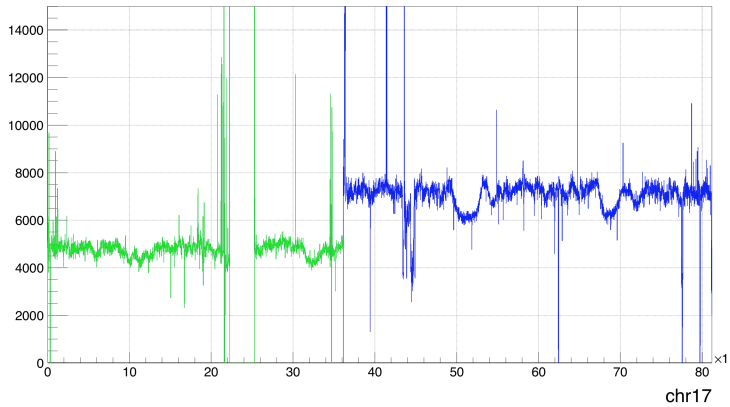

a
Barcode
0
10
20
30
40
50
60
70
248.70
248.65
248.60
248.55
248.50
chr1 (mb)
168.85
168.80
168.75
168.70
chr4 (mb)
b
Barcode
0
10
20
30
40
50
60
36.10
36.15
36.20
36.25
chr17 (mb)
113.60
113.55
113.50
113.45
chr6 (mb)
c
Barcode
0
20
40
60
80
55.65
55.70
55.75
55.80
chr18 (mb)
55.95
56.00
56.05
56.10
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)

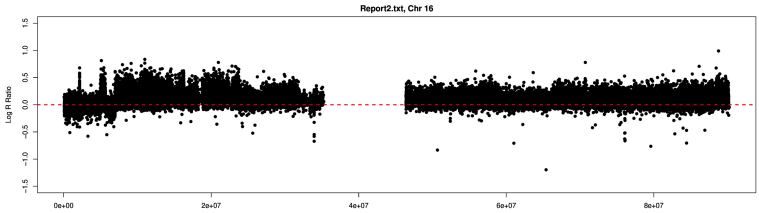

a
Chr5: 57,749,700 - 57,756,900
PLK2
CGI 6693 - 1,463bp
Haplotype 1
PLK2
CGI 6693 - 1,463bp
Haplotype 2
b
Chr17: 59,473,000 - 59,486,900
TBX2
CGI 22251 - 10,206bp
Haplotype 1
5' CGGATTCCTAGGCCGGTACC 3'
TBX2
CGI 22251 - 10,206bp
Haplotype 2
5' ATTCCTAGGCCGGTACACCT 3'
TBX2
CGI 22251 - 10,206bp
Haplotype 2
5' ATTCCTAGGCCGGTACACCT 3'
Methylated CpG
Unmethylated CpG
Allele-specific CRISPR target site
CpG Island
Gene
Transcription on
c
# of RNA−seq reads
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
Methylated
Unmethylated
p = 1.51E-66
150
1
94
1
p = 1.55E-32
431
66
52
94
Hap 1
Hap 2
CGI 6693
Hap 1
Hap 2
CGI 22251

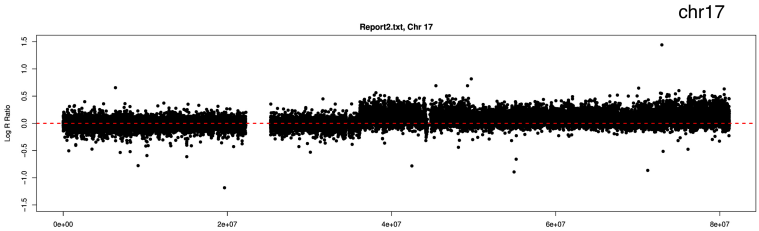

Mate-Pair
Non-reference LINE1 & Alu insertions
WGS
CN
SNVs & Indels (vcf)
GATK haplotypecaller
Linked-reads
Long Ranger
GROC-SVs
LUMPY
BreakDancer
BreakSeq
Pindel
<30kb deletions (vcf)
large SVs (vcf)
SVs (vcf)
SV assemblies (BAM)
deletions, inversions, duplications (vcf)
deletions, inversions, insertions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

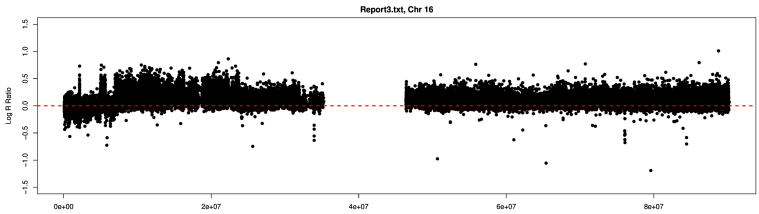

CN4
CN3
CN2
CN1
Coverage
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
20
40
60
80
100
GC content

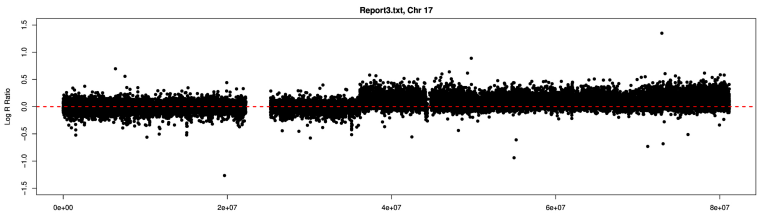

chr16

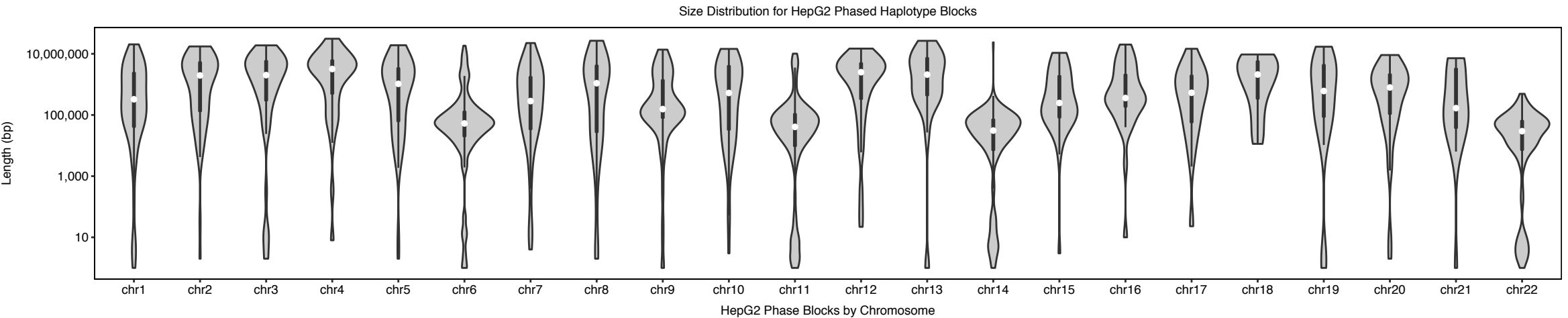

chr17


Report2.txt, Chr 16
Log R Ratio

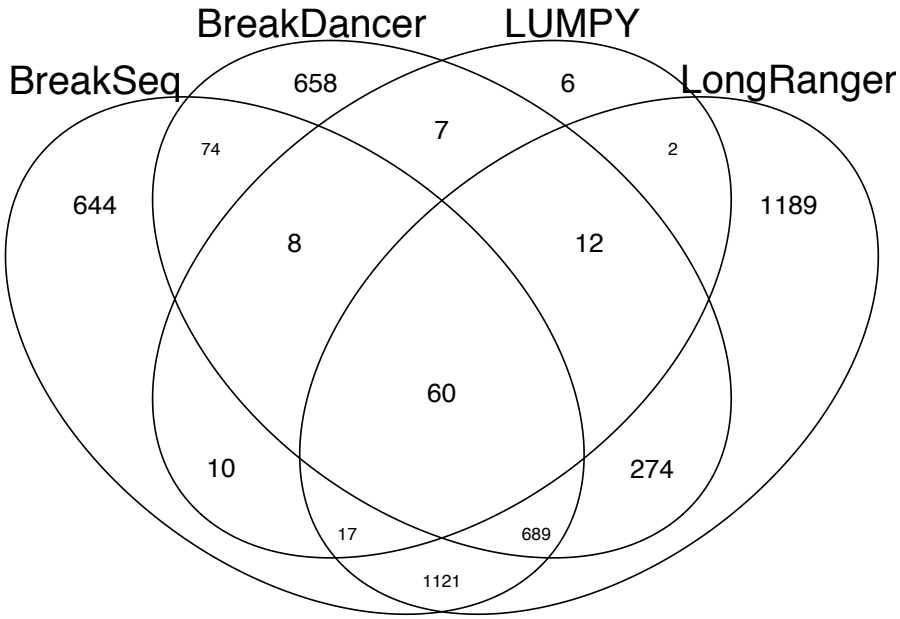

Report2.txt, Chr 17
Log R Ratio


Report3.txt, Chr 16
Log R Ratio

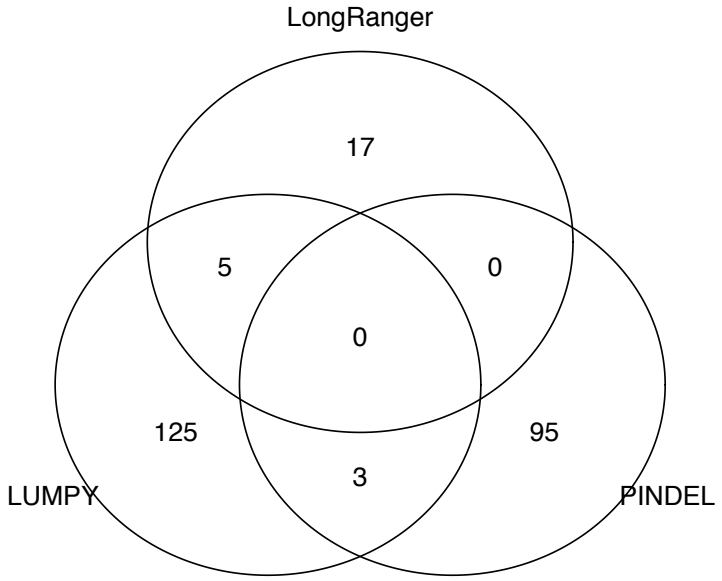

Report3.txt, Chr 17
Log R Ratio


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

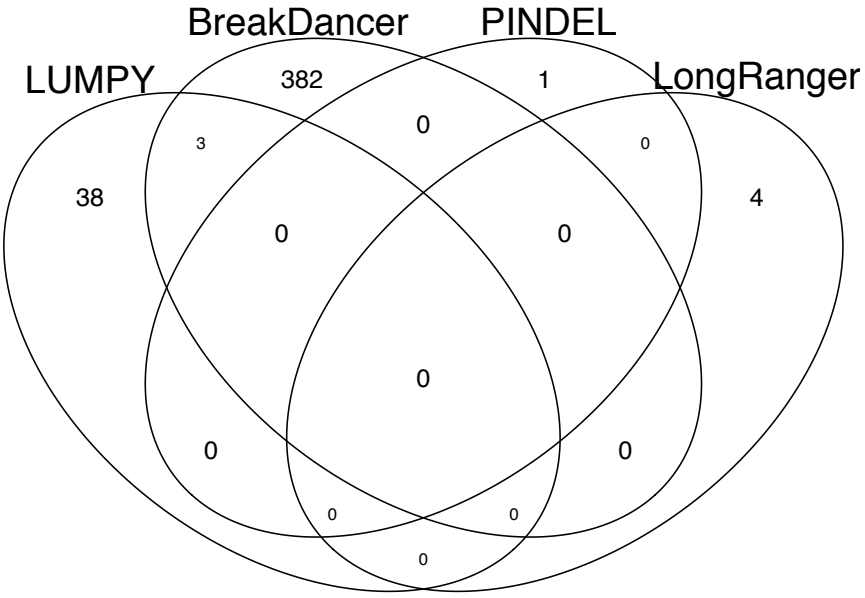
A


BreakDancer
LUMPY
BreakSeq
LongRanger
658
6
7
74
2
644
1189
8
12
60
10
274
17
689
1121

B


LongRanger
17
5
0
0
125
95
3
LUMPY
PINDEL

C


BreakDancer
PINDEL
LUMPY
LongRanger
382
1
0
3
0
38
4
0
0
0
0
0
0
0