Comprehensive, integrated and phased whole-genome analysis of the primary ENCODE cell line K562

Bo Zhou[1,2], Steve Ho[1,2], Xiaowei Zhu[1,2] , Xianglong Zhang[1,2], Noah Spies[2,3,4,5], Seunggyu Byeon[6], Joseph G Arthur[7], Reenal Pattni[1,2], Noa Ben-Efraim[1,2], Michael S Haney[2], Rajini R Haraksingh[1,2,8], Giltae Song[6], Dimitri Perrin[9], Wing H Wong[7], Alexej Abyzov[10], 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]Department of Pathology, Stanford University School of Medicine, Stanford, California 94305, USA
[4]Genome-Scale Measurements Group, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA
[5]Joint Initiative for Metrology in Biology, Stanford, California 94305, USA
[6]School of Computer Science and Engineering, College of Engineering, Pusan National University, Busan 46241, South Korea
[7]Department of Statistics, Department of Biomedical Data Science, Bio-X Program, Stanford University, Stanford, California 94305, USA
[8]Current affiliation: Department of Life Sciences, The University of the West Indies, Saint Augustine, Trinidad and Tobago
[9]Science and Engineering Faculty, Queensland University of Technology, Brisbane, QLD 4001, Australia
[10]Department of Health Sciences Research, Center for Individualized Medicine, Mayo Clinic, Rochester, Minnesota 55905, USA

Corresponding author:

Alexander E Urban, Ph.D
Department of Psychiatry and Behavioral Sciences
Department of Genetics
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: Comprehensive whole-genome analysis of K562

Keywords: ENCODE, K562, structural variation (SV), haplotype phasing, mobile element insertions, CRISPR, allele-specific expression (ASE), allele-specific methylation (ASM)

ABSTRACT

K562 is one of the most widely used human cell lines in biomedical research. It is one of three tier-one cell lines of ENCODE, and one of the cell lines most commonly used for large-scale CRISPR/Cas9 gene-editing screens. Although the functional genomic and epigenomic characteristics of K562 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 K562 renders traditional genome variant analysis methods challenging and partially ineffective. Correct and complete interpretation of the extensive functional genomics data from K562 requires an understanding of the cell line's genome sequence and genome structure. We performed very-deep short-insert whole-genome sequencing, mate-pair sequencing, linked-read sequencing, karyotyping and array CGH, and used a combination of novel and established computational methods to identify and catalog a wide spectrum of genome sequence variants and genome structural features in K562: copy numbers (CN) of chromosome segments, SNVs and Indels (allele frequency-corrected based on copy-number), phased haplotype blocks (N50 = 2.72 Mb), structural variants (SVs) including complex genomic rearrangements, and novel mobile element insertions. A large fraction of SVs were phased, sequence assembled, and experimentally validated. Many chromosomes show striking loss of heterozygosity. To demonstrate the utility of this knowledge, we re-analyzed K562 RNA-Seq and whole-genome bisulfite sequencing data to detect and phase allele-specific expression and DNA methylation patterns, respectively. 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 a phased CRISPR targeting map, i.e. a catalog of loci where CRISPR guide RNAs will bind in an allele-specific manner. This comprehensive whole-genome analysis serves as a resource for future studies that utilize K562 and as the basis of advanced analyses of the rich amounts of the functional genomics data produced by ENCODE for K562. It is also an example for advanced, integrated whole-genome sequence and structure analysis, beyond standard short-read/short-insert whole-genome sequencing, of human genomes in general and in particular of cancer genomes with large numbers of complex sequence alterations.

INTRODUCTION

K562 is an immortalized chronic myelogenous leukemia (CML) cell line derived from a 53-year-old female in 1970 (Lozzio and Lozzio 1975). Since being established, the K562 cell line has been widely used in biomedical research as a "work-horse" cell line, resulting in over 17,000 publications to date. In most cases its use is similar to that of a model organism, contributing to the understanding of basic human biological processes in general and to basic and translational cancer research in particular (Grzanka et al. 2003; Drexler et al. 2004; Butler and Hirano 2014). Along with the H1 human embryonic stem cell line and the GM12878 lymphoblastoid cell line, K562 is one of the three tier-one cell lines of the ENCyclopedia Of DNA Elements Project (ENCODE) (ENCODE Project Consortium 2012), forming the basis of over 1,300 datasets on the ENCODE data portal to date (Sloan et al. 2016). Furthermore, it is also one of a small number of cell lines most commonly used for large-scale CRISPR/Cas9 gene-targeting screens (Wang et al. 2015; Arroyo et al. 2016; Morgens et al. 2016; Han et al. 2017; Adamson et al. 2016; Liu et al. 2017).

Although the functional genomic characteristics of K562 are extensively studied and documented, reflected in close to 600 ChIP-Seq, 400 RNA-Seq, 50 DNase-Seq, and 30 RIP-Seq datasets available through the ENCODE portal (Sloan et al. 2016), the genomic sequence of the K562 cell line and any higher-order structural features of its genome have never been comprehensively characterized, even though past cytogenetic studies conducted on K562 cells using G-banding, fluorescence in situ hybridization (FISH), multiplex-FISH, and comparative genomic hybridization (CGH) have shown that K562 cells contain wide stretches of aneuploidy and multiple gross structural abnormalities (Selden et al. 1983; Wu et al. 1995; Naumann et al. 2001; Gribble et al. 2000), as is not unexpected for a cancer cell line. In other words, the rich amount of K562 functional genomics and epigenomics work conducted to date, in particular the large amounts of integrative analyses that have been carried out using the vast trove of

ENCODE data on K562, were done without taking into account the many differences of the K562 genome relative to the human reference genome. This could potentially lead to skewed interpretation of the epigenomics data and very likely reduce the amount of insights that can be gained from the rich, multi-layered data that has accumulated.

Here we report for the first time a comprehensive characterization of the genome sequence and genomic structure of the K562 cell line. We performed very-deep (>70x genome coverage) short-insert whole-genome sequencing (WGS), long-insert (3 kb) mate-pair sequencing (Korbel et al. 2007a), 10X-Genomics linked-read sequencing (Zheng et al. 2016), array CGH and karyotyping, and used a combination of novel and established computational tools to identify and catalog a wide spectrum of genomic variations and features. We first performed read-depth analysis using the very-deep WGS data to obtain the copy number (CN) or ploidy of chromosome segments, which were validated by array CGH and G-banding karyotype analysis. When identifying single nucleotide variants (SNVs, also including single nucleotide polymorphisms, i.e. SNPs) and small deletions and insertions (Indels), we took the CN of chromosomal regions into account in determining allele frequencies in a manner that lead to a CN-corrected catalog of these variants in the K562 genome. We found that many chromosomes show extensive loss of heterozygosity (LOH). Using linked-read sequencing we then constructed large phased haplotype blocks (N50 >2.72 Mb) and, by combining linked-reads, very-deep WGS and mate-pair data, we identified structural variants (SVs) including complex rearrangements and novel LINE1 and Alu mobile element insertions (MEIs). A large fraction of these SVs were also phased, sequence assembled, and many of the SVs and MEIs were experimentally validated with PCR and Sanger sequencing. Furthermore we took first steps into exploring how knowledge about genome sequence and structural features can influence the interpretation of functional genomics and epigenomics data. We found many examples of allele-specific gene expression. Using our large haplotype blocks as scaffolds, we phased methylated and unmethylated CpG dinucleotides (determined by K562 whole-genome

bisulfite sequencing) and identified loci with allele-specific DNA methylation. We also produced a phased CRISPR map of loci suitable for allele specific-targeted genome editing or screening, and finally, we show two examples of how deeper insights into genomic regulatory complexity can be obtained by integrating genomic sequence and structural context with functional genomics and epigenomics data.

RESULTS

Methods Overview

We combined multiple experimental and computational methods (Figure S1), including karyotyping, array CGH, very-deep (72x non-duplicate coverage) short-insert whole-genome sequencing (WGS), 3 kb-mate-pair sequencing (Korbel et al. 2007a) and 10X-Genomics linked-reads sequencing (Zheng et al. 2016), to comprehensively characterize the genome of the primary ENCODE cell line K562 (Figure 1). The WGS dataset was used to identify CN or ploidy by chromosome segments, SNVs, Indels, novel MEIs (LINE1 and Alu), and SVs such as deletions, duplications, inversions, insertions, and complex rearrangements. This set of SVs from the WGS dataset was identified using an integrated computational approach that includes BreakDancer (Chen et al. 2009), Pindel (Ye et al. 2009), BreakSeq (Lam et al. 2010) and ARC-SV (Arthur et al. 2017). The allele frequencies of heterozygous SNVs and Indels were determined by taking ploidy into account. The 10X-Genomics linked-reads dataset was used to phase SNVs and Indels as well as to identify, assemble, and phase primarily large (>30 kb) and complex SVs (Spies et al. 2017; Zheng et al. 2016). The 3 kb-mate-pair data was used to identify additional structural variants, mostly in the medium size-range (1 kb-100 kb) and also was used to validate large and complex SVs. Functional genomics and epigenomics data from ENCODE was integrated to identify allele-specific RNA expression and allele-specific DNA methylation, and phased variants were used to identify allele-specific CRISPR targets in the K562 genome.

Karyotyping

The K562 cell line was obtained from the Stanford ENCODE Production Center (NGHRI Project 1U54HG006996-01) and exhibits a high degree of aneuploidy (Figure 2A). Analysis of 20 individual K562 cells using the GTW banding method showed that all cells demonstrated a near-triploid karyotype and are characterized by multiple structural abnormalities. The karyotype of our K562 strain is overall consistent (although not identical) with previous published karyotypes of K562 (Selden et al. 1983; Wu et al. 1995; Naumann et al. 2001; Gribble et al. 2000), suggesting that its near-triploid state arose during leukemogenesis or early in the establishment of the K562 cell line and that different K562 cell lines that have been passaged for long times in different laboratories may exhibit some additional karyotypic differences. Although the karyotype for all chromosomes in our K562 cell line was supported by previous karyotype analyses, slight variations do exist among the various published analyses (Table S1) with chromosomes 10, 12, and 21 showing the most variability.

Identification of Genome CN by Chromosome Segments

We used read-depth analysis (Abyzov et al. 2011) to assign a CN to all chromosome segments at 10kb-resolution or entire chromosomes in the K562 genome (Figure 1, Table S2). We first calculated WGS read coverage in 10 kb bins across the genome and plotted the coverage against percent GC content where five distinct clusters were clearly observed (Figure S2). Clusters were designated as corresponding to particular CN based on the mean coverage of each cluster. Such designations revealed that the expected triploid state is the most common in the K562 genome. The CN (or ploidy) assigned to all chromosome segments using this approach are consistent with array CGH (Figure S3) and also with previous CGH analyses (Gribble et al. 2000; Naumann et al. 2001) with minor differences on chromosomes 7, 10, 11, and 20 (Table S3). Notably, while on a very general level the CN identified based on WGS analysis tracks the findings from karyotyping, the sequencing-based analysis reveals the CN changes of many chromosome segments that would not have been apparent from karyotyping alone (Figure 2A and Figure S3). From CN analysis of chromosome segments (Table S2) we

see that 53.5% of the genome has a baseline copy number of three (consistent with the karyotype), 16.9% copy number of four, 1.9% copy number of five, 2.4% copy number of one, and only 30.0% has remained in a diploid state (Figure 2B); several regions in the range from 0.6Mb to 4.7Mb in size residing on chromosomes 4, 6, and 9 have only one copy (Table S2). In addition, two large regions (5.8 Mb and 3.1 Mb in size) on chromosome 9 (20,750,000-26,590,000 and 28,560,000-31,620,000 respectively) were lost entirely (Table S2).

SNVs and Indels

We identified SNVs and Indels and, by taking into account the CN of the chromosomal segments in which they reside, we were able to assign heterozygous allele frequencies to these variants, including non-conventional frequencies (e.g. 0.33 and 0.67 in triploid regions; 0.25, 0.50, and 0.75 in tetraploid regions). Using this approach, we detected and genotyped a total of 3.09 M SNVs (1.45 M heterozygous, 1.64 M homozygous) and 0.70 M Indels (0.39 M heterozygous, 0.31 M homozygous) in K562 (Table 1 and Dataset S1). Interestingly, there are 13,471 heterozygous SNVs and Indels that have more than two haplotypes in chromosomal regions that are more than diploid copies (Dataset 1). Furthermore, chromosomes 3, 9, 13, 14, X, and large contiguous stretches of chromosomes 2, 10, 12, 17, 20, and 22 show striking loss-of-heterozygosity (LOH) (Figure 1 and Table S4). While a normal tissue sample corresponding to K562 is not available for comparative analysis, we overlapped these SNVs and Indels with dbsnp138 (Sherry et al. 2001) and found the proportion of overlap to be ~98% and ~79% respectively (Figure 2C and Dataset S1), suggesting that K562 has accumulated a significant number of somatic SNVs and Indels relative to germline variants present in the population. We found that 424 SNVs and 148 Indels that are present in the K562 genome are private protein-altering (PPA), after filtering for protein-changing SNPs and Indels that overlapped with variants from the 1000 Genomes Project or from the Exome Sequencing Project (Table 1, Dataset S2, and Dataset S3). Furthermore, the overlap between the filtered PPA variants and the Catalogue of Somatic Mutations in Cancer (COSMIC) is 53% and 31% for SNVs and Indels (Table S5).

Eighteen genes that acquired PPA variants in K562 overlap with the Sanger Cancer Gene Census (Table S6), and canonical tumor suppressor genes and oncogenes were notably present, such as *RAD51B*, *TP53*, *PDGFRA*, *RABEP1*, *EPAS1*, and *WHISC1*.

Haplotype Phasing

We resolved the haplotype phase of SNVs and Indels in the K562 genome by performing 10X-Genomics linked-read sequencing (Zheng et al. 2016). K562 genomic DNA was size selected for fragments >30 kb, and 1.06 ng, or approximately 320 genomic equivalents, of high molecular weight (HMW) genomic DNA fragments (mean fragment size = 59 kb, 95.3% >20kb, 11.9% >100 kb), were partitioned into ~1.56 million oil droplets in emulsion, uniquely barcoded (16 bp) within each droplet, and subjected to random priming and isothermal amplification. The emulsion is then broken and barcoded DNA molecules are released and converted to a modified paired-end Illumina sequencing 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. In this manner reads that cover heterozygous SNVs and Indels can be phased by their "HMW fragment barcode". Sequencing (2 x 151 bp) of this library was performed at 59x genome coverage. Half of all reads come from HMW DNA molecules with at least 64 linked-reads (N50 Linked-Reads per Molecule or LPM). We estimate the actual physical coverage (C_F) to be 191x. The overall sequencing coverage is $C = C_R \times C_F = 59x$. The length of actual sequence coverage per 2x151 bp paired-ended read minus 16 bp of "HWM fragment barcode" is 286 bp, thus coverage of the mean of the input HMW genomic DNA (59 kb) by sequencing (C_R) is 18,304 bp (286 bp x 64) or 31.0%. Using the Long Ranger phasing algorithm from 10X-Genomics (Zheng et al. 2016), 1.41 M (97.2%) of heterozygous SNVs and 0.58 M (83.7%) of Indels were successfully phased and distributed over 4,987 haplotype blocks (Figure 1, Table 1**,** and Dataset 4). The longest phased haplotype block is 11.95 Mb (N50 =

2.72 Mb) (Figure 2d, Table 1, and Dataset 4); however, the distribution of haplotype block lengths varies widely across different chromosomes (Figure S4 and Figure 1) with poorly phased regions corresponding to regions with LOH (Figure 1, Table S4, Dataset S1, and Dataset S4).

Identification and Reconstruction of SVs from Linked-Reads

In addition to the phasing of haplotypes, another use for the linked-read sequencing data is to identify breakpoints of large-scale SVs by searching for discordant mapping of groups of linked-reads carrying the same barcodes. The identified SVs can then also be assigned to specific haplotypes if the breakpoint-supporting reads contain phased SNVs or Indels (Zheng et al. 2016). Using this approach, which is also implemented by the Long Ranger software from 10X-Genomics, we identified 186 large SVs larger than 30 kb (98% phased) (Dataset S5) and 3,541 deletions between 50 bp and 30 kb (79% phased) (Dataset S6). The set of large SVs includes deletions, inversions, duplications, and inter- and intra-chromosomal rearrangements (Dataset 5 and Figure 3A-D). As expected, we detected the *BCR/ABL1* gene fusion, a hallmark of K562, as one of the SV calls with highest quality score (Figure 3A), along with two other known gene fusions in K562 (Engreitz et al. 2012): *XKR3/NUP214* between chromosomes 9 and 22 (Figure 3B) and *CDC25A/GRID1* between chromosomes 3 and 10 (Dataset S5).

In addition, we also leveraged the long-range information derived from the 10X-Genomics linked-reads to identify, assemble, and reconstruct SV-spanning breakpoints (including those of complex SVs) in the K562 genome using the recently established computational method Genome-wide Reconstruction of Complex Structural Variants (GROC-SVs) (Spies et al. 2017). In this method, long DNA fragments that span breakpoints are statistically inferred and refined by quantifying barcode similarity between pairs of genomic regions, similar to Long Ranger (Zheng et al. 2016). Sequence reconstruction is then performed by assembling the relevant linked-reads around the identified breakpoints from which complex SVs are then automatically reconstructed. Only SV breakpoints that also have supporting

evidence from the K562 3 kb-mate-pair dataset (see Methods) were retained as high-confidence events. GROC-SVs called a total of 161 high-confidence breakpoints including 12 inter-chromosomal events (Figure 1, Dataset S7 and Figure 4A, B); each event is accompanied with visualization (Dataset S7); 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 (Dataset 7). A striking example of assembly by GROC-SVs is a homozygous complex event on chromosome 6 that involves an inversion of ~40 kb flanked by ~50 kb deletions (Figure 4A); another notable example is a complex intra-chromosomal rearrangement that has occurred on chromosome 13 (Figure 4B).

Complex Rearrangements from Deep WGS

Additional complex structural variants (Figure 5A-E), most of them of a smaller size than is typically covered by other approaches, were identified using a novel algorithm called Automated Reconstruction of Complex Structural Variants (ARC-SV) (Arthur et al. 2017) using the very-deep paired-end WGS reads. ARC-SV detects both simple variants and complex rearrangements through a two-stage approach. Complex events are defined as genomic rearrangements that contain breakpoints not explained by simple events meaning deletion, insertion, tandem duplication, and inversion. In ARC-SV analysis, structural variants are identified through a two-stage approach in which candidate breakpoints in the reference are generated using soft-clipped and split-read alignments from WGS, and nearby breakpoints are clustered as dictated by the data. The clusters typically span 1-100 kb and contain several to a dozen breakpoints. Within each breakpoint cluster, variants are called by rearranging the intervening genomic segments and scoring each configuration according to its likelihood under a generative statistical model for the paired-end alignments to that region. For complex events, each breakpoint is required to have either split-read support or to have >95% of overlapping reads with a mapping quality score >20. Using this approach and after filtering out events less than 50 bp in size or with breakpoints that reside in simple repeats, low complexity regions,

satellite repeats, or segmental duplications, we identified 122 complex events (accompanied with schematic visualizations), 2,235 deletions, 320 tandem duplications, and 6 inversions (Dataset S8). These complex variants cannot be identified using other methods that we are aware of. Examples of such events consist of dispersed duplications (Figure 5A), sometimes with inversions of the inserted sequence and possibly a deletion at the insertion site (Figure 5B, C), inversions flanked on one or both sides by deletions (Figure 5D), duplications that involve multiple non-exact copies, as well as deletion, inversion, and multiple duplications residing at the same locus (Figure 5E). Eight out of ten breakpoints from five complex events identified by ARC-SV were successfully validated individually by using PCR and Sanger sequencing (Table S7).

SVs from Mate-Pair Sequencing Analysis

In order to increase the sensitivity for detecting medium-sized variants (1 kb-100 kb) we constructed a 3 kb-mate-pair library for the K562 cell line and obtained a 2x151 bp sequencing coverage of 6.9x after duplicate removal. Coverage of each 3 kb insert by sequencing (C_R) is 300bp or 10%, which translates to a physical coverage (C_F) of 68.5x. SVs (deletions, inversions, and in particular tandem duplications) from 3 kb-mate-pair libraries were identified by clustering discordant read pairs and split-reads using LUMPY (Layer et al. 2014). Only variants that have support from both discordant-reads and split-reads were retained. Overall, we identified 270 deletions, 35 inversions, and 124 tandem duplications using this approach (Dataset S9). Approximately 83% of these variants are between 1 kb-10 kb, and 88% are between 1 kb-100 kb (Dataset S9). A set of deletion (*n*=12) and tandem duplication calls (*n*=5) from the 3 kb-mate-pair dataset were randomly selected for PCR and Sanger sequencing validation (Table S7). The validation rates were 83% and 80% for deletions and tandem duplications, respectively.

Non-Complex SVs from Deep WGS

Non-complex SVs (simple deletions, inversions or insertions, and simple tandem duplications) were called from the K562 WGS dataset with a combination of established

computational tools, namely Pindel (Ye et al. 2009), BreakDancer (Chen et al. 2009), and BreakSeq (Lam et al. 2010). These SVs were combined with those that were identified previously using ARC-SV, LUMPY and Longranger, where variants (*n*=2,665) with support from multiple methods by more than 50% reciprocal overlap had been merged. In total, 9,082 non-complex structural variant were obtained from all methods, including 5,490 deletions, 531 duplications, 436 inversions, and 2,602 insertions (we note that only BreakDancer (Chen et al. 2009) was designed to call insertions) (Dataset S10). Consistent with previous analyses (e.g. as in (Lam et al. 2012)), deletions show the highest number of concordant calls compared to duplications and inversions, across the various computational tools (Figure S5 and Dataset S10). 16/18 (89%) randomly selected deletions larger than 1 kb and with split-read support, and 13/18 (72%) randomly selected tandem duplications with split-read support were successfully validated using PCR and Sanger sequencing (Table S7).

Mobile Element Insertions

Lastly we also analyzed the number of non-reference LINE1 and Alu mobile element insertions in the K562 genome using the very-deep short-insert WGS data and a modified RetroSeq (Keane et al. 2013) approach. LINE1 and Alu mobile element insertion events were identified from paired-end reads that have one of the reads mapped to the human reference genome and the other read mapped to the Alu or LINE1 consensus sequence in full or split fashion (see Methods). Mobile element insertion events with greater than five supporting reads were deemed as high confidence calls (Table S8). Using this approach, we identified 1,147 novel Alu insertions and 85 novel LINE1 insertions in the genome of K562 (Figure 1). Nine Alu and ten LINE1 insertion events with split-read support were randomly chosen for PCR and Sanger sequencing validation and were successfully validated, at 88% and 100% respectively (Table S9). PCR primers for these validation experiments were designed such that one of the primers was bound within the mobile element sequence and the other primer bound in the unique reference sequence surrounding the given predicted mobile element insertion site.

Allele-Specific Gene Expression

Using the copy numbers of the SNV alleles that we identified (Dataset S1), we re-analyzed two replicates of polyA-mRNA RNA-Seq data (experiment ENCSR000AEM from the ENCODE Portal (Sloan et al. 2016) to identify allele-specific gene expression in K562. We identified 5,980 and 6,190 genes that show allele-specific expression ($p < 0.05$) in replicates one and two, respectively **(**Figure 1, Table S10**)**. We also identified 2,359 and 2,643 genes that would have been falsely identified to have allele-specific expression and 1,808 and 2,063 genes that would not have been identified to have allele-specific expression in replicates one and two, respectively, if the copy numbers of SNV allele frequencies were not taken into consideration **(**Table S11**)**.

Allele-Specific DNA methylation

Using the phased heterozygous SNVs of K562, we identified 228 CpG islands (CGIs) that exhibit allele-specific DNA methylation. We obtained K562 whole-genome bisulfite sequencing reads (2x100 bp, library ENCLB742NWU) from the ENCODE Portal (Sloan et al. 2016) and aligned the reads to hg19 using Bismark (Krueger and Andrews 2011), where 76.9% of reads were uniquely mapped and 26.2% of cytosines were methylated in a CpG context. We then used reads that overlap both phased heterozygous SNVs (Dataset 4) and CpGs to phase the methylated and unmethylated CpGs to their respective haplotypes. Fisher's exact test (taking the CN of a given chromosomal segmental into consideration) was used to evaluate allele-specific methylation. Of the total 234,031,434 CpGs in the human genome, we phased 47,515,608 (20.3%), of which 28,204 CpGs exhibited allele-specific methylation ($p < 0.05$ with Fisher's exact test) (Table S12). We then grouped the phased individual CpGs into CGIs and found 228 CGIs to be methylated in an allele-specific manner (Figure 1, Table S12). Of these 228 CGIs, 68 reside within promoter regions (here defined as 1 kb upstream of a gene); 172 are intragenic, and 56 lie within 1 kb downstream of 223 different genes. The following 8 genes are

within 1 kb of a differentially methylated CGI and overlap with the Sanger Cancer Gene Census: *ABL1*, *AXIN2*, *CCND1, EXT2*, *HOXD11*, *KDR*, *PRDM16,* and *ZNF331*.

Allele-Specific CRISPR Targets

We identified a total of 28,511 targets in the K562 genome suitable for allele-specific CRSIPR targeting (Figure 1, Table S13). Sequences (including reverse complement) of phased variants that differ by more than one base pair between the alleles were extracted to find all possible CRISPR targets by searching for the pattern [G, C, or A]N_{20}GG (see Methods). Using a selection method previously described and validated (Sunagawa et al. 2016), only conserved high-quality targets were retained. We also took gRNA function and structure into consideration and performed furthering filtering of CRISPR targets. Targets with multiple exact matches, extreme GC content, and those containing TTTT (which might break the secondary structure of gRNA), were removed. We also used the Vienna RNA-fold package (Lorenz et al. 2011) to compute gRNA secondary structure and eliminated all targets for which the stem loop structure (for Cas9 recognition) could not form (Nishimasu et al. 2014). Finally, we calculated the off-target risk score by using the tool as described in (Ran et al. 2013). To ensure that all targets are as reliable and as specific as possible, we chose a very strict threshold and rejected candidates with a score below 75.

Genomic Structural Context Provides Insight into Regulatory Complexity

We show two examples of how deeper insights into gene regulation and regulatory complexity can be obtained by integrating genomic structural contexts with functional genomics and epigenomics data (Figure 6A-D). One example is the allele-specific RNA expression and allele-specific DNA methylation in K562 at the *HOXB7* locus on chromosome 17 (Figure 6A). By incorporating the genomic context in which *HOXB7* is expressed in K562 cells, we see that *HOXB7* RNA is only expressed from the two copies of Haplotype 1 (p = 0.007) in which the CGI near its promoter is completely unmethylated (p = 3.18E-18) (Figure 6 A, C). The second example is allele-specific RNA expression and allele-specific DNA methylation of the *HLX* gene

in K562 (Figure 6B). The *HLX* locus on chromosome 1 is in a tetraploid region, and we see that *HLX* is only expressed from Haplotype 1 which has three copies and not expressed in Haplotype 2 (p = 0.043) (Figure 6B, D). The CGI in Haplotype 2 of the *HLX* locus is unmethylated but highly methylated on Haplotype 1 (p = 5.14E-15) (Figure 6B, C). There is also an allele-specific CRISPR targeting site for both haplotypes near the *HLX* gene (Figure 6B).

In addition, we performed Pearson correlation analysis between our deep K562 WGS reads and K562 POLR2A ChIP-Seq reads (previously released on the ENCODE data portal) to determine whether changes in K562 genome CN or ploidy affected binding of the polymerase molecule to genomic DNA in a large-scale fashion (Figure S6). The two sets of data are very well correlated (r=0.51, p<2.2E-16) suggesting that RNA polymerase activity is generally influenced by ploidy in the K562 genome. Furthermore we also correlated the K562 POLR2A ChIP-Seq reads with the FPKM values of four independent K562 polyA RNA-Seq experiments (also previously released on the ENCODE portal) and find that these datasets are consistently very well correlated as well (r=0.46, p<2.2E-16; r=0.58, p<2.2E-16; r=0.47, p<2.2E-16; r=0.46, p<2.2E-16) (Figure S7A-D).

DISCUSSION

K562 is one of the most widely used laboratory “work-horse” cell lines in the world. Among the three tier-one cell lines of ENCODE, K562 has by far the most functional genomics and epigenomics data generated. Furthermore, K562 is also one of the most commonly used cell lines for large-scale CRISPR/Cas9 gene-targeting screens (Wang et al. 2015; Arroyo et al. 2016; Morgens et al. 2016; Han et al. 2017; Adamson et al. 2016; Liu et al. 2017). Yet, despite its wide usage and impact on biomedical research, its genome sequence and genomic structural features have never been comprehensively characterized, beyond its karyotype (Selden et al. 1983; Gribble et al. 2000; Wu et al. 1995; Naumann et al. 2001) and SNPs called from 30x WGS using GATK Haplotypecaller but without taking aneuploidy or copy number into consideration (Cavalli et al. 2016). Analysis, integration, and interpretation of the extensive

collection of functional genomics and epigenomics datasets for K562 has so far relied on the human reference genome. Here, we present the first detailed and comprehensive characterization of the genome of the K562 cell line. By performing very-deep short-insert WGS, 3 kb-insert mate-pair sequencing, 10X-Genomics linked-reads sequencing, array CGH, karyotyping, and using a compendium of novel and established computational analysis methods, as well as integrating the findings from the various approaches, we produced a comprehensive spectrum of genomic structural features (Figure 1) for the K562 cell line that includes SNVs (Dataset S1), Indels (Dataset S1), CN or ploidy by chromosome segments at 10 kb resolution (Table S2), phased haplotype blocks (Dataset S4), phased CRISPR targets (Table S13), novel mobile element insertions (Table S8), and SVs (Dataset S10) including simple deletions, duplications, and inversions, and those that are the result of complex genomic rearrangements (Dataset S7 and Dataset S8). Many of these identified SVs are also phased, assembled, and experimentally verified (Dataset S4, Dataset S7, Table S7, and Table S9).

Chromosomal aneuploidy is a hallmark of many cancers. Previous studies have detected and subsequently confirmed a near triploid karyotype of the K562 cell line (Selden et al. 1983; Gribble et al. 2000; Wu et al. 1995; Naumann et al. 2001). In our analysis this general finding of near-triploidy also held but there are considerable portions of the K562 genome where we found the situation to be much more varied than what had previously been reported. One main reason for this is that, even at the multi-million basepair windows that constitute the main output of our read depth-based analysis of the copy number of entire chromosomal segments, our analysis is of much higher resolution than a karyotype. In addition, it has to be taken into account that if a cell line that has been passaged for as long and in as many different laboratories as K562 has, there are opportunities for additional genome variation to occur. In light of this, it is reassuring that the overall picture presented by the karyotype has not changed much over several decades. However, researchers should still keep this aspect in mind, in particular when working with a version of K562 that has been passaged many times and

separately from the main ENCODE K562 production line we used for our analysis here. Our expectation would be that the vast majority of sequence and structural variants that we describe here can be found across different versions of K562, but for individual variants there is always the chance that different lines of K562 cells may have slightly diverged from each other (Table S1 and Table S3). This is also an important aspect to consider, for example, when studying the various functional genomic datasets available for K562 that have accumulated over the years on ENCODE, in particular when following up on findings for individual loci and using a different K562 line for such follow-up work. A first step that should always be considered in such work would be to experimentally validate the presence of the particular genomic variant of interest in that particular version of K562. For analyses that are carried out on global levels, interrogating and piecing together network and multi-omics interactions, we would consider that the vast majority of the genomic variants described here will exist in the vast majority of K562 sub-lines, and therefore such global analyses should be well feasible and can be expected to yield substantial insights. Even though the high degree of aneuploidy in K562 renders the design and interpretation of K562 genomic and epigenomic studies more challenging, the information we provide in this study enables researchers to continue using this cell line to investigate the effects of different types of genome sequence variation on the multiple levels of functional genomics activity and regulation for which ENCODE data already exists or continues to be produced. Thus analysis of K562 data should not only be more complex and challenging but also potentially much more insightful and rewarding when taking the complex genome structure into account.

Sensitive and accurate identification of SNVs and Indels requires relatively deep WGS coverage (>33x and >60x respectively) (Bentley et al. 2008; Fang et al. 2014). From our >70x coverage WGS we identified large numbers of SNVs and Indels that we could subsequently correct for their allele frequencies according to ploidy. In addition to being essential for correct haplotype identification, these ploidy-corrected variants are also needed for functional genomics

or epigenomics analyses such as the determination of allele-specific gene expression or of allele-specific transcription factor binding in K562 (Cavalli et al. 2016). A statistically significant increase in transcription or transcription factor binding signal at one allele compared to the other at a heterozygous locus, in RNA-Seq or ChIP-Seq data, may be identified 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 consideration and the RNA-Seq or ChIP-Seq signals normalized by ploidy, the case identified might be a result of increased copy number rather than the preferential activation of one allele over the other on the epigenomic level. Indeed, in our re-analysis of two replicates of K562 RNA-Seq data, we identified 2,359 and 2,643 genes that would have been falsely identified to have allele-specific expression in addition to 1,808 and 2,063 genes that would not have been identified to have allele-specific expression in replicates one and two, respectively, if haplotype ploidy was not taken into consideration (Table S11).

The haplotype phase of genomic sequence variants is an essential aspect of human genetics, but current standard WGS approaches entirely fail to resolve this aspect. We performed linked-read sequencing of K562 genomic DNA using the Chromium System from 10X-Genomics (Zheng et al. 2016). After size-selecting for genomic DNA fragments >30kb, 300 genomic equivalents of HMW DNA were partitioned into more than one million oil droplets, uniquely barcoded within each droplet, and subjected to random priming and amplification. Sequencing reads that originate from the same HMW DNA molecule can be identified by their respective droplet barcodes and linked together to produce virtual long reads. Then, by looking for virtual long reads that overlap a previously called set of heterozygous haplotypes (Dataset S1), the phase information of the heterozygous haplotypes was determined and the virtual long reads were constructed into contiguous fragments of phased variant blocks with N50 > 2.72 Mb (Dataset S4, Figure 2D). Chromosomes 3, 9, 13, 14, X, and large portions of chromosomes 2, 10, 12, 20, 22 were especially difficult to phase, resulting in comparatively shorter phased

blocks (Dataset S4, Figure 1, Figure S4). This is not surprising since these chromosomes and chromosomal regions exhibit a very high degree of LOH (Figure 1 and Table S4). Heterozygous loci in aneuploidy regions with more than two haplotypes were excluded from phasing linked-read analysis due to software and algorithmic limitations (Zheng et al. 2016). However, the phase information of these loci could be resolved from our linked-read data in principle, should new algorithms become available.

It has been shown previously that integrating orthogonal methods and signals improves SV-calling sensitivity and accuracy (Mohiyuddin et al. 2015; Layer et al. 2014). Here, we combined very-deep short-insert WGS, long-insert 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 complex rearrangements. The non-complex SV calls were merged and those identified by more than one method were indicated accordingly (Dataset S10). Overall, we see significant overlap as well as variant calls that are specific to each method for deletions (Figure S5A), but overlap is less pronounced for duplications (Figure S5B) and inversions (Figure S5C). This is consistent with previous analysis (Lam et al. 2012) as inversions and duplications are more difficult in principle to accurately resolve. Such results are also expected since each method is designed to utilize different types of signals for SV calling and also optimized to identify different classes of SVs. Again, further investigations of particular SVs not experimentally verified in this study should be preceded by additional experimental validation.

The complex rearrangements identified by ARC-SV from short-insert WGS (Figure 5A-E) and by GROC-SVs from linked reads with assembly (Figure 4A, B) are especially interesting because they are a class of SV that cannot be easily identified and automatically reconstructed using previously existing SV analysis methods. Before the existence of linked-read sequencing, constructing haplotype blocks and resolving large SVs (>30kb) relied heavily on 40kb-fosmid

libraries (Snyder et al. 2015; Williams et al. 2012; Kitzman et al. 2011; Cao et al. 2015; Adey et al. 2013) which were laborious, costly, time consuming, and much less efficient.

The hallmark of CML is the Philadelphia rearrangement t(9; 22)(q34; q11) which results in the fusion of the *ABL1* and *BCR* genes (Heisterkamp et al. 1985; de Klein et al. 1982; Groffen et al. 1984). This gene fusion is known to be extensively amplified in the K562 genome by tandem duplication (Wu et al. 1995). FISH hybridization analysis have shown that *BCR/ABL1* gene fusion fluorescent signals almost always concentrate on a single marker chromosome (Tkachuk et al. 1990; Wu et al. 1995; Gribble et al. 2000). This is also consistent with our data as the linked-reads that support the *BCR/ABL1* gene fusion do not share overlapping barcodes with linked-reads that align to elsewhere in the genome, and the *BCR* and *ABL1* gene regions where the fusion occurs show a >2.8x increase in depth of coverage relative to average genome coverage.

All data and results generated from this comprehensive whole-genome analysis of K562 is available through the ENCODE portal (Sloan et al. 2016). We envision that this analysis will serve as a valuable reference for further understanding the vast amount of existing ENCODE data available for the K562 cell line, such as determining whether a potential or known regulatory sequence element has been altered by SNVs or SNPs, Indels, mobile element insertions, a gain or loss of copies of that given element, or allele-specific regulation. As an initial demonstration of the power of how integrating the genomic context of K562 can yield further understanding of existing ENCODE data, we showed the complex gene regulatory scenario at the *HOXB7* and *HLX* loci in K562 as examples. Hox genes are known to have important roles in hematopoiesis and oncogenesis (Argiropoulos and Humphries 2007; Shah and Sukumar 2010; Eklund 2011) where HOXB7 transcription factor mediates lymphoid development, hematopoietic differentiation and leukemogenesis (Giampaolo et al. 1995; Carè et al. 1999). *HOXB7* overexpression has been reported in leukemia (Raval et al. 2007) as well as many other cancers (Caré et al. 1996; Wu et al. 2006; Yamashita et al. 2006; Shiraishi et al.

2007; Chen et al. 2008; Storti et al. 2011). It is directly upstream of *HOXB8*, which is the first Hox gene found to be an oncogene in leukemia (Blatt et al. 1988). *HLX* has also been suggested to play oncogenic roles in leukemia (Deguchi et al. 1992; Deguchi and Kehrl 1993; Jawad et al. 2006; Fröhling 2012). By integrating the genomic context of *HOXB7* and *HLX* in K562 with RNA-Seq and WGBS data, we see that the RNA of both genes are expressed from haplotypes that exhibit aneuploidy and in an allele-specific manner (Figure 6A,B,D). The allele-specific methylation of the CGIs near these two genes is associated with active transcription in the case of *HLX* and silencing of transcription in the case of *HOXB7* (Figure 6A-C). Such insights into gene regulation cannot be obtained by analyzing functional genomics and epigenomics data alone, after mapping data only onto the reference genome sequence. Furthermore, we also observed that the K562 POLR2A ChIP-Seq signal is very well correlated with both polyA RNA-Seq experiments and genome-segment copy-numbers as determined from our deep WGS sequencing dataset, suggesting an association between both polymerase binding and active transcription and ploidy (Figure S6, S7).

Our work here will also serve to guide future study designs that utilize the K562 "workhorse" cell line, such as CRISPR screens where knowledge of the SNPs can extend or modify the number of editing targets (Table S13) while knowledge of aberrant copy numbers will allow confident interpretation of non-diploid regions. To give an example, in a recent study that uses CRISPRi to screen and elucidate the function of long non-coding RNAs in human cells, out of the seven cell types studied, the number of gRNA hits varied considerably among the various cell types, with 89.4% of hits unique to only one cell type and none in more than five cell types (Liu et al. 2017). Although a large portion of the phenomenon can very likely be attributed to cell-specific effects, it is still quite possible that many of the gRNA hit differences can be the result of differences in genome sequence, and therefore differences in CRISPR targeting sites, rather than differences between cell lines. Our list of allele-specific CRISPR targets (Table S13) will allow for a separation between these two potential reasons for differences in CRISPR

effects during screens and should be particularly valuable for future large-scale CRISPR screens that utilize K562. Lastly, this study may serve as a technical example for advanced, integrated, and comprehensive analysis of genomic sequence and structural variants for other heavily utilized cell lines and genomes in biomedical research such as HepG2.

MATERIALS & METHODS

Genomic DNA extraction

The K562 cell line was obtained from the Stanford ENCODE Product Center. Genomic DNA was extracted using the DNeasy Blood & Tissue Kit (Cat No./ID: 69504) from QIAGEN following standard manufacturer's protocols. K562 genomic DNA concentration was quantified using the Qubit dsDNA HS Assay Kit (Invitrogen Life Technologies, Waltham, MA). DNA was then verified to be pure (OD260/280 > 1.8; OD260/230 > 1.5) and of high molecular weight (>30 kb) using field-inversion gel electrophoresis on the Pippin Pulse System (Sage Science, Beverly, MA, USA).

Karyotyping

K562 cells were sent to the Stanford Cytogenetics Laboratory at the Stanford University Medical Center (Palo Alto, CA, USA) for karyotyping, where twenty metaphase cells were analyzed using the GTW banding method.

Array CGH analysis

Raw data was generated using the NimbleGen aCGH platform and the NimbleGen 2.1 M array from Roche NimbleGen, Inc. (Madison, WI, USA) (Urban et al. 2006; Korbel et al. 2007b, 2009). Experiments were performed following the manufacturer's protocol and also described in (Haraksingh et al. 2011). Two technical replicates were performed. K562 genomic DNA was used as the test sample and a pool of female genomic DNA from Promega (Madison, WI, USA) was used as the control. In brief, K562 genomic DNA was labeled with cy3 and the control pool DNA with cy5. Equal amounts of the K562 and control genomic DNA were hybridized to the arrays for 72 hours. The arrays were washed and scanned in an ozone free environment using

a Roche MS200 scanner. Images were analyzed using NimbleScan 2.6 software (Roche NimbleGen, Inc., Madison, WI 53719, USA). Segtable files were generated using NimbleScan 2.6 with default settings (min segment difference = 0.2, min number of probes per segment = 2). Segments with −0.25<Log R<0.25 were removed. Segments with <5 probes per segment were also removed. The 2.1 M designs are based on hg18 coordinates. These CNV coordinates were converted to hg19 using the UCSC LiftOver tool (Hinrichs et al. 2006).

Illumina short-insert whole-genome sequencing

K562 genomic DNA was sent to Macrogen (Rockville, MD, USA) for standard Illumina (San Diego, CA) short-insert whole-genome sequencing library preparation and sequencing. Two sequencing libraries were prepared and sequenced at 2x151 bp read-length on two lanes of the Illumina HiSeq X to achieve deep (>70x) genomic sequence coverage. Reads were aligned to the human reference genome (hg19) using BWA-MEM version 0.7.5 (Li and Durbin 2009) followed by marking of duplicates using Picard tools (version 1.129; *doc_min_qual=20, doc_window_width=20*) (http://broadinstitute.github.io/picard/) and local Indel realignment and base quality score recalibration (*maximum_cycle_value=500, cov={ReadGroupCovariate, QualityScoreCovariate, CycleCovariate, ContextCovariate}*) using Genome Analysis Tool Kit (GATK) (McKenna et al. 2010; DePristo et al. 2011).

Determining CN by chromosome segments and allele frequencies of SNVs and Indels

Read coverage along the K562 genome was calculated in 10 kb bins genome-wide. Distribution of read coverage in 10 kb bins was plotted against the GC content of each bin to verify the existence of distinct clusters. Copy number (CN) was assigned to clusters based on the ratio of mean coverage. To give 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 ratio of mean coverage of the five distinct clusters observed corresponded almost perfectly to CN1, CN2, CN3, CN4, and CN5. Afterwards, read coverage across the genome was examined visually to assign changes in genome-wide copy numbers. A switch in copy number was

assigned if there was a sharp change in read coverage with adjacent regions that correspond to the designated copy numbers from cluster analysis.

Afterwards, genomic regions were partitioned according to copy number (ploidy), and SNVs and Indels were called in each chromosomal region separately by GATK Haplotypecaller, by specifying the ploidy of the genomic region (stand_emit_conf=0.1, variant_index_type=LINEAR, variant_index_parameter=128000, ploidy={copy number}). The resulting raw GATK Haplotypecaller outputs were then concatenated, and variant quality scores were recalibrated using training datasets (dbSNP 138, HapMap 3.3, Omni 2.5 genotypes, 1000 Genomes Phase 1) as recommended by the GATK Best Practices (Van der Auwera et al. 2013; DePristo et al. 2011) and filtered accordingly (tranche = 99.0). SNVs and Indels were annotated with dbSNP138 (Sherry et al. 2001) with GATK followed by annotation with 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-affecting variants were intersected with the variants from the 1000 Genomes Project and the Exome Sequencing Project (1000 Genomes Project Consortium et al. 2015; http://evs.gs.washington.edu/EVS/) and overlapping variants removed from the callset using Bedtools (version 2.26) (Quinlan and Hall 2010). The resulting variant calls were compared against the Catalogue of Somatic Mutations in Cancer and Sanger Cancer Gene Census (Forbes et al. 2015; Futreal et al. 2004).

Identification of regions exhibiting 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 unphased SNV calls that: (1) were recalibrated and "PASS" filtered from VQSR, (2) overlapped with 1000 Genomes Project variants. The genome was split into 40 kb bins, heterozygous and homozygous SNV calls were tallied for each bin, and bins with less than 12 calls were removed. If a bin's callset has greater than or equal to 50% heterozygous calls, it was classified as heterozygous,

otherwise it was classified as homozygous. The homozygous and heterozygous classifications were used as the model's emission sequence. The model was initialized with equal initiation probabilities and transition probabilities of 10E-8 (Adey et al., 2013), and the Viterbi algorithm was used to estimate a best path. Adjacent LOH intervals were subsequently merged.

Haplotype phasing and variant calling from 10X-Genomics linked-reads

Genomic DNA fragments (~35 kb-80 kb) were first size selected on the BluePippin instrument from Sage Science (Beverly, MA, USA) using the manufacturer's U1 Maker 30 kb High Pass protocol and then diluted to 1 ng/µl to be used as input for the 10X-Genomics (Pleasanton, CA, USA) Chromium reagent delivery system (Zheng et al. 2016). A 10X-Genomics linked-read library was made following standard manufacturer's protocol with 8 cycles of PCR amplification. Libraries were then diluted to 5nM and sent to Macrogen (Rockville, MD, USA) for sequencing (2x151 bp) on two lanes of the Illumina HiSeq X to achieve ~60x coverage. Paired-end linked-reads (median insert size 385 bp, duplication rate 6.19%, Q30 Read1 88.7%, Q30 Read2 63.8%) were aligned to the human genome reference assembly hg19 (alignment rate 90.1%, mean coverage 59.0x, zero coverage 1.14%) and phased using the Long Ranger Software (version 2.1.5) from 10X-Genomics (Pleasanton, CA, USA). Regions with gaps in the reference, with assembly issues such as unplaced contigs, regions that are highly polymorphic (as compiled by 10X-Genomics http://cf.10xgenomics.com/supp/genome/hg19/sv_blacklist.bed) and regions containing segmental duplications (http://cf.10xgenomics.com/supp/genome/hg19/segdups.bedpe) were excluded from the analysis. ENSEMBL annotations were used for genes and exons (http://cf.10xgenomics.com/supp/genome/gene_annotations.gtf.gz). Phasing was done by specifying the set of pre-called and filtered K562 heterozygous SNVs and Indels from GATK (see above) and formatted using the mkvcf tool Long Ranger Software Suite (version 2.1.5). Heterozygous SNVs and Indels with more than two haplotypes were excluded from analysis. Variants were identified using both the Long Ranger *wgs* module with the "somatic" option on,

and GROC-SVs (Spies et al. 2017). Variants from Long Ranger were "PASS" filtered according to default settings. Variants from GROC-SVs were only retained if there was supporting evidence from mate-pair reads (see below).

Allele-specific gene expression

Two replicates of K562 polyA mRNA RNA-Seq (bam files ENCFF412EYU & ENCFF037AFT from experiment ENCSR000AEM) were downloaded from the ENCODE portal (Sloan et al. 2016). 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 the heterozygous SNVs. Binomial tests were performed to test if the fraction of RNA-Seq reads that mapped to the alternative allele is significantly different ($p < 0.05$) from 0.5 or the expected frequency from WGS data for each heterozygous SNV. Only the SNVs with coverage >10 in RNA-Seq data were included in the analysis.

Allele-specific DNA methylation

One replicate of K562 whole-genome shotgun bisulfite sequencing data (fastq files ENCLB542OXH from experiment ENCSR765JPC) was downloaded from the ENCODE portal. Bisulfite reads were aligned to the human reference genome (hg19) using Bismark (version 0.16.3) (Krueger and Andrews 2011) resulting in 28.6x non-duplicate coverage. Paired-ended reads that contain cytosines (with 10-200x sequencing coverage) in a CpG dinucleotide context were pulled out if they also overlapped with phased heterozygous SNVs (Dataset 4). The haplotypes in which cystosines (methylated to unmethylated) belonged were then assigned based on the phased heterozygous SNVs. The numbers of reads that contain methylated or unmethylated cytosines on each haplotype were summed for each CpG locus. Fisher's exact test was applied at each CpG locus to see if the fraction of methylated to unmethylated reads was significantly different between haplotype phases 1 and 2 ($p < 0.05$). CpGs that belonged to CGIs were grouped; reads that contain methylated or unmethylated cytosines on each haplotype were summed for each CGI, and Fisher's exact test applied again ($p < 0.05$).

Allele-specific CRISPR targets

To identify allele-specific CRISPR targets, we started by extracting variants that satisfy the following properties from Dataset 4:

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 distinct (e.g. "0|1" or "1|0", but not "1|1")
4. The difference between the alleles is not a single point mutation (because the targets identified would not be specific enough)

For the 272,027 variants that satisfy these four properties, we extracted the two haplotype sequences (maximum length: 572). 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 'Ref' field was sequence "GTA", the 'Alt' field was "TA", and phasing was "0|1", the sequence containing "GTA" was tagged "0" and the sequence containing "TA" was tagged "1".
- If the 'Ref' field was sequence "GTA", the 'Alt' field was "TA", and phasing was "1|0", the sequence containing "GTA" was tagged "1" and the sequence containing "TA" was tagged "0".
- 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 532,013 candidates, which were then filtered to only retain 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 quality candidate gRNA needs to target a unique site. All the candidates that have multiple exact matches in the genome (irrespective of location) are identified using Bowtie2 (Langmead and Salzberg 2012) and removed from the list. We also removed targets with an extreme GC content (>80% or <20%), and targets that contained 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 for Cas9 recognition could not fold (Nishimasu et al. 2014), except if the folding energy was above -18 (indicating that the 'wrong' structure was very unstable). Finally, we evaluated the off-target risk 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 a score below 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 (using the tags '0', '1' or '2' from the extraction step), the target sequence, its position relative to the start of the variant, its orientation, and its off-target score. 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.

Structural variant identification from deep short-insert WGS

Structural variants from deep short-insert WGS were identified 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 pre-filtered calls. All variant calls were required to be >50 bp in size. We filtered out BreakDancer calls with less than 2 supporting read pairs and confidence scores of less than 90. Pindel calls were filtered for variants with quality scores of greater than 400. No further filtering was performed for BreakSeq calls.

Structural variant calls were also obtained from ARC-SV (Arthur et al, 2017). Candidate breakpoints in the reference are generated using soft-clipped and split-read alignments, and

nearby breakpoints are clustered as dictated by the data. These clusters typically span 1-100 kb and contain several to a dozen breakpoints. Within each breakpoint cluster, variants are called by rearranging the intervening genomic segments and scoring each configuration according to its likelihood under a generative statistical model for the paired-end alignments to that region. The ARC-SV method resolves two (maybe identical) alleles within each region as is appropriate for diploid genomes, thus performing SV calling and genotyping simultaneously. We note that, for complex cancer genomes with duplicated chromosomes, intermediate levels of heterozygosity will not be captured by the diploid genotypes.

There are filters for SVs having breakpoints in hard-to-align places: simple repeats, low complexity regions and satellite repeats -- all from RepeatMasker (repeatmasker.org). We also filter out breakpoints within segmental duplications (hg19) downloaded from the UCSC Genome Browser (Kent et al. 2002; Karolchik et al. 2004). There is also a filter for insertions (and complex events containing insertions) as the calls are not reliable. The BP_UNCERTAINTY filter is due to current technical limitations and should mask just a few putative tandem duplications. Specialized filtering was performed on tandem duplications and complex events. For tandem duplications, we require -either- 95% of reads overlapping the duplicated sequence to have mapq 20 or higher -or- at least 2 supporting split-reads (SR tag in VCF INFO column). For complex events, we require that each breakpoint is supported by a split-read or has >95% of overlapping reads with mapq 20. When part of a "complex event" (see above) fails one of these filters, all variants in the same sequence segment (both alleles) get subjected to the EVENT filter. Finally, we manually removed simple SVs that are too small (< 50 bp) and are better left to Indel calling. For tandem duplications, events must have at least 2 supporting split-reads or 95% of reads overlapping the duplicated sequence must have a mapping quality score >20.

Mate-pair library construction and variant calling

Long-insert mate-pair libraries (3 kb insert length) were constructed using 2.5 µg of high molecular weight genomic DNA (> 30 kb) as input for the Nextera Mate Pair Library Prep Kit (FC-132-1001) from Illumina following the standard protocol. Library insert size selection (2.7 kb to 3.3 kb) was carried out on the BluePippin instrument from Sage Science (Beverly, MA) using the S1 selection marker. Libraries were PCR amplified for 10 cycles before sequencing (2X151 bp) on the Illumina NextSeq 500. Each library was sequenced twice using the NextSeq Mid-Output Kit (FC-404-2003) to achieve ~10X coverage. Sequencing reads were subsequently combined for analysis. Illumina external read adapters and the reverse complement of the Nextera circularized single junction adapter sequence (AGATGTGTATAAGAGACAG) were trimmed from the 3' end of sequences followed by another trimming of the Nextera circularized single junction adapter sequence (CTGTCTCTTATACACATCT) using the FASTQ Toolkit (version 1.0) application on the Illumina Basespace platform (basespace.illumina.com). Trimmed reads were aligned to the hg19 human reference 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 and base scores were recalibrated using the Genome Analysis Tool Kit following the standard best practices workflow (Van der Auwera et al. 2013; DePristo et al. 2011). Afterwards, variant 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 https://github.com/arq5x/lumpy-sv (Layer et al. 2014). The *lumpyexpress* command was issued to obtain pre-filtered variant 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. Variant calls less than 50 bp in size were filtered out. To select for high-confidence calls, only variants

that have a minimum of 5 supporting reads as well as both discordant and split-read support were retained.

Non-Reference Mobile Element 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: (a) split-reads (part of the read maps to uniquely-mapped hg19 genomic sequence which is not normally adjacent to an MEI, and the remainder maps to a library of active transposon consensus sequences from Repbase (Bao et al. 2015)); and/or (b) paired-end reads (one read maps to unique genomic sequence, and the paired read maps to the transposon library rather than to the unique sequence that is normally found in relation to the first read). The default mapping quality filter between the reads and transposon consensus is >85% identity as set in RetroSeq (Keane et al. 2013). We then set the threshold at six or more supporting reads, split or paired-end, to filter out low-confidence calls. The boundaries for the transposon insertion (Supplementary Table 5) are a conservative estimate for the insertion junction. The “left boundary” is the left-most coordinate of the upstream supporting reads, or 1 kb upstream to the downstream supporting reads if “left boundary” is less than 1 kb away from the “right boundary”. The “right boundary” is the right-most coordinate of the downstream supporting reads or 1 kb downstream to the upstream supporting reads if the “right boundary” is less than 1 kb away from the “left boundary”.

Experimental validation of variant calls

Random sets of variant calls from short-insert WGS and mate-pair sequencing were selected from PCR validation. These sets include complex rearrangements (from ARC-SV), mobile element insertions, deletions (size >1 kb) and tandem duplications. PCR primers were designed to span the breakpoints of the variants and produce a PCR amplicon 200-500 bp in size. In the case of complex rearrangement variant calls, pairs of PCR primers were designed to span all breakpoints. For mobile element insertion variants, a random set of 9 Alu and 10 LINE1 non-reference insertion events with split-read support were chosen for validation. PCR primers

were designed to amplify products ranging from 65-150 bp with one primer annealing to the unique sequence of the genome and the other annealing to the mobile element sequence. All PCR amplicons were gel purified and verified using Sanger sequencing.

K562 POLR2A ChIP-Seq and RNA-Seq datasets

Four replicates of K562 polyA mRNA RNA-Seq transcript quantification data (tsv files ENCFF381QQP & ENCFF705JDM from experiment ENCSR000AEM; tsv file ENCFF928EIW from experiment ENCSR000AEO; tsv file ENCFF225LEY from experiment ENCSR545DKY) and K562 POLR2A ChIP-Seq alignment files (ENCFF000YWP & ENCFF000YWR) were downloaded from the ENCODE portal (Sloan et al. 2016). Values were binned in 1 Mbp windows. K562 POLR2A ChiP-Seq signals from the two replicates were summed for the Pearson correlation analysis.

Visualization

Genomic sequence and structural features of the K562 genome were plotted using Circos (Krzywinski et al. 2009).

DATA AVAILABILITY

Most raw and processed data files are publicly released on the ENCODE portal (Sloan et al. 2016). Experiment accessions: ENCSR711UNY and ENCSR025GPQ. 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/l39bka6w1sndetdqu7ouzmo9e326kftx.

ACKNOWLEDGEMENTS

We thank Aditi Narayanan, Dr. Carrie Davis, and Dr. Cricket Sloan for data organization and upload to the ENCODE portal. Dr. Athena Cherry and the Stanford Cytogenetics Laboratory for karyotype analysis. Arineh Khechaduri for performing genomic DNA preparation. 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.

AUTHOR CONTRIBUTIONS

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

DECLARATION OF INTERESTS

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.

Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, Chen Y, Villalta JE, Gilbert LA, Horlbeck MA, Hein MY, et al. 2016. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**: 1867–1882.e21. http://www.ncbi.nlm.nih.gov/pubmed/27984733.

Adey A, Burton JN, Kitzman JO, Hiatt JB, Lewis AP, Martin BK, Qiu R, Lee C, Shendure J. 2013. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**: 207–211. http://www.nature.com/doifinder/10.1038/nature12064.

Argiropoulos B, Humphries RK. 2007. Hox genes in hematopoiesis and leukemogenesis. *Oncogene* **26**: 6766–6776. http://www.nature.com/doifinder/10.1038/sj.onc.1210760.

Arroyo JD, Jourdain AA, Calvo SE, Ballarano CA, Doench JG, Root DE, Mootha VK. 2016. A Genome-wide CRISPR Death Screen Identifies Genes Essential for Oxidative Phosphorylation. *Cell Metab* **24**: 875–885. http://www.ncbi.nlm.nih.gov/pubmed/27667664.

Arthur JG, Chen X, Zhou B, Urban AE. 2017. Detection of complex structural variation from paired-end sequencing data. *bioRxiv* 1–32.

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.

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.

Blatt C, Aberdam D, Schwartz R, Sachs L. 1988. DNA rearrangement of a homeobox gene in myeloid leukaemic cells. *EMBO J* **7**: 4283–90. http://www.ncbi.nlm.nih.gov/pubmed/2907477.

Butler MO, Hirano N. 2014. Human cell-based artificial antigen-presenting cells for cancer immunotherapy. *Immunol Rev* **257**: 191–209. http://www.ncbi.nlm.nih.gov/pubmed/24329798.

Cao H, Wu H, Luo R, Huang S, Sun Y, Tong X, Xie Y, Liu B, Yang H, Zheng H, et al. 2015. De

novo assembly of a haplotype-resolved human genome. *Nat Biotechnol* **33**: 617–22. http://www.ncbi.nlm.nih.gov/pubmed/26006006 (Accessed May 9, 2016).

Caré A, Silvani A, Meccia E, Mattia G, Stoppacciaro A, Parmiani G, Peschle C, Colombo MP. 1996. HOXB7 constitutively activates basic fibroblast growth factor in melanomas. *Mol Cell Biol* **16**: 4842–51. http://www.ncbi.nlm.nih.gov/pubmed/8756643.

Carè A, Valtieri M, Mattia G, Meccia E, Masella B, Luchetti L, Felicetti F, Colombo MP, Peschle C. 1999. Enforced expression of HOXB7 promotes hematopoietic stem cell proliferation and myeloid-restricted progenitor differentiation. *Oncogene* **18**: 1993–2001. http://www.ncbi.nlm.nih.gov/pubmed/10208421.

Cavalli M, Pan G, Nord H, Wallerman O, Wallén Arzt E, Berggren O, Elvers I, Eloranta M-L, Rönnblom L, Lindblad Toh K, et al. 2016. Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression. *Hum Genet* **135**: 485–497. http://link.springer.com/10.1007/s00439-016-1654-x.

Chen H, Lee JS, Liang X, Zhang H, Zhu T, Zhang Z, Taylor ME, Zahnow C, Feigenbaum L, Rein A, et al. 2008. Hoxb7 inhibits transgenic HER-2/neu-induced mouse mammary tumor onset but promotes progression and lung metastasis. *Cancer Res* **68**: 3637–44. http://www.ncbi.nlm.nih.gov/pubmed/18463397.

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.

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.

de Klein A, van Kessel AG, Grosveld G, Bartram CR, Hagemeijer A, Bootsma D, Spurr NK, Heisterkamp N, Groffen J, Stephenson JR. 1982. A cellular oncogene is translocated to the Philadelphia chromosome in chronic myelocytic leukaemia. *Nature* **300**: 765–7. http://www.ncbi.nlm.nih.gov/pubmed/6960256.

Deguchi Y, Kehrl JH. 1993. High level expression of the homeobox gene HB24 in a human T-cell line confers the ability to form tumors in nude mice. *Cancer Res* **53**: 373–7. http://www.ncbi.nlm.nih.gov/pubmed/8093351.

Deguchi Y, Kirschenbaum A, Kehrl JH. 1992. A diverged homeobox gene is involved in the proliferation and lineage commitment of human hematopoietic progenitors and highly expressed in acute myelogenous leukemia. *Blood* **79**: 2841–8. http://www.ncbi.nlm.nih.gov/pubmed/1375114.

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.

Drexler HG, Matsuo Y, MacLeod RAF. 2004. Malignant hematopoietic cell lines: in vitro models for the study of erythroleukemia. *Leuk Res* **28**: 1243–51. http://www.ncbi.nlm.nih.gov/pubmed/15475063.

Eklund E. 2011. The role of Hox proteins in leukemogenesis: insights into key regulatory events in hematopoiesis. *Crit Rev Oncog* **16**: 65–76. http://www.ncbi.nlm.nih.gov/pubmed/22150308.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74. http://www.ncbi.nlm.nih.gov/pubmed/22955616.

Engreitz JM, Agarwala V, Mirny LA. 2012. Three-Dimensional Genome Architecture Influences Partner Selection for Chromosomal Translocations in Human Disease ed. S. Ahmed. *PLoS*

One **7**: e44196. http://dx.plos.org/10.1371/journal.pone.0044196.

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.

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.

Fröhling S. 2012. Widespread over-expression of the non-clustered homeobox gene HLX in acute myeloid leukemia. *Haematologica* **97**: 1453. http://www.ncbi.nlm.nih.gov/pubmed/23053668.

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.

Giampaolo A, Pelosi E, Valtieri M, Montesoro E, Sterpetti P, Samoggia P, Camagna A, Mastroberardino G, Gabbianelli M, Testa U. 1995. HOXB gene expression and function in differentiating purified hematopoietic progenitors. *Stem Cells* **13 Suppl 1**: 90–105. http://www.ncbi.nlm.nih.gov/pubmed/7488973.

Gribble SM, Roberts I, Grace C, Andrews KM, Green AR, Nacheva EP. 2000. Cytogenetics of the Chronic Myeloid Leukemia-Derived Cell Line K562. *Cancer Genet Cytogenet* **118**: 1–8. http://linkinghub.elsevier.com/retrieve/pii/S0165460899001697.

Groffen J, Stephenson JR, Heisterkamp N, de Klein A, Bartram CR, Grosveld G. 1984. Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22. *Cell* **36**: 93–9. http://www.ncbi.nlm.nih.gov/pubmed/6319012.

Grzanka A, Grzanka D, Orlikowska M. 2003. Cytoskeletal reorganization during process of apoptosis induced by cytostatic drugs in K-562 and HL-60 leukemia cell lines. *Biochem Pharmacol* **66**: 1611–7. http://www.ncbi.nlm.nih.gov/pubmed/14555241.

Han K, Jeng EE, Hess GT, Morgens DW, Li A, Bassik MC. 2017. Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nat Biotechnol* **35**: 463–474. http://www.ncbi.nlm.nih.gov/pubmed/28319085.

Haraksingh RR, Abyzov A, Gerstein M, Urban AE, Snyder M. 2011. Genome-Wide Mapping of Copy Number Variation in Humans: Comparative Analysis of High Resolution Array Platforms ed. G. Ast. *PLoS One* **6**: e27859. http://dx.plos.org/10.1371/journal.pone.0027859.

Heisterkamp N, Stam K, Groffen J, de Klein A, Grosveld G. 1985. Structural organization of the bcr gene and its role in the Ph' translocation. *Nature* **315**: 758–61. http://www.ncbi.nlm.nih.gov/pubmed/2989703.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**: D590-8. http://www.ncbi.nlm.nih.gov/pubmed/16381938.

Jawad M, Seedhouse CH, Russell N, Plumb M. 2006. Polymorphisms in human homeobox HLX1 and DNA repair RAD51 genes increase the risk of therapy-related acute myeloid leukemia. *Blood* **108**: 3916–8. http://www.ncbi.nlm.nih.gov/pubmed/16902145.

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.

Kitzman JO, MacKenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, Ng SB, Alkan C, Qiu R, Eichler EE, et al. 2011. Haplotype-resolved genome sequencing of a Gujarati Indian

individual. *Nat Biotechnol* **29**: 59–63. http://www.nature.com/doifinder/10.1038/nbt.1740.

Korbel JO, Tirosh-Wagner T, Urban AE, Chen X-N, Kasowski M, Dai L, Grubert F, Erdman C, Gao MC, Lange K, et al. 2009. The genetic architecture of Down syndrome phenotypes revealed by high-resolution analysis of human segmental trisomies. *Proc Natl Acad Sci* **106**: 12031–12036. http://www.pnas.org/cgi/doi/10.1073/pnas.0813248106.

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

Korbel JO, Urban AE, Grubert F, Du J, Royce TE, Starr P, Zhong G, Emanuel BS, Weissman SM, Snyder M, et al. 2007b. Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. *Proc Natl Acad Sci* **104**: 10110–10115. http://www.pnas.org/cgi/doi/10.1073/pnas.0703834104.

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

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. *Genome Res* **19**: 1639–1645. http://genome.cshlp.org/cgi/doi/10.1101/gr.092759.109.

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.

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.

Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, Attenello FJ, Villalta JE, Cho MY, Chen Y, et al. 2017. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science (80-)* **355**: eaah7111. http://www.sciencemag.org/lookup/doi/10.1126/science.aah7111.

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.

Lozzio CB, Lozzio BB. 1975. Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. *Blood* **45**: 321–34. http://www.ncbi.nlm.nih.gov/pubmed/163658.

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.

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.

Morgens DW, Deans RM, Li A, Bassik MC. 2016. Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes. *Nat Biotechnol* **34**: 634–6. http://www.ncbi.nlm.nih.gov/pubmed/27159373.

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.

Naumann S, Reutzel D, Speicher M, Decker HJ. 2001. Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization. *Leuk Res* **25**: 313–22. http://www.ncbi.nlm.nih.gov/pubmed/11248328.

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.

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.

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.

Raval A, Tanner SM, Byrd JC, Angerman EB, Perko JD, Chen S-S, Hackanson B, Grever MR, Lucas DM, Matkovic JJ, et al. 2007. Downregulation of death-associated protein kinase 1 (DAPK1) in chronic lymphocytic leukemia. *Cell* **129**: 879–90. http://www.ncbi.nlm.nih.gov/pubmed/17540169.

Selden JR, Emanuel BS, Wang E, Cannizzaro L, Palumbo A, Erikson J, Nowell PC, Rovera G, Croce CM. 1983. Amplified C lambda and c-abl genes are on the same marker chromosome in K562 leukemia cells. *Proc Natl Acad Sci U S A* **80**: 7289–92. http://www.ncbi.nlm.nih.gov/pubmed/6580644.

Shah N, Sukumar S. 2010. The Hox genes and their roles in oncogenesis. *Nat Rev Cancer* **10**: 361–371. http://www.nature.com/doifinder/10.1038/nrc2826.

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.

Shiraishi K, Yamasaki K, Nanba D, Inoue H, Hanakawa Y, Shirakata Y, Hashimoto K, Higashiyama S. 2007. Pre-B-cell leukemia transcription factor 1 is a major target of promyelocytic leukemia zinc-finger-mediated melanoma cell growth suppression. *Oncogene* **26**: 339–48. http://www.ncbi.nlm.nih.gov/pubmed/16862184.

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.

Snyder MW, Adey A, Kitzman JO, Shendure J. 2015. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat Rev Genet* **16**: 344–358. http://www.nature.com/doifinder/10.1038/nrg3903.

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.

Storti P, Donofrio G, Colla S, Airoldi I, Bolzoni M, Agnelli L, Abeltino M, Todoerti K, Lazzaretti M, Mancini C, et al. 2011. HOXB7 expression by myeloma cells regulates their pro-angiogenic properties in multiple myeloma patients. *Leukemia* **25**: 527–537. http://www.nature.com/doifinder/10.1038/leu.2010.270.

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.

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.

Tkachuk DC, Westbrook C a, Andreeff M, Donlon T a, Cleary ML, Suryanarayan K, Homge M, Redner a, Gray J, Pinkel D. 1990. Detection of bcr-abl fusion in chronic myelogeneous leukemia by in situ hybridization. *Science* **250**: 559–562.

Urban AE, Korbel JO, Selzer R, Richmond T, Hacker A, Popescu G V, Cubells JF, Green R, Emanuel BS, Gerstein MB, et al. 2006. High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *Proc Natl Acad Sci U S A* **103**: 4534–9. http://www.ncbi.nlm.nih.gov/pubmed/16537408.

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.

Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM. 2015. Identification and characterization of essential genes in the human genome. *Science* **350**: 1096–101. http://www.ncbi.nlm.nih.gov/pubmed/26472758.

Williams LJS, Tabbaa DG, Li N, Berlin AM, Shea TP, MacCallum I, Lawrence MS, Drier Y, Getz G, Young SK, et al. 2012. Paired-end sequencing of Fosmid libraries by Illumina. *Genome Res* **22**: 2241–2249. http://genome.cshlp.org/cgi/doi/10.1101/gr.138925.112.

Wu SQ, Voelkerding K V, Sabatini L, Chen XR, Huang J, Meisner LF. 1995. Extensive amplification of bcr/abl fusion genes clustered on three marker chromosomes in human leukemic cell line K-562. *Leukemia* **9**: 858–62. http://www.ncbi.nlm.nih.gov/pubmed/7769849.

Wu X, Chen H, Parker B, Rubin E, Zhu T, Lee JS, Argani P, Sukumar S. 2006. HOXB7, a homeodomain protein, is overexpressed in breast cancer and confers epithelial-mesenchymal transition. *Cancer Res* **66**: 9527–34. http://www.ncbi.nlm.nih.gov/pubmed/17018609.

Yamashita T, Tazawa S, Yawei Z, Katayama H, Kato Y, Nishiwaki K, Yokohama Y, Ishikawa M. 2006. Suppression of invasive characteristics by antisense introduction of overexpressed HOX genes in ovarian cancer cells. *Int J Oncol* **28**: 931–8. http://www.ncbi.nlm.nih.gov/pubmed/16525643.

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).

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).

FIGURE LEGENDS

Figure 1. Comprehensive View of the K562 Genome

Circos (Krzywinski et al. 2009) visualization of K562 genome variants with the following tracks in inward concentric order: human genome reference track (hg19); CN or ploidy by chromosome segement; merged structural variant density (deletions, duplications,

inversions) in 1.5 Mb windows, called with ARC-SV (Arthur et al. 2017), BreakDancer (Chen et al. 2009), BreakSeq (Lam et al. 2010), LUMPY (Layer et al. 2014), Pindel (Ye et al. 2009), and/or Long Ranger (Zheng et al. 2016); phased haplotype blocks (blocks alternate with color); SNV density in 1 Mb windows; Indel density in 1 Mb windows; dominant zygosity in 1 Mb windows (heterozygous or homozygous > 0.5); regions with loss of heterozygosity; RNA-Seq reads for loci that exhibit allele-specific expression; differentially methylated CpG islands (CgI); histogram of differentially methylated (log-scale) CpG dinucleotides in 50 kb windows; novel Alu and LINE-1 mobile element insertions; allele-specific CRISPR target sites; large-scale structural variants detected by Long Ranger (light blue: intrachromosomal: dark blue: interchromosomal); large-scale structural variants detected by GROC-SVs (Spies et al. 2017) (light-gray: intrachromsomal; dark-gray: interchromosomal).

Figure 2. K562 Karyogram and Callset Overview

(A) Representative karyogram of K562 cells produced with the GTW banding method showing multiple numerical and structural chromosomal abnormalities and an overall near triploid karyotype. ISCN 2013 description in relationship to a triploid karyotype [<3n>]: 53~70<3n>,XX,-X or Y,-3,?dup(6)(p21p25),+7,?inv(7)(p13p22),add(7)(q32),-9,add(9)(p24),del(9)(p13),add(10)(q22),-13,add(13)(p11),-14,add(17)(p11.2)x2,add(18)(q23),-20,der(21)t(1;21)(q21;p11),-22,+4~7mar[cp20] (B) Distribution of CN or ploidy across the K562 genome. (C) Percentage of K562 SNVs and Indels that are novel and known in dbSNP (Sherry et al. 2001). (D) Violin plot with overlaid boxplot of phased haplotype block sizes (Y-axis, log-scaled). N50 represented as a dashed line (N50 = 2,721,866 bp).

Figure 3. Large SVs in the K562 Genome Resolved Using Linked-Read Sequencing

Heat maps of overlapping barcodes for structural variants in K562 called from linked-reads using Long Ranger(Zheng et al. 2016). (A) *BCR*/*ABL1* translocation between chromosomes 9 and 22. (B) *XKR3*/*NUP214* translocation between chromosomes 9 and 22. (C) Duplication within *GPHN* on chromosome 14. (D) Deletion that partially overlaps *ZRANB1* and *CTB2* on chromosome 10.

Figure 4. Complex Structural Variants Reconstructed and Assembled Using GROC-SVs in K562

Each line depicts a fragment inferred from 10X-Genomics data based on clustering of reads with identical barcodes identified from GROC-SVs (Spies et al. 2017). Each row shows fragments within the same barcode across the three sub-panels, indicating fragments spanning the complex variant. Fragments are phased locally with respect to surrounding SNVs and colored cyan for haplotype 1, orange for haplotype 2, and black when no informative SNVs are found nearby. Gray lines indicate portions of fragments that do not support the current breakpoint. (A) Inversion flanked by two deletions in haplotypes 1 and 2 on chromosome 6. The fragments end abruptly at 157.56 Mb on chromosome 6, indicating a breakpoint (deletion), and then pick up again at 157.64 Mb in the inverted orientation (inversion), end abruptly again at 157.61 Mb (deletion) and continue finally at 157.69 Mb. The decreasing coordinates of the center panel indicate

that the minus strand of the genome is depicted. Read-depth coverage is calculated from the standard short-fragment Illumina data and is shown in the three sub-panels in the bottom half of the panel to support the interpretation of a copy number neutral inversion of the middle region concurrent with deletions of the flanking sequences. A high coverage region downstream of the third breakpoint is due to a repetitive element and is not involved in the SV. The decreasing coordinates of the center panel indicate that the minus strand of the genome is depicted. (B) Complex rearrangement occurring on chromosome 13 with informative reads from only one haplotype (region with loss-of-heterozygosity). Fragments end abruptly at 81.47 Mb, indicating a breakpoint, picking up again at 81.09 Mb and continuing to 81.11 Mb where they end abruptly, then picking up again at 90.44 Mb. Coverage from 81.12 Mb to 81.20 Mb are from reads with different sets of linked-read barcodes and thus not part of this fragment set.

Figure 5. Small-scale Complex SVs in K562 Resolved Using ARC-SV
Examples of complex genomic rearrangements called from ARC-SV (Arthur et al. 2017) in the K562 WGS datasets. (A) Deletion of Block C and duplication of Block E between Blocks B and D on chromosome 20 (135,111-136,565). This variant has been validated by PCR. (B) Deletion of Block B and inverted, duplication of Block D between Blocks A and C on chromosome 1 (81,660,347-81,661,554). (C) Deletion of Blocks B and D and inversion of Block C on chromosome 5 (147,553,038-147,554,778) inside *SPINK14* (coding for a serine peptidase inhibitor). (D) Duplication and inversion of Blocks B, C, and D between Blocks B and D on chromosome 3 (158,795,874-158,795,955) inside gene *IQCJ-SCHIP1*. (E) Deletion of Block G, duplications of blocks I, D, and E, and inverted duplication of Block B between Blocks F and H on chromosome 10 (127,190,417-127,201,193).

Figure 6. Genomic Structural Context Provides Insight into Regulatory Complexity
(A) Chr17:46,687,000-46,700,000 locus (triploid in K562) containing genes *HOXB7* and *HOXB8* and CpG Island (CGI) 22086 (1,203 bp) where phased Haplotype 1 has two copies and Haplotype 2 has one copy. Allele-specific expression (transcription) of *HOXB7* on Haplotype 1 and Haplotype 2 has no expression. CpGs in CGI 22086 are unmethylated in Haplotype 1 and methylated in Haplotype 2. (B) Chr1:221,052,000-221,059,000 locus (tetraploid in K562) containing the *HLX* gene and CGI 2209 (294 bp) where phased Haplotype 1 has three copies and Haplotype 2 has one copy. Allele-specific RNA expression of *HLX* on Haplotype 1. CpGs in CGI 2209 are unmethylated in Haplotype 2 and highly methylated in Haplotype 1. Allele-specific CRISPR targeting site 797 bp inside the 5' end of the *HLX* for both Haplotypes. (C) Number of methylated and unmethylated phased whole-genome bisulfite-sequencing reads for Haplotypes 1 and 2 in CGI 22086 and CGI 2209 where both CGIs exhibit allele-specific DNA methylation (p=3.18E-18 and p=5.14E-15 respectively). (D) Number of allele-specific RNA-Seq reads in Haplotypes 1 and 2 for *HLX* and *HOXB7* where both genes exhibit allele-specific RNA expression (p=0.007 and p=0.043 respectively).

Figure S1. Illustration of the Sequencing-Based Methodologies Used

Very-deep short-insert WGS (72x non-duplicate coverage), 3 kb-mate-pair sequencing (Korbel et al. 2007a), and 10X-Genomics linked-reads sequencing (Zheng et al. 2016) were used to comprehensively characterize the genome of K562. The WGS dataset was used to obtain CN or ploidy by chromosomal 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 (MEIs), and SVs, such as deletions, duplications, inversions, insertions, and complex rearrangements, using BreakDancer (Chen et al. 2009), BreakSeq (Lam et al. 2010), Pindel (Ye et al. 2009), and/or ARC-SV (Arthur et al. 2017). MEI output is in bed format; SVs, SNV, and Indel outputs are in standard vcf format. The 10X-Genomics linked-reads dataset was used to phase heterozygous SNVs and Indels as well as to identify, assemble, and phase primarily large (>30 kb) and complex SVs using Long Ranger (Zheng et al. 2016) and GROC-SVs (Spies et al. 2017). Deletions <30 kb were identified by Long Ranger. Output of SVs and phased variants identified using Long Ranger and GROC-SVs are in standard vcf format. The SV assembly file from GROC-SVs is in BAM format. The 3 kb-mate-pair dataset was used to call additional structural variants, mostly in the medium size-range (1 kbp-100 kb) using LUMPY (Layer et al. 2014) (vcf output), and also used to validate large and complex SVs. The union of non-complex SVs identified merged into a single combined vcf.

Figure S2. Plot of GC Bias and K562 WGS Coverage
K562 very-deep WGS read coverage calcuated in 10 kb bins across the genome (X-axis: % GC content. Y-axis: coverage). Clusters correspond to CN (or ploidy).

Figure S3. Comparison CN Assigned to Chromosome Segments Using Read-Depth Analysis with Array CGH in the K562 Genome (Chromosomes 1-22, X)
Upper panel: WGS coverage plot. X-axis genomic coordinate in kb. Y-axis: WGS coverage. Red: CN5, Purple: CN4, Blue: CN3, Green: CN2, Black: CN1, Gold: CN0. Lower panel: Y-axis: array probe signal intensity. X-axis: Genomic coordinate.

Figure S4. Size Distribution of Phased Haplotype Blocks by Chromosome
Violin plots of K562 phased haplotype blocks separated by chromosomes with overlaid boxplot. Y-axis: size in log-scale.

Figure S5. Overlap Between SV Callers
Venn diagram of overlaps (>50% reciprocal) between K562 SVs identified in WGS using ARC-SV (Arthur et al. 2017), BreakDancer (Chen et al. 2009), and BreakSeq (Lam et al. 2010), 3 kb mate-pair sequencing using LUMPY (Layer et al. 2014), and 10X-Genomics linked-read sequencing using Long Ranger (Zheng et al. 2016) for (A) deletions (>50 bp), (B) tandem duplications, and (C) inversions.

Figure S6. K562 POLR2A ChIP-Seq and WGS Pearson Correlation
Pearson correlation (r=0.51, p<2.22E-16) between K562 POLR2 ChIP-Seq reads (Y-axis) and WGS reads (X-axis).

Figure S7. K562 POLR2A ChIP-Seq and RNA-Seq Pearson Correlation

Pearson correlation between K562 POLR2 ChIP-Seq reads (Y-axis) and RNA-Seq FPKM across four independent replicates (X-axis). (A) r=0.46, p<2.22E-16. (B) r=0.58, p<2.22E-16. (C) r=0.47, p<2.22E-16. (D) r=0.46, p<2.22E-16.

Table 1. Summary of K562 SNVs and Indels

Table S1 - Karyotype comparisons with previously published data

Table S2 - Ploidy or Copy Number (CN) by Chromosome Segments in K562

Table S3 - CGH comparisons with previously published data

Table S4 – Loss-of-heterozygosity regions

Table S5 - COSMIC overlap

Table S6 - CENSUS overlap

Table S7 - PCR validations

Table S8 - MEI predictions

Table S9 - MEI validation

Table S10 - Allele-Specific Expression

Table S11 - ASE positives & negatives

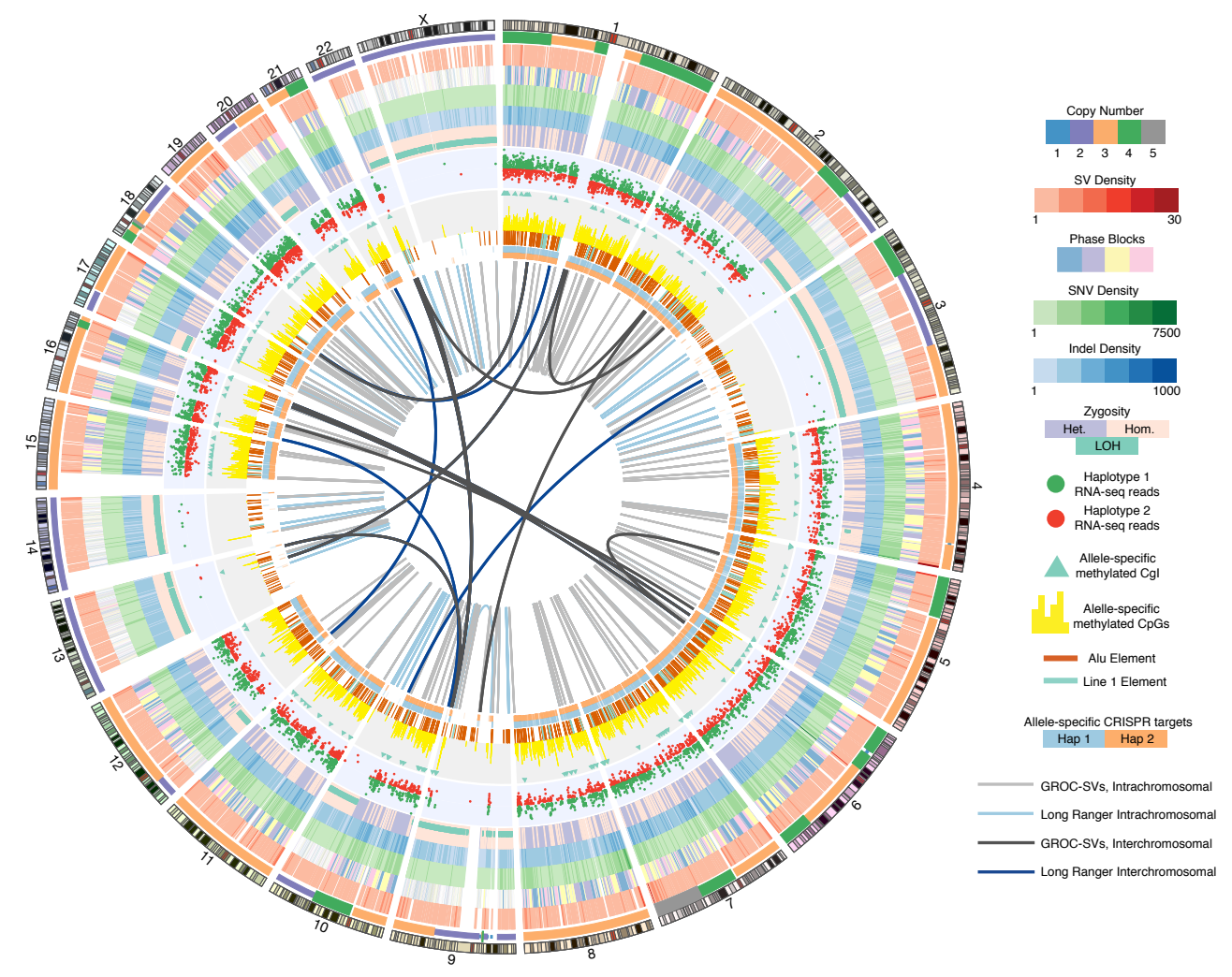
Table S12 - Allele-specific Methylation

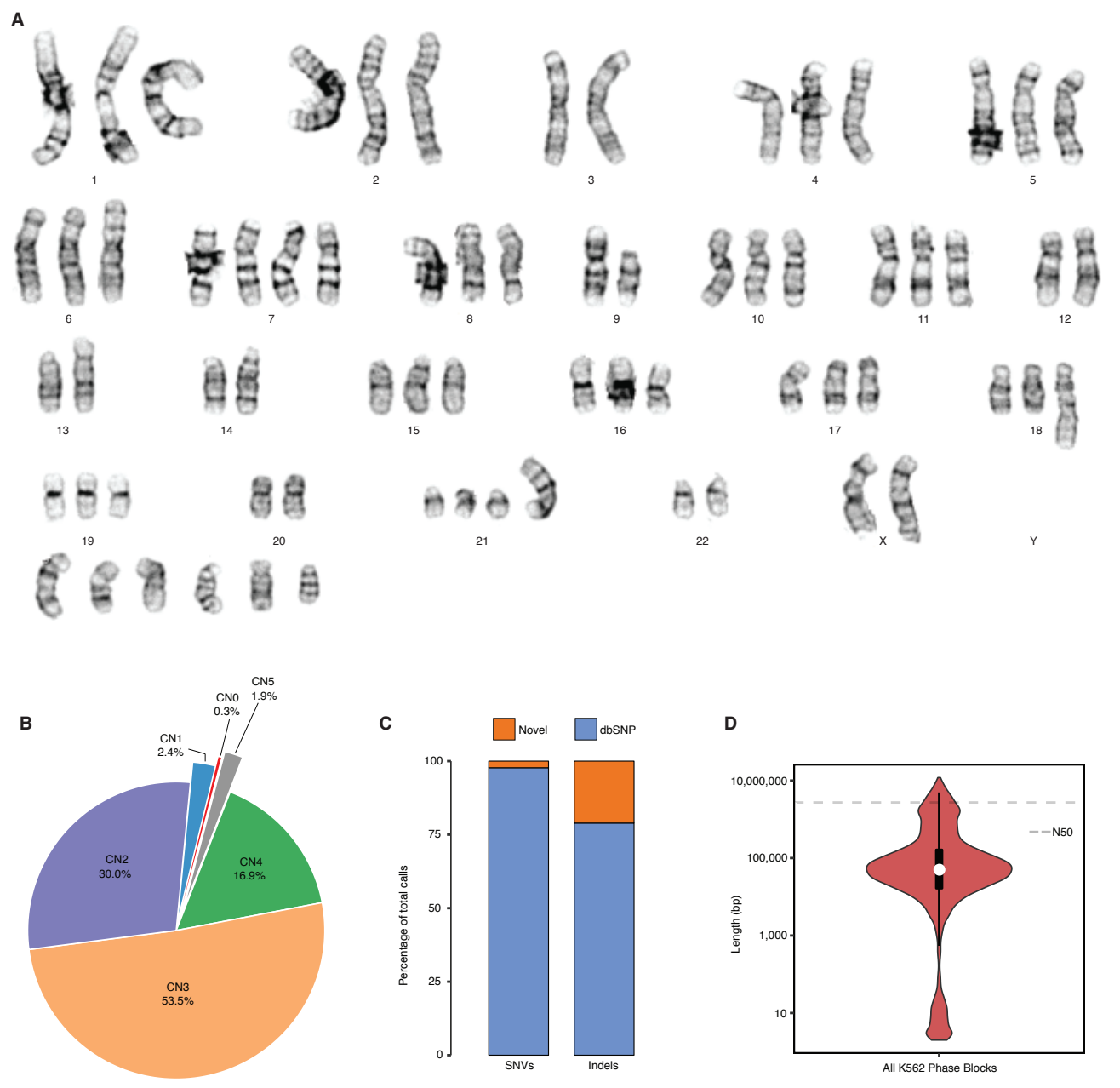
Table S13 - Allele-specific CRISPR targets

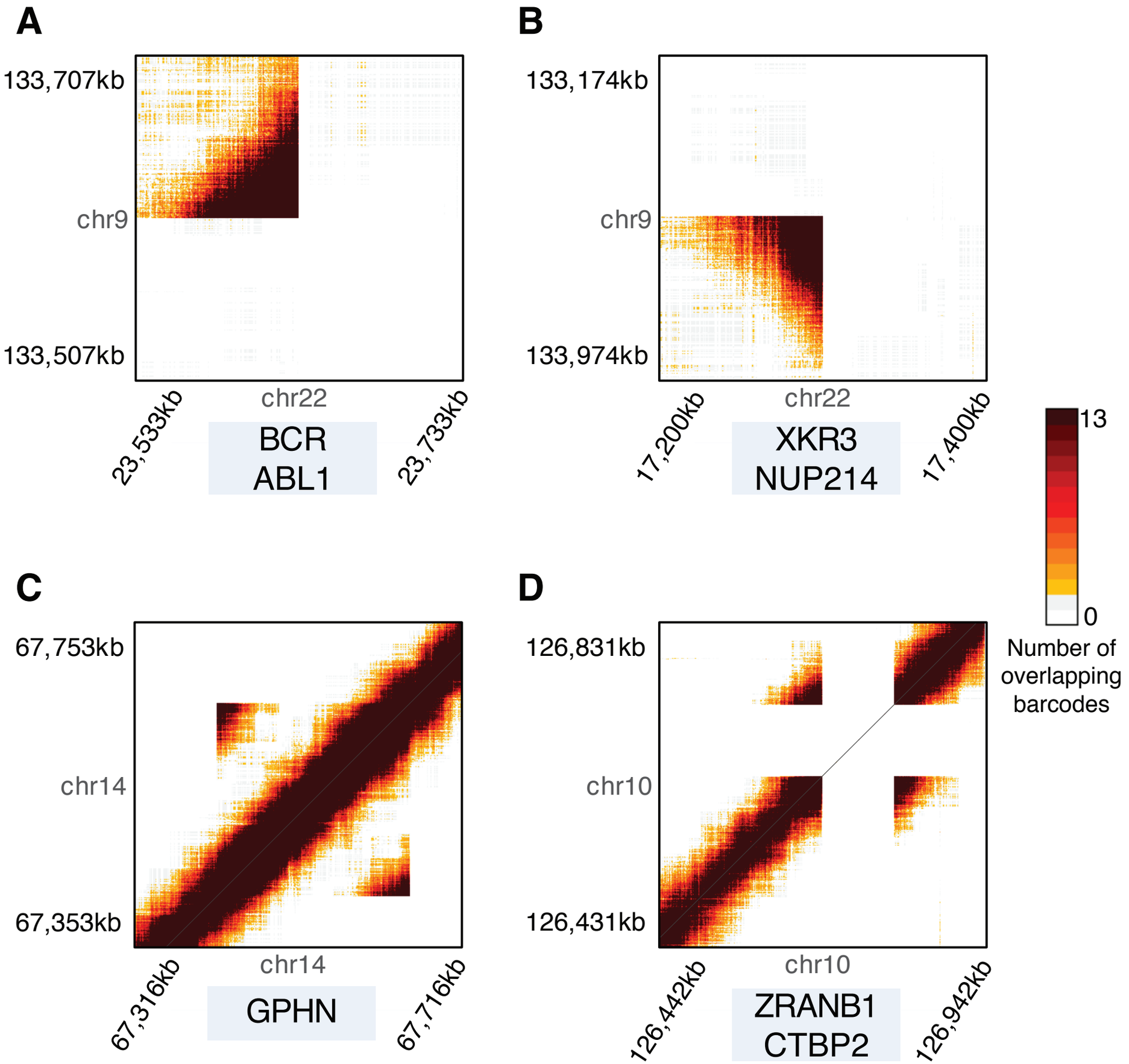
Table 1. Summary of K562 SNVs and Indels

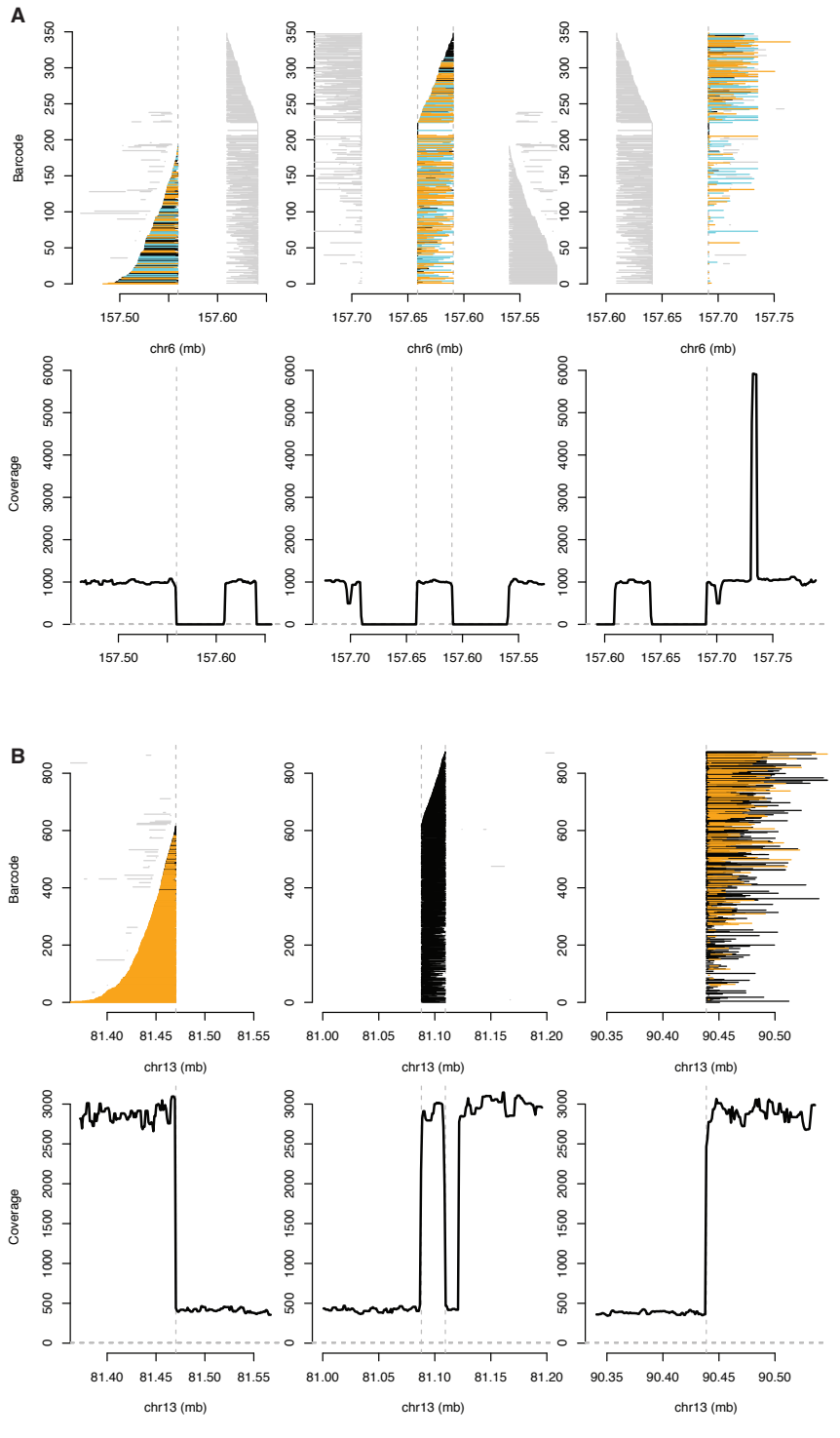
Small Variant Calls	SNVs	Indels
All	3,088,312	702,787
Heterozygous/homozygous	1,451,017/163,7295	393,632 / 309,155
Protein altering	10,831 (0.4%)	1,118 (0.2%)
dbSNP138	3,020,306 (98%)	558,637 (79%)
Heterozygous/homozygous	1,389,196 / 1,629,672	294,850 / 260,606
Novel	69,553 (2%)	149,055 (21%)
Heterozygous/homozygous	61,821 / 7,623	98,782 / 48,548
1000 Genomes Project & Exome Sequencing Project overlap (with protein alerting variants)	970 (96%)	10,407 (87%)
Novel protein altering	424	148
COSMIC overlap	227 (53%)	46 (32%)

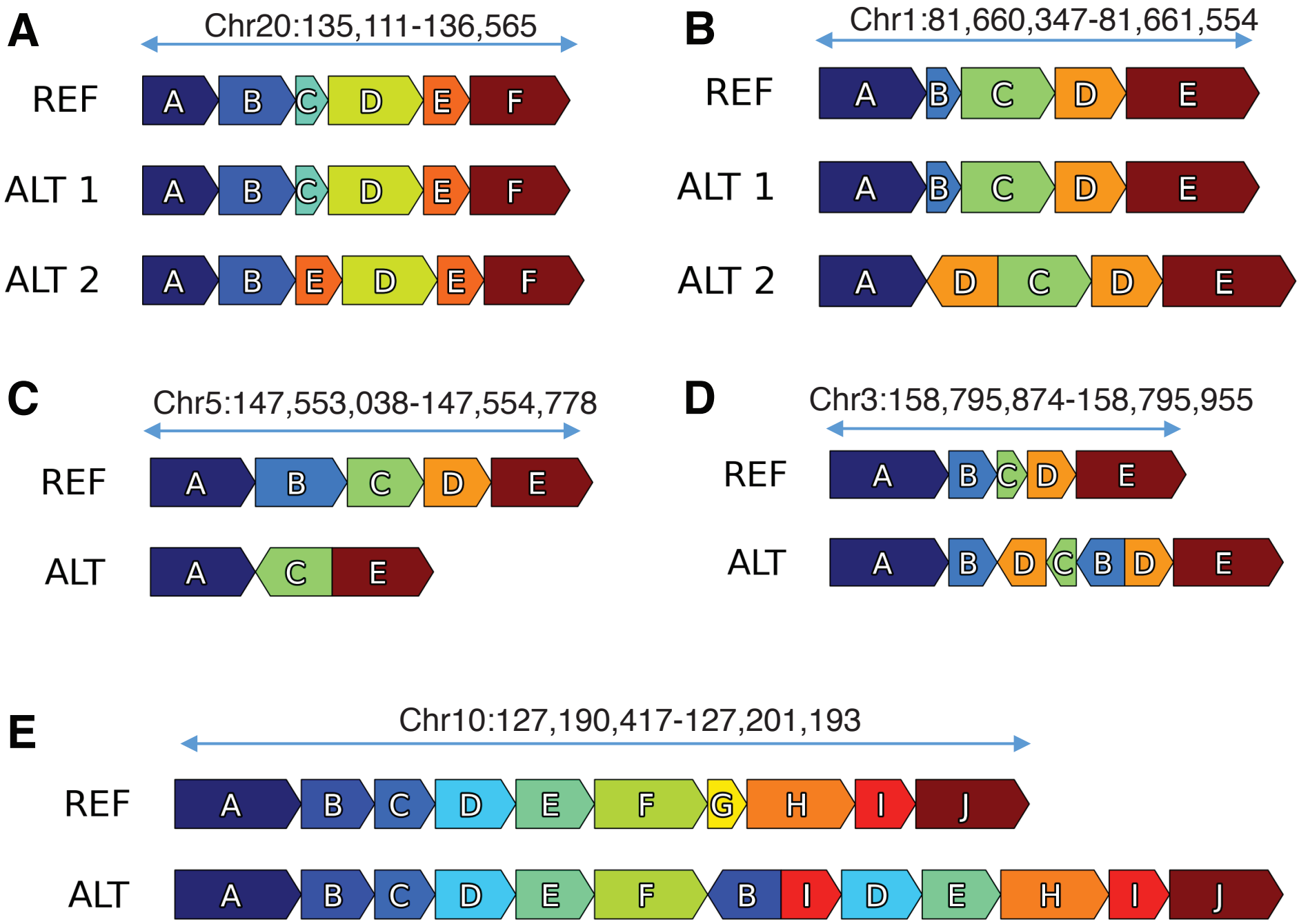
Phased WGS	
% phased heterozygous SNVs	97
% phased heterozygous Indels	84
Longest phase block	1,195,3412
Number of phase blocks	4,987
N50 phase block	2,721,866
N50 linked-reads per molecule	64
Barcodes detected	1,562,771
Mean DNA per barcode (bp)	456,351

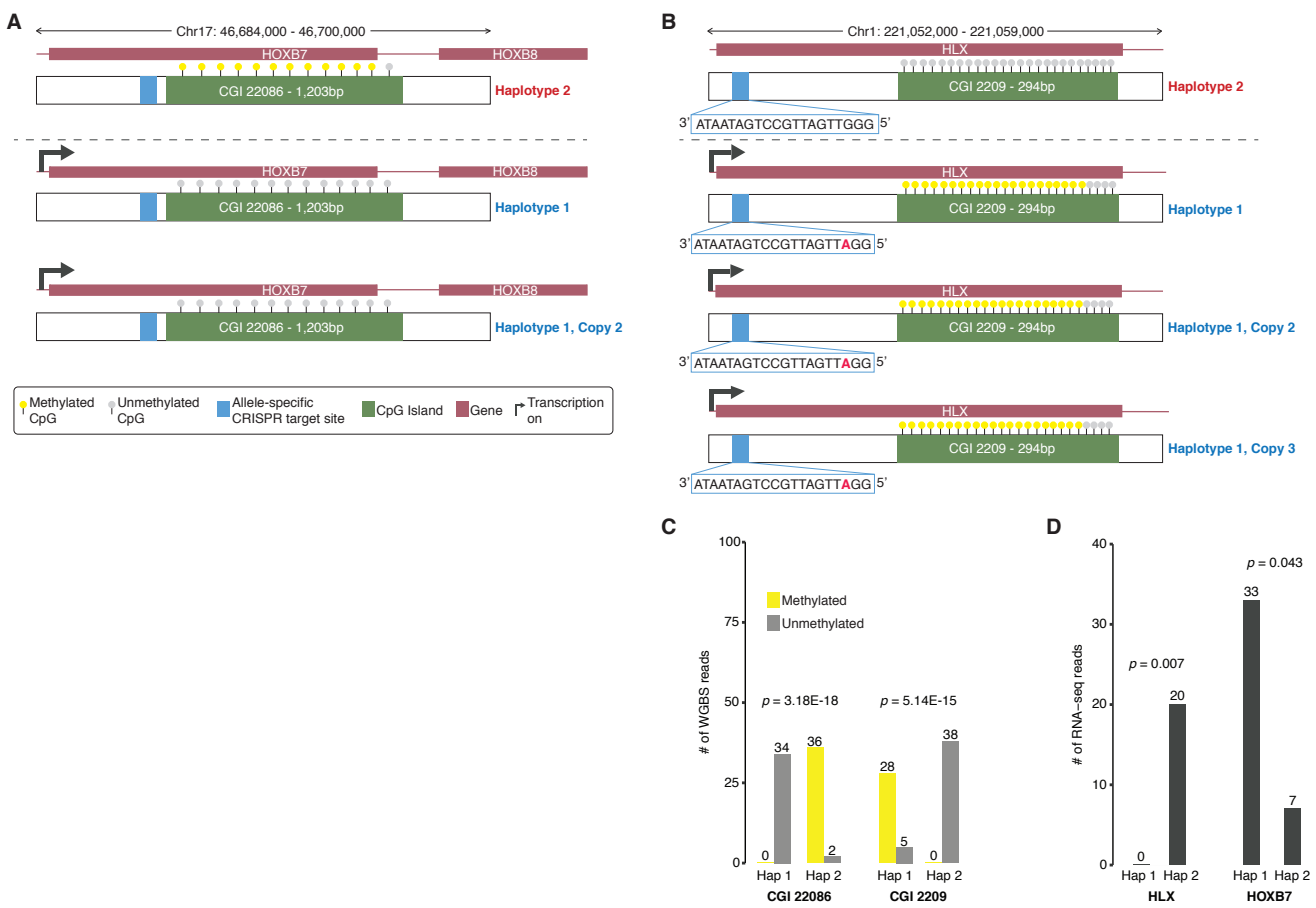

X
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Copy Number
1 2 3 4 5
SV Density
1
30
Phase Blocks
SNV Density
1
7500
Indel Density
1
1000
Zygosity
Het.
Hom.
LOH
Haplotype 1
RNA-seq reads
Haplotype 2
RNA-seq reads
Allele-specific
methylated CgI
Alelle-specific
methylated CpGs
Alu Element
Line 1 Element
Allele-specific CRISPR targets
Hap 1
Hap 2
GROC-SVs, Intrachromosomal
Long Ranger Intrachromosomal
GROC-SVs, Interchromosomal
Long Ranger Interchromosomal


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
CN5
1.9%
CN0
0.3%
CN1
2.4%
CN2
30.0%
CN4
16.9%
CN3
53.5%
C
Novel
dbSNP
Percentage of total calls
100
75
50
25
0
SNVs
Indels
D
10,000,000
100,000
1,000
10
Length (bp)
N50
All K562 Phase Blocks


A
133,707kb
chr9
133,507kb
chr22
23,533kb
23,733kb
BCR
ABL1
B
133,174kb
chr9
133,974kb
chr22
17,200kb
17,400kb
XKR3
NUP214
13
0
Number of overlapping barcodes
C
67,753kb
chr14
67,353kb
chr14
67,316kb
67,716kb
GPHN
D
126,831kb
chr10
126,431kb
chr10
126,442kb
126,942kb
ZRANB1
CTBP2


A
Barcode
0
50
100
150
200
250
300
350
157.50
157.60
chr6 (mb)
157.70
157.65
157.60
157.55
chr6 (mb)
157.60
157.65
157.70
157.75
chr6 (mb)
Coverage
0
1000
2000
3000
4000
5000
6000
157.50
157.60
157.70
157.65
157.60
157.55
157.60
157.65
157.70
157.75
B
Barcode
0
200
400
600
800
81.40
81.45
81.50
81.55
chr13 (mb)
81.00
81.05
81.10
81.15
81.20
chr13 (mb)
90.35
90.40
90.45
90.50
chr13 (mb)
Coverage
0
500
1000
1500
2000
2500
3000
81.40
81.45
81.50
81.55
chr13 (mb)
81.00
81.05
81.10
81.15
81.20
chr13 (mb)
90.35
90.40
90.45
90.50
chr13 (mb)


A
Chr20:135,111-136,565
REF
A B C D E F
ALT 1
A B C D E F
ALT 2
A B E D E F
B
Chr1:81,660,347-81,661,554
REF
A B C D E
ALT 1
A B C D E
ALT 2
A D C D E
C
Chr5:147,553,038-147,554,778
REF
A B C D E
ALT
A C E
D
Chr3:158,795,874-158,795,955
REF
A B C D E
ALT
A B D C B D E
E
Chr10:127,190,417-127,201,193
REF
A B C D E F G H I J
ALT
A B C D E F B I D E H I J


A
Chr17: 46,684,000 - 46,700,000
HOXB7
HOXB8
CGI 22086 - 1,203bp
Haplotype 2
Haplotype 1
Haplotype 1, Copy 2
Methylated CpG
Unmethylated CpG
Allele-specific CRISPR target site
CpG Island
Gene
Transcription on
B
Chr1: 221,052,000 - 221,059,000
HLX
CGI 2209 - 294bp
Haplotype 2
3' ATAATAGTCCGTTAGTTGGG 5'
Haplotype 1
3' ATAATAGTCCGTTAGTTAGG 5'
Haplotype 1, Copy 2
Haplotype 1, Copy 3
C
# of WGBS reads
Methylated
Unmethylated
p = 3.18E-18
p = 5.14E-15
34
36
2
28
5
38
0
0
Hap 1
Hap 2
CGI 22086
CGI 2209
D
# of RNA-seq reads
p = 0.007
p = 0.043
0
20
33
7
HLX
HOXB7