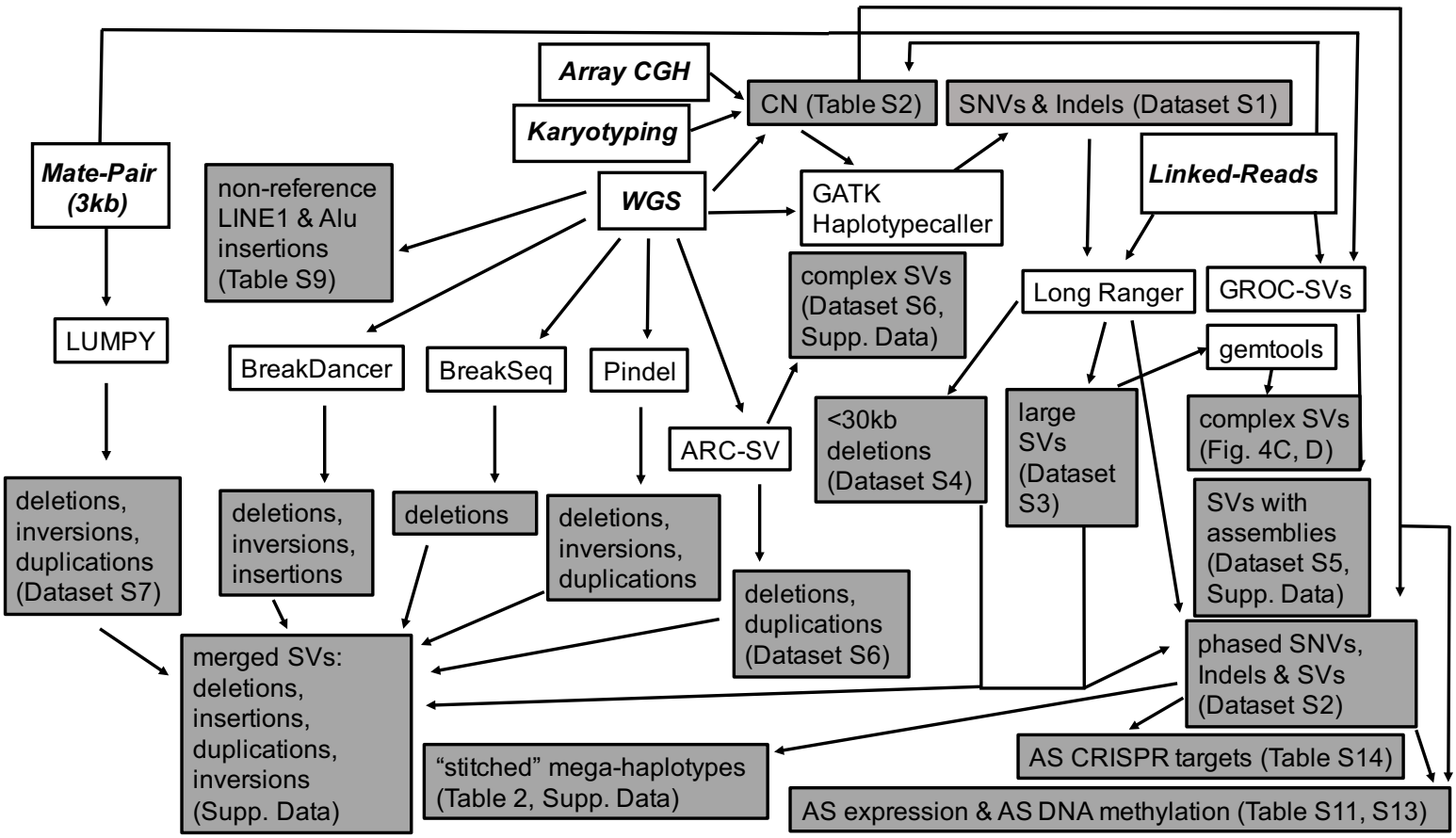
Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562

Supplemental Material

1. Supplemental Figure S1
    - Detailed ENCODE experimental and file accessions of all raw and processed data including Datasets S1-S7
    - Description of Supplemental Data
2. Supplemental Figure S2
3. Supplemental Figure S3
4. Supplemental Figure S4
5. Supplemental Figure S5
6. Supplemental Figure S6
7. Supplemental Figure S7
8. Supplemental Discussion
9. Supplemental Methods

A. Illustration of Method



B. ENCODE accession numbers of datasets and other resources generated for K562

Dataset S1. SNVs & Indels (corrected for ploidy) identified from WGS using GATK Haplotypecaller and Table S2 (ENCFF606RIC VCF format)
Dataset S2. Phased variants (SNVs, Indels, and SVs) generated from integrating linked-reads and Dataset S1 using Long Ranger (ENCFF241LKI VCF format)
Dataset S3. Large SVs (>30 kb) identified from linked-reads using Long Ranger (ENCFF657ZQS VCF format)
Dataset S4. Deletions (<30 kb) identified using Long Ranger identified from linked-reads using Long Ranger (ENCFF344VXT VCF format)
Dataset S5. SVs identified and assembled using GROC-SVs (ENCFF872PDV VCF format, ENCFF266ZFI BAM format). PDF images of SVs in Supplemental Data
Dataset S6. SVs identified from WGS using ARC-SV (ENCFF131ESS VCF format)
Dataset S7. SVs (deletions, tandem duplications, inversions identified mate-pair sequencing using LUMPY (ENCFF823IED VCF format)
Supplemental Data. Merged SV calls from Break Dancer, PINDEL, BreakSeq, ARC-SV, LUMPY, and Long Ranger (VCF format)
Table S2. High-resolution CN (or ploidy) of K562 chromosome segments identified from WGS using read-depth analysis.
Table S8. non-reference LINE1 & Alu insertions identified from WGS using RetroSeq
Table S11. Allele-specific RNA expression, two replicates of ENCODE K562 RNA-Seq data reanalyzed taking ploidy into account
Table S13. Phased and allele-specific DNA methylation, ENCODE K562 whole-genome bisulfite sequencing data reanalyzed taking ploidy and CpG phasing into account.
Table S14. Allele-specific CRISPR targeting sites.
Figure 2A. Karyotyping by GTW banding
Supplemental Data. Mega-base-scale SNV haplotypes "stitched" from Dataset S2 by leveraging the allelic imbalance in aneuploid regions
Supplemental Data. NimbleGen array CGH data with comparisons to WGS read-depth signal (PDF images, array CGH data,)

C. Raw sequencing read and alignment data generated for K562

- Short-insert WGS
  - ENOCDE accession: ENCLB544CFT (sequencing), ENCFF691CKS (hg19 alignment)
  - Link: https://www.encodeproject.org/experiments/ENCSR711UNY

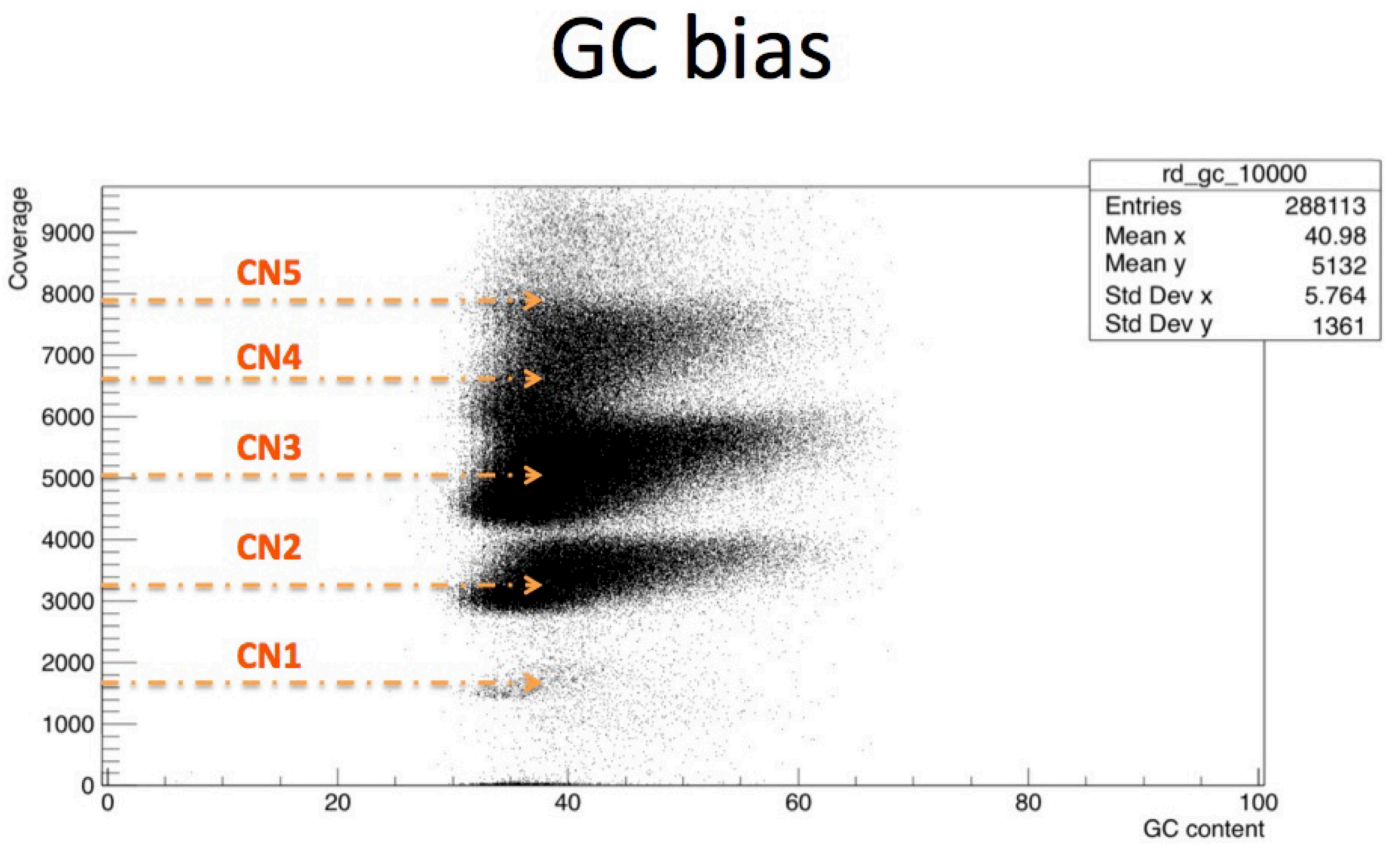
- 3kb-mate-pair sequencing
  - ENCODE accession: ENCLB531FQO (sequencing), ENCFF462TFC (hg19 alignment)
  - Link: https://www.encodeproject.org/experiments/ENCSR025GPQ/

- Linked-read sequencing
  - ENCODE accessions: ENCLB557IGA (sequencing), ENCFF287PIC (hg19 alignment)
  - Link: https://www.encodeproject.org/experiments/ENCSR053AXS/

D. Description of Supplemental Data

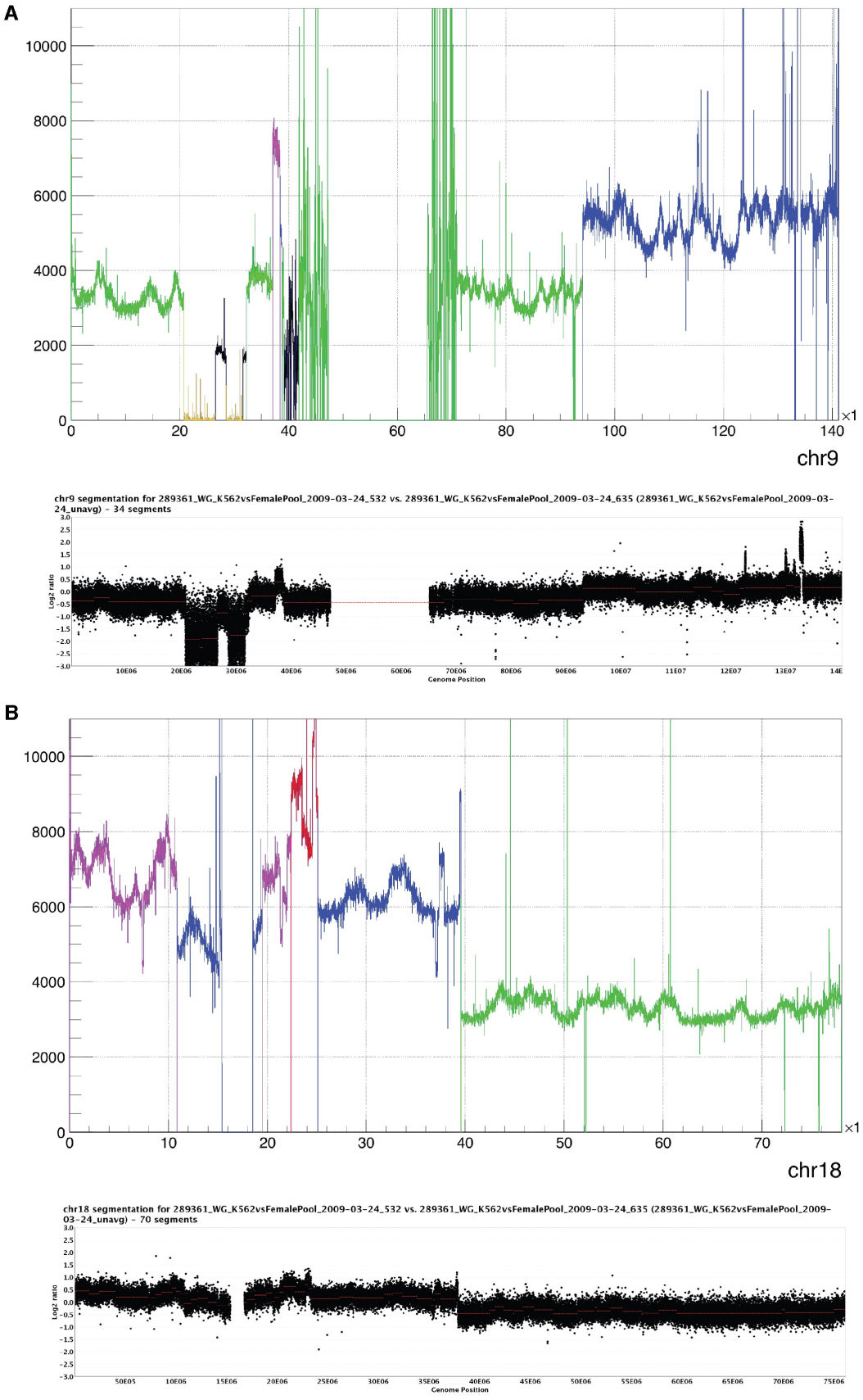
Comparison of CN identified from WGS read-depth analysis with NimbleGen aCGH.zip
PDF images. Upper panel: WGS coverage plot. X-axis genomic coordinate in kb. Y-axis: WGS coverage. Red: CN5, Purple: CN4, Blue: CN3, Green: CN2, Black: CN1, Gold: CN0. Lower panel: Y-axis: array probe signal intensity. X-axis: genomic coordinate. Array CGH data of K562 vs. female pool DNA with dye swap.
GROC-SVs_visualizations.zip
Visualizations of SVs in Dataset S5 identified using GROC-SVs. Each line depicts a fragment inferred from linked-read data based on clustering of reads with identical barcodes identified from GROC-SVs (Spies et al. 2017). 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. Read-depth coverage is calculated from the short-insert WGS dataset.
ARC-SV_visualizations.zip
Visualizations of SVs in Dataset S6 identified from short-insert WGS using ARC-SV.
K562_SVs_merged_vcf.zip
K562 merged SV calls from Break Dancer, PINDEL, BreakSeq, ARC-SV (non-complex SVs), LUMPY, and Long Ranger (deletions <30kb). VCF format.
K562_Mega_Haplotypes.zip
Haplotypes blocks of 100 or more phased SNVs from Dataset S2 "stitched" into mega-haplotypes by leveraging the haplotype imbalance in aneuploid regions using method described in (Bell et al., 2017). Columns: chromosome, position, minor SNV allele, major SNV allele. Only SNVs present in both K562 and NA12878 were included in the mega-haplotype blocks (see Supplemental Methods).
K562_megahaplotyping_scripts.zip
scripts for running mega-haplotyping based on Bell et al 2017. Step by step instructions are provided in proc_steps.pair.sh

Supplemental Figure S1. Illustration of method and description of resources generated for K562
(*A*) Short-insert WGS at 72x non-duplicate coverage, 3 kb-mate-pair sequencing (Korbel et al. 2007a) at 69x physical coverage, linked-read sequencing (Zheng et al. 2016) at 59x coverage, array CGH, and karyotyping were used to comprehensively characterize the genome of K562. WGS was used to obtain high-resolution CN by chromosomal segments using read-depth analysis (Supplemental Table S2) (Abyzov et al. 2011), SNVs and Indels using GATK Haplotypecaller (McKenna et al. 2010) with CN taken into account (Dataset S1), non-reference LINE-1 and Alu insertions (Supplemental Table S9), and SVs (Datasets S6, Supplemental Data), such as deletions, duplications, inversions, insertions, and complex SVs, using BreakDancer (Chen et al. 2009), BreakSeq (Lam et al. 2010), Pindel (Ye et al. 2009), and ARC-SV (Arthur et al. 2017). CN of chromosome segments were orthogonally validated with array CGH and karyotyping (Fig. 2A, Supplemental Fig. S3, Supplemental Data). A statistical approach that counts unique linked-read barcodes from the major and minor alleles also confirmed aneuploid CN>2 regions (Bell et al. 2017). The linked-reads were used to phase SNVs and Indels as well as to identify, assemble, and phase primarily large (>30 kb) and complex SVs using Long Ranger (Marks et al. 2018), gemtools (Greer et al. 2017), and GROC-SVs (Spies et al. 2017) (Datasets S2, S3, S5, Supplemental Data, Figure 4C, D). Deletions <30 kb were also identified by Long Ranger (Dataset S4). Haplotype blocks from Long Ranger in Dataset S2 with ≥100 heterozygous phased SNVs in aneuploid chromosomal regions were "stitched" to mega-haplotypes by leveraging haplotype imbalance (Bell et al. 2017) (Table 2, Fig. 3, Supplemental Data). CN and phased SNVs were integrated to identify allele-specific RNA expression and allele-specific DNA methylation (Supplemental Tables S11, S13). The 3 kb-mate-pair dataset was used to call additional SVs (Dataset S7), mostly in the medium size-range (1 kbp-100 kb), using LUMPY (Layer et al. 2014) and also used to validate large and complex SVs identified from linked-read data. Small-scale complex SVs were identified using ARC-SV (Dataset S6, Supplemental Data). The union of non-complex SV calls from LUMPY, Break Dancer, BreakSeq, Pindel, ARC-SV, and Long Ranger are listed in Supplemental Data where SVs overlapping by ≥50% reciprocally were merged (Supplemental Data). (*B*) Description and ENCODE accession numbers for Datasets S1-S7 and other resources generated for K562. Datasets can be downloaded from https://www.encodeproject.org (Sloan et al. 2016). (*C*) ENCODE accession numbers and download links for sequencing and alignment data for K562. (*D*) Description of Supplemental Data files included with this manuscript.

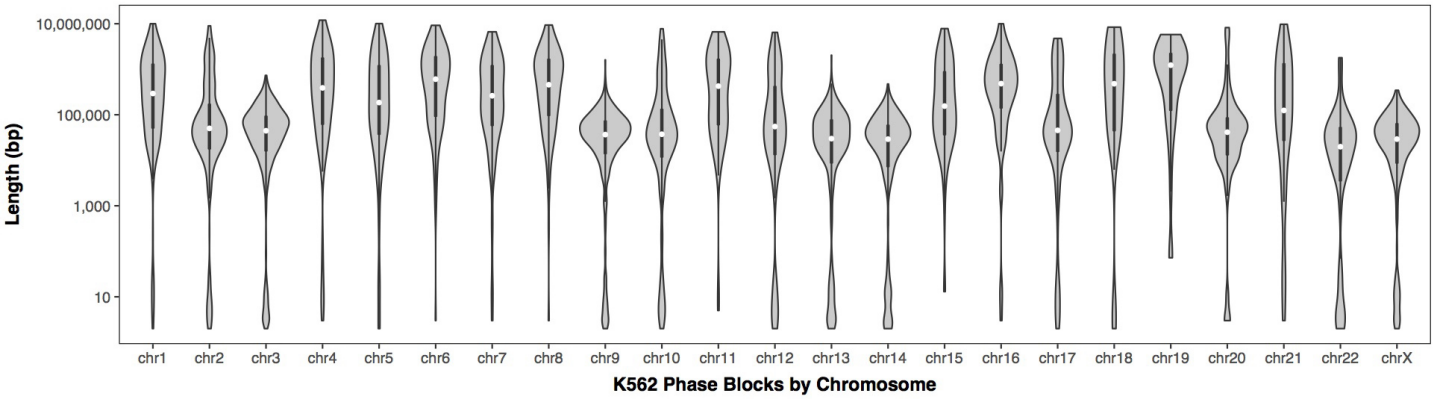


Supplemental Figure S2. K562 WGS coverage vs. % GC content
Y-axis: K562 WGS coverage in 10 kb bins across the genome; X-axis: % GC content of bins. coverage). Clusters correspond to CN (i.e. ploidy).

Supplemental Figure S3

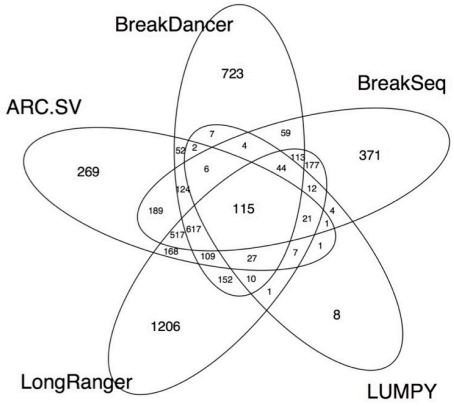


Supplemental Figure S3. Validation of CN determined from using read-depth analysis with array CGH. Upper panel: WGS coverage plot. X-axis genomic coordinate in kb. Y-axis: WGS coverage. Red: CN5, Purple: CN4, Blue: CN3, Green: CN2, Black: CN1, Gold: CN0. Lower panel: Y-axis: array probe signal intensity. X-axis: Genomic coordinate. (A) chromosomes 9. (B) chromosome 18. For complete chromosomes (1-22, X), see Supplemental Data.

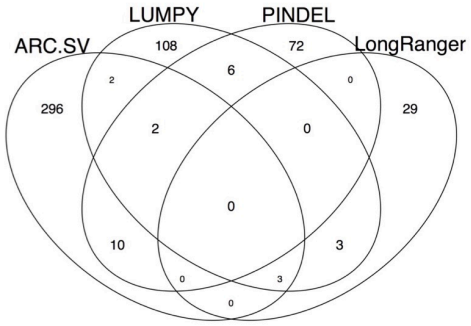


Supplemental Figure S4. Size distributions of haplotype blocks by each chromosome
Violin plots, with overlaid boxplot, of K562 phased haplotype blocks (Dataset S2) for each chromosome. Y-axis: size in log-scale.

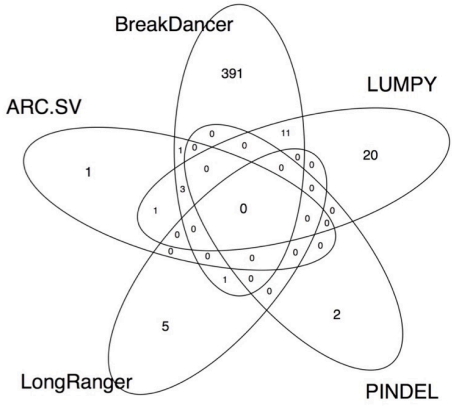
A. Deletions



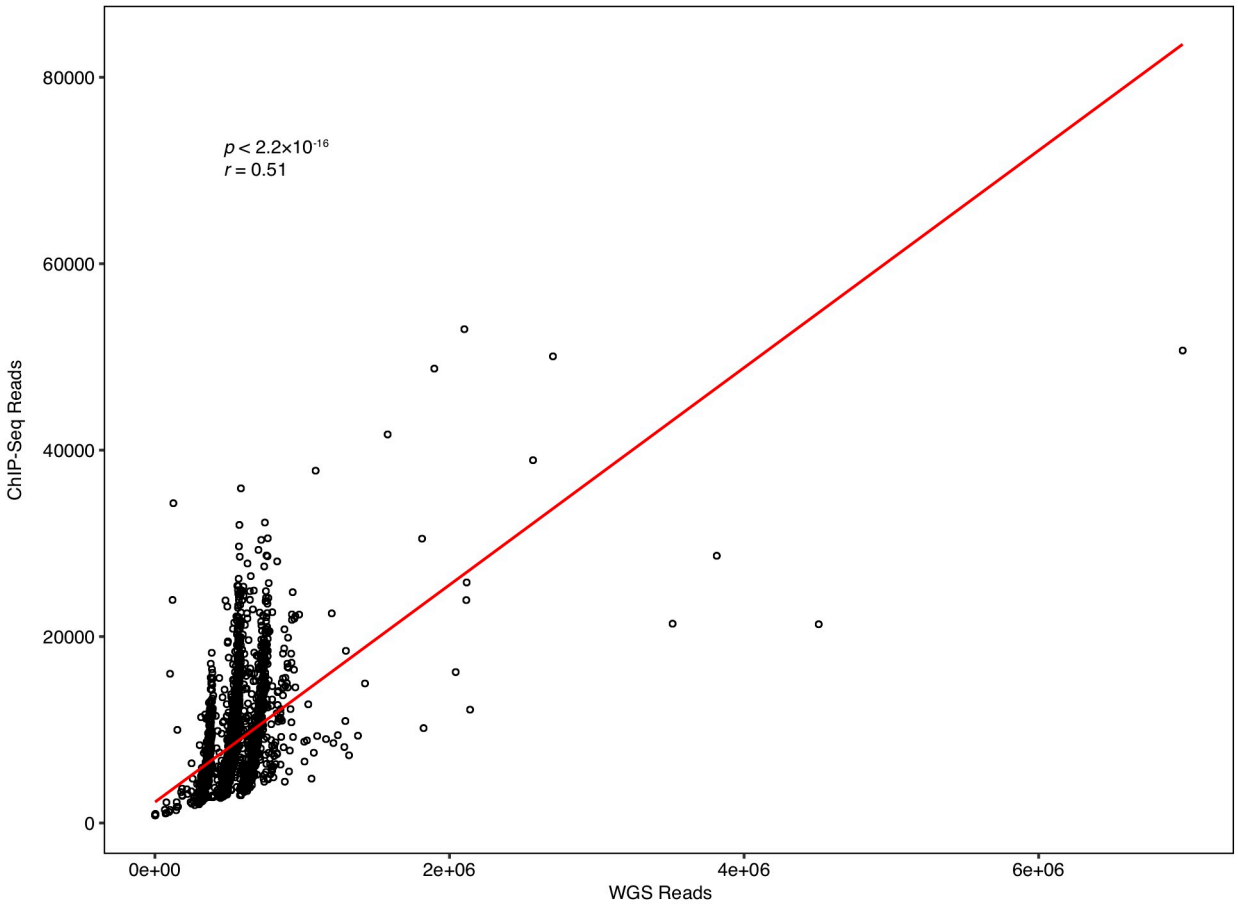
B. Duplications



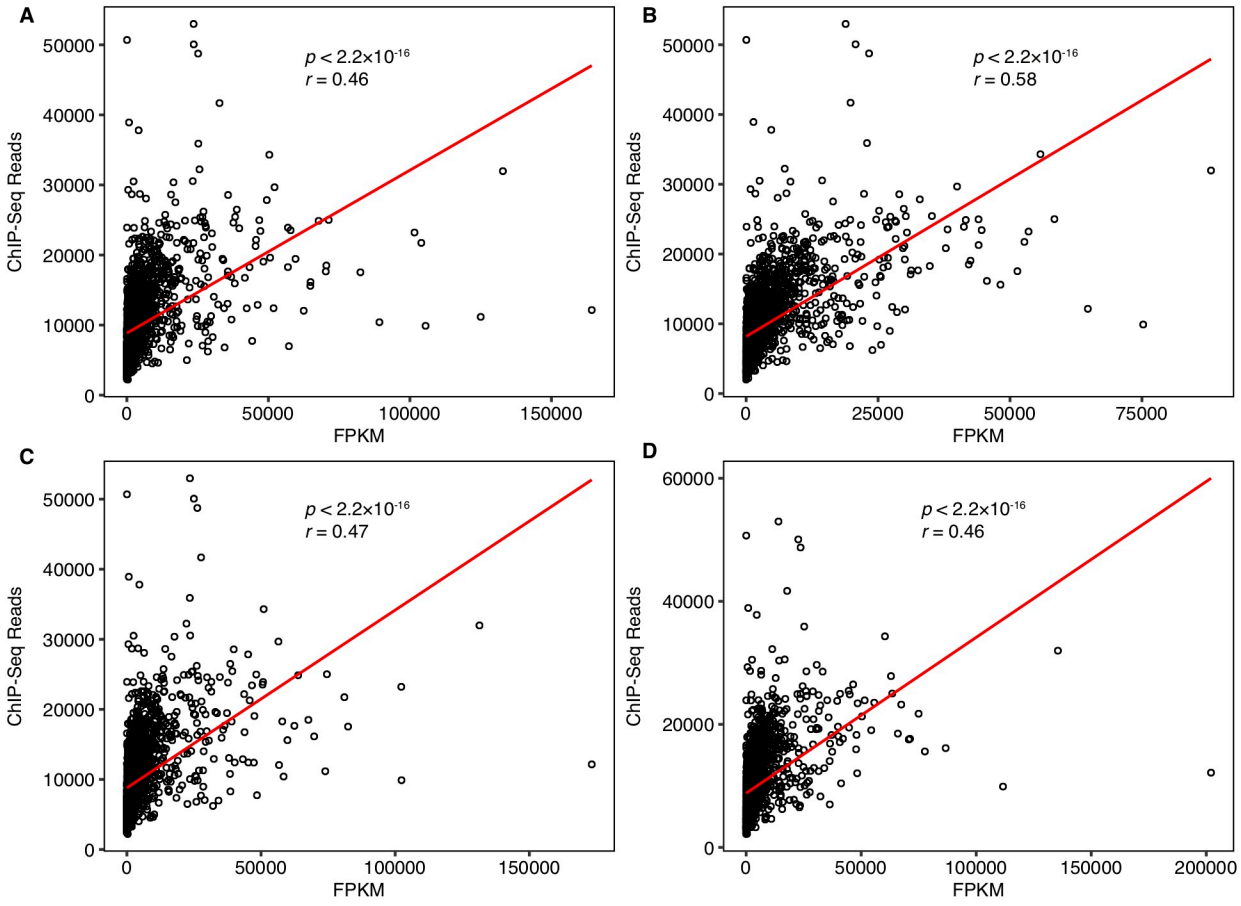
C. Inversions



Supplemental Figure S5. Overlap of SV Calls
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; Marks et al. 2018) for (A) deletions (>50 bp), (B) tandem duplications, and (C) inversions.



Supplemental Figure S6. K562 POLR2A ChIP-seq and WGS Pearson correlation
Pearson correlation between K562 POLR2 ChIP-seq reads (Y-axis) and WGS reads (X-axis).



Supplemental 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 (X-axis) across four independent replicates (A-D).

SUPPLEMENTAL DISCUSSION

K562 Karyotype

It should be taken into consideration that for a widely used cell line with decades of history, such as K562, genome variation is expected. In light of this, it is reassuring that the overall karyotype has not changed much over the years. However, researchers should still keep this aspect in mind when working with a version of K562 that has been separated from the main ENCODE K562 line used here. We expect that the vast majority of genomic variants that we describe here to be universal for K562, but for individual variants, it is possible that different lines of K562 may have slightly diverged from each other (Supplemental Table S1, Supplemental Table S3). When using a different K562 line and following up on findings for individual loci of interest, a first step should always be to experimentally validate their presence. When incorporating these genomic variants for global analyses, such as interrogating network interactions, the vast majority of them will exist, thus such global analyses are expected to yield substantial insights. Even though the pervasive aneuploidy in K562 renders the design and interpretation of K562 studies more challenging, the information we provide here enables researchers to continue the use of this cell line to investigate the effects of genetic 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 its complex genome structure into account.

Haplotype Phasing

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 the K562 genome and used Long Ranger to perform haplotype phasing (Zheng et al. 2016; Marks et al. 2018). 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. Implemented by Long Ranger (Marks et al. 2018), 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 phased haplotype blocks with N50 > 2.72 Mb (Dataset S2, Fig. 2D). Chromosomes 3, 9, 13, 14, X, and large portions of Chromosomes 2, 10, 12, 20, 22 were difficult to phase, resulting in comparatively shorter phased blocks (Dataset S2, Fig. 1, Supplemental Fig. S4). This is not surprising since these chromosomes and chromosomal regions exhibit a very high degree of LOH (Fig. 1 and Supplemental 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.

We extended on the already-impressive phasing capabilities of Long Ranger and constructed mega-haplotypes in K562—often spanning entire chromosome arms (Fig. 3, Table 2, Supplementary Data)—by leveraging the haplotype imbalance in aneuploid chromosomes using a recently developed method for which its effectiveness in cancer has already been demonstrated (Bell et al. 2017). Since gene dosage is a fundamental component of genome biology and for which aneuploidy contribute large effects in terms of amplification and reduction, the ability to haplotype across long stretches of aneuploidy is essential for understanding of the genetic regulations of cancer and an important component for developing genetically targeted cancer treatment.

Large SVs Identified from Linked-Read Sequencing

FHIT is frequently seen to harbor deletions in many types of human cancers, most commonly of epithelial origin, such as lung, stomach, cervix, head and neck, breast, and kidney

(Lubinski et al. 1994; Ohta et al. 1996; Huebner et al. 1998; Ingvarsson 2001). Reduction or absence in its protein expression occurs in nearly 50% of all cancers (Huebner et al. 1998; Waters et al. 2014). While LOH and allele-specific deletions within *FHIT* have been previously reported (Wistuba et al. 1997; Li et al. 2016), to our knowledge, this is the first discovery of phased and allele-specific tandem duplications within *FHIT* and the first report of *FHIT* mutations for CML. Curiously, all previous reports of deletions within *FHIT* for various cancer types (not including CML) were all centered on and include exon 5 (Durkin et al. 2008), whereas exon 5 is duplicated in K562. Deletions of all three *FHIT* exons 6, 7, and 8 (Fig. 4C) are less frequent but have been reported for lung cancer and esophageal adenocarcinoma (Sozzi et al. 1996; Dagmar et al. 1997).

We identified highly allele-specific RNA expression for *ORC6* ($p<1.58 \times 10^{-8}$) and *MYLK3* ($p<1.93 \times 10^{-17}$) in K562 (Supplemental Table S11), which is likely contributed by the allele-specific complex intra-chromosomal rearrangement residing on the non-duplicated haplotype of 16q11.2 (triploid) (Fig. 4D). *ORC6* codes for origin recognition protein complex subunit 6 and is essential for coordinating DNA replication, chromosome segregation, and cytokinesis (Prasanth et al. 2002). One allele of *ORC6* is "deleted" by one of the inversion breakpoints of this rearrangement and maintains allele-specific expression on the other allele (Fig. 4D). The origin recognition complex, in which Orc6 is a subunit, serves as a "landing pad" for bringing together components of the pre-replicative complex required for DNA replication. Reduction of Orc6 dosage by small interfering RNA results in decreased DNA replication, aberrant mitosis, and the formation of multiple nuclei and multipolar spindles in cells; a long period of this reduction increases cell death (Prasanth et al. 2002). Such effects are not observed in K562 cells even though RNA expression from one allele of *ORC6* is "depleted" by this rearrangement ($p<1.58 \times 10^{-8}$, Supplemental Table S11). This is likely because the other "normal" *ORC6* allele was duplicated, rendering this locus in K562 triploid and thus maintain a "diploid" gene dosage. Perhaps more importantly, this insight also raises important questions

regarding the history of mutation as well as selective pressures that occurred within K562 cells. It is possible that one copy of Chromosome 16 was first duplicated, freeing this locus from selective pressures, allowing it to acquire new mutations, since “diploid” copies are still maintained in the genome. It is also conceivable that duplication of this locus is disadvantageous for cell proliferation, and K562 cells that acquired this rearrangement after the duplication of Chromosome 16 also acquired a selective advantage since they now reverted this locus back to “diploid” copies. However, it is also possible, though perhaps less likely, that this rearrangement occurred before Chromosome 16 duplication, putting a negative selection pressure on K562 cells, and that this pressure is released by duplication of the other copy. This rearrangement also inverts an intact copy of *MYLK3* (Fig. 4D), which was identified to encode a novel cardiac-specific myosin light chain kinase (Seguchi et al. 2007). Since its expression is expected to be normally repressed except in heart cells, the allele-specific RNA expression of *MYLK3* (p<1.93E-17, Supplemental Table S11) from this inverted allele suggests that this inversion activated the ectopic expression of this gene in K562, possibly by disrupting or disconnecting it from its promoter or proximal enhancer elements that impose repressive regulatory mechanisms. Finally, this complex rearrangement on Chromosome 16 of K562 also duplicates *NETO2* in a tandem fashion, also on the same allele (Fig. 4D). *NETO2* codes for a single-pass membrane protein neuropilin and tolloid-like 2 (Stöhr et al. 2002). Its expression is frequently up-regulated in many types of human cancers including lung, cervical, colon, and renal carcinomas and has been suggested as a potential genetic marker for cancer (Oparina et al. 2012). Its up-regulation also correlates with the progression and poor prognosis of colorectal carcinoma (Hu et al. 2015). It is conceivable that the increase in *NETO2* gene dosage due to duplication in K562 cells contributes to their efficient proliferation in culture and that, at least in some cancers, the frequent up-regulations observed for *NETO2* are also contributed by this similar mechanism of allele-specific tandem duplication.

The hallmark of CML is the Philadelphia rearrangement t(9; 22)(q34; q11) which results in the fusion of *ABL1* and *BCR* (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 analysis showed that fluorescent signals from the *BCR/ABL1* gene fusion 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 elsewhere in the genome, and the *BCR* and *ABL1* gene regions where the fusion occurs show a >2.8× increase in sequencing coverage relative to average sequencing coverage across the genome.

Complex Gene Regulation at the *HOXB7* and *HLX* Loci in K562

Hox genes are known to have important roles in hematopoiesis and oncogenesis (Argiropoulos and Humphries 2007; Shah and Sukumar 2010; Eklund 2011). The 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 in 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 (Fig. 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* (Fig. 6A-C). Such insights into potential oncogene regulation cannot be

obtained by analyzing functional genomics and epigenomics data alone without genome structural information i.e. correct genomic context.

Combining Linked-Reads with Short-Insert WGS Reads for CN Analysis

While it is reasonable to assume that combining linked-read data with short-insert WGS reads should yield higher resolution for read-depth-based CN analysis, we combined our K562 linked-reads and short-insert WGS reads (total combined sequencing coverage >131×) for analysis of CN changes across the K562 genome. While doing so, we found that the genome coverage from linked-reads is more biased than that of WGS. This is not entirely surprising since the linked-read library preparation has a proprietary isothermal amplification step where the input is only 1 nanogram of DNA in addition to PCR. After performing this analysis, we found that higher-resolution is not achieved by including linked-reads data. Thus, we used only our deep-coverage short-insert WGS (~72x non-duplicate coverage) for high-resolution read-depth analysis of CN or ploidy changes across chromosome segments (Table S2).

Comparison of SV-Calling with Linked-Reads Vs. with Standard Paired-End WGS

By comparing the SVs called from using 10x-Genomics linked-read sequencing (e.g. produces Illumina reads, but considered as "third generation technology") with those identified from standard Illumina paired-end (short-insert or mate-pair), we see that 1,263 out of 3,360 SVs or 38% are not ascertainable by standard Illumina paired-end sequencing methods. We also would like to point out that this can also be interpreted as the majority (62%) of SVs identified using 10x-Genomics linked-read sequencing can also be identified by standard Illumina paired-end sequencing. However, this interpretation is not accurate since the vast majority of SVs detected by only linked-read sequencing (the ones not ascertainable by standard Illumina paired-end sequencing) are large SVs (>30 kb) and may contain multiple breakpoints. We see that 2,097 out of the total 4,831 SVs (43%) identified using only standard Illumina paired-end sequencing (excluding complex SVs identified from ARC-SV) were also identified using linked-read sequencing. Linked-reads, if sequenced to the same coverage

depths, already possess most of the information obtained from standard short-insert Illumina paired-end sequencing, and thus can also be analyzed as such by not taking into account the high-molecular weight DNA barcodes that 'link' these reads together. Therefore, the more accurate interpretation is that 43% of the SVs -- identified by combining standard short-insert and mate-pair Illumina sequencing -- can already be identified by using 10x-Genomics linked-read sequencing alone in the K562 cancer genome.

In addition, we see that 170 out of 220 (77%) of the large or complex SVs identified from linked-read sequencing using GROC-SVs have at least some supporting evidence (i.e. one or more supporting paired-end reads) from standard Illumina mate-pair sequencing; 142 of these were also identified as SVs using standard Illumina mate-pair sequencing. However, for 162 out of these 170 SVs (>95%) with supporting mate-pair evidence, the breakpoints were either more precisely defined, haplotype-resolved, or sequence assembled from the linked-read data which cannot be done so from the standard Illumina mate-pair data.

REFERENCES

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.

Bell JM, Lau BT, Greer SU, Wood-Bouwens C, Xia LC, Connolly ID, Gephart MH, Ji HP. 2017. Chromosome-scale mega-haplotypes enable digital karyotyping of cancer aneuploidy. *Nucleic Acids Res* **45**: e162. http://www.ncbi.nlm.nih.gov/pubmed/28977555.

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.

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.

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.

Dagmar M, Beer DG, Wilke CW, Miller DE, Glover TW. 1997. Frequent deletions of FHIT and FRA3B in Barrett's metaplasia and esophageal adenocarcinomas. *Oncogene* **15**: 1653–1659. http://www.nature.com/articles/1201330.

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.

Durkin SG, Ragland RL, Arlt MF, Mulle JG, Warren ST, Glover TW. 2008. Replication stress induces tumor-like microdeletions in FHIT/FRA3B. *Proc Natl Acad Sci* **105**: 246–251. http://www.pnas.org/cgi/doi/10.1073/pnas.0708097105.

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.

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.

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.

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.

Hu L, Chen H-Y, Cai J, Yang G-Z, Feng D, Zhai Y-X, Gong H, Qi C-Y, Zhang Y, Fu H, et al. 2015. Upregulation of NETO2 expression correlates with tumor progression and poor prognosis in colorectal carcinoma. *BMC Cancer* **15**: 1006. http://www.ncbi.nlm.nih.gov/pubmed/26699544.

Huebner K, Garrison PN, Barnes LD, Croce CM. 1998. The role of the FHIT/FRA3B locus in cancer. *Annu Rev Genet* **32**: 7–31. http://www.ncbi.nlm.nih.gov/pubmed/9928473.

Ingvarsson S. 2001. FHIT alterations in breast cancer. *Semin Cancer Biol* **11**: 361–366. http://linkinghub.elsevier.com/retrieve/pii/S1044579X01903918.

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.

Li Y, Zhou S, Schwartz DC, Ma J. 2016. Allele-Specific Quantification of Structural Variations in Cancer Genomes. *Cell Syst* **3**: 21–34. http://linkinghub.elsevier.com/retrieve/pii/S240547121630182X.

Lubinski J, Hadaczek P, Podolski J, Toloczko A, Sikorski A, McCue P, Druck T, Huebner K. 1994. Common regions of deletion in chromosome regions 3p12 and 3p14.2 in primary clear cell renal carcinomas. *Cancer Res* **54**: 3710–3. http://www.ncbi.nlm.nih.gov/pubmed/8033088.

Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney J, Fehr A, et al. 2018. Resolving the Full Spectrum of Human Genome Variation using Linked-Reads. Preprint at. https://www.biorxiv.org/content/early/2018/01/09/230946.

Ohta M, Inoue H, Cotticelli MG, Kastury K, Baffa R, Palazzo J, Siprashvili Z, Mori M, McCue P, Druck T, et al. 1996. The FHIT gene, spanning the chromosome 3p14.2 fragile site and renal carcinoma-associated t(3;8) breakpoint, is abnormal in digestive tract cancers. *Cell* **84**: 587–97. http://www.ncbi.nlm.nih.gov/pubmed/8598045.

Oparina NY, Sadritdinova AF, Snezhkina A V., Dmitriev AA, Krasnov GS, Senchenko VN, Melnikova N V., Belenikin MS, Lakunina VA, Veselovsky VA, et al. 2012. Increase in NETO2 gene expression is a potential molecular genetic marker in renal and lung cancers. *Russ J Genet* **48**: 506–512. http://link.springer.com/10.1134/S1022795412050171.

Prasanth SG, Prasanth K V, Stillman B. 2002. Orc6 involved in DNA replication, chromosome segregation, and cytokinesis. *Science* **297**: 1026–31. http://www.ncbi.nlm.nih.gov/pubmed/12169736.

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.

Seguchi O, Takashima S, Yamazaki S, Asakura M, Asano Y, Shintani Y, Wakeno M, Minamino T, Kondo H, Furukawa H, et al. 2007. A cardiac myosin light chain kinase regulates sarcomere assembly in the vertebrate heart. *J Clin Invest* **117**: 2812–24. http://www.ncbi.nlm.nih.gov/pubmed/17885681.

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.

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.

Sozzi G, Veronese ML, Negrini M, Baffa R, Cotticelli MG, Inoue H, Tornielli S, Pilotti S, De Gregorio L, Pastorino U, et al. 1996. The FHIT Gene at 3p14.2 Is Abnormal in Lung Cancer. *Cell* **85**: 17–26. http://linkinghub.elsevier.com/retrieve/pii/S0092867400810788.

Stöhr H, Berger C, Fröhlich S, Weber BHF. 2002. A novel gene encoding a putative transmembrane protein with two extracellular CUB domains and a low-density lipoprotein class A module: isolation of alternatively spliced isoforms in retina and brain. *Gene* **286**: 223–31. http://www.ncbi.nlm.nih.gov/pubmed/11943477.

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.

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.

Waters CE, Saldivar JC, Hosseini SA, Huebner K. 2014. The FHIT gene product: tumor

suppressor and genome “caretaker.” *Cell Mol Life Sci* **71**: 4577–4587. http://link.springer.com/10.1007/s00018-014-1722-0.

Wistuba II, Virmani AK, Gazdar AF, Lam S, LeRiche J, Behrens C, Fong KM, Samet JM, Srivastava S, Minna JD. 1997. Molecular Damage in the Bronchial Epithelium of Current and Former Smokers. *JNCI J Natl Cancer Inst* **89**: 1366–1373. https://academic.oup.com/jnci/article-lookup/doi/10.1093/jnci/89.18.1366.

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.

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).

SUPPLEMENTAL METHODS

Analysis Overview

We combined multiple experimental and analysis methods (Supplementary Fig. S1A), including karyotyping, array CGH, deep (72× non-duplicate coverage) short-insert whole-genome sequencing (WGS), 3 kb-mate-pair sequencing (Korbel et al. 2007) and 10x Genomics linked-reads sequencing (Zheng et al. 2016; Marks et al. 2018), to comprehensively characterize the genome of the primary ENCODE cell line K562 (Fig. 1). The WGS dataset was used to identify CN i.e. ploidy by chromosome segments, SNVs, Indels, non-reference LINE1 and *Alu* insertions (Lupski 2010; Sudmant et al. 2015), and SVs such as deletions, duplications, inversions, insertions, and small-scale complex SVs. These SVs were identified using an integrated 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 linked-reads were used to phase SNVs and Indels as well as to identify, phase, reconstruct, and assemble primarily large (>30 kb) and complex SVs (Greer et al. 2017; Spies et al. 2017; Marks et al. 2018), though additional small-scale deletions were also identified and phased (Zheng et al. 2016). SVs and REIs were experimentally validated with PCR and Sanger sequencing. Phased SNV haplotype blocks in aneuploid regions were “stitched” to mega-haplotypes by leveraging haplotype imbalance (Bell et al. 2017). The 3 kb-mate-pair data was used to identify additional SVs and was also used to validate large and complex SVs identified from linked-reads. Functional genomics and epigenomics datasets from ENCODE were integrated with CN and phasing information to identify allele-specific RNA expression and allele-specific DNA methylation. Phased variants were also used to identify allele-specific CRISPR targets in the K562 genome.

Array CGH analysis

Array hybridization was performed using the HD2 080131_HG18_WG_CGH_v2DCR_HX1 array CGH platform from Roche NimbleGen, Inc. (Madison, WI, USA) (Urban et al. 2006; Korbel et al.

2007, 2009). This array contains 2.1 million probes tilting hg18 with a median probe spacing of 1169bp and probe length of 60bp. We followed the protocol in "NimbleGen Array User's Guide – CGH Analysis" from the manufacturer. Briefly, genomic DNA from K562 cells was labeled with cy3 dye and genomic DNA from a pool of 7 females (Promega) was labeled with cy5 dye. DNA was not sonicated as recommended by NimbleGen Technical Support. Equal amounts of the labeled K562 and pooled-control DNA were hybridized to the arrays for 72 hours. A dye swap was also performed. Arrays were washed and scanned (PMT gain: 532nm, 635nm) with using the Axon 4200A scanner from Genepix (San Jose, CA, USA). The images were analyzed using the SegMNT algorithm in the NimbleScan 2.5 software suite (Roche NimbleGen, Inc) with default settings. Genomic coordinates were then converted from hg18 to hg19 using the UCSC LiftOver tool (Hinrichs et al. 2006).

Illumina short-insert WGS

K562 genomic DNA was sent to Macrogen (Rockville, MD, USA) for standard Illumina short-insert library preparation and WGS. Two independent libraries were prepared and sequenced (2x151 bp) on two lanes of the Illumina HiSeq X to achieve >70x genomic coverage. Reads were aligned to the hg19 using BWA-MEM version 0.7.5 (Li and Durbin 2009) followed by marking of duplicates using Picard tools (version 1.129) (http://broadinstitute.github.io/picard/) Local Indel realignment and base quality recalibration 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

WGS sequencing coverage was calculated in 10 kb bins across the genome and plotted against the % GC content of each bin to verify the existence of discrete clusters corresponding to discrete CNs (Supplementary Fig. S2). CN was assigned to a cluster based on the ratio of its mean coverage to that of the lowest cluster. For an example, the cluster with the lowest mean coverage was assigned CN1, and the cluster with twice as much mean coverage was assigned CN2 and so forth. The ratios for the five discrete clusters observed corresponded almost

perfectly to CN1, CN2, CN3, CN4, and CN5. WGS coverage across the genome and across each chromosome was examined visually to assign CN for different chromosome segments or entire chromosomes based on the cluster analysis where adjacent chromosomal segments with different CNs could be identified by the clearly visible sharp and steep changes in sequencing coverage (Supplementary Fig. S3, Supplemental Data). For each chromosome segment, SNVs and Indels were called using by GATK Haplotypecaller (version 3.7) (McKenna et al. 2010) by specifying the CN or ploidy of that chromosome segment (*stand_emit_conf=0.1, variant_index_type=LINEAR, variant_index_parameter=128000, ploidy={CN}*). The resulting Haplotypecaller outputs from all chromosome segments were then concatenated, and variant quality scores were recalibrated using GATK VQSR with training datasets (dbSNP 138, HapMap 3.3, Omni 2.5 genotypes, 1000 Genomes Phase 1) as recommended by GATK Best Practices (Van der Auwera et al. 2013; DePristo et al. 2011) and filtered with the setting *tranche = 99.0*. SNVs and Indels were annotated using dbSNP138 (Sherry et al. 2001) followed by SnpEff (version 4.3; *canonical transcripts*) (Cingolani et al. 2012a) and then filtered for protein altering variants using SnpSift (version 4.3; *'HIGH' and 'MODERATE' putative impact*) (Cingolani et al. 2012b). Protein-altering variants were intersected with the variants from the 1000 Genomes Project (The 1000 Genomes Project Consortium et al. 2015) and the Exome Sequencing Project (http://evs.gs.washington.edu/EVS/) where overlapping variants were removed using Bedtools (version 2.26) (Quinlan and Hall 2010). The resulting PPA variant calls were overlapped against the Catalogue of Somatic Mutations in Cancer (Forbes et al. 2015) and Sanger Cancer Gene Census (Futreal et al. 2004).

Identification of 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 SNVs that were (1) recalibrated and "PASS"-filtered from GATK VQSR and (2) overlapped 1000 Genomes Project variants (Sudmant et al. 2015). The genome was split into 40 kb bins; heterozygous and

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

10x Genomics linked-read library and sequencing

K562 genomic DNA sample (~35 kb-80 kb) was size selected on the BluePippin instrument (Sage Science, Beverly, MA, USA) using the manufacturer's U1 Maker 30 kb High Pass protocol and then diluted to 1 ng/µl and used as input for the Chromium reagent delivery system (Zheng et al. 2016; Marks et al. 2018) from 10x Genomics (Pleasanton, CA, USA) where HMW DNA fragments are partitioned into >1 million droplets, uniquely barcoded (16 bp) within each droplet, and subjected to random priming and isothermal amplification following standard manufacturer's protocol. Afterwards, the emulsion was broken, and the barcoded DNA molecules were released and converted to a Chromium linked-read library in which each library molecule retains its "HMW fragment barcode". Read-pairs generated in this manner, i.e. linked-reads, that come from the same HMW DNA fragment, can be identified by their "HMW fragment barcode" and can be used to construct a virtual long-read that is representative of the sequence of the original HMW genomic DNA fragment. Reads that cover heterozygous SNVs and Indels can be phased by their "HMW fragment barcode". The final library (8 cycles of PCR amplification) was diluted to 5 nM and sent to Macrogen (Rockville, MD, USA) for sequencing (2x151 bp) on two lanes of the Illumina HiSeq X to achieve ~60x genomic coverage. We estimate the actual physical coverage (C_F) to be 191×. The overall sequencing coverage is C = C_R x C_F = 59×. The length of sequence coverage per 2 × 151 bp paired-ended read minus 16 bp of "HWM fragment barcode" is 286 bp, thus coverage (C_R) of the average input HMW genomic DNA (59 kb) is 18,304 bp (286 bp x 64 linked-reads) or 31.0% of 50 kb.

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

Paired-end linked-reads (median insert size 385 bp, duplication rate 6.19%, Q30 Read1 88.7%, Q30 Read2 63.8%) were aligned to hg19 (alignment rate 90.1%, mean coverage 59.0x, zero coverage 1.14%) and analyzed using the Long Ranger Software (version 2.1.5) from 10x Genomics (Zheng et al. 2016; Marks et al. 2018) (Pleasanton, CA, USA). Segmental duplications, reference gaps, unplaced contigs, regions with assembly issues, and highly polymorphic sites (http://cf.10xgenomics.com/supp/genome/hg19/sv_blacklist.bed, http://cf.10xgenomics.com/supp/genome/hg19/segdups.bedpe) were excluded from the analysis. ENSEMBL annotations (http://cf.10xgenomics.com/supp/genome/gene_annotations.gtf.gz) were used for genes and exons. Phasing was performed by specifying the set of pre-called and filtered K562 heterozygous SNVs and Indels from GATK (see above) and formatted using *mkvcf* from Long Ranger (version 2.1.5). Heterozygous SNVs and Indels with more than two types of alleles in ploidy>2 regions were excluded from analysis. Large (>30 kb) SVs and large-scale complex rearrangements were identified using both the Long Ranger *wgs* module with the "*--somatic*" option, GROC-SVs (default settings with breakpoint assembly) (Spies et al. 2017), and gemtools (Greer et al. 2017). The *"--somatic*" option increases the sensitivity of the large-scale SV caller for somatic SVs by allowing the detection of sub-haplotype events and does not affect small-scale variant calling. Variants from Long Ranger analysis indicated as "PASS" were retained. SV breakpoints identified using GROC-SVs were also analyzed for supporting evidence from mate-pair reads (see below).

Mega-haplotype analysis

Haplotype-blocks from Long Ranger were "stitched" to mega-haplotype blocks by leveraging the haplotype imbalance in aneuploid regions using the methods described in (Bell et al. 2017). Briefly, we counted the number of linked-read barcodes for each phased heterozygous SNVs assigned to haplotype blocks that contain ≥100 phased SNVs (Dataset S2).

Since each barcode is specific to a given HMW DNA molecule, the total number of unique barcodes is directly associated with the number of individual HMW DNA molecules sequenced. In other words, the counting of unique barcode associated with a particular sequence gives the fractional representation of that sequence (or genomic locus). Thus, for each phased haplotype in aneuploid regions with CN>2, major and minor haplotypes can be assigned according to the number of barcodes associated with each haplotype (Fig. 3), where the major haplotype simply has more associated unique barcodes than the minor. In diploid regions, the two haplotypes are expected to have similar barcode counts. A matched normal control genome is required in order to confidently discriminate between the major and minor haplotypes in a case genome (Bell et al. 2017). Because K562 has no matching normal sample, we used a female genome of the same ethnicity (NA12878) for which linked-read data is publicly available (https://support.10xgenomics.com/genome-exome/datasets/2.2.1/NA12878_WGS_v2). Only SNVs present in both K562 and NA12878 were included in the mega-haplotype blocks (Supplemental Data, Table 2). After verifying aneuploidy (or haplotype imbalance) by barcode counting and performing the normalization procedures and statistical tests as described in (Bell et al. 2017), we then "stitched" together contiguous phased haplotype blocks based on the imbalance between the major and minor haplotypes.

Allele-specific RNA expression

Two replicates of K562 poly(A) mRNA RNA-seq (bam files ENCFF412EYU & ENCFF037AFT from experiment ENCSR000AEM) were downloaded from the ENCODE portal (Sloan et al. 2016). Replicates were analyzed separately. Samtools mpileup (version 0.1.19) (Li et al. 2009) and BCFtools (version 0.1.19) (Narasimhan et al. 2016) were used to count the number of reads mapped to each allele of heterozygous SNVs; SNVs with coverage <10 in RNA-seq data were filtered out. The binomial test was used to determine if the fraction of RNA-seq reads that mapped each allele of a particular SNV significantly ($p < 0.05$) deviated from the

expected value of 0.5 for diploid regions or the expected values in aneuploid regions (e.g. 0.33, 0.67 in triploid regions) for each heterozygous SNV.

Allele-specific DNA methylation

One replicate of K562 WGBS data (library 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; *--dovetail, --bowtie2, -I 100, -X 500, --unmapped, --ambiguous*) (Krueger and Andrews 2011) resulting in 28.6x non-duplicate coverage. Paired-end reads that contain cytosines (with 10-200x sequencing coverage) in a CpG dinucleotide were selected if they also overlapped with phased heterozygous SNVs (Dataset S2). The haplotypes in which cytosines (methylated or unmethylated) belonged were then assigned based on phased heterozygous SNVs on that same read. CpG containing reads that overlap SNVs on different haplotypes within the same read (<1% of cases) were filtered out. CpGs were then grouped into CGIs. The phased reads overlapping each CGI were then grouped by haplotype, and the number of reads containing methylated or unmethylated cytosines were normalized by the CN of their corresponding haplotypes. Fisher's exact test was applied at each CGI to determine if the fraction of methylated to unmethylated reads was significantly different between the alleles. The *p*-values of each CGI from the Fisher's exact tests were used as input for the q-value package (https://github.com/StoreyLab/qvalue) in R (https://www.R-project.org) to see which Fisher's exact tests were significant given a target false discovery rate (FDR) of 10%. Since each CGI is overlapped by a different set of reads, we expect that each Fisher's exact test performed is independent of the others, thus satisfying the assumptions required for conservative control of the FDR (Storey and Tibshirani 2003).

Allele-specific CRISPR targets

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

1. They passed quality control (VCF field 'Filter' is equal to "PASS")

2. They are phased (VCF field 'GT' uses "|" as separator rather than "/")
3. The alleles are heterozygous (e.g. "0|1" or "1|0", but not "1|1")
4. The difference between the alleles is > 1 bp (This is required to ensure target specificity. If there were target sites with two or more SNVs difference between the alleles, then these target sites were included.)

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 sequence in the 'Ref' field was "GTA", the 'Alt' field sequence was "TA", and phasing was "0|1" i.e. Haplotype_1|Haplotype_2", the sequence containing "GTA" was tagged "1" for Haplotype 1 and the sequence containing "TA" was tagged "2" for Haplotype 2.
- If the 'Ref' field was sequence "GTA", the 'Alt' field was "TA,GCTA", phasing was "1|2", the sequence containing "TA" is tagged "1", the sequence containing "GCTA" was tagged "2" (and the sequence containing "GTA" was not used)

A regular expression was used to extract all potential CRISPR targets from these sequences (i.e. all sequences that matched a [G, C, or A]N_{20}GG pattern and those for which the reverse-complement matched this pattern). This yielded 532,013 candidates, which were then filtered to retain only high-quality targets. The process is adapted from a selection method previously described and validated (Sunagawa et al. 2016) and has already been used for more than 20 genes (Tatsuki et al. 2016). A high-quality candidate gRNA needs to target a unique site. All the candidates that have multiple exact matches in hg19 (irrespective of location) were identified using Bowtie2 (Langmead and Salzberg 2012) and removed. We also removed targets with extreme GC content (>80% or <20%), and targets that contain TTTT, which tends to break the gRNA's secondary structure. We also used the Vienna RNA-fold package (Lorenz et al. 2011)

to compute the gRNA's secondary structure. We eliminated all candidates for which the stem loop structure cannot fold correctly for Cas9 recognition (Nishimasu et al. 2014), except if the folding energy was above -18 (indicating that the "incorrectly-folded" structure is very unstable). Finally, we evaluated the off-target risk score using our own implementation of the Zhang tool (Ran et al. 2013). To ensure that all targets are as reliable and specific as possible, we used a very strict threshold and rejected candidates with an off-target risk score <75. Candidates that satisfy all these requirements are considered high quality. For each candidate, we report the location of the variant (chromosome and position), the haplotype ("1" or "2" from the "Phased Genotype" column i.e. "Haplotype_1 | Haplotype_2"), the gRNA target sequence, its position relative to the start of the variant, its orientation, its off-target score, and the genomic element targeted (gene or enhancer). Note that the position relative to the start of the variant is for the 5'-most end of the target respective to the genome: if the target is 5'-3', it is its 5' end; if a target was extracted on the reverse complement, it is its 3' end. List of protein-coding and RNA genes and enhancers were obtained from GENCODE ("comprehensive gene annotation", https://www.gencodegenes.org/releases/19.html) and VISTA Enhancer Browser (https://enhancer.lbl.gov) (Visel et al. 2007).

SV identification from deep-coverage short-insert WGS

SVs from deep-coverage short-insert WGS were identified using BreakDancer (version 1.4.5) (Chen et al. 2009), Pindel (version 0.2.4t) (Ye et al. 2009), BreakSeq (version 2.0) (Lam et al. 2010), and ARC-SV (Arthur et al, 2017) with default settings to obtain pre-filtered calls. All SV calls were required to be >50bp. We filtered out BreakDancer calls with <2 supporting paired-end reads and confidence scores <90. Pindel calls were filtered for quality scores >400. No further filtering was performed for BreakSeq calls. For ARC-SV calls, SVs with breakpoints in simple repeats, low complexity regions and satellite repeats (repeatmasker.org) as well as segmental duplications (hg19) downloaded from the UCSC Genome Browser (Kent et al. 2002; Karolchik et al. 2004) were filtered out. Insertion calls and complex SV calls containing

insertions were also filtered out as these calls are not reliable. The *BP_UNCERTAINTY* filter was added due to current technical limitations of the software package and masks just a few putative tandem duplications (Arthur et al. 2017). Specialized filtering was performed on tandem duplications and complex SVs. For tandem duplications, we required either 95% of reads overlapping the duplicated sequence to have mapq ≥20 or -or- ≥2 supporting split-reads (*SR* tag in VCF *INFO* column). For complex SVs, we required that each breakpoint is supported by a split-read or has >95% of overlapping reads with mapq ≥20. When part of a complex SV fails one of these filters, nearby breakpoints that were part of the same ARC-SV call i.e. those that were genotyped simultaneously with the complex SV (*EVENT* filter) were also removed.

Mate-pair library construction and sequencing

K562 mate-pair library (3 kb insert) was constructed using 2.5 μg of high molecular weight genomic DNA (mean >30 kb) as input for the Nextera Mate Pair Library Prep Kit (FC-132-1001) from Illumina (San Diego, CA, USA) following the manufacturer's protocol with 10 cycles of PCR. Insert-size selection (2.7 kb to 3.3 kb) was carried out on the BluePippin instrument from Sage Science (Beverly, MA, USA) using the S1 selection marker. Sequencing (2x151 bp) was performed twice on the Illumina NextSeq 500 using the NextSeq Mid-Output Kit (Illumina catalogue # FC-404-2003). Data from the two sequencing runs were combined for analysis. Illumina external read adapters and the reverse complement sequence of the Nextera circularized single junction adapter sequence (AGATGTGTATAAGAGACAG) were trimmed from the 3' end of reads followed by another trimming of the Nextera circularized single junction adapter (CTGTCTCTTATACACATCT) using the FASTQ Toolkit (version 1.0) application on Illumina Basespace (basespace.illumina.com). Trimmed reads were aligned to hg19 using BWA-MEM (version 0.7.12-r1039) with the "*-M*" option (Li and Durbin 2009). Duplicates were marked and removed using Samtools (version 1.2) (Li et al. 2009) followed by Picard tools version 1.52 (http://broadinstitute.github.io/picard/). Indels were then locally realigned against the Mills & 1000 Genomes Gold Standard Indels (Mills et al. 2011; The 1000 Genomes Project

Consortium et al. 2012) and base scores were recalibrated using GATK following the standard Best Practices workflow (Van der Auwera et al. 2013; DePristo et al. 2011).

SV calling from mate-pair sequencing

SV calls were made using LUMPY (version 0.2.11) (Layer et al. 2014). Split-reads and discordantly-mapped reads were first extracted and sorted from the processed alignment file as described in github.com/arq5x/lumpy-sv (Layer et al. 2014). The *lumpyexpress* command was issued to obtain pre-filtered SV calls. Segmental duplications and reference gaps (hg19) downloaded from the UCSC Genome Browser (Kent et al. 2002; Karolchik et al. 2004) were excluded from the analysis through the "*-x*" option. SV calls <50 bp were filtered out. To select for high-confidence calls, only SVs that have ≥5 supporting reads as well as both discordant and split-read support were retained.

Non-Reference Retrotransposon Insertions (REI)

We adapted the RetroSeq package (Keane et al. 2013) to call non-reference LINE1 and Alu insertions. The calling depends on (a) split-reads where part of the read maps to uniquely hg19 sequence not normally adjacent to a retrotransposon sequence and the remainder maps to a catalogue of active LINE1 and Alu consensus sequences from Repbase (Bao et al. 2015) and/or (b) paired-end reads where one read maps to unique sequence and the paired read maps to LINE1 or Alu consensus sequences and not to unique hg19 sequence. The mapping quality was required to be >85% identity as set in RetroSeq (Keane et al. 2013). We also required ≥6 supporting reads (split or paired-end) in order to filter out low-confidence calls. The boundaries for the transposon insertions (Supplementary Table S9) are conservative estimates for the insertion junctions. 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 <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 <1 kb away from the "left boundary".

K562 POLR2A ChIP-seq and RNA-seq analysis

Four replicates of K562 poly(A) mRNA RNA-seq transcript quantification data (ENCFF381QQP & ENCFF705JDM from experiment ENCSR000AEM; ENCFF928EIW from experiment ENCSR000AEO; ENCFF225LEY from experiment ENCSR545DKY) and K562 POLR2A ChIP-seq alignments (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 Pearson correlation analysis.

REFERENCES

Arthur JG, Chen X, Zhou B, Urban AE. 2017. Detection of complex structural variation from paired-end sequencing data. *bioRxiv* 200170.

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.

Bell JM, Lau BT, Greer SU, Wood-Bouwens C, Xia LC, Connolly ID, Gephart MH, Ji HP. 2017. Chromosome-scale mega-haplotypes enable digital karyotyping of cancer aneuploidy. *Nucleic Acids Res* **45**: e162. http://www.ncbi.nlm.nih.gov/pubmed/28977555.

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.

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.

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.

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.

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

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.

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.

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, Grubert F, Du J, Royce TE, Starr P, Zhong G, Emanuel BS, Weissman SM, Snyder M, et al. 2007. 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.

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.

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.

Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney J, Fehr A, et al. 2018. Resolving the Full Spectrum of Human Genome Variation using Linked-Reads. *bioRxiv* 230946. https://www.biorxiv.org/content/early/2018/01/09/230946.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–303. http://www.ncbi.nlm.nih.gov/pubmed/20644199.

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

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

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

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.

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.

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

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

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

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

Sunagawa GA, Sumiyama K, Ukai-Tadenuma M, Perrin D, Fujishima H, Ukai H, Nishimura O, Shi S, Ohno R-I, Narumi R, et al. 2016. Mammalian Reverse Genetics without Crossing Reveals Nr3a as a Short-Sleeper Gene. *Cell Rep* **14**: 662–677. http://www.ncbi.nlm.nih.gov/pubmed/26774482.

Tatsuki F, Sunagawa GA, Shi S, Susaki EA, Yukinaga H, Perrin D, Sumiyama K, Ukai-Tadenuma M, Fujishima H, Ohno R, et al. 2016. Involvement of Ca(2+)-Dependent Hyperpolarization in Sleep Duration in Mammals. *Neuron* **90**: 70–85. http://www.ncbi.nlm.nih.gov/pubmed/26996081.

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

The 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang

HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74. http://www.ncbi.nlm.nih.gov/pubmed/26432245.

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.

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

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).