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

Zhou, Bo; Ho, Steve S.; Greer, Stephanie U.; Zhu, Xiaowei; Bell, John M.; Arthur, Joseph G.; Spies, Noah; Zhang, Xianglong; Byeon, Seunggyu; Pattni, Reenal; Ben-Efraim, Noa; Haney, Michael S.; Haraksingh, Rajini R.; Song, Giltae; Ji, Hanlee P.; Perrin, Dimitri; Wong, Wing H.; Abyzov, Alexej; Urban, Alexander E.

Published in:
Genome Research

Published: 01/03/2019

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

License:
CC BY-NC

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1101/gr.234948.118

Publication details:

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.
Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562

Bo Zhou,1,2 Steve S. Ho,1,2 Stephanie U. Greer,3 Xiaowei Zhu,1,2 John M. Bell,4 Joseph G. Arthur,5,13 Noah Spies,2,6,7,14 Xianglong Zhang,1,2 Seunggyu Byeon,8 Reenal Pattni,1,2 Noa Ben-Efraim,1,2 Michael S. Haney,1,2 Rajini R. Haraksingh,1,2,15 Giltae Song,8 Hanlee P. Ji,3,4 Dimitri Perrin,9 Wing H. Wong,5,10 Alexej Abyzov,11 and Alexander E. Urban1,2,12

1Department of Psychiatry and Behavioral Sciences, 2Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA; 3Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, California 94305, USA; 4Stanford Genome Technology Center, Stanford University, Palo Alto, California 94304, USA; 5Department of Statistics, Stanford University, Stanford, California 94305, USA; 6Department of Pathology, Stanford University School of Medicine, Stanford, California 94305, USA; 7Genome-Scale Measurements Group, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA; 8School of Computer Science and Engineering, College of Engineering, Pusan National University, Busan 46241, South Korea; 9Science and Engineering Faculty, Queensland University of Technology, Brisbane, QLD 4001, Australia; 10Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, California 94305, USA; 11Department of Health Sciences Research, Center for Individualized Medicine, Mayo Clinic, Rochester, Minnesota 55905, USA; 12Tashia and John Morgridge Faculty Scholar, Stanford Child Health Research Institute, Stanford, California 94305, USA

K562 is widely used in biomedical research. It is one of three tier-one cell lines of ENCODE and also most commonly used for large-scale CRISPR/Cas9 screens. Although its functional genomic and epigenomic characteristics have been extensively studied, its genome sequence and genomic structural features have never been comprehensively analyzed. Such information is essential for the correct interpretation and understanding of the vast troves of existing functional genomics and epigenomics data for K562. We performed and integrated deep-coverage whole-genome (short-insert), mate-pair, and linked-read sequencing as well as karyotyping and array CGH analysis to identify a wide spectrum of genome characteristics in K562: copy numbers (CN) of aneuploid chromosome segments at high-resolution, SNVs and indels (both corrected for CN in aneuploid regions), loss of heterozygosity, megabase-scale phased haplotypes often spanning entire chromosome arms, structural variants (SVs), including small and large-scale complex SVs and nonreference retrotransposon insertions. Many SVs were phased, assembled, and experimentally validated. We identified multiple allele-specific deletions and duplications within the tumor suppressor gene FHIT. Taking aneuploidy into account, we reanalyzed K562 RNA-seq and whole-genome bisulfite sequencing data for allele-specific expression and allele-specific DNA methylation. We also show examples of how deeper insights into regulatory complexity are gained by integrating genomic variant information and structural context with functional genomics and epigenomics data. Furthermore, using K562 haplotype information, we produced an allele-specific CRISPR targeting map. This comprehensive whole-genome analysis serves as a resource for future studies that utilize K562 as well as a framework for the analysis of other cancer genomes.

[Supplemental material is available for this article.]

Although the functional genomic characteristics of K562 have been extensively studied and documented, reflected in close to 600 ChIP-seq, 400 RNA-seq, 50 DNase-seq, and 30 RIP-seq data sets available through the ENCODE portal (Sloan et al. 2016), the sequence and structural features of the K562 genome have never been comprehensively characterized, even though past cytogenetic studies using G-bandning, fluorescence in situ hybridization (FISH), multiplex-FISH, and comparative genomic hybridization (CGH) showed that K562 cells contain pervasive aneuploidy and multiple gross structural abnormalities (Selden et al. 1983; Wu et al. 1995; Gribble et al. 2000; Naumann et al. 2001), 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, integrative analyses that have been carried out using the vast troves of K562 ENCODE data—were done without taking into account the many differences of the K562 genome relative to the human reference genome. This leads to skewed interpretations and reduces the amount of knowledge that can be gained from the rich, multilayered ENCODE data sets that continue to accumulate.

Here, we report for the first time a comprehensive characterization of the K562 genome (Fig. 1; Supplemental Fig. S1A) that includes copy numbers (CNs) of chromosome segments at high-resolution; single-nucleotide variants (SNVs, also including single-nucleotide polymorphisms [SNPs]); and small insertions and deletions (indels) with allele frequencies corrected by CN in aneuploid regions, loss of heterozygosity, megabase-scale phased haplotypes often spanning entire chromosome arms, and structural variants (SVs), including small and large-scale complex SVs with phasing. We then took first steps into exploring how knowledge

![Figure 1](https://genome.cshlp.org/)

**Figure 1.** Comprehensive overview of the K562 genome. Circos (Krzywinski et al. 2009) visualization of the K562 genome with the following tracks in inward concentric order: chromosomes; CN, i.e., ploidy by chromosome segment; merged SV density in 1.5-Mb windows of deletions, duplications, and inversions identified using ARC-SV (Arthur et al. 2018), BreakDancer (Chen et al. 2009), BreakSeq (Lam et al. 2010), LUMPY (Layer et al. 2014), Pindel (Ye et al. 2009), and Long Ranger (Zheng et al. 2016; Marks et al. 2018); phased haplotype blocks (demarcated with four colors for clearer visualization); SNV density in 1-Mb windows; indel density in 1-Mb windows; dominant zygosity in 1-Mb windows (heterozygous or homozygous >50%) with regions exhibiting loss of heterozygosity (LOH) indicated; RNA-seq reads for loci exhibiting allele-specific expression; CpG islands (CGI) exhibiting allele-specific methylation; histogram (log-scale) of allele-specifically methylated CpGs in 50-kb windows; nonreference Alu and LINE-1 insertions; allele-specific CRISPR target sites; large-scale rearrangements detected by Long Ranger (Zheng et al. 2016; Marks et al. 2018) (light blue: intrachromosomal; dark blue: interchromosomal); and by GROC-SVs (Śpies et al. 2017) (light gray: intrachromosomal; dark gray: interchromosomal).
about genome sequence and structural features can influence the interpretation of functional genomics and epigenomics data and show examples of how deeper insights into genome regulatory complexity can be obtained by integrating genomic context. These insights also shed light on important questions regarding cancer evolution.

**Results**

**Karyotyping**

The K562 cell line exhibits pervasive aneuploidy (Fig. 2A). Analysis of 20 individual K562 cells using GTW banding showed that all cells demonstrated a near-triploid karyotype and are characterized by multiple structural abnormalities. The karyotype of our line of K562 cells is overall consistent (although not identical) with previously published karyotypes (Selden et al. 1983; Wu et al. 1995; Gribble et al. 2000; Naumann et al. 2001), suggesting that its near-triploid state arose during leukemogenesis or early in the establishment of the cell line. It also suggests that different K562 cell lines kept and passaged 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 (Supplemental Table S1) with Chromosomes 10, 12, and 21 showing the most variability.

![Figure 2. K562 ploidy and haplotypes. (A) Representative karyogram of K562 cells produced by GTW banding 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,-3,7dup(6)(p21p25),+7,7inv(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)(q22),-20,der(21)t(1;21)(q21;p11),-22,+4∼7mar[cp20]. (B) CN (i.e., ploidy) by percentage 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 box plot, of phased haplotype block sizes (y-axis, log-scaled) in which the dashed line represents the N50 value (2,721,866 bp).](image)
Identification of copy number (CN) by chromosome segments

We used read-depth analysis (Abzyov et al. 2011) to assign a CN, that is, ploidy to all chromosome segments at 10-kb resolution or entire chromosomes in the K562 genome (Fig. 1; Supplemental Table S2). We first calculated WGS coverage in 10-kb bins and plotted it against %GC content where five distinct clusters were clearly observed (Supplemental Fig. S2). Clusters were designated as corresponding to particular CNs based on the mean coverage of each cluster (Supplemental Methods). Such designations confirm that the triploid state is the most common in the K562 genome. The CNs assigned to all chromosome segments using this approach are consistent with array CGH (Supplemental Fig. S3; Supplemental Data) and also with previous CGH analyses (Gribble et al. 2000; Naumann et al. 2001) with minor differences on Chromosomes 7, 10, 11, and 20 (Supplemental Table S3). Although on a general level, the CNs identified based on read-depth analysis track the findings from karyotyping, read-depth analysis reveals the CNs of many chromosome segments that would not have been apparent from karyotyping alone (Supplemental Fig. S3; Supplemental Data; Supplemental Table S2). We see that 53.5% of the K562 genome has a baseline CN of three (consistent with the karyotype) (Fig. 2A), 16.9% CN of four, 1.9% CN of five, 2.4% CN of one, and only 30.0% has remained in a diploid state (Fig. 2B). In addition, two large regions (5.8 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 (Supplemental Table S2).

SNVs and indels

We identified SNVs and indels in the K562 genome. We assigned heterozygous allele frequencies to these variants by taking into account ploidy in which nonconventional frequencies are included (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 million SNVs (1.45 million heterozygous, 1.64 million homozygous) and 0.70 million indels (0.39 million heterozygous, 0.31 million homozygous) (Table 1; Supplemental Data Set S1). There are 13,471 heterozygous SNVs and indels that have more than two haplotypes in aneuploid regions where CN is >2 (Supplemental Data Set S1). Furthermore, Chromosomes 3, 9, 13, 14, and X along with large stretches of Chromosomes 2, 10, 12, 17, 20, and 22 show loss of heterozygosity (LOH) (Fig. 1; Supplemental Table S4). Although a normal tissue sample corresponding to K562 is not available for comparative analysis, we overlapped these SNVs and indels with those in dbSNP138 (Sherry et al. 2001) and found the overlap to be 98% and 79%, respectively (Fig. 2C; Supplemental Data Set S1), suggesting an accumulation of a significant number of K562-specific SNVs and indels relative to germline variants present in the population. After filtering for protein-altering SNVs and indels in K562 that overlap with those identified from the 1000 Genomes Project or from the Exome Sequencing Project (http://evs.gs.washington.edu/EVS/), we found that 424 SNVs and 148 indels are private protein-altering (PPA) (Table 1; Supplemental Table S5). Furthermore, the overlap between the PPA variants and the Catalogue of Somatic Mutations in Cancer (COSMIC) is 53% and 31% for SNVs and indels, respectively (Supplemental Table S6). Eighteen genes that acquired PPA variants overlap with the Sanger Cancer Gene Census; canonical tumor suppressor genes and oncogenes such as RAD51B, TP53, PDGFRα, RABEPI, EPAS1, and WHSC1 are notably present among them (Supplemental Table S7).

Haplotype phasing

We performed haplotype phasing for the K562 genome by performing 10x Genomics linked-read library preparation and sequencing (Zheng et al. 2016; Marks et al. 2018). This library was sequenced (2 × 151 bp) to 59× genome coverage. Post-sequencing quality-control analysis showed that 1.06 ng, or about 320 genome equivalents, of high molecular weight (HMW) K562 genomic DNA fragments (average fragment size = 59 kb, 95.3% >20 kb, 11.9% >100 kb) were partitioned into 1.56 million oil droplets for unique barcoding. Half of all reads come from HMW DNA molecules with at least 64 linked reads (N50 Linked Reads per Molecule or LPM) (Table 1). We estimate the actual physical coverage (Cp) to be 191 × (Supplemental Methods). Using Long Ranger (Marks et al. 2018), 1.41 million (97.2%) of heterozygous SNVs and 0.58 million (83.7%) of indels (previously identified) (Supplemental Data Set S1) were successfully phased into 4987 haplotype blocks (Fig. 1; Table 1; Supplemental Data Set S2). The longest is 11.95 Mb (N50 = 2.72 Mb) (Fig. 2D; Table 1; Supplemental Data Set S2); however, haplotype block lengths vary widely across different chromosomes (Supplemental Fig. S4; Fig. 1) with poorly phased regions corresponding to regions with LOH (Fig. 1; Supplemental Table S4; Supplemental Data Set S2).

Mega-haplotypes encompassing entire chromosome arms

Leveraging the haplotype imbalance in aneuploid regions, we constructed mega-haplotypes (Table 2; Fig. 3; Supplemental Data) often encompassing entire K562 chromosome arms, by “stitching” the phased haplotype blocks obtained from Long Ranger (Supplemental Data Set S2) that contain ≥100 phased heterozygous SNVs

<table>
<thead>
<tr>
<th>Small variant calls</th>
<th>SNVs</th>
<th>Indels</th>
<th>Phasing</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>3,088,312</td>
<td>702,787</td>
<td>Percentage of phased heterozygous SNVs 97</td>
</tr>
<tr>
<td>Heterozygous/homozygous</td>
<td>1,451,017/1,637,295</td>
<td>393,632/309,155</td>
<td>Percentage of phased indels 84</td>
</tr>
<tr>
<td>Protein altering</td>
<td>10,831 (0.4%)</td>
<td>1118 (0.2%)</td>
<td>Longest phase block 11,953,412</td>
</tr>
<tr>
<td>dbSNP138</td>
<td>3,020,306 (98%)</td>
<td>558,637 (79%)</td>
<td>Number of phase blocks 11,953,412</td>
</tr>
<tr>
<td>Heterozygous/homozygous</td>
<td>1,389,196/1,629,672</td>
<td>294,850/260,606</td>
<td>N50 phase block 2,721,866</td>
</tr>
<tr>
<td>Novel</td>
<td>69,553 (2%)</td>
<td>149,053 (21%)</td>
<td>N50 linked reads per molecule 64</td>
</tr>
<tr>
<td>Heterozygous/homozygous</td>
<td>61,821/7623</td>
<td>98,782/48,548</td>
<td>Barcodes detected 1,562,771</td>
</tr>
<tr>
<td>The 1000 Genomes Project and Exome Sequencing Project overlap (with protein-altering variants)</td>
<td>10,407 (96%)</td>
<td>970 (87%)</td>
<td>Mean DNA per barcode (bp) 456,351</td>
</tr>
<tr>
<td>Novel protein altering</td>
<td>10,407 (96%)</td>
<td>970 (87%)</td>
<td>Mean DNA per barcode (bp) 456,351</td>
</tr>
<tr>
<td>COSMIC overlap</td>
<td>227 (53%)</td>
<td>46 (32%)</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Haplotypes constructed in aneuploid regions by leveraging haplotype imbalance

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Start</th>
<th>End</th>
<th>Chromosome arm</th>
<th>Arm covered (%)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19,708,577</td>
<td>21,759,128</td>
<td>1p</td>
<td>2</td>
<td>3.00 × 10⁻⁹</td>
</tr>
<tr>
<td>3</td>
<td>40,634,625</td>
<td>42,102,575</td>
<td>1p</td>
<td>1</td>
<td>0.45</td>
</tr>
<tr>
<td>1</td>
<td>54,592,451</td>
<td>108,745,206</td>
<td>1p</td>
<td>97</td>
<td>2.20 × 10⁻⁹</td>
</tr>
<tr>
<td>2</td>
<td>21,888</td>
<td>89,128,628</td>
<td>2p</td>
<td>98</td>
<td>2.20 × 10⁻⁹</td>
</tr>
<tr>
<td>2</td>
<td>98,318,199</td>
<td>133,102,616</td>
<td>2q</td>
<td>37</td>
<td>2.28 × 10⁻⁷</td>
</tr>
<tr>
<td>4</td>
<td>186,265</td>
<td>1,265,477</td>
<td>4p</td>
<td>2</td>
<td>2.97 × 10⁻⁸</td>
</tr>
<tr>
<td>1</td>
<td>4,013,687</td>
<td>49,037,941</td>
<td>4q</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>52,684,820</td>
<td>190,151,131</td>
<td>4q</td>
<td>99</td>
<td>2.20 × 10⁻⁹</td>
</tr>
<tr>
<td>5</td>
<td>50,641,459</td>
<td>180,442,383</td>
<td>5q</td>
<td>99</td>
<td>2.20 × 10⁻⁹</td>
</tr>
<tr>
<td>6</td>
<td>329,512</td>
<td>55,484,834</td>
<td>6p</td>
<td>94</td>
<td>3.36 × 10⁻⁷</td>
</tr>
<tr>
<td>6</td>
<td>54,577,331</td>
<td>136,247,472</td>
<td>6q</td>
<td>66</td>
<td>5.82 × 10⁻⁸</td>
</tr>
<tr>
<td>7</td>
<td>41,888</td>
<td>56,879,588</td>
<td>7p</td>
<td>98</td>
<td>1.59 × 10⁻¹²</td>
</tr>
<tr>
<td>7</td>
<td>77,575,701</td>
<td>81,218,337</td>
<td>7q</td>
<td>4</td>
<td>7.40 × 10⁻¹⁰</td>
</tr>
<tr>
<td>8</td>
<td>420,276</td>
<td>41,300,905</td>
<td>8p</td>
<td>93</td>
<td>1.29 × 10⁻⁸</td>
</tr>
<tr>
<td>8</td>
<td>49,341,541</td>
<td>146,298,338</td>
<td>8q</td>
<td>97</td>
<td>2.20 × 10⁻¹⁶</td>
</tr>
<tr>
<td>10</td>
<td>66,397</td>
<td>38,815,636</td>
<td>10p</td>
<td>99</td>
<td>2.99 × 10⁻⁸</td>
</tr>
<tr>
<td>11</td>
<td>192,155</td>
<td>51,581,408</td>
<td>11p</td>
<td>100</td>
<td>3.14 × 10⁻¹²</td>
</tr>
<tr>
<td>11</td>
<td>54,794,727</td>
<td>114,705,705</td>
<td>11q</td>
<td>75</td>
<td>2.20 × 10⁻¹⁶</td>
</tr>
<tr>
<td>11</td>
<td>119,488,464</td>
<td>134,944,160</td>
<td>12q</td>
<td>29</td>
<td>1.62 × 10⁻⁴</td>
</tr>
<tr>
<td>12</td>
<td>22,658,188</td>
<td>32,747,964</td>
<td>12p</td>
<td>12q</td>
<td>2.20 × 10⁻¹⁶</td>
</tr>
<tr>
<td>12</td>
<td>39,187,395</td>
<td>133,501,212</td>
<td>12q</td>
<td>98</td>
<td>2.20 × 10⁻¹⁶</td>
</tr>
<tr>
<td>13</td>
<td>23,617,885</td>
<td>102,306,088</td>
<td>12q</td>
<td>95</td>
<td>2.20 × 10⁻¹⁶</td>
</tr>
<tr>
<td>13</td>
<td>69,820</td>
<td>32,652,191</td>
<td>15q</td>
<td>92</td>
<td>6.56 × 10⁻⁹</td>
</tr>
<tr>
<td>14</td>
<td>46,534,541</td>
<td>90,163,275</td>
<td>16q</td>
<td>92</td>
<td>9.75 × 10⁻¹¹</td>
</tr>
<tr>
<td>15</td>
<td>25,268,060</td>
<td>80,982,386</td>
<td>17q</td>
<td>100</td>
<td>2.20 × 10⁻¹⁶</td>
</tr>
<tr>
<td>19</td>
<td>488,930</td>
<td>24,601,177</td>
<td>19p</td>
<td>98</td>
<td>2.62 × 10⁻⁸</td>
</tr>
<tr>
<td>19</td>
<td>27,829,851</td>
<td>59,096,950</td>
<td>19q</td>
<td>100</td>
<td>2.62 × 10⁻¹²</td>
</tr>
<tr>
<td>20</td>
<td>29,804,208</td>
<td>62,917,729</td>
<td>20q</td>
<td>99</td>
<td>2.40 × 10⁻⁹</td>
</tr>
</tbody>
</table>

Derived from linked reads using a recently published method (Supplemental Methods; Bell et al. 2017). Using this approach, a total of 31 autosomal mega-haplotypes were constructed (Table 2; Supplemental Data), 15 of which encompass entire (or >95%) chromosome arms such as 19p, 19q, 10p, 7p, and 5q (Fig. 3). The average mega-haplotype is 50.7 Mb or roughly four times longer than the longest phased haplotype block from Long Ranger (Fig. 2D; Tables 1, 2; Supplemental Data Set S2). The longest mega-haplotype is approximately 137 Mb long (4q). In this approach, smaller phase blocks (less than 100 SNVs) from Long Ranger are not included in the mega-haplotype assembly. Thus, these mega-haplotypes (referred to as such hereafter) do not directly supplant the Long Ranger phase blocks (Supplemental Data Set S2) in terms of detailed local phasing information.

Identification and reconstruction of structural variants (SVs) from linked reads

In addition to phasing, another use for the linked-read sequencing data is to identify breakpoints of large-scale SVs by searching for the discordant mapping of clusters 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 >30 kb (98% phased) (Supplemental Data Set S3) and 3541 deletions between 50 bp and 30 kb (79% phased) (Supplemental Data Set S4). The large SVs include deletions, inversions, duplications, and inter- and intrachromosomal rearrangements (Supplemental Data Set S3; Fig. 4A). As expected, we detected the BCR/ABL1 gene fusion, a hallmark of K562, as one of the SV calls with highest quality score (Fig. 4A; Supplemental Data Set S3), along with two other known gene fusions in K562 (Engreitz et al. 2012): XKR3/NUP214 between Chromosomes 9 and 22 (Fig. 4A), and CDC25A/GRID1 between Chromosomes 3 and 10 (Supplemental Data Set S5; Supplemental Data).

We also leveraged the long-range information derived from the linked reads to identify, assemble, and reconstruct SV-spanning breakpoints (including those of large-scale complex rearrangements) in the K562 genome using the recently established 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 (Marks et al. 2018). Sequence reconstruction is then performed by assembling the relevant linked reads around the identified breakpoints from which complex SVs are then automatically reconstructed. The breakpoints that have supporting evidence from the K562 3-kb mate-pair data set (Supplemental Methods) were determined as high-confidence events (Supplemental Data Set S5). GROC-SVs identified a total of 161 high-confidence breakpoints including 12 interchromosomal events (Figs. 1, 4B; Supplemental Data Set S5); each event is accompanied with visualization (Supplemental Data); and 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 (Supplemental Data Set S5). A notable example of assembly by GROC-SVs is a complex intrachromosomal rearrangement on Chromosome 13 (Fig. 4B).

Using gemtools as described (Greer et al. 2017), we identified phased structural rearrangements (multiple deletions and tandem
duplications) within the tumor suppressor gene *FHIT* on 3p14.2 (Fig. 4C; Waters et al. 2014). Because K562 exhibits LOH on Chromosome 3, the SVs within *FHIT* were phased using linked-read barcodes instead of heterozygous SNVs. The hemizygous deletion between 59.74 and 60.08 Mb of 3p14.2 results in the loss of *FHIT* exons 6, 7, and 8. For the two phased tandem duplications on the same allele, one is intronic, and the other duplicates exon 5 (Fig. 4C). The two deletions downstream from the phased duplications are on two different alleles of *FHIT*. Another allele-specific, complex, intrachromosomal rearrangement in K562 spans ~0.5 Mb on 16q11.2 and 16q12.1 (Fig. 4D), involving two overlapping inversions (62 and 125 kb) and a tandem duplication (163 kb). These events affect *ORC6*, *MYLK3*, *RHBDF1* (previously known as *C16orf8*), and *NETO2*, which has recently been identified as a cancer marker gene (Oparina et al. 2012; Hu et al. 2015). This rearrangement resides on the nonduplicated haplotype of this tripliod region. *ORC6* is located entirely within the more centromeric inversion of this locus on 16q11.2 and is “deleted” by the left breakpoint of the more telomeric inversion, which also “deletes” *RHBDF1* and inverts *MYLK3*, possibly disrupting its promoter region or proximal enhancers or disconnecting MYLK3 from their regulation (Fig. 4D).

Small-scale complex SVs from deep-coverage WGS

Small-scale complex SVs (Fig. 5A–E) as well as noncomplex SVs were identified using a novel algorithm called Automated Reconstruction of Complex Structural Variants (ARC-SV) (Arthur et al. 2018) from deep-coverage WGS data (Supplemental Data Set S6; Supplemental Data). These small-scale complex SVs are defined as genomic rearrangements with multiple breakpoints that cannot be explained by one well-defined (noncomplex) SV type such as deletions, insertions, tandem duplications, or inversions. After filtering out SVs <50 bp or with breakpoints that reside in simple repeats, low complexity regions, satellite repeats, or segmental duplications, we identified 122 complex SVs (accompanied with schematic visualizations), 2235 deletions, 320 tandem duplications, and 6 inversions (Supplemental Data Set S6). Examples of complex SVs include dispersed duplications in which duplicated sequences are inserted elsewhere in the genome in a nontandem fashion (Fig. 5A). These dispersed duplications sometimes involve inversions of the inserted sequence and deletions at the insertion site (Fig. 5B, C). Other examples include inversions flanked on one or both sides by deletions (Fig. 5D), duplications that involve multiple nonexact copies, as well as deletion, inversion, and multiple duplications residing at the same locus (Fig. 5E). Eight of 10 breakpoints from five complex SVs were successfully validated by PCR and Sanger sequencing (Supplemental Table S8).

SVs from mate-pair sequencing analysis

To increase the sensitivity of detecting medium-sized SVs (1–100 kb) in K562, we constructed a 3-kb mate-pair library and sequenced (2×151 bp) to 6.9× nonduplicate coverage. The sequence coverage (C<sub>n</sub>) of each 3-kb insert is 302 bp or 10%, which translates to a physical coverage (C<sub>p</sub>) of 68.5×. From the mate-pair library, SVs (deletions, inversions, and tandem duplications) were identified by clustering discordant read-pairs and split-reads using LUMPY (Layer et al. 2014). Only SVs that have both discordant read-pair and split-read support were retained. Overall, we identified 270 deletions, 35 inversions, and 124 tandem duplications using this approach (Supplemental Data Set S7). Approximately 83% of these SVs are between 1 and 10 kb, and 88% are between 1 and 100 kb (Supplemental Data Set S7). Twelve deletions and five tandem duplications were randomly selected for PCR and Sanger sequencing validation (Supplemental Table S8). The validation rates were 83% and 80%, respectively.

Noncomplex SVs from deep-coverage WGS

Noncomplex SVs (deletions, inversions, insertions, and tandem duplications) in K562 were called from deep-coverage WGS data using a combination of established methods, namely Findel (Ye et al. 2009), BreakDancer (Chen et al. 2009), and BreakSeq (Lam et al. 2010). These SVs were combined with those of the same SV type that were identified using ARC-SV, LUMPY, and Long Ranger, in which SVs (n = 2665) with support from multiple methods by ≥50% reciprocal overlap were merged. Through this combination of methods, a total of 9082 noncomplex SVs were identified in the K562 genome, including 5490 deletions, 531 duplications, 436 inversions, and 2602 insertions (Supplemental Data). (We note that only BreakDancer [Chen et al. 2009] was designed to call insertions.) Consistent with previous analyses (e.g., Lam et al.

---

Figure 3. Mega-haplotypes of entire K562 chromosome arms: (x-axis) chromosome coordinate (Mb); (y-axis) difference in unique linked-read barcode counts between major and minor haplotypes, normalized for SVN density. Haplotype blocks from normal control sample (NA12878) in blue and from K562 in dark gray. Density plots on the right reflect the distribution of the differences in haplotype-specific barcode counts for the control sample (blue) and K562 (dark gray). These density distributions are used for testing of significant difference (P < 0.001) using a one-sided t-test. Significant difference in haplotype-specific barcode counts indicates aneuploidy and haplotype imbalance. Haplotype blocks (with 100 or more phased SNVs) generated from Long Ranger (Supplemental Data Set S2) for the major and minor haplotypes were then "stitched" to mega-haplotypes encompassing the entire chromosome arm: (A) 5q (triploid); (B) 7p (tetraploid).
Figure 4. K562 SVs including large complex rearrangements resolved using linked-read sequencing. (A) Heat maps of overlapping barcodes for SVs in K562 resolved from linked-read sequencing using Long Ranger (Zheng et al. 2016; Marks et al. 2018). BCR/ABL1 translocation between Chromosomes 9 and 22. XKR3/NUP214 translocation between Chromosomes 9 and 22. Duplication within GPHN on Chromosome 14. Deletion that partially overlaps ZRANB1 and CTB2 on Chromosome 10. (B) Large complex rearrangement occurring on Chromosome 13 with informative reads from only one haplotype (region with loss of heterozygosity). Each line depicts a fragment inferred from linked reads based on clustering of identical barcodes (y-axis) using GROC-SVs (Spies et al. 2017). Abrupt endings (vertical dashed lines) of fragments indicate locations of breakpoints of this complex rearrangement. Fragments are phased locally with respect to surrounding SNVs (colored orange for same haplotype and black when no informative SNVs are found nearby). Gray lines indicate portions of fragments that do not support the current breakpoint. Fragments end abruptly at 81.47 Mb, indicating a breakpoint, picking up again at 81.10 Mb and continuing to 81.11 Mb where they end abruptly, then picking up again at 90.44 Mb. Coverage from 81.12 to 81.20 Mb are from reads with different sets of linked-read barcodes and thus are not part of this fragment set. (C, D) Complex rearrangements involving multiple haplotype-resolved SVs. Using gemtools (Greer et al. 2017), each SV is identified from linked reads grouped by identical barcodes (i.e., SV-specific barcodes, y-axis) indicative of single HMW DNA molecules (depicted by each row) that span the breakpoints. SVs are represented in different colors. The x-axis shows the hg19 genomic coordinate. Dotted lines represent individual breakpoints with schematic diagram of the rearranged structures drawn below the plot. (C) Multiple SVs within FHIT on 3p14.2. Deletion (DEL; red) (59.74–60.08 Mb) results in the loss of multiple exons. Two overlapping duplications (DUP; blue and green) in cis orientation (same allele of FHIT) indicated by the presence of HMW molecules spanning both DUPS. Two adjacent DELS (pink and purple) in trans (different alleles of FHIT), indicated by the absence of shared SV-specific barcodes for the HMW molecules spanning each DEL. SV haplotypes analyzed using SV-specific barcodes (not enough informative SNVs due to LOH). (D) Complex, intrachromosomal rearrangement spanning approximately 0.5 Mb on 16q11.2 and 16q12.1 that involve two overlapping inversions, 63 kb (red) and 125 kb (blue), and a 163-kb tandem duplication (green). This rearrangement resides on the nonduplicated haplotype of this triploid region. ORC6 is located entirely within the 63-kb inversion on 16q11.2 and is “deleted” by the left breakpoint of the 125-kb inversion, which also inverts MYLK3. C16orf8 on the same haplotype is also partially “deleted” by the 125-kb inversion (blue); NETO2 is duplicated by the 163-kb tandem duplication (green). (Inset) MYLK3 and ORC6 show allele-specific expression (Supplemental Table S12). MYLK3 is only expressed from this rearranged allele (Haplotype 2); ORC6 is expressed from the non-rearranged “diploid” allele (Haplotype 1).
PCR and Sanger sequencing. The validation rates were 88% and split-read support were randomly chosen for validation using and C on Chromosome 1 (81,660,347) and duplication of Block E between Blocks B and D on Chromosome 20 (135,111,978) between Blocks B and D on Chromosome 3 (158,795,874).

Figure 5. Small-scale complex SVs in K562 resolved using ARC-SV. Examples of small-scale complex SVs resolved using ARC-SV (Arthur et al. 2018) from the K562 WGS data set. (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) Duplication and inversion of Blocks B, C, and D between Blocks B and D on Chromosome 3 (158,795,874–158,799,955) overlapping IQC1-SCIPI1. (D) Inversion of Block C flanked by deletions of Blocks C and D on Chromosome 5 (147,554,038–147,554,778) inside SPINK7 (coding for a serine peptidase inhibitor). (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).

LINE-1 and Alu insertions

We identified nonreference LINE-1 and Alu retrotransposon insertions (REIs) in the K562 genome from our deep-coverage short-insert WGS data using a modified RetroSeq (Keane et al. 2013) approach (Supplemental Methods). Nonreference REIs were identified from paired-end reads that have one of the paired reads mapping to the human reference genome and the other read mapping to either the Alu or LINE-1 consensus sequence in a full or split-read fashion (Methods). We identified 1147 nonreference Alu insertions and 85 nonreference LINE-1 insertions in K562 (Supplemental Table S9; Fig. 1). Nine Alu and 10 LINE-1 insertions with split-read support were randomly chosen for validation using PCR and Sanger sequencing. The validation rates were 88% and 72%, respectively (Supplemental Table S8).

Allele-specific gene expression

Integrating CN information (i.e., allele frequencies) of the heterozygous SNVs (Supplemental Data Set S1), we reanalyzed two replicates of ENCODE poly(A)-mRNA RNA-seq data to identify allele-specific gene expression in K562. We identified 5053 and 5149 genes that show allele-specific expression (P < 0.05) in replicates one and two, respectively (Fig. 1; Supplemental Table S11). We also identified 2342 and 2176 genes that would have been falsely identified to have allele-specific expression and 1641 and 1710 genes that would not have been identified to have allele-specific expression in replicates one and two, respectively, if the allele frequencies of heterozygous SNVs in aneuploid regions were not taken into consideration (Supplemental Table S12).

Allele-specific DNA methylation

By integrating CN and phase information of heterozygous SNVs of K562, we identified 110 CpG islands (CGIs) that exhibit allele-specific DNA methylation (Fig. 1; Supplemental Table S13). We obtained K562 whole-genome bisulfite sequencing (WGBS) reads from ENCODE (Sloan et al. 2016) and aligned the reads to hg19 using Bismark (Krüger and Andrews 2011), in which 76.9% of reads were uniquely mapped and 26.2% of cytosines were methylated in a CpG context (Supplemental Methods). We then used reads that overlap both phased heterozygous SNVs (Supplemental Data Set S2) and CpGs to phase the methylated and unmethylated CpGs to their respective haplotypes. We then grouped the phased individual CpGs into CGIs. Fisher’s exact test (taking the CN of a given CGI locus into account) was used to evaluate allele-specific methylation (P < 0.05), and significant results were selected using a target false discovery rate of 10%. Of these 110 CGIs, 35 reside within promoter regions (here defined as 1 kb upstream of a gene), 83 are intragenic, and 28 lie within 1 kb downstream from 113 different genes. The following six genes are within 1 kb of a differentially methylated CGI and overlap with the Sanger Cancer Gene Census: ABL1, AXIN2, CCND1, HOXD11, KDR, and PRDM16.

Allele-specific CRISPR targets

We identified a total of 28,511 targets in the K562 genome suitable for allele-specific CRISPR targeting (Fig. 1; Supplemental Table S14). Sequences (including reverse complement) of phased variants that differ by more than 1 bp between the alleles were extracted to find all possible CRISPR targets by searching for the pattern [G, C, or A]N20GG. Using a selection method previously described and validated (Sunagawa et al. 2016), only conserved high-quality targets were retained (Supplemental Methods). Of the 28,511 allele-specific CRISPR target sites, 15,488 are within an annotated protein-coding or noncoding RNA transcript, 705 within an exon, and 13 targets are within an experimentally validated enhancer (Supplemental Table S14; Visel et al. 2007).

Genomic structural context provides insight into regulatory complexity

We show examples of how deeper insights into gene regulation and regulatory complexity can be obtained by integrating genomic structural contexts with functional genomics and epigenomics data (Fig. 6A–D). One example is the allele-specific RNA expression and allele-specific DNA methylation in K562 at the HOXB7 locus on Chromosome 17 (Fig. 6A). By incorporating the
genomic context of HOXB7 in K562, we see that HOXB7 exhibits highly preferential RNA expression from the two copies of Haplotype 1 (\(P=0.007\)) in which the CGI near its promoter is completely unmethylated (\(P=3.18 \times 10^{-18}\)) (Fig. 6A,C). The second example is allele-specific RNA expression and allele-specific DNA methylation of the HLX gene in K562 (Fig. 6B). The HLX locus on Chromosome 1 is tetraploid, and we see that HLX is only expressed from Haplotype 1, which has three copies, and is not expressed in Haplotype 2 (\(P=0.043\)) (Fig. 6B,D). The CGI of the HLX locus is unmethylated in Haplotype 2 but highly methylated on Haplotype 1 (\(P=5.14 \times 10^{-15}\)) (Fig. 6B,C). There is also an allele-specific CRISPR targeting site 797 bp inside the 5′ end of the HLX for both Haplotypes. (C) Number of methylated and unmethylated phased WGBS reads for Haplotypes 1 and 2 in CGI 22086 and CGI 2209 in which both CGIs exhibit allele-specific DNA methylation. (D) Number of RNA-seq reads for Haplotypes 1 and 2 of HLX and HOXB7 in which both genes exhibit allele-specific RNA expression.

Discussion

Despite its wide usage and impact on biomedical research, K562’s genomic sequence and structural features have never been comprehensively characterized, beyond its karyotype (Selden et al. 1983; Wu et al. 1995; Gribble et al. 2000; Naumann et al. 2001) and SNPs called from 30×-coverage WGS but without taking
aneuploidy or CN into consideration (Cavalli et al. 2016). Analysis, integration, and interpretation of the extensive collection of functional genomics and epigenomics data sets for K562 had so far relied solely on the human reference genome. Here, we present the first detailed and comprehensive characterization of the K562 genome. In summary, by performing deep-coverage short-insert WGS, 3-kb-insert mate-pair sequencing, deep-coverage linked reads sequencing, array CGH, karyotyping, and integrating a compendium of novel and established analysis methods (Supplemental Fig. S1A), we produced a comprehensive spectrum of genomic structural features (Fig. 1) for K562 that includes SNVs (Supplemental Data Set S1), indels (Supplemental Data Set S1), ploidy by chromosome segments at 10-kb resolution (Supplemental Data), phased haplotypes (Supplemental Data Set S2; Supplemental Data)—often of entire chromosome arms (Table 2; Supplemental Data)—phased CRISPR targets (Supplemental Table S14), nonreference REIs (Supplemental Table S9), and SVs (Supplemental Data) including deletions, duplications, and inversions, and complex SVs (Supplemental Data Sets S6, S7). Many SVs were also phased, assembled, and experimentally verified (Supplemental Data Sets S2–S5; Supplemental Tables S8, S10). Of the 3,784,863 variants that were haplotype-phased in the K562 genome (Supplemental Data Sets S2–S5), 3,088,185 (81.6%) are SNVs; 692,998 (18.31%) are indels; 3451 are deletion SVs (51 bp to 30 kb; 0.1%); and 229 are large SVs. We used the hg19 genome build for this study in order to keep the data sets consistent for analysis and integration since the data generation phase started before the GRCh38 genome release. Due to the genome-wide nature of the study, realigning the sequencing reads to GRCh38 will not significantly affect the results.

Pervasive aneuploidy is a characteristic of many cancers. Previous studies have confirmed the near-triploid karyotype of K562 (Selden et al. 1983; Wu et al. 1995; Gribble et al. 2000; Naumann et al. 2001). In our analysis, however, we also found considerable portions of the K562 genome to be much more varied than what had previously been reported. This is because by leveraging deep-coverage WGS, the CN across different chromosome segments (Fig. 1; Supplemental Table S2; Supplemental Data), as determined by our read-depth analysis, is of much higher resolution than karyotyping. Furthermore, the identified chromosome segments with aneuploidy (CN>2) and orthogonally supported by karyotyping and array CGH were further validated (Supplemental Fig. S3; Supplemental Data), also orthogonally, from a statistical approach in which significant differences in unique linked-read barcode counts between the major and minor haplotypes were determined using a one-sided t-test (P<0.001) (Fig. 3; Supplemental Data; Bell et al. 2017). In addition, it has to be taken into consideration that for a widely used cell line with decades of history such as K562, additional genome variation is expected (Supplemental Discussion).

Sensitive and accurate identification of SNVs and indels requires relatively deep WGS coverage (>33× and >60×, respectively) (Bentley et al. 2008; Fang et al. 2014). From our greater than 70× nonduplicate coverage WGS data, 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). From RNA-seq or ChIP-seq data analysis, a statistically significant increase in transcription or transcription factor binding signal in one allele compared to the other at a heterozygous locus, may be identified as a case of allele-specific expression or allele-specific transcription-factor binding, which usually suggests allele-specific gene regulation at this locus. However, if aneuploidy can be taken into consideration and the signals normalized by ploidy, the case identified might be a result of increased CN rather than the preferential activation of one allele over the other on the epigenomic level. Indeed, in our reanalysis of two replicates of ENCODE K562 RNA-seq data, we identified 2359 and 2643 genes that would have been falsely identified to have allele-specific expression in addition to 1808 and 2063 genes that would not have been identified to have allele-specific expression in replicates one and two, respectively, if ploidy was not taken into consideration (Supplemental Table S12).

It was previously shown that integrating orthogonal methods and signals improves SV-calling sensitivity and accuracy (Layer et al. 2014; Mohiyuddin et al. 2015). Here, we combined deep-coverage short-insert WGS, mate-pair sequencing, linked-read sequencing, and several SV-calling methods to identify many noncomplex SVs. To obtain the union set of noncomplex SV calls from the various methods, the SVs identified by multiple methods were merged and indicated accordingly (Supplemental Data). For deletions (Supplemental Fig. S5A), we see strong overlap for the various methods, but this overlap is less pronounced for duplications (Supplemental Fig. S5B) and inversions (Supplemental Fig. S5C). This is consistent with previous analysis (Lam et al. 2012) as inversions and duplications are more difficult in principle to accurately resolve (Lin et al. 2015; Sudmant et al. 2015). We also expect the detection of many SVs to be method-specific, since each method is designed to utilize different types of signals and also optimized to identify different classes of SVs (Pabinger et al. 2014; Lin et al. 2015). Again, if particular SVs are of interest for follow-up studies, they should first be experimentally validated.

The complex rearrangements identified by using ARC-SV from short-insert WGS (Fig. 5A–E; Supplemental Data Set S6; Supplemental Data) and by using GROC-SVs from linked reads (Fig. 4B; Supplemental Data Set S5; Supplemental Data) are classes of SVs that could not be easily identified and automatically reconstructed using previously existing methods. The small-scale complex SVs that were identified by using ARC-SV and experimentally validated (Fig. 5; Supplemental Table S8; Supplemental Data) describe a subtle class of complex rearrangements in cancer genomes that have been relatively understudied (Perry et al. 2008; Quinlan and Hall 2012; Collins et al. 2017). Detecting and automatically reconstructing these small-scale complex SVs, especially in a “hay” of canonical SVs and in highly rearranged cancer genomes, has remained an unsolved problem for many years. In other words, our results reveal a class of previously overlooked complex SVs in cancer that can now be identified from standard short-insert WGS data and elucidated further. They have clear implications for the conventional models of cancer evolution which often assume gradual, step-by-step mutations; however, these complex SVs support a form of punctuated genome evolution (Davis et al. 2017). A major unsolved question still is how complex SVs arise mechanistically for which there are general models: template switching during replication (Lee et al. 2007; Hastings et al. 2009) and chromothripsis (Stephens et al. 2011). Furthermore, the functional consequences of these small-scale complex SVs are also unknown. These important questions remain unsolved mainly due to the lack of data and examples. It is possible that this mutational complexity contributes to genome innovation, at least in cancer, or is just a curious sideshow (Quinlan and Hall 2012; Collins et al. 2017).
2012). Only the accumulation of such examples and data will allow researchers in fields such as cancer evolution to begin to address these important questions.

Before the existence of linked-read sequencing, haplotype phasing and resolving large SVs (>30 kb) relied heavily on fosmid libraries (Kitzman et al. 2011; Williams et al. 2012; Adey et al. 2013; Cao et al. 2015; Snyder et al. 2015), which were laborious, costly, time consuming, and much less efficient. Using linked-read sequencing and gemtools (Greer et al. 2017), we phased and resolved complex SVs that are especially compelling on 3p14.2 (within the tumor suppressor gene *FHIT*) and on 16q11.2 and 16q12.1 of the K562 genome (Fig. 4C,D; Supplemental Discussion).

Data generated from this comprehensive whole-genome analysis of K562 is available through the ENCODE portal (Supplemental Fig. S1B,C; Sloan et al. 2016). We envision that this analysis will serve as a valuable resource for further understanding the vast troves of ENCODE data available for K562, such as determining whether a potential or known regulatory sequence element has been altered by SNVs or SNPs, indels, retrotransposon insertions, a gain or loss of copies of that given element, or allele-specific regulation and between polymerase binding and ploidy (Supplemental Discussion). In addition, we also observed that the K562 POLR2A ChIP-seq signals in both replicates are very well correlated with poly(A) RNA-seq signal and with WGS coverage, suggesting an association between polymerase binding and active transcription and between polymerase binding and ploidy (Supplemental Figs. S6, S7).

Our work here serves to guide future studies that utilize the K562 “workhorse” cell line, such as CRISPR screens where knowledge of the sequence variants can extend or modify the number of editing targets (Supplemental Table S13) and knowledge of aberrant CN will allow for much more confident data interpretation. To give an example, in a recent study that uses CRISPRi to screen and elucidate the function of long noncoding RNAs in human cells, 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 are very likely explained by cell-specific effects, it is still quite possible that many of the gRNA hit differences were the result of differences in genome sequence or ploidy. Our list of allele-specific CRISPR targets (Supplemental Table S13) will allow for the discernment between these two potential reasons for differences in CRISPR effects and should be particularly valuable for future large-scale CRISPR screens that utilize K562. Lastly, this study also serves as a technical example for the advanced, integrated, and comprehensive analyses of other heavily utilized cell lines and genomes in biomedical research such as HepG2.

**Methods**

**Genomic DNA extraction and karyotyping**

K562 cells were obtained from the Stanford ENCODE Product Center (NHGRI Project 1U54HG006996). Genomic DNA was extracted using the QIAGEN DNeasy Blood and Tissue Kit (Catalog No. 69504) and quantified using the Qubit dsDNA HS Assay Kit (Invitrogen). DNA was then verified to be pure (OD260/280 > 1.8; OD260/230 > 1.5) using NanoDrop (Thermo Fisher Scientific) and high molecular weight (mean >30 kb) using field-inversion gel electrophoresis on the Pippin Pulse System (Sage Science). K562 cells were sent to the Cytogenetics Laboratory at Stanford University (http://cytogenetics.stanford.edu) for karyotyping where 20 metaphase cells were analyzed using GTW banding.

**Data generation and analysis**

Illumina short-insert WGS, mate-pair WGS, 10x Genomics linked-read WGS, and array CGH were all performed using standard experimental techniques. Genomic sequence and structural features of the K562 genome were plotted using Circos (Krzywinski et al. 2009). For full descriptions of experimental and computational procedures (including analysis code), see Supplemental Methods.

**Experimental validation**

Random sets of SV calls from short-insert WGS and mate-pair sequencing were selected from PCR validation. These sets include complex SVs (from ARC-SV), deletions (>1 kb), and tandem duplications. PCR primers were designed such that the amplicons span the breakpoints and produce products between 200 and 500 bp. In the case of complex SVs, pairs of PCR primers were designed to validate multiple breakpoints. Ten Ali and 10 LINE-1 events with split-read support were randomly chosen for validation. PCR primers were designed to amplify products ranging from 65 to 150 bp in which one primer anneals to unique sequence and the other anneals to the retrotransposon sequence. All PCR amplicons were gel purified and Sanger sequenced.

**Data access**

All resources generated for K562 in this study are listed with detailed descriptions in Supplemental Figure S1. All data (raw sequences, processed data, and Supplemental Data Sets S1–S7) generated in this study have been submitted to ENCODE (https://www.encodeproject.org) under accession number ENCSR0806UYV. Analysis code is available as Supplemental Material. Accession numbers for individual data files such as Supplemental Data Sets S1–S7 and experiments are listed in Supplemental Figure S1B,C.

**Acknowledgments**

We thank Ms. Aditi Narayanan, Dr. Idan Gabdank, Mr. Nathaniel Watson, Dr. Carrie Davis, Ms. Kathirina Onate, Mr. Zachary Myers, Ms. Khine Z. Lin, and Dr. Cricket Sloan for assistance with data release on the ENCODE portal. We thank Dr. Athena Cherry and the Stanford Cytogenetics Laboratory for karyotype analysis and Ms. Arineh Khechaduri for performing genomic DNA preparation. We thank Dr. Minyi Shi for providing K562 cells. A.E.U. is a Tashia and John Morgridge Faculty Scholar of the Stanford Child Health Research Institute. A.E.U. was supported by National Institutes of Health (NIH), National Human Genome Research Institute (NHGRI) grant 5P0-HG007735 and the Stanford Medicine Faculty Innovation Program. B.Z. was additionally supported by NIH training grant T32-HL110952. W.H.W. received support from NIH/NHGRI grants R01-HG007734 and P05-HG007735. J.G.A. received funding from NIH training grant T32-GM096982 and National Science Foundation Graduate Fellowship DGE-114747. A.A. was funded by NIH grant U24CA220242. H.P.J. was funded by the Intermountain Healthcare Research Award and NIH/NHGRI grants R01-HG006137 and P01-HG00205.


References


Received January 22, 2018; accepted in revised form December 28, 2018.


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

Bo Zhou, Steve S. Ho, Stephanie U. Greer, et al.

*Genome Res.* 2019 29: 472-484 originally published online February 8, 2019
Access the most recent version at doi:10.1101/gr.234948.118

**Supplemental Material**
http://genome.cshlp.org/content/suppl/2019/02/08/gr.234948.118.DC1

**References**
This article cites 57 articles, 9 of which can be accessed free at:
http://genome.cshlp.org/content/29/3/472.full.html#ref-list-1

**Creative Commons License**
This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

**Email Alerting Service**
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](https://genome.cshlp.org/subscriptions).