Published online 16 August 2017

Nucleic Acids Research, 2017, Vol. 45, No. 19 ***e162***
doi: 10.1093/nar/gkx712

Chromosome-scale mega-haplotypes enable digital karyotyping of cancer aneuploidy

John M. Bell[1], Billy T. Lau[1], Stephanie U. Greer[2], Christina Wood-Bouwens[2], Li C. Xia[2], Ian D. Connolly[3], Melanie H. Gephart[3] and Hanlee P. Ji[1,2,*]

[1]Stanford Genome Technology Center, Stanford University, Palo Alto, CA 94304, USA, [2]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA and [3]Department of Neurosurgery, Stanford University Hospital and Clinics, Stanford, CA 94305, USA

Received April 28, 2017; Revised July 20, 2017; Editorial Decision August 02, 2017; Accepted August 05, 2017

ABSTRACT

Genomic instability is a frequently occurring feature of cancer that involves large-scale structural alterations. These somatic changes in chromosome structure include duplication of entire chromosome arms and aneuploidy where chromosomes are duplicated beyond normal diploid content. However, the accurate determination of aneuploidy events in cancer genomes is a challenge. Recent advances in sequencing technology allow the characterization of haplotypes that extend megabases along the human genome using high molecular weight (HMW) DNA. For this study, we employed a library preparation method in which sequence reads have barcodes linked to single HMW DNA molecules. Barcode-linked reads are used to generate extended haplotypes on the order of megabases. We developed a method that leverages haplotypes to identify chromosomal segmental alterations in cancer and uses this information to join haplotypes together, thus extending the range of phased variants. With this approach, we identified mega-haplotypes that encompass entire chromosome arms. We characterized the chromosomal arm changes and aneuploidy events in a manner that offers similar information as a traditional karyotype but with the benefit of DNA sequence resolution. We applied this approach to characterize aneuploidy and chromosomal alterations from a series of primary colorectal cancers.

INTRODUCTION

Genomic instability is a common feature of cancer. In some tumors, structural alterations such as duplications and deletions alter genomic segments up to the size of entire chromosomal arms. Resolving these somatic chromosomal events can delineate critical drivers with both prognostic and therapeutic predictive information. For example, the routine evaluation of acute myelogenous leukemias involves karyotyping to quantitatively determine the number and type of aneuploidy events (1) and to stratify patients by recurrence risk (2,3). Many other diseases and congenital disorders demonstrate large chromosomal structural and aneuploidy changes; thus, large chromosome structural events are frequent and important genomic features of many human disorders.

Current methods for determining rearrangements and aneuploidy include standard karyotyping, spectral karyotyping (SKY), genomic microarrays and more recently, whole genome sequencing (WGS) (4–9). Standard karyotyping methods use a variety of dyes to stain banding patterns on individual, intact chromosomes while SKY uses molecular cytogenetic techniques to visualize individual chromosomes with different fluorescent colors. As a result, these methods provide images of chromosome-scale events. In general, however, karyotyping is a slow, arduous methodology because it requires cell-by-cell extraction of chromosomes, interpretive expertise and generates low-resolution determination of chromosome changes. Microarray analysis offers more granular information than karyotyping via genotyping single nucleotide polymorphism (SNPs) (10–12). However, all microarrays have a limited number of features that restrict the resolution of chromosomal events that can be identified as well as inability to detect some somatic structural changes, such as rearrangements. Conventional WGS with short DNA insert libraries (<1.0 kb) and short sequence reads (<300 bases) provides copy number variation (CNV) information that can be used to identify large chromosomal aberrations, but even with WGS, accurate determination of aneuploidy events from cancer genomes is a challenge. CNV analysis does not generally discriminate between changes in the maternal and paternal chromosomes. Even with allele-specific copy number analysis, it is difficult to determine which parental chromosome underwent a change, given the limited haplotype information available from WGS.

*To whom correspondence should be addressed. Tel: +1 650 721 1503; Fax: +1 650 725 1420; Email: genomics_ji@stanford.edu

Recent advances in genome sequencing technology enable the characterization of high molecular weight (HMW) DNA molecules and identification of blocks of *cis*-related heterozygous variants, also referred to as haplotypes, which are located on either the paternal or maternal chromosome. The process of identifying these blocks is referred to as phasing. To generate very large haplotypes, a number of novel sequencing technologies either generate kilobase-length sequence reads or trace barcode sequences back to individual HMW molecules (13–19). This enables the phasing of contiguous single nucleotide variants (SNVs) that originate from the same HMW DNA molecule. When extended to large segments, this phased variant information can be used to derive megabase (Mb)-scale haplotypes. For example, Adey *et al*. observed that it is possible to use the distribution of allele frequencies across haplotypes to generate extended scaffolds of phase blocks that are larger than the individual blocks (20). In some cases, one can deconvolute apparently complex alterations via haplotypes when rearrangements, such as overlapping deletion and amplification segments, occur on separate haplotypes.

In this study, we describe a method that uses barcode-linked sequence reads from primary cancer genomes to generate phase blocks that can encompass entire chromosome arms (Figure 1A). We described previously a new technology that uses 1 ng of input DNA separated into as many as a million or more partitions with non-overlapping barcodes (15). Individual HMW DNA molecules are associated with barcodes. After sequencing with an Illumina system, the sequencing read-barcode associations allow individual HMW molecules to be partially reconstructed. The overlapping reconstructed sequences in turn allow the generation of Mb-scale haplotypes, based on the variants identified in the HMW DNA. This method has advantages over other long read sequencing approaches (i.e. Pacific Biosciences, Oxford Nanopore) because short read sequencing has fewer errors, the DNA input is minimal (1 ng) and it allows sequencing to 30× coverage at lower cost and with more accurate variant detection.

By utilizing this new technology, we developed an approach that integrates initial haplotype blocks into even larger haplotypes that cover entire chromosome arms (Figure 1B). As we demonstrate, our method includes more than 90% of the variants per arm. We use this haplotype information to generate sequencing-based digital karyotypes that cover Mbs of the genome. When applied to cancer samples, this method provides information similar to standard karyotyping or fluorescence *in situ* hybridization but also has the advantage of resolving high resolution haplotype information with specific phased variants. Moreover, our approach uses experimental sequencing data and, as a result, does not require population-derived haplotypes (3,21). We applied this approach to a series of primary colorectal cancers and generated extremely large haplotypes, some covering entire chromosomes.

MATERIALS AND METHODS

Genomic DNA samples and preparation

The Institutional Review Board at Stanford University School of Medicine approved the study and informed consent was obtained. Samples were obtained from the Stanford Cancer Center Tissue Bank. The tissue samples were collected at the time of surgical resection and flash frozen. We analyzed three sets (Patients 232, 1465 and 1532) of three matched tissue samples that included a primary colorectal adenocarcinoma, a dysplastic lesion that did not demonstrate fully invasive carcinoma and matched normal tissue. Also, we analyzed a metastatic colorectal carcinoma (Patient 5378) resected from the brain as well as matched normal brain tissue from one individual. Genomic DNA was extracted with the E.Z.N.A. SQ DNA/RNA Protein Kit (Omega Bio-Tek). DNA was quantified with Life Technologies Qubit. The colorectal cancer cell line LS411N was obtained from ATCC (ATCC no. CRL-2159). To karyotype the cell line, we used the University of Cincinnati Children's Hospital karyotyping service. The fetal trisomy cell line NA03576 was obtained from Coriell (Coriell ref. no NA03576).

Phased genome sequencing with barcode reads

A 10× Genomics chromium instrument (Pleasanton, CA) was used for sample preparation according to the manufacturer's protocol (15). Genomic DNA was processed and barcoded libraries were prepared without size selection. Barcoded sequencing libraries were generated with either the GemCode or Chromium kits per the manufacturer's protocol (Table 1). Phased genome sequencing was conducted with an Illumina HiSeq 2500. Reads were either 2 × 98 or 2 × 93. For paired-end data processing, read 1 was truncated by 10 bases in order to remove primer extension artifacts. All samples also had an 8-base sample index and 14-base (GemCode) or 16-base (Chromium) barcode sequence, following 10× Genomics standard assay preparation. After basecalling and demultiplexing, Long Ranger version 1.2 (10× Genomics) was used to process the barcode sequence data (Step 1 of Supplementary Figure S1). Long Ranger was run using the NCBI build 37 human genome as the reference and with the Supplementary Variant Call Format (VCF) setting (–vc_mode = precalled). Supplementary VCFs were generated from conventional WGS, as described below.

Conventional whole genome sequencing

Conventional WGS was performed, which provided variants for the supplementary phasing mode in the Long Ranger software and allowed validation of our haplotype results. We define conventional WGS as sequencing where the library generation involves fragmenting DNA into low molecular weight species, as is generally required for Illumina sequencing. Libraries were generated with an Illumina TruSeq kit and run on a HiSeq 2500 with 2 × 100 cycle reads that included indexing. All samples were aligned against the NCBI build 37 human reference genome using bwa version 0.7.5a-r405 (22) with commands aln and sampe. After alignment, duplicates were marked using Picard MarkDuplicates and the Genome Analysis Toolkit (GATK) variant calling pipeline steps (23) were followed using standard best practices, including IndelRealigner and BaseRecalibrator. Samples were merged at the indel realignment step. After the

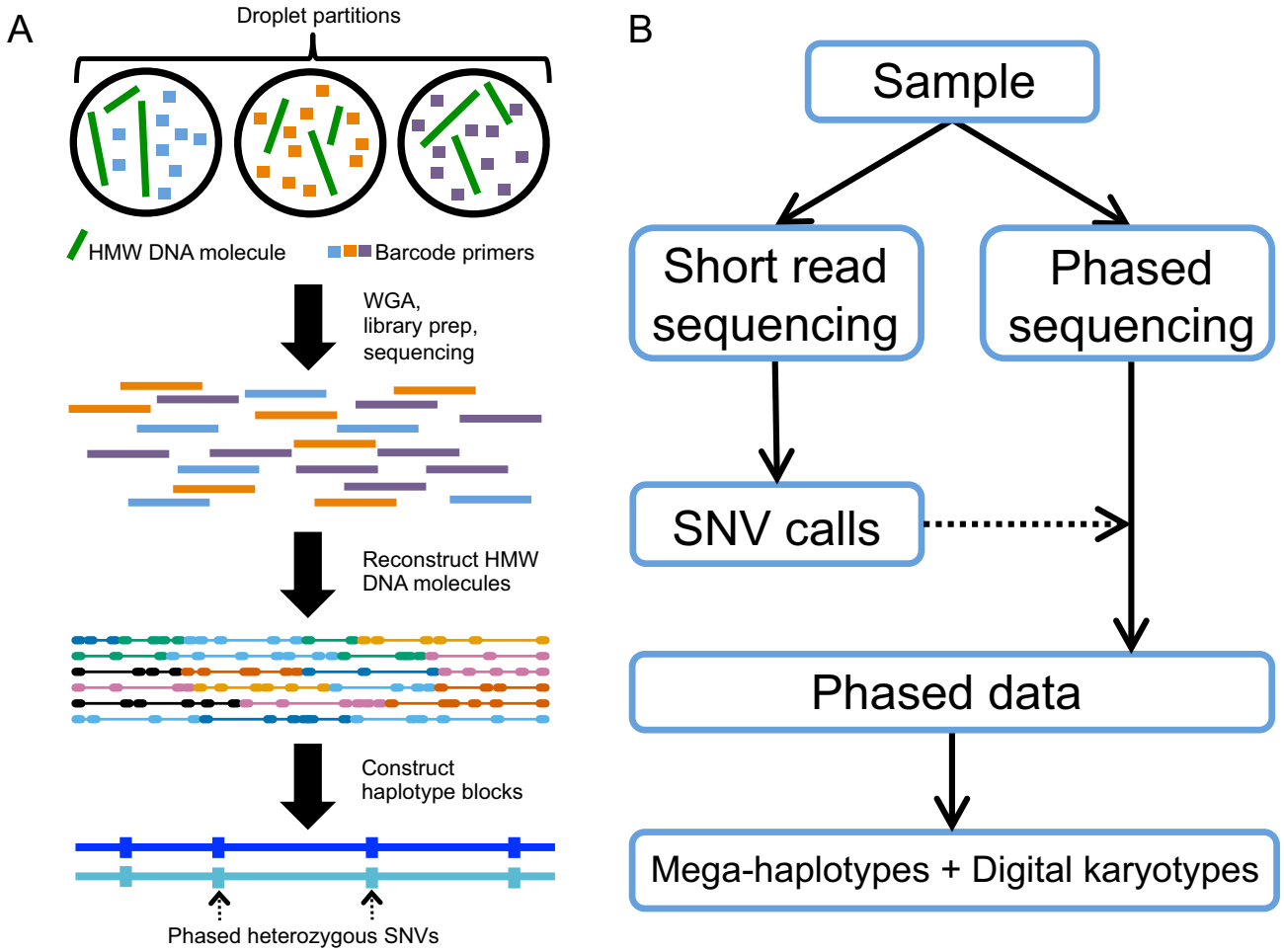

Figure 1. Overview of linked read sequencing and mega-haplotype analysis. (**A**) HMW DNA molecules are partitioned into droplets; within the droplets the molecules are associated with unique barcodes and amplified. Then, the partitions are dissolved, sequencing libraries are prepared and Illumina sequencing is performed as usual. After base-calling, the HMW molecules can be reconstructed via the barcodes and then scaffolded together to produce phase blocks. (**B**) Samples were first sequenced and aligned using conventional short read paired-end sequencing and variant calling was performed with GATK to generate a list of high-quality single nucleotide variant (SNV) calls. The same samples then underwent library preparation with GemCode or Chromium Genome kits (Table 1) and were sequenced. The Long Ranger software was run, using the previously generated SNV calls as reference variants. The phased SNV calls generated by Long Ranger served as input for the mega-haplotyping method.

Table 1. Summary of mega-haplotype results

Colorectal cancer		Independent conventional WGS	Phase block size N50 (Mb)	# Imbalanced arms[a]	Mean size of mega-haplotype (Mb)	Mean percentage of arm in mega-haplotype	Total Phased Hetozygotes in mega-haplotypes
ID	sample						
P00232	primary tumor	Yes	1.089	15	60.5	96.3%	501 058
	dyplasia	Yes	1.158	11	56.1	96.2%	323 094
P01465	primary tumor	Yes	0.575	1	48.2	53.2%	26 066
	dyplasia	Yes	1.099	1	48.2	53.2%	26 066
P01532	primary tumor	Yes	1.019	9	57.4	84.7%	265 579
P05378	metastasis	No	0.679	37	65.8	94.7%	1 264 976
LS411N	colon cancer cell line	No	0.855	21	65.8	82.9%	147 764
NA03576	trisomy cell line	No	0.990	3	83.6	77.8%	34 339

UnifiedGenotyper was run, VariantRecalibration and ApplyRecalibration were run to produce filtering information. Only SNVs passing these filters and having a total allelic depth of at least ten were used to generate a VCF file for the Long Ranger Supplementary phasing mode processing described above.

The GATK DepthOfCoverage (23) tool was used to determine sequencing coverage depth. To ensure that sequencing metrics were as congruent as possible, we removed contig-related information and only used depth information from autosomes and sex chromosomes, including chromosome Y if the patient was male (Individual 232, Individual 1465, cell lines LS411N and NA03576). Thus, the total number of positions used to calculate sequencing depth were the same for conventional WGS and the linked read WGS for each patient, but varied depending on the sex of the patient. Normal tissue and tumor/dysplastic samples from conventional WGS were compared to identify somatic cancer-specific variants. Only somatic variants passing quality filter and with an overall allelic depth >= 10 were included.

Determination and barcode counts of shared phase blocks between samples

Our analysis relies on the initial haplotype blocks provided by the Long Ranger software. These results are then used to 'stitch' together haplotypes that extend across chromosome arms that have undergone allelic imbalances. To process the phased variant data, we wrote a series of Perl, Python and R scripts that are available as part of the Supplementary Data and Notes. A stepwise list for the script command lines are provided in Supplementary Note 1. For this processing, the Long Ranger software provides a VCF file of phased variant calls for each sample. These Long Ranger VCFs were used as the initial input for the creation of phase blocks. First, the phased VCF file for each sample was filtered to include only variant calls passing quality filter and located on autosomes (and the X chromosome if the patient was female). Next, the files were numerically sorted. The phased variant information was simplified to include the genotype (GT), the phase sequence id number unique to each phase block (PS) and the list of barcodes that support the assignment of each allele to the phase block (BX)**.** The VCF processing steps are summarized in Step 2 of Supplementary Figure S1. Next, the filtered, re-sorted and simplified VCF files for each sample were combined to generate one VCF file per patient, including only positions in which a variant call was identified in all samples (Step 3 of Supplementary Figure S1).

The combined VCF file was then parsed to a simpler format to enable extraction of phase blocks as described in Step 4 of Supplementary Figure S1. The beginning and end coordinates of the phase blocks were determined according to the phase sequence id numbers denoted by the code 'PS' in the VCF file. Then, phase blocks were compared among the matched samples of a patient (normal, dysplastic and tumor). We denoted a common phase block among the samples if they all had a phase block that shared at least two phased heterozygotes. For example, if one sample had phased heterozygote calls at positions 15, 20 and 25 that shared a common phase sequence id (PS), while the other sample had calls at 20, 25 and 30 with a common phase sequence id, then the common phase block between the two samples in this region would be from positions 20–25.

Within each common phase block for a given sample, the barcodes supporting each phased heterozygote SNV were counted. Multiple occurrences of the same barcode within the phase block were ignored (i.e. each barcode was counted once), in order to minimize library amplification effects and eliminate duplicates. Thus, the barcode counts are close estimates of the number of individual DNA molecules sequenced in a given region (15).

Analysis of the distribution of phased SNVs showed that 90–95% of all phased heterozygotes in the intersected file were included in the analysis when only phase blocks containing 100 or more phased heterozygotes were included, indicating that a small number of very large blocks can contain a large proportion of the total number of heterozygotes (Supplementary Table S1). Our results showed that phase blocks containing at least 100 heterozygotes had the smallest overall proportion of ambiguous blocks. We defined an ambiguous block as >10% and <90% of the phased heterozygotes in a block matching in GTs between the samples being compared. Ambiguity is related to switch errors but is a better indicator of the overall consistency of the phase block. The relative rareness of large ambiguous blocks implies that large blocks were of higher quality than blocks with fewer phased heterozygotes (Supplementary Table S1). For these reasons, we used only phase blocks of 100 or more phased heterozygotes in the haplotyping analysis.

General description of determining meta-blocks and mega-haplotypes

To increase the statistical power for detecting significant differences between haplotypes, we developed a normalization procedure to account for variation in SNV frequency across chromosomal regions (Figure 3A and Step 5 of Supplementary Figure S1). Our procedure generated joined phased variant segments which we refer to as 'meta-blocks'. Haplotypes are based on the meta-blocks representing phased variants. For this procedure, these meta-blocks can cross an entire chromosome arm, hence the term 'mega-haplotype'.

For a given sample and a specific phase block, we performed the following steps: (i) the number of unique barcodes in each haplotype was divided by the number of phased heterozygotes; (ii) the total number of unique barcodes in both haplotypes was summed across all phase blocks; (iii) the sum of the barcodes from the normal sample (Step b) was divided by the sum of the barcodes from the tumor or dysplastic sample (Step b); (iv) the vector of normalized ratios from (Step a) were multiplied by the factor in (Step c), generating normalized results as illustrated in Figure 3B.

Based on the normalized data for a given phase block, the haplotype with more barcodes was denoted as major and the one with less barcodes was denoted as minor. Then, the difference between the normalized ratios of major and minor haplotypes of each block, *D,* was determined, as shown in Figure 3C and Step 6 of Supplementary Figure S1. For a given sample, the set of *D* differences across a

certain region—specifically, a chromosome arm—was assumed to follow a normal distribution. Imbalanced regions could then be detected by comparing the distribution of each tumor or dysplastic sample's D values to the distribution of the normal sample. We used a one-sided t-test to compare the density distributions for each chromosome arm of the two samples (matched normal tissue versus dysplastic or tumor) (Figure 3D and Step 7 in Supplementary Figure S1). Tests were performed on the distribution of D values across entire chromosome arms. We rejected the null hypothesis (that the samples came from the same distribution, implying no significant allelic imbalance was detected) for those cases with a P-value < 0.001. Our P-value included a Bonferroni multiple test correction (from 0.05) for the number of chromosome arms analyzed per set of matched samples.

On the chromosome arms that passed the t-test, we sought to group together all the phase blocks displaying large variation between case and control data. For each chromosome arm, we determined the value $v > 97.5\%$ of the D values from the normal sample (D_n) as described in Step 7 in Supplementary Figure S1. This result defined the extent of variation in D_n from which non-normal (dysplastic or tumor) samples could be compared. For any chromosome arm that passed the t-test, the D value of each non-normal phase block (dysplastic: D_d or tumor: D_t) was tested against v and an indicator function marked whether it was greater or lesser. Finally, all contiguous phase blocks marked as $> v$ were combined into a single meta-block (Figure 3E and Step 8 in Supplementary Figure S1).

SNVs in imbalanced regions

Phased SNVs in a joined meta-block of a total size >1 Mb were defined as a mega-haplotype. For each chromosome arm that we analyzed, the total number of phased heterozygotes found in meta-blocks were counted using the phased heterozygotes listed in the combined intersected VCF file. Only the SNVs found in the contiguous blocks described above and with GTs 0|1 and 1|0 were included in the SNV totals. In addition to counts of SNVs within meta-blocks, haplotypes were generated for the meta-blocks representing phased variants across entire chromosome arms, hence the term 'mega-haplotype'.

We identified the mega-haplotype phase of somatic variants by integrating somatic variants detected from conventional WGS with mega-haplotype information as determined above. Somatic variants derived from conventional WGS were intersected with the coordinates of the mega-haplotyped variants, then the number of phased major and minor haplotype somatic calls were counted for each meta-block.

Allelic imbalance analysis of a colorectal cell line and fetal trisomy cell line

We analyzed a set of samples with demonstrated karyotype abnormalities as a test of our method. For the analysis of the colorectal cell line LS411N and the fetal trisomy cell line NA03576, we used for a normal control the results of phased sequencing of NA12877, which had been sequenced previously (15). These samples have no shared ancestors to the best of our knowledge, but nevertheless shared enough variants to allow allelic imbalance analysis to be performed.

Copy number and allelic imbalance analysis from conventional whole genome sequencing

We determined copy number alterations from conventional WGS to identify imbalanced or otherwise aberrant regions. For CNV analysis, we used the program BICseq with the λ parameter set at 30 to provide smoothing of genomic regions which demonstrated a significant change (24,25). To generate larger contiguous segments of allelic imbalance, we joined adjacent segments if they were <1 Mb apart. In addition, adjacent segments were combined if one of the following conditions was met: (i) the difference in the $\log_2$ ratio for a given segment was <0.25 and demonstrated a copy number change; (ii) both segments were either amplified and deleted and the weighted size (size * absolute value of $\log_2$ ratio over normal) of one was 10× the other or greater; or (iii) the weighted size of one was 20× greater than the other, as long as neither sequence was 5 Mb or longer and neither sequence was within 0.1 of a $\log_2$ ratio of 0. These conditions eliminated segments that did not significantly influence the copy number of the chromosomes, while avoiding regions that demonstrated diploid copy number.

After the segments were joined, the $\log_2$ ratios were converted to copy number values. A CNV was not called if the copy number was between 1.85 and 2.15. We did not consider copy number alteration segments that were <7.5 Mb. We chose the genomic segment size of 7.5 Mb because it is half the length of the shortest chromosome arm (18p) and our analysis of mega-haplotypes focused exclusively on large chromosomal alterations.

Deriving haplotypes from allelic imbalances using conventional WGS

We conducted a comparison between haplotyping based on phased sequencing (i.e. barcoded reads) versus haplotyping based on conventional WGS. Conventional WGS required us to use somatic allelic imbalances in chromosomal segments to identify variant allelic fractions that were offset from what one would expect in a normal diploid genome. To determine the accuracy of haplotypes generated using conventional WGS with a variant allele frequency (VAF) haplotyping method, we used variant calls that had a minimum depth of 10 and positions where both the normal and tumor samples were heterozygote. Variant call files containing dysplastic versus normal and tumor versus normal were filtered separately to minimize the effects of filtering. We determined haplotypes as follows: the lower depth allele was assigned to the minor haplotype and the higher depth allele was assigned to the major haplotype.

The VCF files were then merged and intersected phase blocks were determined according to the requirement that they contain at least 100 phased heterozygote calls, in order to match the blocks generated by our linked read method. The results were filtered to only include imbalanced chromosome arms.

Comparison of mega-haplotype analysis between phased versus conventional WGS

To compare the accuracy of the haplotyping method from phased versus conventional sequencing, we assessed the type of switch errors found in large blocks by comparing the haplotypes for a given sample using both methods. For describing switch errors, we refer to the term 'matching' for the case when haplotypes from a matched sample pair (i.e. tumor versus normal) have a phased GT such as '0|1' at the same position and the term 'opposing' or 'non-matching' when one sample has, e.g. a '0|1' GT and the other has a '1|0' GT. The haplotypes of samples from the same patient should be the same in general: two phase blocks over the same area should have all matching or opposing. There should not be a mixture of the two, seeing that a mixture indicates switch errors. The number of GT switches relative to the normal sample was counted for each phase block that had been called as imbalanced in the phased samples and in the VAF-haplotyped (conventional) samples.

Determination of switch errors in phase blocks

When phase blocks were intersected between sample types for the same patient, the overall extent of congruency between the haplotypes of each phase block was calculated. Twenty-five cases of blocks with >100 phased heterozygotes and a congruency of <90% between normal and malignant samples in Patient 1532 were inspected visually. The cases were classified into those with large gaps in SNVs proximal to the switch (8/25), those with loss of heterozygosity proximal to the switch (12/25), those with both (2/25) and those with different phased GTs appearing throughout the phase block (3/25). The three latter cases were further analyzed in terms of the proportions of each type of GT pairing. The types were defined as unphased (either of the two GTs was not phased), homozygote (either of the two GTs was homozygote), matching (the GTs were the same and were phased), opposite (the GTs were not the same and were phased). The proportions for each of these types were calculated.

RESULTS

Haplotype-based characterization of chromosomal alterations

Recently, we developed and applied a droplet technology for whole genome and exome sequencing preparation that makes use of barcodes to identify partitions and reconstruct large molecules of DNA (Figure 1A) (15). The amount of input DNA is 1 ng, representing ∼300 genome equivalents. Also, the input DNA is not sheared and remains intact with molecular sizes upward of 20 kb (Supplementary Table S2). With this method, a small amount of input DNA is divided between a large number of droplet partitions (>1 000 000), such that only a small number of HMW DNA molecules, typically three to five, are present in each droplet, following a Poisson distribution. Thus, the large number of droplet partitions limits the amount of sequence overlap among DNA molecules and so for a given partition, there is a <1% chance that the molecules originated from overlapping segments of the genome. As a result, sequencing these barcode libraries provides information about the DNA sample under investigation at a near-single-molecule level. Furthermore, since the input DNA is not sheared, this process retains the genomic contiguity of the HMW DNA molecules for phasing variants and generating extended haplotypes.

With our approach, we attempt to extend the haplotypes from phased sequencing across even larger regions of the genome. Our method relies on the initial haplotype blocks provided by the Long Ranger software. These results are then used to link haplotypes that extend across chromosome arms that have undergone allelic imbalances. As a result, we can generate much larger haplotypes than are provided initially by the Long Ranger software. A general description of our approach is detailed in Figure 1B. Specific data processing details are noted in Figure 2 and Supplementary Figure S1. Descriptions of the statistics and programming script execution are described in Supplementary Notes 1 and 2. After initial data processing, the phasing information associated with each heterozygous SNV was used to generate phase blocks of contiguous segments. As noted above, each barcode is specific to a given droplet partition and is associated with the sequence from an individual HMW DNA molecule (Figure 1A). For a given genomic segment, the count of unique barcodes associated with aligned sequence reads directly represents the number of individual DNA molecules sequenced. Thus, counting barcodes provides the fractional representation of a given genomic segment. In the case of a normal diploid genome, the number of barcodes for each haplotype (i.e. two per a given genomic segment) should be approximately equivalent for either parental chromosome segments. In cases where a chromosomal alteration occurred with a loss or gain of a chromosomal segment, haplotypes within phase blocks were assigned as either the major or the minor haplotype, depending on their barcode ratios. The major haplotype contains more barcodes while the minor haplotype contains fewer. In this study, the concurrent analysis of a matched normal genome was a critical step for the determination of haplotype imbalance.

We used the barcode counts for phased heterozygous variants assigned to a specific phase block and determined the presence of large chromosomal changes and aneuploidy ('Materials and Methods' section). First, we normalized barcode counts at the phase block level. This procedure involved dividing the number of barcodes by the number of phased heterozygotes assigned to a phase block. Second, we carried out normalization using the total number of barcodes in each sample. Normalization of the SNV distribution reduced the variance of the blocks with respect to other blocks from the same haplotype in the same region (Figure 2A and B; Supplementary Figure S1). This step improved the discrimination of statistically significant differences between the normal and non-normal (dysplastic or tumor) samples (Figure 2C and D).

For meta-block generation, we used phase blocks that were shared across all of the matched samples (i.e. normal, dysplastic and tumor) from an individual patient. We required that the phase blocks contained at least 100 phased heterozygous variants; since most variants were contained within larger phase blocks, we retained 90% of all phased heterozygotes for analysis (Supplementary Table S1). We

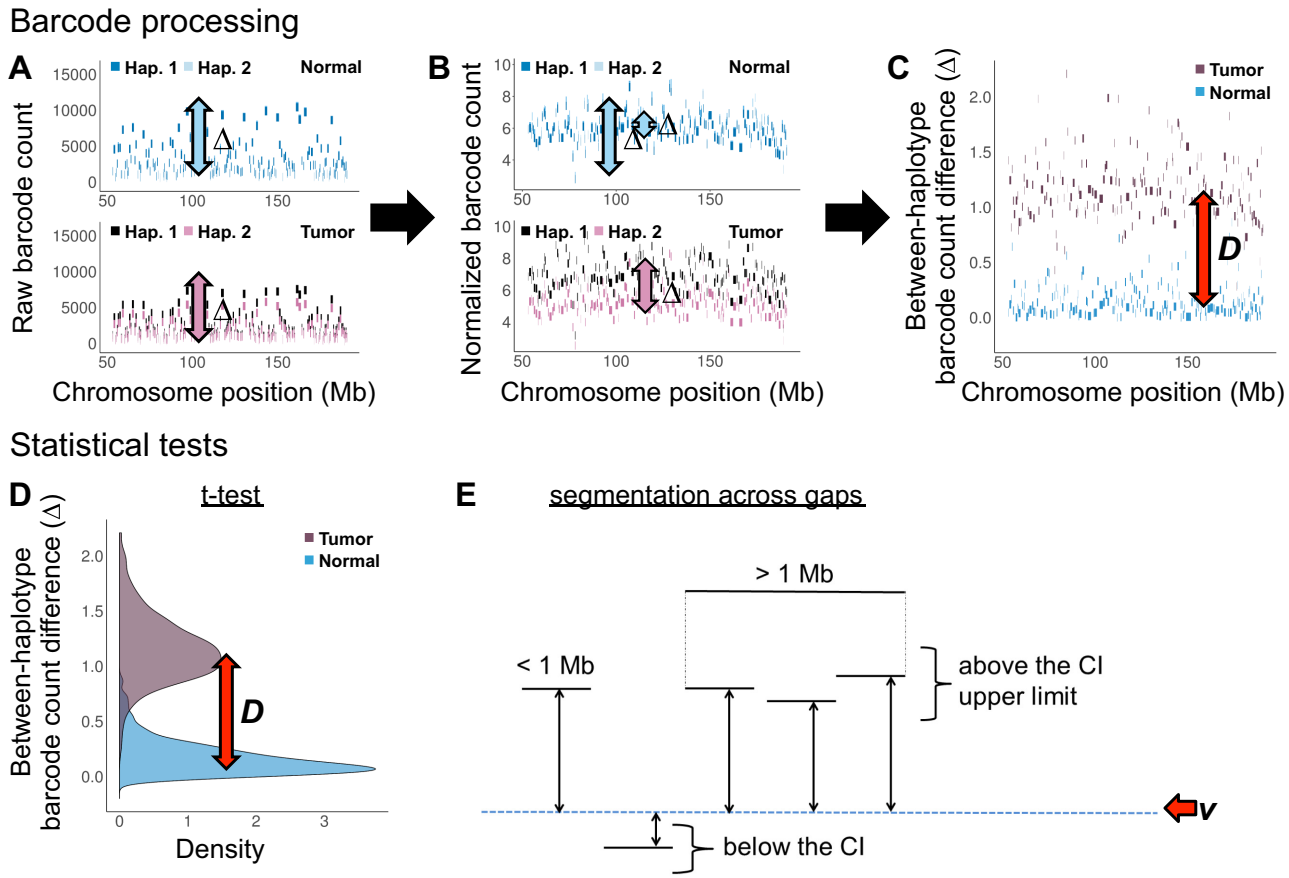

Figure 2. Summary of analysis method. After sequencing, variant files were processed to produce phase blocks (with at least 100 heterozygous phased SNVs). (**A**) Unique barcodes were counted for each haplotype of each phase block across the chromosome arm. (**B**) Blocks were normalized by dividing the number of unique barcodes per block by the number of single nucleotide polymorphism (SNPs) per block and non-normal samples were normalized by multiplying each block total by (total unique barcodes in normal sample)/(total unique barcodes in non-normal sample). (**C**) For each block in each sample, the differencebetween blocks in major and minor haplotypes was calculated. (**D**) Density distributions of Δ are used to perform a one-sided *t*-test (with a Bonferroni-adjusted *P*-value of .001) between normal and tumor, represented by *D*. (**E**) For all chromosome arms with $P < .001$, non-normal blocks are tested against the 97.5% upper confidence limit (v) of the normal Δ distribution. If they fall below this limit, they are not called. For regions >1Mb that pass these tests, the haplotypes are combined across the large blocks to produce mega-haplotypes.

also noted that switch errors—locations where one or more variants were incorrectly phased relative to the variants preceding or following them—were minimized in blocks of 100–200 phased heterozygotes (Supplementary Table S1), suggesting that blocks with higher numbers of variants provided more accurate haplotypes than smaller blocks.

We analyzed each chromosome arm for allelic imbalances between the major and minor haplotypes. For each chromosome arm, we used a one-sided *t*-test to detect the presence of statistically significant allelic imbalances in a given dysplastic or tumor sample as determined by differences in normalized barcode counts (Figure 2D). First, we generated an empirical null distribution derived from barcode count differences between the major and minor haplotypes in normal samples. From this, we determined a 97.5% threshold above which we assume that a difference found in the non-normal sample is caused by a true chromosomal alteration. We joined all of the contiguous phase blocks of 100 or more phased heterozygotes whose normalized barcode counts were above this limit. We refer to these aggregations of phase blocks as meta-blocks. Afterward, we generated chromosome-scale mega-haplotypes, which consist of the two haplotypes that exist within the coordinates of each meta-block.

Mega-haplotypes generated from a fetal trisomy cell line

We validated our first method by applying it to a cell line with known chromosomal abnormalities. Here, we performed barcoded linked read sequencing on the fetal trisomy cell line NA03576 (Table 1). The aneuploid characteristics noted by Coriell included an additional copy of chromosome 2 and of chromosome 21 (Coriell ref. no NA03576).

Because this germline-based cell line would not have a matched sample, and because it came from a male subject, we used phased genome data from NA12877 for comparison ('Materials and Methods' section). As noted above, the use of unrelated samples may affect the success of the method since it requires a sufficient number of overlapping SNV positions to generate large phase blocks. For this pair, we found 1 907 889 overlapping variants in the autosomal chromosomes of NA03576 and NA12877, including

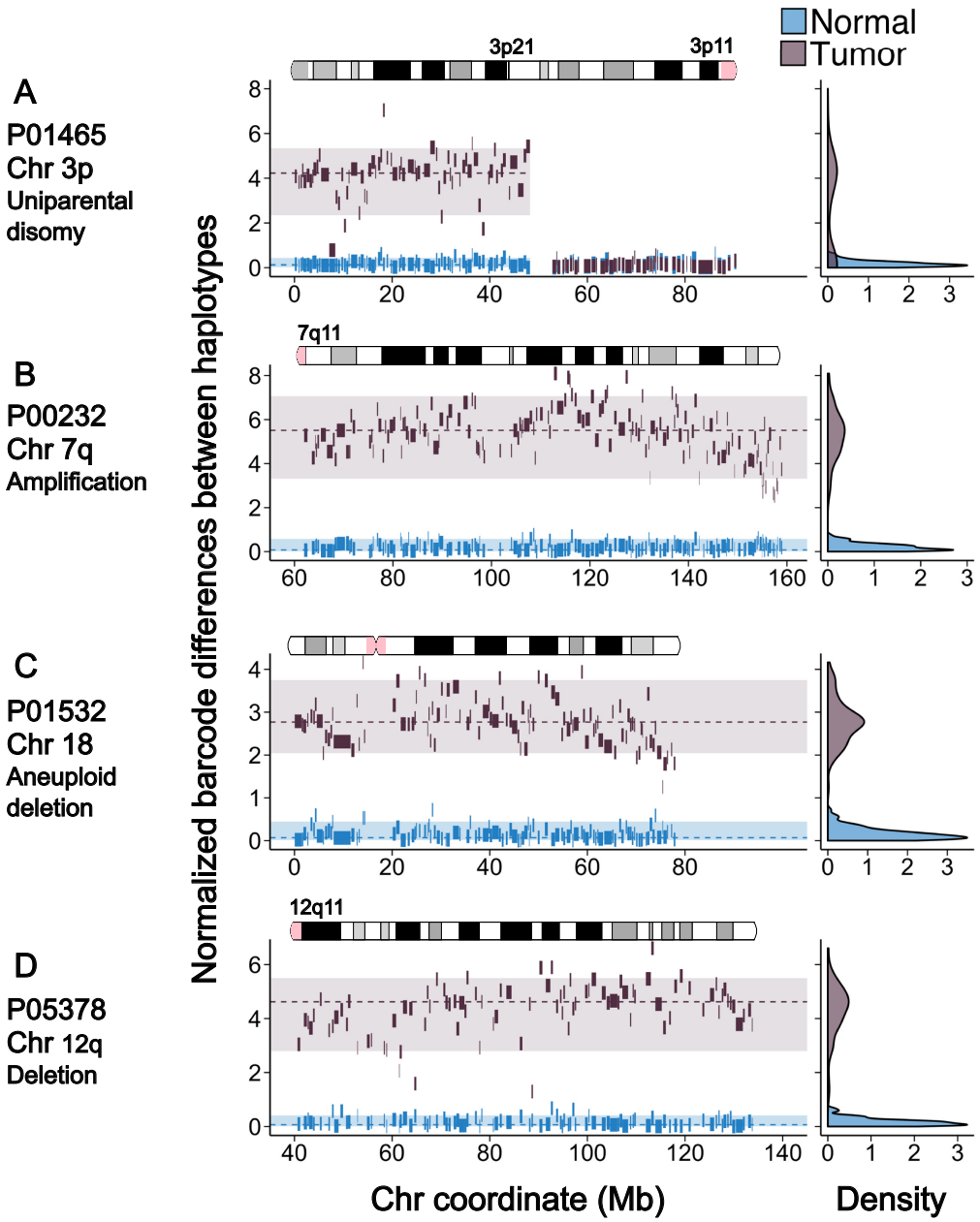

Figure 3. Mega-haplotypes of different samples. The x-axis denotes megabases; the y-axis shows the difference between major and minor haplotypes for each block, normalized for SNV density. The blue blocks indicate the difference between major and minor haplotypes in the normal sample; the dark blocks indicate the difference in the malignant sample. For each sample shown, the density plot to the right reflects the distribution of the haplotype differences. These density distributions are used for the *t*-test of significant differences. (**A**) Difference between haplotypes across the only multiple-megabase imbalanced region in Patient 1465. No copy number variation (CNV) is detected in this sample, showing this to be a case of uniparental disomy. (**B**) Mega-haplotype of the 7q region of Patient 232. Here the imbalance reflects an amplification in Patient 232's malignant lesion, and the mega-haplotype extends across the entirety of the chromosome arm. (**C**) Difference between haplotypes across chromosome 18 of Patient 1532. This is a case of aneuploidy as the entire chromosome has been deleted in the tumor. (**D**) A deletion in the 12q arm of Patient 5378, from a sample of a brain metastasis of a colorectal cancer.

60.7% of the NA03576 calls (3 142 116 SNVs total) and 59.6% of the NA12877 calls (3 202 638 total). We determined that there was sufficient SNV overlap to generate mega-haplotypes for all chromosomes.

For NA03576, we found five chromosome arms that included a mega-haplotype of >1 Mb. Our method identified three chromosome arms (2p, 2q and 21) that showed significant differences when subjected to a single sided *t*-test ($P < 0.001$), showing perfect concordance to the three which showed imbalances per traditional karyotyping (Supplementary Table S3). From CNV calls derived from conventional WGS, we also observed some subclonal variation in chromosome 19, but which was not major enough to pass the above test for allelic imbalance (Supplementary Table S4). The use of NA12877 conventional WGS sequence data, which had different read length conditions, led to small

CNV calls for some chromosomes; these were corrected by re-normalization by a constant factor, resulting in complete CNV concordance with karyotyping and linked read data.

Mega-haplotypes generated from an aneuploid colorectal cancer cell line

As an initial test of our method for the determination of large chromosomal aberrations and aneuploidy events, we analyzed a cancer cell line (LS411N) that had been previously characterized with karyotyping and SKY analysis (26). We repeated the standard karyotyping and demonstrated that LS411N genome was generally triploid among a total of 20 cells. We also note some clonal variation among the different cells that were examined, with, for instance, 3/24 chromosomes showing two distinct karyotype values. Given the clonal variation seen on karyotype, we used the mode value for making comparisons between our method and the karyotype. Among the cell karyotyped, 13/24 chromosomes had a karyotype mode value of three (Supplementary Table S5). This result was consistent with previous reports (26). The overall number of arms that showed imbalances was 26.

We sought to determine whether our digital karyotyping method would produce results concordant with the traditional karyotyping results. As the LS411N cell line does not have a matched normal, we used phased genome data from a diploid sample (NA12877) for comparison ('Materials and Methods' section). The use of unrelated samples could potentially complicate our analysis since our karyotyping method requires a sufficient number of overlapping SNV positions between samples to generate large phase blocks. In this case, we found that there were 1 431 779 overlapping variants in the autosomal chromosomes of LS411N and NA12877, including 55.6% of the LS411N calls (2 574 621 SNVs total) and 44.7% of the NA12877 calls (3 202 638 SNVs total). We determined that there was sufficient SNV overlap to generate mega-haplotypes for all chromosomes except for 17 and 18, which had lower degrees of SNV overlap compared to other chromosomes.

In total, we identified 35 chromosome arms with mega-haplotypes for LS411N, including 26 of the arms which showed imbalances per traditional karyotyping. Among the autosomes, our method identified 24 chromosome arms that showed significant differences when subjected to a single sided *t*-test ($P < 0.001$). Thus, 92.3% (24/26) of the imbalanced chromosome arms were detected by our digital karyotyping method (Supplementary Table S5). In general, the CNVs as determined with conventional WGS were concordant with our results (Supplementary Table S6).

Mega-haplotype characterization of chromosomal aberrations in colorectal cancer

We analyzed a series of matched colorectal tumor samples taken directly from surgical resections. We obtained sets of matched samples from three patients (232, 1465 and 1532), including: (i) dysplastic tissue (*in situ* carcinoma) not demonstrating invasive cancer; (ii) fully invasive, primary carcinoma and (iii) matched normal tissue. For these samples, we performed conventional WGS alongside barcode-linked WGS for phasing to provide a supplementary list of germline and somatic variants (Table 1 and Supplementary Tables S7, 8 and 9). Additionally, we analyzed a brain metastasis from a colorectal carcinoma (Patient 5378).

We applied our method to these samples. For each patient, we reported a meta-block if it covered more than half the chromosome arm (Supplementary Tables S10 and 11). Thus, we generated individual meta-blocks (containing mega-haplotypes) on the order of tens of megabases, demonstrating a substantial improvement over existing phasing methods. In cases where the dysplastic and tumor sample both showed imbalance in the same region, the outermost beginning and ending positions between the two samples were listed. In most cases, gaps between individual phase blocks making up the meta-block constituted a small fraction of the entire meta-block size.

We observed that the level of chromosomal instability varied substantially among these patients with colorectal cancer. From our analysis of dysplastic tissue and a primary colorectal adenocarcinoma from Patient 1465, the tumor sample had only low levels of genomic instability as determined by the extent of allelic imbalance. There was a single extended genomic segment across 48.16 Mb (53.2%) of the p arm of chromosome 3 in both the dysplastic and tumor sample (Figure 3A and Supplementary Table S10). Of the 28 178 heterozygous SNVs in this region called with a single alternate across all three samples, we were able to resolve 92.5% (26 066) of the variants into a single mega-haplotype. Notably, CNV analysis using only conventional WGS data detected no copy number change in this arm (Supplementary Table S12), implying that the region had undergone a segmental uniparental disomic conversion.

Our analysis of dysplastic tissue and primary colorectal adenocarcinoma from Patient 232 identified multiple large-scale chromosomal alterations in both samples. Also, evidence of shared somatic chromosomal changes in both samples supported a common clonal derivation. Fifteen chromosome arms originating from 11 chromosomes (including acrocentric ones) demonstrated an allelic imbalance in the primary tumor, as did 11 arms of 7 chromosomes in the dysplastic sample (Supplementary Table S10). Chromosomes with imbalances in both samples included 7, 8, 13, 14, 17, 18 and 20. Chromosomes 4, 15 and 21 demonstrated an allelic imbalance in only the primary tumor and the dysplastic sample contained only one unique chromosomal aberration on chromosome 20. The majority of the chromosomes with allelelic imbalances in Patient 232 were also found to harbor significant arm-level alterations that were observed in The Cancer Genome Atlas (TCGA) analysis of 257 colorectal tumors (27).

The regions of allelic imbalance for the chromosome arms in Patient 232 were extensive, enabling us to identify extremely large mega-haplotypes. For example, the 7q arm in both the tumor and dysplastic samples contained allelic imbalances spanning almost 100 Mb, thus covering >90% of the length of the chromosome arm (Figure 3B). In the primary tumor sample, we obtained a 96.8 Mb mega-haplotype that constituted 98.7% of the total 7q arm with no significant gaps in the haplotype coverage of the affected segment. Of 57 819 heterozygous SNVs across the 7q arm in the primary tumor sample called with a single alternate allele, 51 306 (88.7%) were successfully assigned within this

mega-haplotype (Supplementary Table S10). Chromosome 7q is frequently amplifed in colorectal cancer (CRC), leading to copy number gains of known cancer drivers including *MET* (7q31.2) and *WNT2* (7q31) (28,29). These represent important events that are involved in colorectal cancer biology and therapeutic resistance as is frequently seen with the *MET* gene.

We analyzed dysplastic tissue and a primary colorectal adenocarcinoma from Patient 1532. While we found no evidence of significant genomic instability in the dysplastic sample, the primary tumor included nine imbalanced arms across eight chromosomes (Figure 3C and Supplementary Table S10). For example, chromosome 18 showed significant allelic imbalances across both p and q arms, together covering 95.1% of the total chromosome. On the p arm, 9830 of 10 367 heterozygous variants (94.8%) were grouped into a single mega-haplotype, while 31 545 of the total 34 184 (92.3%) heterozygous variants were grouped in a single mega-haplotype on the q arm. Chromosome arms 18p and 18q were deleted in 66% of 257 CRC tumors analyzed in a TCGA study (27), causing a loss of the known tumor suppressor gene *SMAD4* (18q21.2) (30,31). After integration with CNV calls derived from conventional WGS, we observed that the cancer genome of Patient 1532 also demonstrated uniparental disomy in both 3 and 5q, covering ∼68.5 and 80.3% of the arms, respectively. While these two regions showed no copy number alteration, the minor and major allele frequencies across the regions were drastically different.

For Patient 5378, we analyzed a brain metastasis of a colorectal carcinoma. This tumor sample showed an extreme degree of genomic instability with apparent aneuploidy in 11 chromosomes (Figure 3D and Supplementary Table S13). Here, our analysis yielded mega-haplotypes for 37 of 41 total chromosome arms, covering ∼88.5% of the total genome (Figure 4 and Supplementary Table S11). Remarkably, we achieved mega-haplotypes spanning over 90% of the arm length in 29 of the 37 aberrant chromosome arms. In total, ∼92% of all phased SNVs common to both the normal and cancer genomes were mega-haplotyped at a chromosome-arm or near-chromosome-arm level. Previous studies have also reported high levels of genomic instability in brain metastases of primary colorectal tumors, including chromosome level gains and losses (32,33).

Haplotyping comparison of conventional versus phased WGS

We investigated whether it is possible to generate Mb-scale haplotypes with conventional WGS using only allele frequencies derived from sequence depth. Presumably, variants that are represented on a genomic segment with increased copy number can be used for phasing and generation of phase blocks. For this comparison, we used conventional WGS data ('Materials and Methods' section) generated from the samples of Patient 232, and assigned the allele with greater read depth to one haplotype and the allele with lesser read depth to the other haplotype, across the chromosome regions that had been detected by our mega-haplotyping analysis. We then assessed whether the haplotypes generated from conventional WGS were correct by comparing them to the haplotypes generated from phased WGS using the barcode sequence reads.

We considered the conventional WGS data to be correct if the GT assignments of the heterozygous variants within a phase block agreed with the barcoded phased data >90% (matching) or <10% (opposing) of the time ('Materials and Methods' section). Our analysis foucsed on the altered regions, as reported in Supplementary Table S10, of the 232 tumor sample. We found that only 55.8% (732/1312) of the phase blocks compared between the conventional versus phased sequencing had congruent phased GTs (Supplementary Table S14).

In the dysplastic sample, 22.0% (186/845) of intersected blocks were congruent. We also compared the switch error rates as previously defined for these regions (13,34). As a percentage of phased heterozygotes, conventional haplotyping of the malignant sample yielded substantially higher short and long switch errors of 5.2 and 3.4%, respectively; barcoded phased haplotyping yielded 0.2 and 0.1% short and long switch errors, respectively (Supplementary Table S14). The results were similar for the other samples. Overall, the use of standard WGS for haplotyping regions with allelic imbalances generated a high frequency of switch errors compared to those generated with phased genome sequencing, indicating that phased genome sequencing is necessary for the accuracy of our mega-haplotyping approach.

Assigning somatic variants to chromosomal arm-scale haplotypes

We placed somatic heterozygous variants in the context of specific mega-haplotypes—this enabled us to understand how cancer-specific SNVs are distributed among large-scale genomic aberrations. Because the generation of large-scale phase blocks used only variants found in all of the matched samples from each patient (i.e. all variants were present in the matched normal), we determined the assignment of somatic variants to major and minor haplotypes after establishing a specific mega-haplotype. This included placement of heterozygous mutations to the mega-haplotypes derived from chromosomal arm changes and aneuploidy events. Because somatic variants can be phased within a sample, they can be assigned to major and minor mega-haplotypes in the context of chromosome-scale aberrations (Supplementary Tables S10 and 11).

The distribution of the somatic mutations between the major and minor haplotypes within a mega-haplotype generally followed one of two patterns. The SNVs were either split nearly evenly between the haplotypes, as in chromosome arm 4q of Patient 232 (Major hap: 435 SNVs; Minor hap: 424 SNVs) (Supplementary Figure S2a and Table S7), or were associated mainly with the major haplotype, as in 8p of Patient 232 (Major hap: 88 SNVs; Minor hap: 31 SNVs) (Supplementary Table S10).

As an example of the distribution of somatic variants with respect to haplotype, we analyzed the haplotype distribution of somatic variants within a 30 Mb region (positions 10–40 Mb) of a large amplification on chromosome 7 in the malignant sample of Patient 232 (Supplementary Figure S2b). Of the 98 somatic mutations in the region, 57 belonged to the major haplotype and 41 to the minor haplotype. Mutations assigned to the major haplotype had a mean barcode proportion of 0.48; that is, averaged across

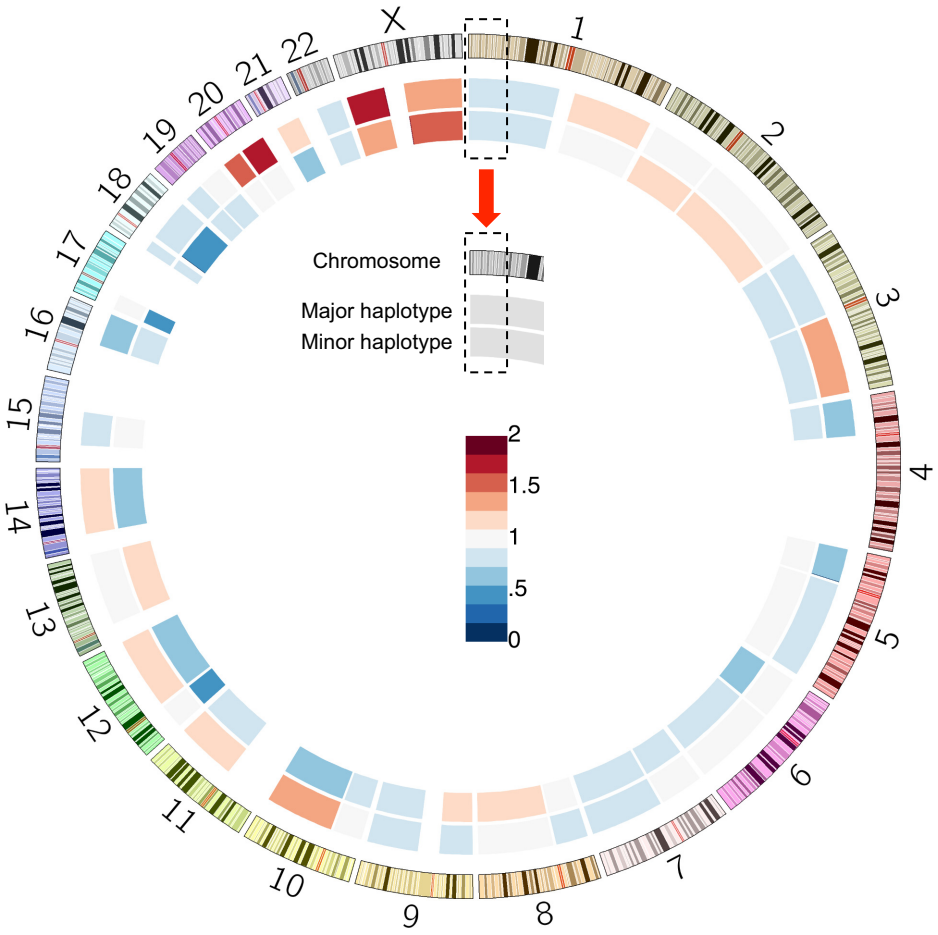

Figure 4. Imbalance in terms of copy number in a brain metastasis. This circos plot shows chromosome arms where a large imbalance (>50% of the arm) was called by our method. The X chromosome is included as Patient 5378 was female. The colors reflect the proportion of the haplotype relative to the the entire genome for each haplotype independently. For instance, chromosome X shows amplifications in both haplotypes although the more affected haplotype varies between arms. Chromosome 18 shows deletions of both haplotypes and both arms, with the greatest effect on the minor haplotype of the *q* arm.

the 57 variants, 48% of the barcodes that supported those positions were associated with the alternate allele of the somatic variant on the major haplotype, rather than the wild-type allele on the minor haplotype. In other words, the barcode distribution showed no significant bias towards the major haplotype variant. In contrast, the alternate alleles of somatic variants assigned to the minor haplotype were found to have a lower mean barcode proportion of 0.236. Thus, the distribution of somatic variant barcodes was concordant with our findings in the CNV analysis and the general allelic imbalance analysis.

DISCUSSION

We have established a method for generating digital karyotypes and multi-megabase haplotypes on the imbalanced regions of chromosomally unstable genomes. We demonstrate the effectiveness of this method in digital karyotyping of colorectal cancer genomes from various sources. Multiple regions of chromosome imbalance are characteristic of many cancers. Because altered chromosomes may have a large effect on gene dosage levels both in amplification and deletion, the ability to haplotype across very large regions of these chromosomes is an important element in the development of effective genetically targeted cancer treatment. We determined the haplotypes of chromosome arms up to 135 Mb (Patient 232 4q) and 146 Mb (Patient 5378 2q) as compared to, e.g. the hybrid method described by Mostovoy *et al.*, which gives a longest scaffold of 99.96 Mb (18).

The approach can be applied to any results from a variety of experimental haplotype methods; the data need not be of exceptional quality to give large scale haplotypes, as long as there are allelically imbalanced chromosome regions. For this study, we used a method that relies on droplet partition barcodes that are related to the number of unique molecules sequenced (15). Any barcode, however, that denotes unique molecules or can be reliably associated with them can serve as the input to this method.

The input is a VCF file with initial haplotypes generated from the Long Ranger software. Afterward, this approach uses a combination of perl, python and R scripts to produce results and graphics. This method's utility in haplotyping

benefits from increased copy number of large chromosomal segments. Because the signal used by the method is directly proportional to allelic imbalance, the more highly amplified one chromatid is, the more easily can mega-haplotypes be constructed. However, because barcodes are counted on both major and minor haplotypes, it is possible to detect homozygous gains and losses via barcode counts, as on Chromosome 20 of Patient 232 (Supplementary Figure S3). Thus this method is likely to be more convenient than karyotyping in the characterization of non-diploid genomes.

Our method allows for quick and easy karyotyping as well as haplotyping over very long stretches of cancer genomes. Therefore, we believe it will be a useful tool in the analysis of cancer genomes and those of other genetic diseases, especially in examining large chromosomal alterations and the determination of *cis*- and *trans*- relations in gene regulation.

AVAILABILITY

The dbGAP accession number for cancer sample data is phs001400. The scripts used in the analysis method are available as a 30 KB tar.gz Supplementary file.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Grace Zheng, Alex Wong and Michael Schnall-Levin from 10× Genomics for advice about the barcode sequencing technology.

FUNDING

National Institutes of Health (NIH) [NHGRI P01HG000205 to B.T.L., J.M.B., H.P.J.; NHGRI R01HG006137 (to L.C.X., H.P.J.); NCI R33CA174575 (to J.M.B., H.P.J.); NCI R21CA193046 (to I.D.C., M.H.G.)]; Stanford Cancer Institute Translational Research Award (to H.P.J., J.M.B.); American Cancer Society Research Scholar Grant [RSG-13–297-01-TBG] (to S.G., H.P.J); Doris Duke Charitable Foundation, Clayville Foundation, Seiler Foundation and Howard Hughes Medical Institute (to H.P.J.). Funding for open access charge: NIH.
Conflict of interest statement. None declared.

REFERENCES

1. Sudmant,P.H., Rausch,T., Gardner,E.J., Handsaker,R.E., 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.
2. Grimwade,D., Walker,H., Oliver,F., Wheatley,K., Harrison,C., Harrison,G., Rees,J., Hann,I., Stevens,R., Burnett,A. *et al.* (1998) The importance of diagnostic cytogenetics on outcome in AML: analysis of 1,612 patients entered into the MRC AML 10 trial. *Blood*, **92**, 2322–2333.
3. Slovak,M.L., Kopecky,K.J., Cassileth,P.A., Harrington,D.H., Theil,K.S., Mohamed,A., Paietta,E., Willman,C.L., Head,D.R., Rowe,J.M. *et al.* (2000) Karyotypic analysis predicts outcome of preremission and postremission therapy in adult acute myeloid leukemia: a Southwest Oncology Group/Eastern Cooperative Oncology Group study. *Blood*, **96**, 4075–4083.
4. Knutsen,T., Padilla-Nash,H.M., Wangsa,D., Barenboim-Stapleton,L., Camps,J., McNeil,N., Difilippantonio,M.J. and Ried,T. (2010) Definitive molecular cytogenetic characterization of 15 colorectal cancer cell lines. *Genes chromosomes Cancer*, **49**, 204–223.
5. Reddy,U.M., Page,G.P., Saade,G.R., Silver,R.M., Thorsten,V.R., Parker,C.B., Pinar,H., Willinger,M., Stoll,B.J., Heim-Hall,J. *et al.* (2012) Karyotype versus microarray testing for genetic abnormalities after stillbirth. *N. Engl. J. Med.*, **367**, 2185–2193.
6. Schröck,E., Manoir,S.d., Veldman,T., Schoell,B., Wienberg,J., Ferguson-Smith,M.A., Ning,Y., Ledbetter,D.H., Bar-Am,I., Soenksen,D. *et al.* (1996) Multicolor Spectral Karyotyping of Human Chromosomes. *Science*, **273**, 494–497.
7. Ning,Y., Liang,J.C., Nagarajan,L., Schröck,E. and Ried,T. (1998) Characterization of 5q deletions by subtelomeric probes and spectral karyotyping. *Cancer Genet. Cytogenet.*, **103**, 170–172.
8. Morelli,S.H., Deubler,D.A., Brothman,L.J., Carey,J.C. and Brothman,A.R. (1999) Partial trisomy 17p detected by spectral karyotyping. *Clin. Genet.*, **55**, 372–375.
9. Dennis,T.R. and Stock,A.D. (1999) A molecular cytogenetic study of chromosome 3 rearrangements in small cell lung cancer: consistent Involvement of chromosome band 3q13.2. *Cancer Genet. Cytogenet.*, **113**, 134–140.
10. Pollack,J.R., Perou,C.M., Alizadeh,A.A., Eisen,M.B., Pergamenschikov,A., Williams,C.F., Jeffrey,S.S., Botstein,D. and Brown,P.O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, **23**, 41–46.
11. Pollack,J.R., Sørlie,T., Perou,C.M., Rees,C.A., Jeffrey,S.S., Lonning,P.E., Tibshirani,R., Botstein,D., Børresen-Dale,A.-L. and Brown,P.O. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. U.S.A.*, **99**, 12963–12968.
12. Cho,E.K., Tchinda,J., Freeman,J.L., Chung,Y.J., Cai,W.W. and Lee,C. (2006) Array-based comparative genomic hybridization and copy number variation in cancer research. *Cytogenet. Genome Res.*, **115**, 262–272.
13. Kuleshov,V., Xie,D., Chen,R., Pushkarev,D., Ma,Z., Blauwkamp,T., Kertesz,M. and Snyder,M. (2014) Whole-genome haplotyping using long reads and statistical methods. *Nat. Biotechnol.*, **32**, 261–266.
14. Kaper,F., Swamy,S., Klotzle,B., Munchel,S., Cottrell,J., Bibikova,M., Chuang,H.-Y., Kruglyak,S., Ronaghi,M., Eberle,M.A. *et al.* (2013) Whole-genome haplotyping by dilution, amplification, and sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 5552–5557.
15. Zheng,G.X.Y., Lau,B.T., Schnall-Levin,M., Jarosz,M., Bell,J.M., Hindson,C.M., Kyriazopoulou-Panagiotopoulou,S., Masquelier,D.A., Merrill,L., Terry,J.M. *et al.* (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.*, **34**, 303–311.
16. Borgström,E., Redin,D., Lundin,S., Berglund,E., Andersson,A.F. and Ahmadian,A. (2015) Phasing of single DNA molecules by massively parallel barcoding. *Nat. Commun.*, **6**, e7173.
17. Pendleton,M., Sebra,R., Pang,A.W.C., Ummat,A., Franzen,O., Rausch,T., Stutz,A.M., Stedman,W., Anantharaman,T., Hastie,A. *et al.* (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods*, **12**, 780–786.
18. Mostovoy,Y., Levy-Sakin,M., Lam,J., Lam,E.T., Hastie,A.R., Marks,P., Lee,J., Chu,C., Lin,C., Dzakula,Z. *et al.* (2016) A hybrid approach for de novo human genome sequence assembly and phasing. *Nat. Methods*, **13**, 587–590.
19. Seo,J.-S., Rhie,A., Kim,J., Lee,S., Sohn,M.-H., Kim,C.-U., Hastie,A., Cao,H., Yun,J.-Y., Kim,J. *et al.* (2016) De novo assembly and phasing of a Korean human genome. *Nature*, **538**, 243–247.
20. Adey,A., Burton,J.N., Kitzman,J.O., Hiatt,J.B., Lewis,A.P., Martin,B.K., Qiu,R., Lee,C. and Shendure,J. (2013) The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*, **500**, 207–211.
21. Dong,H., Zhang,H., Liang,J., Yan,H., Chen,Y., Shen,Y., Kong,Y., Wang,S., Zhao,G. and Jin,W. (2011) Digital karyotyping reveals probable target genes at 7q21.3 locus in hepatocellular carcinoma. *BMC Med. Genomics*, **4**, e60.
22. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

23. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al.* (2010) The Genome Analysis Toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
24. Xi,R., Hadjipanayis,A.G., Luquette,L.J., Kim,T.-M., Lee,E., Zhang,J., Johnson,M.D., Muzny,D.M., Wheeler,D.A., Gibbs,R.A. *et al.* (2011) Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, E1128–E1136.
25. Xi,R., Lee,S., Xia,Y., Kim,T.-M. and Park,P.J. (2016) Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Res.*, **44**, 6274–6286.
26. Abdel-Rahman,W.M., Katsura,K., Rens,W., Gorman,P.A., Sheer,D., Bicknell,D., Bodmer,W.F., Arends,M.J., Wyllie,A.H. and Edwards,P.A.W. (2001) Spectral karyotyping suggests additional subsets of colorectal cancers characterized by pattern of chromosome rearrangement. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 2538–2543.
27. Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330–337.
28. Zeng,Z.-S., Weiser,M.R., Kuntz,E., Chen,C.-T., Sajid,K.A., Nash,G.M., Gimbel,M., Yamaguchi,Y., Culliford,A.T., D'Alessio,M. *et al.* (2008) c-Met gene amplification is associated with advanced stage colorectal cancer and its liver metastases. *Cancer Lett.*, **265**, 258–269.
29. Jung,Y.-S., Jun,S., Lee,S.H., Sharma,A. and Park,J.-I. (2015) Wnt2 complements Wnt/β-catenin signaling in colorectal cancer. *Oncotarget*, **6**, 37257–37268.
30. Alazzouzi,H., Alhopuro,P., Salovaara,R., Sammalkorpi,H., Järvinen,H., Mecklin,J.-P., Hemminki,A., Schwartz,S., Aaltonen,L.A. and Arango,D. (2005) SMAD4 as a prognostic marker in colorectal cancer. *Clin. Cancer Res.*, **11**, 2606–2611.
31. Voorneveld,P.W., Kodach,L.L., Jacobs,R.J., Liv,N., Zonnevylle,A.C., Hoogenboom,J.P., Biemond,I., Verspaget,H.W., Hommes,D.W., de Rooij,K. *et al.* (2014) Loss of SMAD4 alters BMP signaling to promote colorectal cancer cell metastasis via activation of Rho and ROCK. *Gastroenterology*, **147**, 196–208.
32. Petersen,I., Hidalgo,A., Petersen,S., Schlüns,K., Schewe,C., Pacyna-Gengelbach,M., Goeze,A., Krebber,B., Knösel,T., Kaufmann,O. *et al.* (2000) Chromosomal imbalances in brain metastases of solid tumors. *Brain Pathol.*, **10**, 395–401.
33. Gutenberg,A., Gerdes,J.S., Jung,K., Sander,B., Gunawan,B., Bock,H.C., Liersch,T., Brück,W., Rohde,V. and Füzesi,L. (2010) High chromosomal instability in brain metastases of colorectal carcinoma. *Cancer Genet. Cytogenet.*, **198**, 47–51.
34. Amini,S., Pushkarev,D., Christiansen,L., Kostem,E., Royce,T., Turk,C., Pignatelli,N., Adey,A., Kitzman,J.O., Vijayan,K. *et al.* (2014) Haplotype-resolved whole genome sequencing by contiguity preserving transposition and combinatorial indexing. *Nat. Genet.*, **46**, 1343–1349.