Somatic mutation phasing and haplotype extension using linked-reads in multiple myeloma

Steven M. Foltz[1,2], Yize Li[1,2], Lijun Yao[1,2], Nadezhda V. Terekhanova[1,2], Amila Weerasinghe[1,2], Qingsong Gao[1,2], Guanlan Dong[1,2], Moses Schindler[1,2], Song Cao[1,2], Hua Sun[1,2], Reyka G. Jayasinghe[1,2], Robert S. Fulton[2], Catrina C. Fronick[2], Justin King[1], Daniel R. Kohnen[1], Mark A. Fiala[1], Ken Chen[3], John F. DiPersio[1,4], Ravi Vij*[1,4], Li Ding*[1,2,4,5]

1 Department of Medicine, Washington University in St. Louis, St. Louis, MO, 63110, USA
2 McDonnell Genome Institute, Washington University in St. Louis, St. Louis, MO, 63108, USA
3 Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA
4 Siteman Cancer Center, Washington University in St. Louis, St. Louis, MO, 63110, USA
5 Department of Genetics, Washington University in St. Louis, St. Louis, MO, 63110, USA

*Corresponding authors: lding@wustl.edu (L.D.) and rvij@wustl.edu (R.V.)

Abstract

Somatic mutation phasing informs our understanding of cancer-related events, like driver mutations. We generated linked-read whole genome sequencing data for 23 samples across disease stages from 14 multiple myeloma (MM) patients and systematically assigned somatic mutations to haplotypes using linked-reads. Here, we report the reconstructed cancer haplotypes and phase blocks from several MM samples and show how phase block length can be extended by integrating samples from the same individual. We also uncover phasing information in genes frequently mutated in MM, including *DIS3*, *HIST1H1E*, *KRAS*, *NRAS*, and *TP53*, phasing 79.4% of 20,705 high-confidence somatic mutations. In some cases, this enabled us to interpret clonal evolution models at higher resolution using pairs of phased somatic mutations. For example, our analysis of one patient suggested that two *NRAS* hotspot mutations occurred on the same haplotype but were independent events in different subclones. Given sufficient tumor purity and data quality, our framework illustrates how haplotype-aware analysis of somatic mutations in cancer can be beneficial for some cancer cases.

Introduction

Human genomes are diploid with two copies of each autosomal chromosome. Homologous chromosomes are distinct because they represent unique patterns of germline variation inherited from each parent. While genotypes represent the alleles at a specific locus, haplotypes are defined as groups of alleles across many loci separated according to which homolog they come from. Variant phasing and haplotype reconstruction may be achieved through technological and computational methods with a variety of data types and integration strategies from large public databases and individual samples.[1-24]

Determining the haplotype of cancer-associated mutations informs our understanding of the oncogenic process, but that information is typically lost with next-generation bulk sequencing.[25,26] Linked-read sequencing overcomes that limitation by labelling DNA from the same haplotype with the same barcode. Zheng *et al.* described this linked-read approach, accurately modeling fusion breakpoints and revealing biallelic *TP53* inactivation by phasing a mutation and hemizygous deletion to opposite haplotypes.[27] Marks *et al.* established the accuracy and reliability of linked-reads and explored the impact of variant density and heterozygosity on phasing performance.[28] Linked-reads have impacted cancer study design and are especially well-suited for structural variant detection.[29-38] Greer, et al. compared gastric cancer metastases and delineated a complex structural variant leading to *FGFR2* amplification.[39] Viswanathan, et al. determined the order of events in a cohort of prostate cancer patients, showing androgen receptor (*AR*) gene duplications and *CDK12* inactivation, phasing somatic mutations if the reads supporting it were assigned to a haplotype and phase block, and developing allele-specific copy number detection methods.[40,41] Sereewattanawoot, et al. matched *cis*-acting regulatory variants with allele-specific expression in lung cancer cell

lines.[42] ENCODE cell lines K562 and HepG2 have been used for deeply-integrated linked-read investigations.[43,44]

In this study, we analyzed 23 samples from a cohort of 14 multiple myeloma patients using linked-read whole genome sequencing (lrWGS) generated using the 10X Genomics Chromium System. Multiple myeloma (MM) is the second most common form of blood cancer and has a median 5-year survival around 50%.[45] MM is caused by clonal proliferation of plasma cells in the bone marrow. Primary genetic aberrations include hyperdiploidy and translocations that join the highly expressed IGH locus (chr14) with oncogenes, including t(11;14) (*CCND1*), t(4;14) (*WHSC1*), t(6;14) (*CCND3*), and t(14;20) (*MAFB*). Secondary events include *MYC* translocations and driver mutations. MAPK is the most commonly mutated pathway in MM, including somatic mutations in *KRAS*, *NRAS*, and *BRAF*.[45] Better appreciation of the haplotype context of these events, both driver mutations and structural variations, is necessary to improve targeted therapies and understanding of myelomagenesis. We created a framework for systematically phasing somatic mutations to haplotypes, allowing for deeper interpretation of tumor evolution in some cases. We also illustrate the concept of extending phase blocks using shared germline information across samples from the same individual. Our cohort represents a large resource of multiple myeloma lrWGS data and improves our understanding of human haplotype and cancer haplotype analysis.

Results

Haplotype-aware methods build on phasing information to analyze somatic mutations

The advantage of lrWGS over traditional WGS is that reads mapping to the same genomic region with the same barcode most likely originated from the same piece of high molecular weight (HMW) DNA (Fig. 1a).[27] The Long Ranger pipeline (10X Genomics) aligns reads, calls and phases variants, reports structural variants (SVs), and produces phasing quality metrics. With enough sequencing depth and allelic heterogeneity, Long Ranger is able to phase variants and reads. Variants and reads are grouped into phase blocks, defined as genomic ranges in which haplotype assignments are consistent. Within a phase block, all variants assigned to a certain haplotype are thought to have originated from the same biological haplotype. The haplotype order may switch in another phase block, so haplotype assignments cannot be compared between phase blocks. Long Ranger phasing is designed to work with germline variants and does not distinguish between germline variants and somatic mutations in cancer. Phasing performance may be suboptimal for somatic mutations with low variant allele frequency (VAF), in regions of copy number variation, and in tumor samples with low purity or heterogeneous clonal structure. Specific methods are necessary to overcome this limitation.[46]

To enable further downstream processing of lrWGS data, we developed additional methods that use Long Ranger output to further analyze single nucleotide variant (SNV) mutations collectively referred to as SomaticHaplotype (Fig. 1b) (see **Methods** and **Code Availability**). Given the phased variant call format (VCF) file and phased bam file produced by Long Ranger, the *phaseblock* module constructs PhaseBlock and Variant objects with information derived from reads and variant calls for use by later modules. The *summarize* module reports summary information about each phase block, including genomic range and number of variants, and global statistics like phase block length N50. The *somatic* module uses two complementary approaches to assign high-confidence somatic mutations to haplotypes and then analyzes the haplotype relationship between proximal pairs of events. The *extend* module utilizes germline

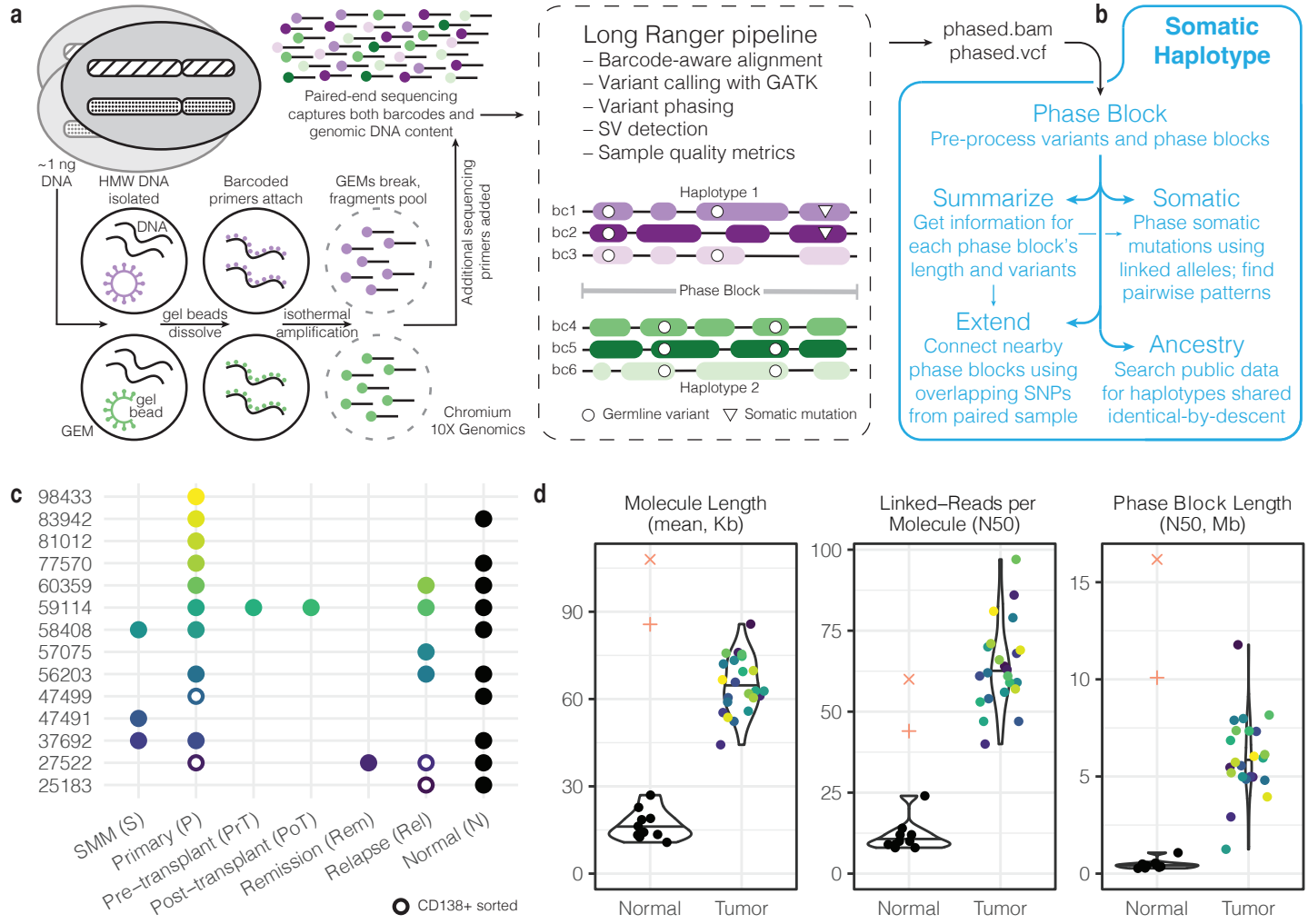


Figure 1. Linked-read data generation and analysis pipeline. a. The 10X Genomics Chromium platform tags large DNA molecules with barcodes such that reads originating from the same molecule have the same barcode. The Long Ranger pipeline aligns reads and phases variants. b. SomaticHaplotype builds upon Long Ranger output with several modules, including phaseblock, summarize, somatic, extend, and ancestry. c. Our cohort comprises 14 multiple myeloma patients across several disease stages for a total of 23 tumor samples. d. Quality control measures for our tumor and normal samples plus 1000 Genomes samples NA12878 (+) and NA19240 (x). Violin plots defined as: center line, median; violin limits, minimum and maximum values; points, every observation. Molecule Length (mean, Kb): length-weighted mean input DNA length in kilobases. Linked-Reads per Molecule (N50): N50 of read-pairs per input DNA molecule. Phase Block Length (N50, Mb): N50 length of phase blocks in megabases.

variation from matched samples to bridge gaps between phase blocks and suggests how to make neighboring phase blocks have consistent haplotype assignments. The *ancestry* module augments lrWGS data with information from large-scale phased resources, like the 1000 Genomes Project.

Our data set comprises lrWGS data from 14 patients diagnosed with multiple myeloma (Fig. 1c). Longitudinal samples were taken from the premalignant smoldering multiple myeloma (S), primary diagnosis (P), pre-transplant (PrT), post-transplant (PoT), remission (Rem), and relapse (Rel) stage. In total, 23 tumor samples and 10 skin normal samples were processed with lrWGS. Four tumor samples were CD138+ sorted to enrich for plasma cells, increasing tumor purity. Other samples were not CD138+ sorted and contain varying compositions of microenvironment cells along with tumor plasma cells. In addition, for 9 CD138+ sorted tumor samples with matched lrWGS, we generated whole genome sequencing (WGS) data with increased tumor purity to make high confidence somatic mutation calls (6 samples available at first data freeze) and structural variant calls (9 samples) (Supplementary Table 1; see **Methods**). Please see Supplementary Table 1 for tumor purity estimates of lrWGS samples with matched CD138+ sorted WGS samples (median tumor purity of sorted lrWGS = 0.676, n = 1; median tumor purity of unsorted lrWGS = .202, n = 4).

Cell-type composition, including tumor purity, shapes our interpretation of results from the cohort collectively and from individual samples. CD138+ sorting of four tumor samples selected for tumor-associated plasma cells, increasing tumor purity and our ability to detect interesting somatic mutation events. In unsorted samples comprising many immune and stromal cells not carrying the somatic mutations found in the tumor, we found tumor purity to be an important limiting factor that restricted our ability to more broadly generalize our findings

across the dataset. Instead, we illustrate the types of analysis enabled by our framework by focusing on particular cases with the data quality sufficient for confident interpretation.

Quality control measures of our tumor samples compared well with data from publicly-available gold-standard data from two 1000 Genomes samples (see **Data Availability**) (Fig. 1d, Supplementary Figure 1, Supplementary Table 2). Molecule length refers to the size of the long, HMW DNA fragments. In our tumor samples, the mean molecule length per sample ranged from 44.3 Kb to 85.8 Kb with a median of 62.8 Kb, whereas in our normal skin samples, the median value was 15.3 Kb. Linked-reads per molecule is the number of read pairs originated from each molecule, and the N50 value indicates that half of the molecules have that many reads pairs or more. In our tumor samples, the N50 linked-reads per molecule ranged from 40 to 97 with a median of 62, compared to a median of 10 in our skin samples. Finally, the N50 phase block length in tumor samples ranged from 1.3 Mb to 11.8 Mb with a median of 5.7 Mb, whereas the median was 0.4 Mb in skin samples. Given the consistent lack of informative linked-read information in our skin samples, we excluded them from downstream analysis. The skin samples were only used as a control for somatic mutation calling from our sorted WGS samples. For tumor samples, the median corrected mass of input DNA loaded into the Chromium chip was 1.3 ng, and the median mean sequencing depth was 71.6 reads. The median percentage of single nucleotide variants (SNVs) phased by Long Ranger was 99.2%. See Zhang, et al. for additional quality metrics that may be applied to linked-read data.[47]

Phase block lengths reflect biologically-relevant genomic changes

We examined the distribution of phase block lengths to explore patterns in our data. N50 phase block lengths were consistent across chromosomes, with the median N50 ranging from

4.42 Mb on chr15 to 7.74 Mb on chr18 (Supplementary Figure 2a). Chr1 showed the least variation in N50 phase block length (median 4.52 Mb, standard deviation 1.37 Mb). Chr21 showed the greatest variation (median 5.78 Mb, standard deviation 9.33 Mb) and also had the highest overall values, with 6 samples having N50 phase block lengths above 20 Mb, 4 of which came from Patient 59114. Some samples, such as 25183 (Rel), had consistently higher N50 values across many chromosomes (Supplementary Figure 2b). This may be due to this sample having the highest mean molecule length (85.8 Kb) and percentage of mapped reads (97.7%) of all tumor samples. Another sample, 58408 (P), had consistently shorter phase blocks, but quality control measures did not clearly indicate why.

Chr13 and chr22 from 27522 (P) showed low N50 phase block lengths, and the distribution of phase block lengths from those two chromosomes is strikingly different from that of other chromosomes (Supplementary Figure 2c). The N50 phase block lengths for chr13 and chr22 were 0.42 Mb and 0.38 Mb, respectively, compared to that sample's overall median N50 of 5.9 Mb. Both chr13 and chr22 had a one copy deletion across the entire chromosome, leading to a lack of heterozygosity needed for long phase blocks (Supplementary Figure 3). Hemizygous chr13 and chr22 phase blocks from 27522 (P) are much shorter across the entire chromosome compared to those from the remission sample, which is closer to an overall diploid state with low tumor content (Supplementary Figure 2d). However, we can interpret this sequencing artifact in a biologically meaningful way, and one benefit of homozygosity across an entire chromosome is the potential to resolve the entire chromosome's haplotype structure. Deletion size, tumor purity, and the proportion of tumor cells with copy number loss are important factors determining the ability of deletion regions to be phased.

In total, phase blocks cover 60.6 Gb across our 23 tumor samples (Supplementary Figure 2e), for an average of 2.6 Gb per sample. 72.2% (32,426/44,918 phase blocks) of phase blocks are between 0 and 1 Mb, but those short segments account for only 8.4% (5.1/60.6 Gb) of the total amount of genome covered by phase blocks in these samples. In comparison, 3,776 phase blocks are between 1-2 Mb and cover 5.5 Gb (9.0%). The distribution of genomic coverage by phase blocks of increasing length has a right-skewed long tail distribution. There are 19 phase blocks longer than 30 Mb, and the longest phase block is 59.2 Mb. As expected, there is a strong linear relationship between phase block length and the number of phased heterozygous variants (r^2 = 0.96). Over the 5.0 Mb human leukocyte antigen (HLA) region of chr6 (chr6:28510120-33480577), we observed a median of 4 phase blocks greater than 1kb in length (range 1-13 phase blocks), which covered between 93.5% and 100% of the region (median 98.7%). HLA region haplotyping could help match patients and donors before allogeneic stem cell transplants in limited and specific MM cases.[48]

Somatic mutations can be phased to specific haplotypes using linked alleles

The haplotype context in which somatic mutations occur may be biologically relevant. For example, knowing the phase of two mutations affecting the same gene would indicate whether they cause biallelic inactivation or only alter one copy. However, tumor impurity, heterogeneity, and variable sequencing coverage make somatic mutations harder to identify and phase using standard approaches. To phase somatic mutations, we built upon the strengths of Long Ranger by examining germline variants that occur on each barcode associated with a somatic mutation site (Fig. 2a). We defined linked alleles as alleles co-occurring on the same barcode with either the reference or alternate allele at the somatic mutation site. We know that alleles co-occurring with the same barcode most likely originated from the same molecule of HMW

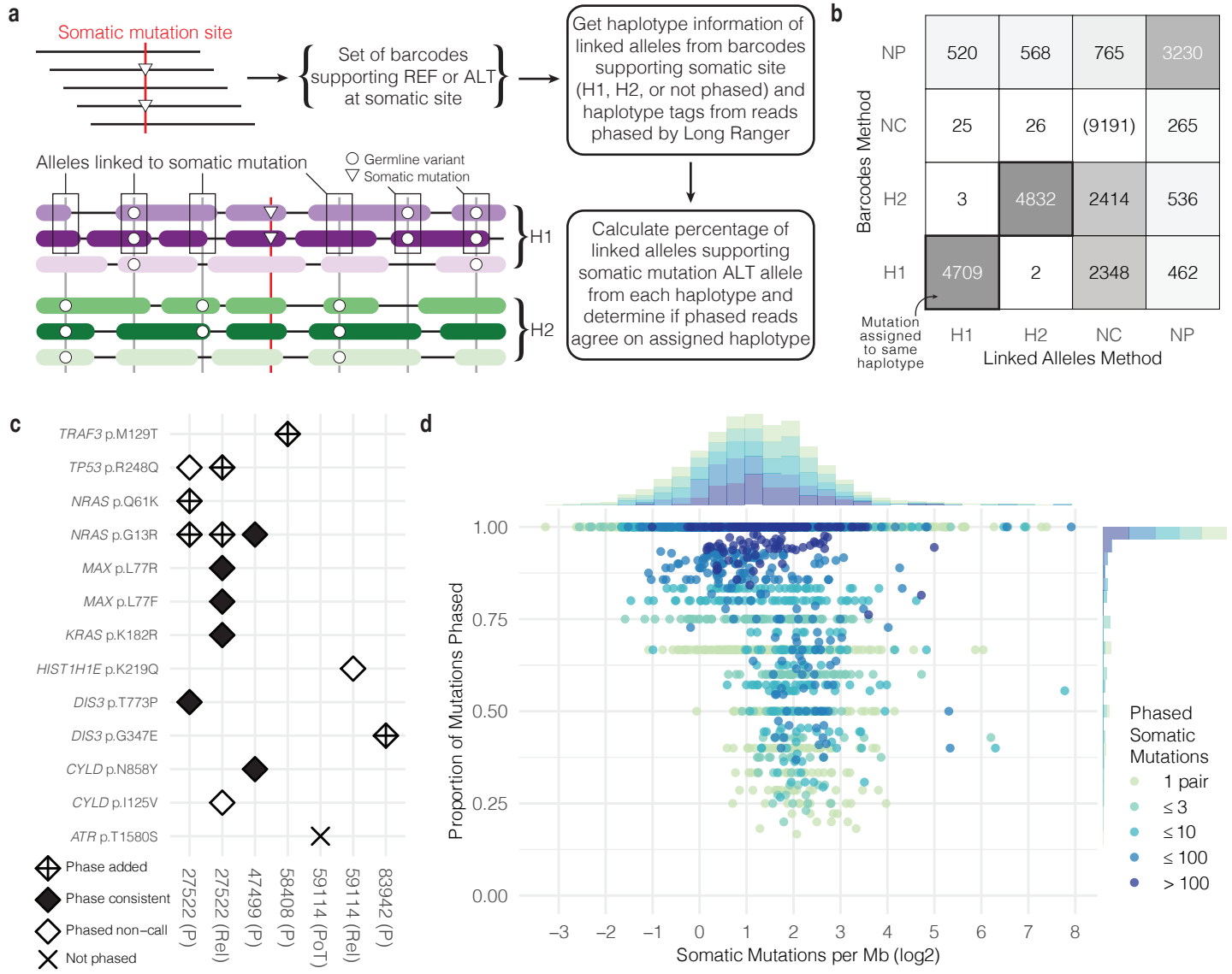


Figure 2. Phasing somatic mutations to haplotypes. a. Overview of methods used to phase somatic mutations. b. Number of somatic mutations phased using two phasing methods (H1 = phased to haplotype 1; H2 = phased to haplotype 2; NC = not enough coverage for phasing; NP = not phased). c. Phasing somatic mutations commonly observed in multiple myeloma. d. Distribution of somatic mutations per phase block and the proportion of mutations phased.

DNA, and we know the haplotype assignment of most (~99%) linked alleles. We developed two methods to phase somatic mutations even if the mutation was not phased by the standard pipeline. In the "linked alleles" approach, if the linked alleles co-occurring with the somatic mutation are consistently phased to the same haplotype, we can infer the haplotype of the somatic mutation since it is most likely the same as the linked alleles. Alternatively, we can also use the "barcodes" approach which leans on the assigned phased of reads supporting the alternate allele as evidence. We required complete agreement of reads with assigned phases to confidently infer the haplotype of somatic mutations. In this approach, we extract the haplotype annotation for each read, which is reported as a tag in the phased bam output from Long Ranger. However, this information is not given for all reads. In our tumor sample data, 71.6% of reads overlapping a somatic mutation site were assigned a haplotype. Combining these two approaches increases phasing power when one approach lacks adequate coverage.

For six lrWGS samples with matched CD138+ sorted WGS, we called high-confidence somatic mutations using the sorted WGS tumor sample (see **Methods**). In total, we detected 32,842 somatic SNVs from our six sorted WGS samples, or 5,474 somatic SNVs per sample. Of those, 29,896 mutations (4,983 per sample) were SNVs with coverage in the matched lrWGS samples, and 20,705 (69.2%) met our minimum coverage requirement of at least 10 linked alleles from barcodes supporting the mutant allele or at least one phased read supporting the mutant allele. To establish a linked allele threshold at which we could confidently phase somatic mutations, we overlapped high-confidence somatic mutations from our WGS calls with phased Long Ranger calls to create a comparison set. Using the phased Long Ranger calls as the gold standard, we found that requiring at least 91% of linked alleles to be from the same haplotype before phasing a mutation led to an optimal balance of precision (0.997) and recall (0.936) (Supplementary Figure 4a) (see **Methods**). Overall, 79.4% (16,440/20,705 mutations) of

somatic mutations with enough coverage were phased using that cutoff. Overall, the linked alleles and barcodes phasing methods were concordant on 99.95% of phasing decisions where both methods made a phasing decision (H1 or H2) (9,541/9,546 calls) (Fig. 2b). The barcodes approach added 5,760 calls where linked alleles did not have enough coverage or did not meet the phasing threshold. The linked alleles approach added 1,139 calls. See Supplementary Figure 4b for an overview of all results by phasing method.

We sought to contextualize the phasing performance of our simple heuristics focused on known somatic mutations within the broader landscape of genome-wide variant phasing software tools. We intersected variant phasing results reported by three tools (Long Ranger (v2.2.2), WhatsHap[49] (v1.1), and HapCUT2[11] (v1.3)) (see **Methods**) with our results to compare when each tool made a confident phasing decision. Of 20,705 variants with enough coverage, 34.0% (7,033/20,705 variants) were reported by each tool and were either phased or not phased. Our targeted, heuristic approach limited to known somatic mutations phased 88.2% (6,203/7,033 variants) in that intersection, while WhatsHap phased 59.3% (4,171/7,033 variants), HapCUT2 phased 52.0% (3,656/7,033 variants), and Long Ranger phased 52.0% (3,654/7,033 variants).

Figure 2c highlights seven samples with somatic mutations commonly associated with multiple myeloma, including mutations in *CYLD*, *DIS3*, *HIST1H1E*, *KRAS*, *NRAS*, and *TP53*.[45] In 9 out of 16 examples shown, we confidently phased somatic mutations that were either not called or were not phased by Long Ranger. One mutation in *ATR* was not called by Long Ranger and was not phased by our approach since the linked alleles did not clearly favor one haplotype over the other (60.2% of phased linked alleles supporting the somatic mutation were phased to Haplotype 1, and 39.8% were phased to Haplotype 2). In 27522 (P), the *NRAS* G13R mutation

was phased by our method to Haplotype 2, but was phased to Haplotype 1 in 27522 (Rel). However, since haplotype numbering is arbitrary, such differences are trivial. Further, we noticed that well-known hotspot *NRAS* mutations G13R and Q61K were both phased to the same haplotype in 27522 (P). Later analysis suggested that these two events occurred independently in separate tumor subclones.

We grouped high-confidence somatic mutations by phase block and found the proportion phased by our approach (Fig. 2d). The number of phased somatic mutations per megabase within each phase block showed a log2-normal distribution ranging from 0.10 to 241.3, with a median of 2.25. One application of phasing somatic mutations is establishing the pairwise haplotype relationship with other somatic mutations. Close to half of phase blocks longer than 1 kb had zero pairs of somatic mutations (44.8%, 2,212/4,941 phase blocks), with 11.1% having zero somatic mutations and 33.6% having only one somatic mutation. But among those 2,729 phase blocks longer than 1 kb with at least one pair of somatic mutations, 33.2% had exactly one pair, 18.0% 2-3 pairs, 20.4% 4-10 pairs, 22.3% 11-100 pairs, and the remaining 6.0% had more than 100 pairs. 64.6% of those phase blocks had every mutation phased, and 77.5% had at least 75% of mutations phased.

Pairs of phased somatic mutations illustrate patterns of clonal evolution

In short read sequencing, if two mutant alleles are called together on the same read or read pair, then we can infer they occurred in the same cell and on the same molecule of DNA. With lrWGS, we have the benefit of more linked-reads to consider when we look for such co-occurring mutations. From six samples with lrWGS as well as high-confidence somatic mutation calls and CNV profiles from WGS, we focused on mutations in copy number neutral

regions with coverage between 10 and 100 phased barcodes at that position and one or more barcodes supporting the alternate allele (Supplementary Figure 5a). We examined 59,063 pairs of mutations and, as expected, the probability of one barcode covering both sites decreases as the distance between sites increases, with 98.4% (54,643/55,559 pairs) of mutation pairs located greater than 62 kb apart sharing no overlap. (62 kb is the median of the mean molecule lengths described in Fig. 1d.) Therefore, we focused on the 3,504 mutation pairs within 62 kb. For the 2,648 mutation pairs within this proximity but greater than 100 bp apart, 13.0% did not share any barcodes, 77.3% shared between 1 and 10 barcodes, 8.3% between 11 and 20 barcodes, and 1.4% greater than 20 barcodes (Fig. 3a). For the 856 mutation pairs located less than 100 bp apart, each pair had at least one shared barcode (Supplementary Figure 5b). Overall, 5.9% (3,504/59,063 pairs) of somatic mutation pairs were within 62 Kb (Fig. 3b). Of those, 90.2% (3,159/3,504 pairs) share at least one barcode in common, and, of those, 64.6% (2,042/3,159 pairs) have a barcode on which one or both somatic mutations is represented, potentially enabling direct observation of mutation patterns in the same cell.

We then considered the observed pairwise relationship of each reference and alternate allele on barcodes covering the two somatic sites (Fig. 3c). Of the 2,042 remaining mutation pairs, most (53.3%) share barcodes that only support either both reference alleles (REF-REF) or both alternate alleles (ALT-ALT). This means they have at least one barcode where both alleles are REF and at least one barcode where both alleles are ALT. Other observed patterns are less common, but include REF-REF with REF-ALT or ALT-REF, in which there is at least one barcode supporting one of the alternate alleles but not both. 6.7% of pairs show barcodes supporting REF-ALT and ALT-REF. In these cases, if the two alternate alleles are phased to the same haplotype in a copy number neutral context, this could indicate that the two mutations occurred on the same haplotype but in different cells. Finally, 7.1% of pairs have a pattern of

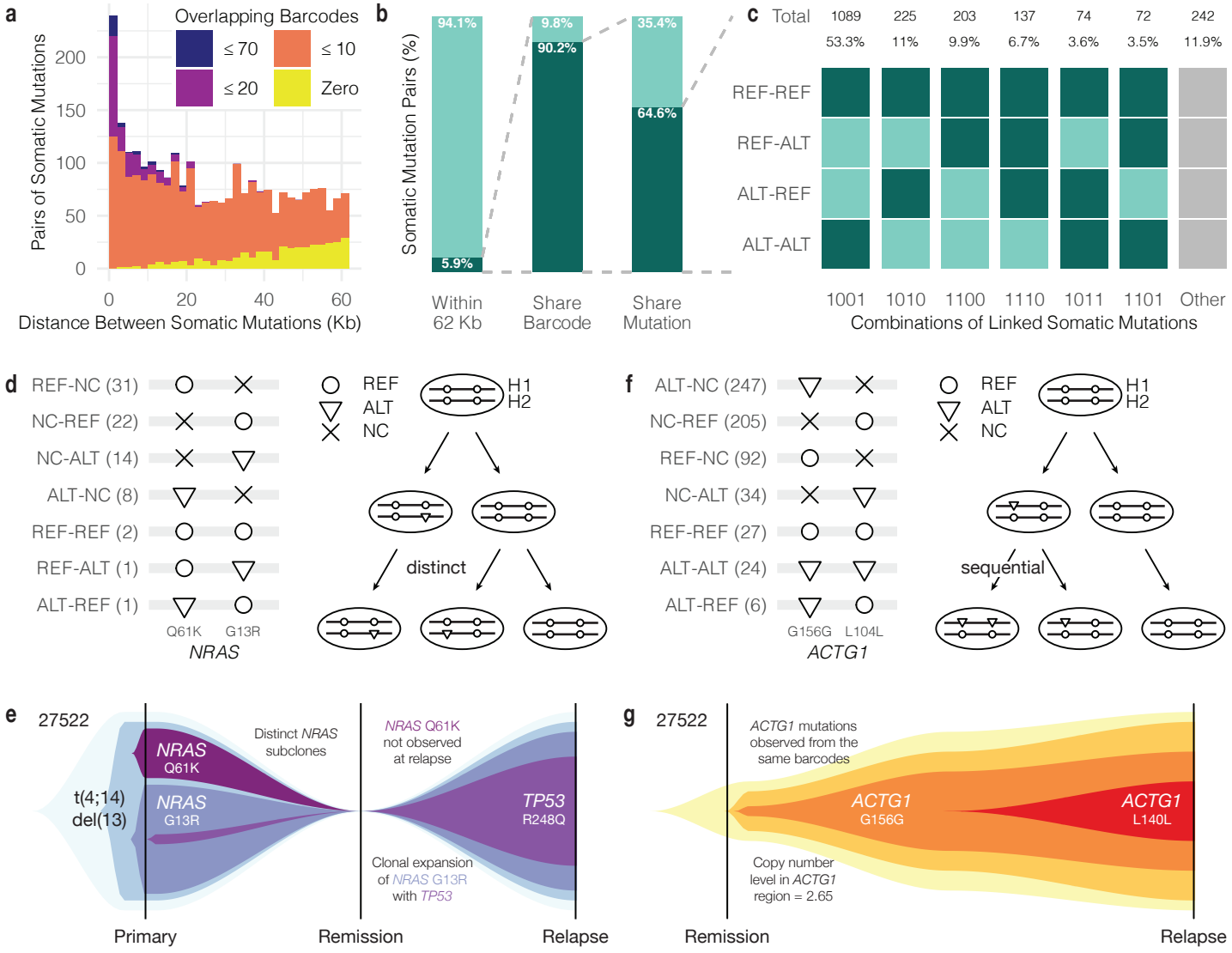


Figure 3. Tumor evolution models derived from mutation pairs. a. Number of overlapping barcodes by distance between somatic mutations. b. Proportion of somatic mutation pairs in close proximity sharing barcodes and mutations. c. Patterns of mutation pairs observed on barcodes (REF = reference allele; ALT = alternate allele). A dark green square indicates that a barcode with that pattern of two alleles was observed. Combinations of patterns can interpreted as evidence of sequential (e.g. 1101, 1011) or distinct (e.g. 1110) mutations. d. NRAS mutation pair observed in 27522 (P) and evolution model (NC = no coverage). e. Interpretation of evolution model observed from NRAS mutation pair in 27522 (P). f. ACTG1 mutation pair observed in 27522 (Rel) and evolution model. g. Interpretation of evolution model observed from ACTG1 mutation pair in 27522 (Rel).

306 REF-ALT or ALT-REF along with ALT-ALT, suggesting a pattern of sequential mutation events.
307 With greater tumor purity, we would expect to see a higher proportion of informative allele
308 patterns with the potential to inform patient-specific models of tumor evolution.
309
310 One such example where the pattern of somatic mutations may be informative for refining
311 tumor phylogenies and may have clinical implications came from CD138+ sorted sample 27522
312 (P). We observed two hotspot mutations in *NRAS* (G13R and Q61K) (Fig. 3d). *NRAS* is a known
313 cancer driver oncogene and mutations may lead to dysregulation of the Ras pathway. We
314 phased both mutations to the same haplotype (H2) (Supplementary Figure 6). We observed 2
315 barcodes supporting REF-REF, 1 barcode supporting REF-ALT, and 1 barcode supporting
316 ALT-REF. Based on sorted lrWGS data, the variant allele frequency (VAF) of the G13R mutation
317 was 35.7% and the Q61K VAF was 22.2% at the primary stage. At relapse, the G13R VAF was
318 20.5% and the Q61K mutation was not detected (VAF 0.0%). Such basic VAF calculations
319 must be interpreted within the context of imperfect tumor cell sorting, tumor heterogeneity with
320 subclonal structure, and potential partial copy number loss on the opposite haplotype
321 (Supplementary Figure 3, Supplementary Figure 6). It may be clinically relevant to know if the
322 two mutations occurred independently or in the same subclone even though multiple activating
323 mutations in the same gene are not necessary for clonal expansion. Without the benefit of
324 phasing, one possible interpretation could be that Q61K occurred in the same clone as G13R,
325 and then the double mutant subclone was eliminated after therapy. However, with linked-
326 reads, we directly observed both mutations occurring without the other, and we never
327 observed them together, guiding the interpretation that these mutations occurred
328 independently in separate subclones and that the Q61K subclone was later lost (Fig. 3e).[50]
329

In another instance, we detected a pair of mutations in *ACTG1* (G156 and L104) that may have occurred in sequential order on the same biological haplotype. Six barcodes demonstrate the ALT-REF pattern, with ALT G156 and REF L104, and 24 barcodes had ALT-ALT with both sites mutated (Fig. 3f). Under a parsimonious model in which the same mutation occurs only once, the G156 mutation must have preceded the L104 mutation. Since there are barcodes supporting both mutant alleles simultaneously, the mutations most likely occur within the same cells, and we interpret this to mean the cells with both mutations form a later subclone within the subclone of cells with only the G156 mutation (Fig. 3g). We also noted elevated copy number in this region (estimated to be 2.65). This would often preclude clonality analysis due its effect on the VAF.[51] However, the combination of alleles present on the same barcodes enables us to interpret a sequential order of events.

Oncogenic IGH translocations in myeloma map to specific haplotypes

Multiple myeloma is characterized by recurrent clonal translocations that take advantage of overexpressed IGH locus by dysregulating oncogene expression. Barwick, et al. analyzed 795 newly-diagnosed multiple myeloma patients from the Multiple Myeloma Research Foundation CoMMpass study (NCT01454297) and reported clonal translocations across the cohort, including 16% of patients with t(11;14) impacting *CCND1*, 11% with t(4;14) (*WHSC1*), 3.3% with t(14;16) (*MAF*), 1.1% with t(6;14) (*CCND3*), and 1.0 % with t(14;20) (*MAFB*).[52] In our cohort of 14 patients, we detected common myeloma translocations from lrWGS using the Long Ranger pipeline and as well as from sorted WGS in 9 matching samples and found t(11;14) in 2 patients and t(4;14) in 1 patient (see **Methods**).[53] After selecting high-confidence events reported from sorted WGS, we found supporting evidence from lrWGS barcodes and mapped those events to haplotypes.

From the 9 matched sorted WGS samples, we identified 88 high-confidence translocations (see **Methods**). We then interrogated matching lrWGS data to find barcodes supporting the event. Of those 88 high-confidence events, 20.5% (18/88 events) had at least two barcodes with a read pattern in support of the translocation. This low rate of support may be attributed to most lrWGS samples not being sorted to select for tumor cells. However, of the 18 events with at least two barcodes, the read haplotype assignment of 94.4% (17/18 events) of translocations showed a consistent haplotype assignment, suggested that using high-confidence SV calls from WGS is a robust prior for haplotype mapping of SVs in high purity lrWGS data.

In Patient 27522, 6 out of 7 SVs detected from both Primary and Relapse samples were also detected from WGS (Supplementary Figure 7a). This patient had a t(4;14) event detected at primary diagnosis present later at relapse which juxtaposed the IGH enhancers with *WHSC1* and *FGFR3*, leading to overexpression of both oncogenes (Supplementary Figures 7b, 8a-b). *WHSC1* overexpression in t(4;14) tumors increases dimethylation of H3K36 and broadly dysregulates the myeloma epigenome.[54] The coverage heat map showing where discordant barcodes map on chr4 and chr14 clearly shows the translocation breakpoint within the first intron of *WHSC1* at chr4:1871962 and near *IGHM* on chr14 and also indicates a deletion proximal to the translocation breakpoint on chr14. We then visualized the coverage pattern of barcodes with reads mapping to both chromosomes in a window around the reported t(4;14) breakpoints (Supplementary Figure 7c). The barcode coverage indicates a reciprocal event leading to two new derived chromosomes der(4) and der(14) with reads from barcodes supporting t(4;14) arbitrarily assigned to H2 on both chromosomes. A pair of events in 27522, t(6;17) and t(4;6), showed similar breakpoints on chr6, approximately 14 kb apart. However, we

did not observe convincing evidence of barcodes with a read coverage pattern linking the three chromosomes, supporting the interpretation that these events occurred independently.

For 77570 (P), Long Ranger reported two t(11;14) events affecting different regions of IGH but with the same breakpoint upstream of *CCND1* (Supplementary Figure 7d-e, 8c-f). One event linked the IGH variable gene region (chr14:106269142) to *CCND1* on chr11. The other at chr14:105741942 linked the coding region of *IGHG1* to the same *CCND1* breakpoint. Barcode coverage analysis suggests these two reported events may actually be one complex reciprocal event with a t(11;14) translocation and deletion on chr14 giving the observed pattern of read coverage upstream and downstream of each breakpoint (Supplementary Figure 7f).

One application of translocation mapping is matching allele-specific expression to translocation events, for example if a germline heterozygous coding variant from the same haplotype of the dysregulating translocation were detected from RNA-seq, then the connection between translocation and expression could be made more explicitly.

Shared germline variants from matched samples enable phase block extension

Phase block boundaries may differ between samples originating from the same patient. However, samples from the same patient do share germline variants, and those germline variants should be phased together in the same groups in both samples.[55] In contrast to previous sections in which somatic mutations from the same sample and same phase block were analyzed together, by comparing the phase of germline variants from overlapping phase blocks from two samples, we can determine if the two phase blocks are oriented the same way, or if one needs to be flipped for them to be consistent. We compared germline variants

from overlapping phase blocks found in two samples, the target sample and the reference sample (Fig. 4a) (see **Methods**). If the shared germline variants were consistently assigned to the same haplotype, the target and reference phase blocks have the same orientation. If they were consistently assigned to opposite haplotypes, they have opposite orientation and the target needs to be switched. If two target phase blocks overlap the same reference phase block, then we can infer the haplotype orientation of the target phase blocks. However, if a switch error occurs in one phase block, that error will propagate as phase blocks are extended.

We analyzed data from 6 patients having multiple tumor samples, with a total of 68,374 overlapping phase blocks from 26 target and reference sample pairs. For example, we examined phase blocks originating from chr1 of 27522 (P) and 27522 (Rel), using 27522 (P) as the reference sample (bottom) and 27522 (Rel) as the target sample (top) (Fig. 4b). Reference phase block 1 (R1) (colored blue) spans multiple target phase blocks (T1-T7). For T1, T2, T3, and T5, there are not enough overlapping variants to draw conclusions about their orientation relative to R1. Phase blocks T4 and T7 must be switched in order to be consistent with R1, and T6 is already consistent with R1. Since T4 and T7 have the same orientation relative to R1, they have the same haplotype orientation and do not need to be switched. However, T6 must be switched to be consistent with T4 and T7. By grouping disconnected phase blocks together, we increase the number of pairs of loci with known haplotype orientation.

In general, at least 10 overlapping phased variants are required before making a switch or no switch recommendation (Fig. 4c). Since the number of shared variants is correlated with the length of the overlap, the length of overlap tends to be greater than 100 kb before a recommendation can be made (Fig. 4d). We were not surprised to find roughly equal proportions of recommendations to switch (28.3%) and not switch (27.6%) given that

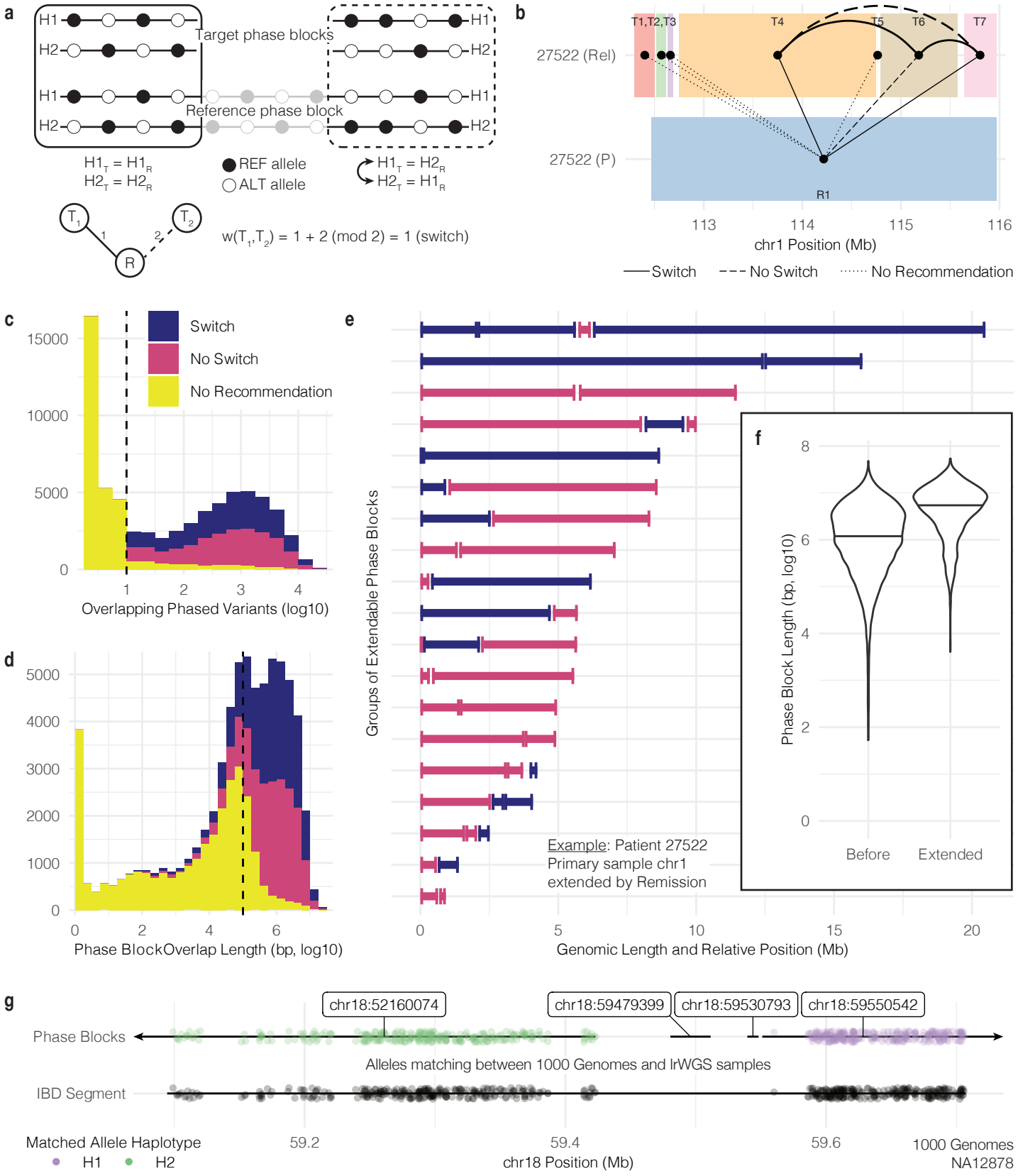


Figure 4. Extension of phase blocks using additional sample information. a. Model for phase block extension using overlap between target and reference phase blocks. b. Data-driven example of phase block overlap between samples. c. Number of phased variants needed for switch/no switch recommendation. d. Length of phase block overlap needed for switch/no switch recommendation. e. Phase block groups extended by overlap with another sample. f. Distribution of phase block lengths before and after extension. Violin plots defined as: center line, median; violin limits, minimum and maximum values; individual points not shown. g. Use of identity-by-descent segments as overlap between phase blocks.

haplotype numbering is random. For the remaining 44.1% of cases, the algorithm was not able to make a recommendation to switch or not switch. For extendable phase blocks from chr1 in target sample 27522 (P) (extended by reference 27522 (Rem)), we found that, before extension, the median phase block length was 1.6 Mb, and after extension, it was 5.7 Mb, a 3.5-fold increase (Fig. 4e). Similarly, from all samples with extendable phase blocks, we found that median phase block length increased from 1.2 Mb (6.1 on log10 bp scale) to 5.5 Mb (6.7 on log10 bp scale), a 4.6 fold increase from before extension to after extension (Fig. 4f).

We also developed methods to leverage publicly available population-scale phased data to learn more about the origin of haplotypes present in our cohort and to improve our lrWGS results. After reporting identical-by-descent (IBD) segments shared between 2,504 individuals from 1000 Genomes data (see **Methods**; see **Data availability**), we identified IBD segments overlapping multiple lrWGS phase blocks in NA12878.[56] Using phased heterozygous variants shared between the 1000 Genomes VCF of this sample and the VCF output from Long Ranger, we found the proportion of IBD alleles matching each haplotype in each phase block. IBD alleles consistently matched one haplotype or the other with the occasional short switch error. For example, NA12878 shares an IBD segment with NA10851 from position 59,094,547 to 59,706,930 on chr18 (LOD score 15.64, 1.576 cM) (Fig. 4g). That IBD segment bridges multiple lrWGS phase blocks. Since the IBD alleles match Haplotype 2 from phase block chr18:52160074 and match Haplotype 1 from chr18:595505042, those two phase blocks may be in opposite orientation.

Discussion

As sequencing technologies evolve and analysis methods more regularly include haplotype phasing, somatic mutation phasing will become a more common practice. The current methodological approaches to haplotype-aware somatic mutation analysis will mature from *ad hoc* investigations to standard pipelines. We have developed a systematic approach to somatic mutation analysis in a cohort of multiple myeloma patients over the course of disease. Our methods build on the backbone of the Long Ranger variant calling and phasing pipeline for linked-read sequencing data. These methods are an opportunity for future development in a climate of rapid technological advances with many applications. We need better understanding of the haplotypes carrying germline variants related to predisposition of many diseases, including cancer, as well as better methods to identify ancestry-specific risk modifiers.[57-63] Biallelic *TP53* inactivation indicates poor prognosis in multiple myeloma[64], and double *PIK3CA* mutations on the same haplotype can be more oncogenic but also more susceptible to PI3Ka inhibitors.[65] Other medical applications of linked-read sequencing include more sensitive prenatal diagnosis [66,67], better predictions about how protein structure may change in response to multiple mutations [68], and more accurate neoepitope prediction.[69] Tools such as HAPDeNovo capitalize on haplotype structures from linked-reads to eliminate false-positive from studies of rare, *de novo* variation.[70]

We noted several limitations in our analysis potentially due to data generation. We observed shorter phase blocks in our skin normal controls samples potentially due to lower input molecule size or sequencing depth. For our somatic analyses, an important caveat was controlling for copy number changes which disrupt the strict two haplotype paradigm of variant phasing. Another limitation of our somatic analysis was low tumor purity. Only 4 of our 23 samples were CD138+ sorted, and two samples in particular gave us the most confident results. Higher tumor purity and lower variability in cell-type composition are likely important for

robust somatic variant haplotype analysis. Calling somatic mutations with low variant allele frequency is a challenge for any mutation caller, especially those like Long Ranger built for germline variant detection. In our case, pairing linked-read data with high-confidence somatic mutation calls from a separate WGS sample was necessary to gain sensitivity. Future analyses using lrWGS in multiple myeloma should include analysis of chromoplexy and chromothripsis as these complex events are important in MM pathogenesis but cannot be fully appreciated using short reads.[71] Additionally, long-range PCR of known somatic variant regions could validate the phasing performance and data interpretations enabled by our framework.

Moving beyond next-generation sequencing to Third Generation and single-cell approaches holds the promise of increased resolution in cancer genome analyses.[72-75] With long reads and linked-reads, we get haplotype resolution. With single-cell RNA-seq, we observe cell-specific patterns of gene expression and copy number and can map coding mutations to specific cells.[76] Single-cell DNA sequencing analyses, including approaches that incorporate haplotypes, offer even deeper resolution of tumor evolution and the ability to optimize treatment strategies.[18,77-84] Methodological integration of single-cell data with the resolution gained from haplotype analysis is a direction for continued research.

Methods

SomaticHaplotype modules

Our framework of interconnected modules builds on the phased bam and variant output files from the Long Ranger pipeline (10X Genomics). Our analysis code was written in python and is freely available under the MIT license (see **Code availability**). Additional inputs to our pipeline may include high-confidence somatic mutation calls and identity-by-descent segments. There are five analysis modules: *phaseblock*, *summarize*, *somatic*, *extend*, and *ancestry*.

Phaseblock: The *phaseblock* module is the first module run on any new data. The inputs are a phased bam and phased variant call format (VCF) file from the Long Ranger pipeline. Given a genomic range of interest, such as an entire chromosome, *phaseblock* constructs PhaseBlock and Variant objects by extracting information from reads and variant calls. PhaseBlock objects collect information about variants common to phase blocks identified by Long Ranger. Variant objects store information about small variants, including genotype and phase, and map to specific PhaseBlock objects based on tags given by Long Ranger. Each Variant also stores the barcodes of reads supporting the reference and alternate allele at that position. The objects are designed with methods for later utility. Dictionaries referencing those objects are stored in an output file used as input to downstream modules.

Summarize: The *summarize* module takes input from *phaseblock* and produces a summary of each phase block and a global summary about phase block lengths. Output from *summarize* is used as input to *somatic* and *extend*.

Somatic: The *somatic* module collects barcode and haplotype information supporting somatic mutation sites. Somatic mutation sites are defined using an input parameter, either as a mutation annotation format (MAF) file or list of mutations. Barcodes supporting somatic mutation sites are extracted by a separately-run submodule called 10Xmapping which mines the bam for reads supporting the mutation site and also from the VCF if the mutation was called by Long Ranger. The 10Xmapping submodule identifies bam reads supporting the reference and alternate alleles at a somatic mutation site and gathers barcode and haplotype information from each read. 10Xmapping is contained as a submodule of in our repository but is also freely available at https://github.com/ding-lab/10Xmapping. Output from *somatic* includes information about every (germline and somatic) variant from barcodes overlapping each somatic mutation site, information necessary for phasing each somatic mutation, barcode sharing analysis of each pair of somatic mutations, and somatic mutation summaries for each phase block.

In later analysis, users interpret output from *somatic* to decide if somatic mutations are phased or not. For example, we combined two approaches to determine the phase of each somatic mutation. In our “linked alleles” approach, we analyzed the proportion of linked alleles mapping to a particular haplotype and found 0.91 (and above) to be an appropriate threshold that balanced phasing decision precision and recall. We combined that with the “barcodes” approach, which relies on the reported haplotype assignment of reads supporting the somatic mutation. We determined a somatic mutation to be phased if at least one barcode supported the mutant allele and all barcodes supporting the mutant allele agreed on the haplotype assignment. For pairs of somatic mutations, the barcode sharing analysis finds barcodes with reads mapping to both somatic mutation sites. For each barcode, the alleles supporting each site are combined as allele pairs (REF-REF, REF-ALT, ALT-REF, and ALT-ALT).

546

547 *Extend*: The extend module combines germline variants from two related samples (e.g. from
548 the same individual) to determine the haplotype orientation between disconnected phase
549 blocks in one of the samples. Once the haplotype orientation between two phase blocks is
550 determined, the phase blocks can be conceptually extended. The two samples are defined as
551 the “target” (with phase blocks to be extended) and the “reference”, which the target is
552 compared against. To determine if the target and reference phase blocks have the same or
553 different haplotype orientation, *extend* compares the haplotype assignments of overlapping
554 germline variants and finds the proportion of target haplotype assignments that need to be
555 switched in order to be consistent with the reference. *Extend* uses a two-sided binomial test
556 (significant number of “switch” or “not switch” given a conservative switch error rate) and a
557 hard cutoff (more than 95% “switch” or less than 5% “switch”) to determine if the target and
558 reference phase blocks have the same or opposite orientation. Then *extend* module then builds
559 a bipartite graph in which nodes are phase blocks and edges connect overlapping target and
560 reference sample phase blocks. Edge weights are defined as 1 if a switch is necessary
561 between the target and reference phase block or 2 if a switch is not necessary. If two target
562 phase blocks overlap the same reference phase block, then there is a connected path between
563 the target phase blocks and we find the sum of the weighted edges connecting them. If the
564 sum (mod 2) is zero, then the two target phase blocks have the same orientation. If the sum
565 (mod 2) is one, then they have opposite orientation. *Extend* output describes the overlap of
566 each target phase block with reference phase blocks and also forms groups of connected
567 target phase blocks that may be extended via this method.

568

569 *Ancestry*: The *ancestry* module uses a similar concept to *extend* but instead relies on output
570 from an identity-by-descent tool such as Refined IBD instead of phase blocks from a related

sample.[56] By examining the haplotype assignment of alleles from overlapping IBD segments and phase blocks defined by Long Ranger, *ancestry* may bridge gaps between phase blocks and find where phase block haplotype orientations are congruent or not. *Ancestry* also assigns population history to portions of phase blocks that overlap IBD segments.

Generation of linked-read whole genome sequencing data

The 10X Genomics Chromium System generates linked-read sequencing data. From a bulk sample of cells, long fragments of DNA, also called high-molecular weight (HMW) DNA, are isolated into an individual gel bead in emulsion (GEM). Each GEM contains a gel bead with primers including a 16-bp DNA barcode unique to that GEM. The gel bead dissolves and releases the barcoded primers, which attach to the DNA and undergo isothermal amplification. Now each short fragment of amplified DNA contains a barcode identifying which GEM it originated from. The GEMs break and the barcoded fragments are pooled together and sequenced.

Patient cohort

Fourteen (10 male, 4 female) patients with multiple myeloma were included in the analysis. The median age at diagnosis was 63 (range 46-69). Eight patients had IgG isotype (4 kappa and 4 lambda), 2 had IgA kappa isotype, 2 had light chain only disease (1 kappa and 1 lambda), and 2 were non-secretory. Five were International Staging System Stage I, 2 were Stage II, 3 were stage III, and 4 were unreported. The median plasma cell burden by flow cytometry in bone marrow at diagnosis was 24% (range 4-63). By standard fluorescence in situ hybridization (FISH), 1 patient had t(4;14), 3 had t(11;14), and 2 showed del(17p). A total of 23 samples were

596 collected from multiple disease stages, including smoldering multiple myeloma (SMM), primary
597 diagnosis, pre- and post-transplant, remission, and relapse.

598 **Sample collection and data generation**

599 Research bone marrow aspirate samples were collected at the time of the diagnostic
600 procedure. Bone marrow mononuclear cells (BMMCs) were isolated using Ficoll-Paque.
601 BMMCs were cryopreserved in a 1:10 mixture of dimethyl sulfoxide and fetal bovine serum.
602 Upon thawing, whole BMMCs were used for linked-read whole genome sequencing. Plasma
603 cells were separated from a sub-aliquot by positive selection using CD138-coated magnetic
604 beads in an autoMACs system (Miltenyi Biotec, CA) and used for whole genome and exome
605 sequencing. Skin punch biopsies were performed at the time of the diagnostic bone marrow
606 collection to serve as normal controls. Although many studies use peripheral blood
607 mononuclear cells (PBMCs) as a control, abnormal B cells and circulating tumor cells
608 frequently contaminate the peripheral blood of patients with multiple myeloma. Therefore,
609 using PBMCs may lead to the omission of genetic events potentially important in disease
610 pathogenesis.

611 *Linked-read whole genome sequencing (lrWGS).* Normal skin samples were processed with a
612 standard Qiagen DNA isolation kit resulting in 10-50Kb DNA fragments. 250K tumor cells were
613 processed with the MagAttract HMW DNA extraction kit (Qiagen) resulting in 100-150Kb DNA
614 fragments. 600-800ng of normal DNA was size selected on the Blue Pippin utilizing the 0.75%
615 Agarose Dye-Free Cassette to attempt to remove low molecular weight DNA fragments. The
616 size selection parameters were set to capture 30-80 Kb DNA fragments (Sage Science). The
617 resulting size selected DNA from the normal samples and the HMW DNA from the tumor cells
618 were diluted to 1ng/μL prior to the v2 Chromium Genome Library prep (10X Genomics).

Approximately 10-15 DNA molecules were encapsulated into nanoliter droplets. DNA molecules within each droplet were tagged with a 16 nucleotide barcode and 6 nucleotide unique molecular identifier during isothermal incubation. The resulting barcoded fragments were converted into a sequence ready Illumina library with an average insert size of 500bp. The concentration of each library was accurately determined through qPCR (Kapa Biosystems) to produce cluster counts appropriate for sequencing on the HiSeqX/NovaSeq6000 platform (Illumina). 2x150 sequence data were generated targeting 30x (normal) and 60x (tumor) coverage providing linked-reads across the length of individual DNA molecules.

Standard whole genome sequencing (WGS). Manual libraries were constructed with 50-2000ng of genomic DNA utilizing the Lotus Library Prep Kit (IDT Technologies) targeting 350bp inserts. Strand-specific molecular indexing is a feature associated with this library method. The molecular indexes are fixed sequences that make up the first 8 bases of read 1 and read 2 insert reads. The concentration of each library was accurately determined through qPCR (Kapa Biosystems). 2x150 paired-end sequence data generated ~200 Gb per tumor sample leading to 60x (tumor) haploid coverage.

lrWGS data processing with Long Ranger

Long Ranger (10X Genomic) performs linked-read alignment, variant calling, and variant phasing. We ran Long Ranger (v2.2.2) to align reads to the human genome reference GRCh38 (GRCh38-2.1.0) and used --vcmode with GATK[85,86] (version 3.7.0-gcfedb67) for variant calling. Long Ranger also produces quality metrics associated with each sample. Publicly-available 1000 Genomes lrWGS samples were processed with Long Ranger (version 2.2.1) and aligned to hg19.

lrWGS data processing with other tools

In addition to Long Ranger, we used WhatsHap[49] (v1.1) and HapCUT2[11] (v1.3) to phase our linked read WGS samples using human genome reference GRCh38 (GRCh38-2.1.0). We applied the additional extractHAIRS and LinkFragments steps to prepare our 10X data for use by HapCUT2.

High-confidence somatic mutation detection

Somatic mutations were called by our SomaticWrapper pipeline, which includes four established bioinformatic tools, namely Strelka[87], Mutect[88], VarScan2[89] (2.3.83), and Pindel[90] (0.2.54). We retained SNVs and INDELs using the following strategy: keep SNVs called by any 2 callers among Mutect, VarScan, and Strelka and INDELs called by any 2 callers among VarScan, Strelka, and Pindel. For these merged SNVs and INDELs, we applied coverage cut-offs of 14X and 8X for tumor and normal, respectively. We also filtered SNVs and INDELs with a high-pass variant allele fraction (VAF) of 0.05 in tumor and a low-pass VAF of 0.02 in normal. The SomaticWrapper pipeline is freely available at https://github.com/ding-lab/somaticwrapper.

Copy number profiling

We used BIC-seq2[91], a read-depth-based CNV calling algorithm to detect somatic copy number variations (CNVs) using standard WGS tumor samples and paired skin linked-read WGS data. The procedure involves 1) retrieving all uniquely mapped reads from the tumor and paired skin BAM files, 2) removing biases by normalization (NBICseq-norm_v0.2.4) 3) detecting CNV based on normalized data (NBICseq-seg_v0.7.2) with BIC-seq2 parameters set as --lambda=90 --detail --noscale --control. We defined copy number neutral regions as having a $\log_2$ copy number ratio between -0.25 and 0.2 in the sorted WGS.

665

666 **Tumor purity estimation**

667

668 We used the R package sciClone[51] (v1.1.0) to estimate tumor purity based on clusters detected
669 using variants from copy number neutral regions. We designated the cluster with the greatest
670 median variant allele frequency (VAF) as the founding clone and doubled its VAF to estimate
671 the sample's tumor purity.

672

673 **Structural variant detection**

674

675 Somatic structural variants (SVs) were detected by Manta[53] using tumor/normal sample pairs of
676 standard WGS and paired skin linked-read WGS. SVs were filtered according to the following
677 guidelines. Record-level filters included a QUAL score < 20; somatic variant quality score < 30;
678 depth greater than 3x the median chromosome depth near one or both variant breakpoints; for
679 variants significantly larger than the paired read fragment size, no paired reads support the
680 alternate allele in any sample. Sample-level filters included a Genotype Quality < 15. This
681 approach optimizes the analysis of somatic variation in tumor/normal sample pairs. In addition
682 to the built-in Manta filters (labeled as PASS), we further prioritized the high-confidence variants
683 by (1) the number of support spanning read pairs >= 5; (2) the coverage at the given
684 breakpoints > 10; (3) events must involve only autosomes and/or sex chromosomes; (4) events
685 passing manual IGV review on the read evidence.

686

687 We also used gemtools[33] (https://github.com/sgreer77/gemtools) and the python package
688 pysam (0.15.3) with samtools[92] (v1.9) to identify reads and barcodes supporting SVs in lrWGS.

689

690 **Identity-by-descent reporting**

691

692 We obtained phased haplotype information for 2,504 individual from the 1000 Genomes and
693 ran Refined IBD with default parameters (refined-ibd.16May19.ad5) (see **Data availability**).[56]

694

695 **Data availability**

696

697 The Washington University Institutional Review Board approved the study protocol, and all
698 relevant ethical regulations, including obtaining informed consent from all participants, were
699 followed. Patients were treated and sampled at Washington University in St. Louis.

700

701 All data and scripts necessary to recreate figures are available at
702 doi.org/10.6084/m9.figshare.12295922.

703

704 Publicly-available 1000 Genomes lrWGS samples can be downloaded from
705 https://support.10xgenomics.com/genome-exome/datasets/2.2.1/NA12878_WGS_v2 and
706 https://support.10xgenomics.com/genome-exome/datasets/2.2.1/NA19240_WGS_v2.

707

708 Phased 1000 Genomes VCFs (2,504 samples) were downloaded from
709 http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

710

711 The remaining data and methods are available in the Article, Supplementary Tables, or are
712 available from the author upon reasonable request.

713

714 **Code availability**

SomaticHaplotype is freely-available at https://github.com/ding-lab/SomaticHaplotype.

References

1 Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-1097, doi:10.1086/521987 (2007).

2 Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153-159, doi:10.1093/bioinformatics/btn298 (2008).

3 Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**, 210-223, doi:10.1016/j.ajhg.2009.01.005 (2009).

4 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

5 Snyder, M. W., Adey, A., Kitzman, J. O. & Shendure, J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat Rev Genet* **16**, 344-358, doi:10.1038/nrg3903 (2015).

6 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

7 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81, doi:10.1038/nature15394 (2015).

8 Mostovoy, Y. *et al.* A hybrid approach for de novo human genome sequence assembly and phasing. *Nat Methods* **13**, 587-590, doi:10.1038/nmeth.3865 (2016).

9 Porubsky, D. *et al.* Direct chromosome-length haplotyping by single-cell sequencing. *Genome Res* **26**, 1565-1574, doi:10.1101/gr.209841.116 (2016).

10 Loh, P. R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443-1448, doi:10.1038/ng.3679 (2016).

11 Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res* **27**, 801-812, doi:10.1101/gr.213462.116 (2017).

12 Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res* **27**, 757-767, doi:10.1101/gr.214874.116 (2017).

13 Zhang, F. *et al.* Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. *Nat Biotechnol* **35**, 852-857, doi:10.1038/nbt.3897 (2017).

14 Bell, J. M. *et al.* Chromosome-scale mega-haplotypes enable digital karyotyping of cancer aneuploidy. *Nucleic Acids Res* **45**, e162, doi:10.1093/nar/gkx712 (2017).

15 Porubsky, D. *et al.* Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat Commun* **8**, 1293, doi:10.1038/s41467-017-01389-4 (2017).

16 Choi, Y., Chan, A. P., Kirkness, E., Telenti, A. & Schork, N. J. Comparison of phasing strategies for whole human genomes. *PLoS Genet* **14**, e1007308, doi:10.1371/journal.pgen.1007308 (2018).

17 Martin, A. R. *et al.* Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland. *Am J Hum Genet* **102**, 760-775, doi:10.1016/j.ajhg.2018.03.003 (2018).

18 Satas, G. & Raphael, B. J. Haplotype phasing in single-cell DNA-sequencing data. *Bioinformatics* **34**, i211-i217, doi:10.1093/bioinformatics/bty286 (2018).
19 Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet* **103**, 338-348, doi:10.1016/j.ajhg.2018.07.015 (2018).
20 Belsare, S. *et al.* Evaluating the quality of the 1000 genomes project data. *BMC Genomics* **20**, 620, doi:10.1186/s12864-019-5957-x (2019).
21 Bansal, V. Integrating read-based and population-based phasing for dense and accurate haplotyping of individual genomes. *Bioinformatics* **35**, i242-i248, doi:10.1093/bioinformatics/btz329 (2019).
22 Yan, Z. *et al.* scHaplotyper: haplotype construction and visualization for genetic diagnosis using single cell DNA sequencing data. *BMC Bioinformatics* **21**, 41, doi:10.1186/s12859-020-3381-5 (2020).
23 Berger, E. *et al.* Improved haplotype inference by exploiting long-range linking and allelic imbalance in RNA-seq datasets. *Nat Commun* **11**, 4662, doi:10.1038/s41467-020-18320-z (2020).
24 Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat Commun* **10**, 5436, doi:10.1038/s41467-019-13225-y (2019).
25 Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics. *Nat Rev Genet* **12**, 215-223, doi:10.1038/nrg2950 (2011).
26 Cooke, D. P., Wedge, D. C. & Lunter, G. A unified haplotype-based method for accurate and comprehensive variant calling. *Nat Biotechnol* **39**, 885-892, doi:10.1038/s41587-021-00861-3 (2021).
27 Zheng, G. X. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34**, 303-311, doi:10.1038/nbt.3432 (2016).
28 Marks, P. *et al.* Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res* **29**, 635-645, doi:10.1101/gr.234443.118 (2019).
29 Spies, N. *et al.* Genome-wide reconstruction of complex structural variants using read clouds. *Nat Methods* **14**, 915-920, doi:10.1038/nmeth.4366 (2017).
30 Elyanow, R., Wu, H. T. & Raphael, B. J. Identifying structural variants using linked-read sequencing data. *Bioinformatics* **34**, 1-16, doi:10.1093/bioinformatics/btx712 (2017).
31 Xia, L. C. *et al.* Identification of large rearrangements in cancer genomes with barcode linked reads. *Nucleic Acids Res* **46**, e19, doi:10.1093/nar/gkx1193 (2018).
32 Luo, R., Sedlazeck, F. J., Darby, C. A., Kelly, S. M. & Schatz, M. C. LRSim: A Linked-Reads Simulator Generating Insights for Better Genome Partitioning. *Comput Struct Biotechnol J* **15**, 478-484, doi:10.1016/j.csbj.2017.10.002 (2017).
33 Greer, S. U. & Ji, H. P. Structural variant analysis for linked-read sequencing data with gemtools. *Bioinformatics* **35**, 4397-4399, doi:10.1093/bioinformatics/btz239 (2019).
34 Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**, 1784, doi:10.1038/s41467-018-08148-z (2019).
35 Rodriguez, O. L., Ritz, A., Sharp, A. J. & Bashir, A. MsPAC: a tool for haplotype-phased structural variant detection. *Bioinformatics* **36**, 922-924, doi:10.1093/bioinformatics/btz618 (2020).
36 Fang, L. *et al.* LinkedSV for detection of mosaic structural variants from linked-read exome and genome sequencing data. *Nat Commun* **10**, 5585, doi:10.1038/s41467-019-13397-7 (2019).
37 Karaoglanoglu, F. *et al.* VALOR2: characterization of large-scale structural variants using linked-reads. *Genome Biol* **21**, 72, doi:10.1186/s13059-020-01975-8 (2020).

38 Nordlund, J. *et al.* Refined detection and phasing of structural aberrations in pediatric acute lymphoblastic leukemia by linked-read whole genome sequencing. *bioRxiv*, 375659, doi:10.1101/375659 (2018).
39 Greer, S. U. *et al.* Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. *Genome Med* **9**, 57, doi:10.1186/s13073-017-0447-8 (2017).
40 Ha, G. *et al.* TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res* **24**, 1881-1893, doi:10.1101/gr.180281.114 (2014).
41 Viswanathan, S. R. *et al.* Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. *Cell* **174**, 433-447 e419, doi:10.1016/j.cell.2018.05.036 (2018).
42 Sereewattanawoot, S. *et al.* Identification of potential regulatory mutations using multi-omics analysis and haplotyping of lung adenocarcinoma cell lines. *Sci Rep* **8**, 4926, doi:10.1038/s41598-018-23342-1 (2018).
43 Zhou, B. *et al.* Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res* **29**, 472-484, doi:10.1101/gr.234948.118 (2019).
44 Zhou, B. *et al.* Haplotype-resolved and integrated genome analysis of the cancer cell line HepG2. *Nucleic Acids Res* **47**, 3846-3861, doi:10.1093/nar/gkz169 (2019).
45 Manier, S. *et al.* Genomic complexity of multiple myeloma and its clinical implications. *Nat Rev Clin Oncol* **14**, 100-113, doi:10.1038/nrclinonc.2016.122 (2017).
46 Darby, C. A. *et al.* Samovar: Single-Sample Mosaic Single-Nucleotide Variant Calling with Linked Reads. *iScience* **18**, 1-10, doi:10.1016/j.isci.2019.05.037 (2019).
47 Zhang, L., Zhou, X., Weng, Z. & Sidow, A. Assessment of human diploid genome assembly with 10x Linked-Reads data. *Gigascience* **8**, doi:10.1093/gigascience/giz141 (2019).
48 Buhler, S. *et al.* High-resolution HLA phased haplotype frequencies to predict the success of unrelated donor searches and clinical outcome following hematopoietic stem cell transplantation. *Bone Marrow Transplant* **54**, 1701-1709, doi:10.1038/s41409-019-0520-6 (2019).
49 Martin, M. *et al.* WhatsHap: fast and accurate read-based phasing. *bioRxiv*, 085050, doi:10.1101/085050 (2016).
50 Miller, C. A. *et al.* Visualizing tumor evolution with the fishplot package for R. *BMC Genomics* **17**, 880, doi:10.1186/s12864-016-3195-z (2016).
51 Miller, C. A. *et al.* SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol* **10**, e1003665, doi:10.1371/journal.pcbi.1003665 (2014).
52 Barwick, B. G. *et al.* Multiple myeloma immunoglobulin lambda translocations portend poor prognosis. *Nat Commun* **10**, 1911, doi:10.1038/s41467-019-09555-6 (2019).
53 Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-1222, doi:10.1093/bioinformatics/btv710 (2016).
54 Kuo, A. J. *et al.* NSD2 links dimethylation of histone H3 at lysine 36 to oncogenic programming. *Mol Cell* **44**, 609-620, doi:10.1016/j.molcel.2011.08.042 (2011).
55 Greer, S. U., Lau, B. T., Nadauld, L. D. & Ji, H. P. Abstract 1280: Chromosome-scale haplotyping enables comprehensive discovery of cancer rearrangements and germline-related susceptibility mutations. *Cancer Research* **78**, 1280-1280, doi:10.1158/1538-7445.Am2018-1280 (2018).
56 Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459-471, doi:10.1534/genetics.113.150029 (2013).

57 Gudmundsson, J. *et al.* Frequent occurrence of BRCA2 linkage in Icelandic breast cancer families and segregation of a common BRCA2 haplotype. *Am J Hum Genet* **58**, 749-756 (1996).
58 Haiman, C. A. *et al.* A comprehensive haplotype analysis of CYP19 and breast cancer risk: the Multiethnic Cohort. *Hum Mol Genet* **12**, 2679-2692, doi:10.1093/hmg/ddg294 (2003).
59 Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet* **100**, 635-649, doi:10.1016/j.ajhg.2017.03.004 (2017).
60 Huang, K. L. *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* **173**, 355-370 e314, doi:10.1016/j.cell.2018.03.039 (2018).
61 Yuan, J. *et al.* Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers. *Cancer Cell* **34**, 549-560 e549, doi:10.1016/j.ccell.2018.08.019 (2018).
62 Ramroop, J. R., Gerber, M. M. & Toland, A. E. Germline Variants Impact Somatic Events during Tumorigenesis. *Trends Genet* **35**, 515-526, doi:10.1016/j.tig.2019.04.005 (2019).
63 Musa, J. *et al.* Cooperation of cancer drivers with regulatory germline variants shapes clinical outcomes. *Nat Commun* **10**, 4128, doi:10.1038/s41467-019-12071-2 (2019).
64 Walker, B. A. *et al.* A high-risk, Double-Hit, group of newly diagnosed myeloma identified by genomic analysis. *Leukemia* **33**, 159-170, doi:10.1038/s41375-018-0196-8 (2019).
65 Vasan, N. *et al.* Double PIK3CA mutations in cis increase oncogenicity and sensitivity to PI3Kalpha inhibitors. *Science* **366**, 714-723, doi:10.1126/science.aaw9032 (2019).
66 Jang, S. S. *et al.* Targeted linked-read sequencing for direct haplotype phasing of maternal DMD alleles: a practical and reliable method for noninvasive prenatal diagnosis. *Sci Rep* **8**, 8678, doi:10.1038/s41598-018-26941-0 (2018).
67 Mortensen, O. *et al.* Using dried blood spot samples from a trio for linked-read whole-exome sequencing. *Eur J Hum Genet* **27**, 980-988, doi:10.1038/s41431-019-0343-3 (2019).
68 Cuppens, T., Ludwig, T. E., Trouve, P. & Genin, E. GEMPROT: visualization of the impact on the protein of the genetic variants found on each haplotype. *Bioinformatics* **35**, 2492-2494, doi:10.1093/bioinformatics/bty993 (2019).
69 Wood, M. A. *et al.* neoepiscope improves neoepitope prediction with multivariant phasing. *Bioinformatics* **36**, 713-720, doi:10.1093/bioinformatics/btz653 (2020).
70 Zhou, X., Batzoglou, S., Sidow, A. & Zhang, L. HAPDeNovo: a haplotype-based approach for filtering and phasing de novo mutations in linked read sequencing data. *BMC Genomics* **19**, 467, doi:10.1186/s12864-018-4867-7 (2018).
71 Maura, F. *et al.* Genomic landscape and chronological reconstruction of driver events in multiple myeloma. *Nat Commun* **10**, 3835, doi:10.1038/s41467-019-11680-1 (2019).
72 Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet* **19**, 329-346, doi:10.1038/s41576-018-0003-4 (2018).
73 van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends Genet* **34**, 666-681, doi:10.1016/j.tig.2018.05.008 (2018).
74 Pan, W., Gong, D., Sun, D. & Luo, H. HICANCER: accurate and complete cancer genome phasing with Hi-C reads. *Sci Rep* **11**, 6609, doi:10.1038/s41598-021-86104-6 (2021).
75 Jeong, H. *et al.* Haplotype-aware single-cell multiomics uncovers functional effects of somatic structural variation. *bioRxiv*, 2021.2011.2011.468039, doi:10.1101/2021.11.11.468039 (2021).

76 Petti, A. A. *et al.* A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nat Commun* **10**, 3660, doi:10.1038/s41467-019-11591-1 (2019).
77 Ross, E. M. & Markowetz, F. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol* **17**, 69, doi:10.1186/s13059-016-0929-9 (2016).
78 Bohrson, C. L. *et al.* Linked-read analysis identifies mutations in single-cell DNA-sequencing data. *Nat Genet* **51**, 749-754, doi:10.1038/s41588-019-0366-2 (2019).
79 Malikic, S., Jahn, K., Kuipers, J., Sahinalp, S. C. & Beerenwinkel, N. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nat Commun* **10**, 2750, doi:10.1038/s41467-019-10737-5 (2019).
80 Ramazzotti, D., Graudenzi, A., De Sano, L., Antoniotti, M. & Caravagna, G. Learning mutational graphs of individual tumour evolution from single-cell and multi-region sequencing data. *BMC Bioinformatics* **20**, 210, doi:10.1186/s12859-019-2795-4 (2019).
81 Malikic, S. *et al.* PhISCS: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome Res* **29**, 1860-1877, doi:10.1101/gr.234435.118 (2019).
82 Zafar, H., Navin, N., Chen, K. & Nakhleh, L. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome Res* **29**, 1847-1859, doi:10.1101/gr.243121.118 (2019).
83 Dou, Y. *et al.* Accurate detection of mosaic variants in sequencing data without matched controls. *Nat Biotechnol* **38**, 314-319, doi:10.1038/s41587-019-0368-8 (2020).
84 Zaccaria, S. & Raphael, B. J. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nat Biotechnol*, doi:10.1038/s41587-020-0661-6 (2020).
85 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
86 Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, 201178, doi:10.1101/201178 (2018).
87 Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591-594, doi:10.1038/s41592-018-0051-x (2018).
88 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219, doi:10.1038/nbt.2514 (2013).
89 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).
90 Ye, K. *et al.* Systematic discovery of complex insertions and deletions in human cancers. *Nat Med* **22**, 97-104, doi:10.1038/nm.4002 (2016).
91 Xi, R., Lee, S., Xia, Y., Kim, T. M. & Park, P. J. 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, doi:10.1093/nar/gkw491 (2016).
92 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

Acknowledgements

This work has been supported by the Paula C. and Rodger O. Riney Blood Cancer Research Initiative Fund to L.D. and R.V. and NCI U24CA211006 and U2CCA233303 funds to L.D.

Author contributions

L.D. and R.V. led project design. S.M.F. led tool development, performed data analysis, wrote manuscript, generated figures. N.V.T., Y.L., Q.G., L.Y., and H.S. ran tools for alignment, mutation, SV, and CNV calling. A.W., Q.G., G.D., M.S., S.C., and R.G.J. contributed to tool development. R.S.F. and C.C.F. led sequencing data generation. J.K., D.R.K., and M.A.F. managed in-house sample collection. R.G.J., K.C., J.F.D., R.V., and L.D. reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Figure legends

Figure 1. Linked-read data generation and analysis pipeline. a. The 10X Genomics Chromium platform tags large DNA molecules with barcodes such that reads originating from the same molecule have the same barcode. The Long Ranger pipeline aligns reads and phases variants. **b.** SomaticHaplotype builds upon Long Ranger output with several modules, including *phaseblock*, *summarize*, *somatic*, *extend*, and *ancestry*. **c.** Our cohort comprises 14 multiple myeloma patients across several disease stages for a total of 23 tumor samples. **d.** Quality control measures for our tumor and normal samples plus 1000 Genomes samples NA12878 (+) and NA19240 (x). Violin plots defined as: center line, median; violin limits, minimum and maximum values; points, every observation. Molecule Length (mean, Kb): length-weighted mean input DNA length in kilobases. Linked-Reads per Molecule (N50): N50 of read-pairs per input DNA molecule. Phase Block Length (N50, Mb): N50 length of phase blocks in megabases.

Figure 2. Phasing somatic mutations to haplotypes. a. Overview of methods used to phase somatic mutations. **b.** Number of somatic mutations phased using two phasing methods (H1 = phased to haplotype 1; H2 = phased to haplotype 2; NC = not enough coverage for phasing; NP = not phased). **c.** Phasing somatic mutations commonly observed in multiple myeloma. **d.** Distribution of somatic mutations per phase block and the proportion of mutations phased.

Figure 3. Tumor evolution models derived from mutation pairs. a. Number of overlapping barcodes by distance between somatic mutations. **b.** Proportion of somatic mutation pairs in close proximity sharing barcodes and mutations. **c.** Patterns of mutation pairs observed on barcodes (REF = reference allele; ALT = alternate allele). A dark green square indicates that a barcode with that pattern of two alleles was observed. Combinations of patterns can interpreted as evidence of sequential (e.g. 1101, 1011) or distinct (e.g. 1110) mutations. **d.** *NRAS* mutation pair observed in 27522 (P) and evolution model (NC = no coverage). **e.** Interpretation of evolution model observed from *NRAS* mutation pair in 27522 (P). **f.** *ACTG1*

mutation pair observed in 27522 (Rel) and evolution model. **g.** Interpretation of evolution model observed from *ACTG1* mutation pair in 27522 (Rel).

Figure 4. Extension of phase blocks using additional sample information. a. Model for phase block extension using overlap between target and reference phase blocks. **b.** Data-driven example of phase block overlap between samples. **c.** Number of phased variants needed for switch/no switch recommendation. **d.** Length of phase block overlap needed for switch/no switch recommendation. **e.** Phase block groups extended by overlap with another sample. **f.** Distribution of phase block lengths before and after extension. Violin plots defined as: center line, median; violin limits, minimum and maximum values; individual points not shown. **g.** Use of identity-by-descent segments as overlap between phase blocks.

Supplementary Figure Legends

Supplementary Figure 1. Phasing performance quality control summary measures for our tumor and normal samples plus 1000 Genomes samples NA12878 (+) and NA19240 (x). Violin plots defined as: center line, median; violin limits, minimum and maximum values; points, every observation. Definitions of metrics may be found here: https://support.10xgenomics.com/genome-exome/software/pipelines/latest/output/metrics.

Supplementary Figure 2. Phase block length distribution. a. Phase block length by chromosome across all samples. Outlier phase blocks from sample 25183 (Rel) circled. Violin plots defined as: center line, median; violin limits, minimum and maximum values; points, every observation. **b.** Phase block length per sample across all chromosomes. **c.** Phase block lengths of chr13, chr22, and others from 27522 (P). Phase blocks less than 1 kb filtered out for plotting. **d.** Chr13 and chr22 phase block boundaries from 27522 (P) and 27522 (Rem). Alternating dark and light boxes indicate adjacent phase blocks. **e.** Total phase block genome coverage from all samples combined, grouped by phase block length.

Supplementary Figure 3. Copy number profile of Patient 27522 at the primary disease stage. Y-axis values are copy number ratios on the log2 scale.

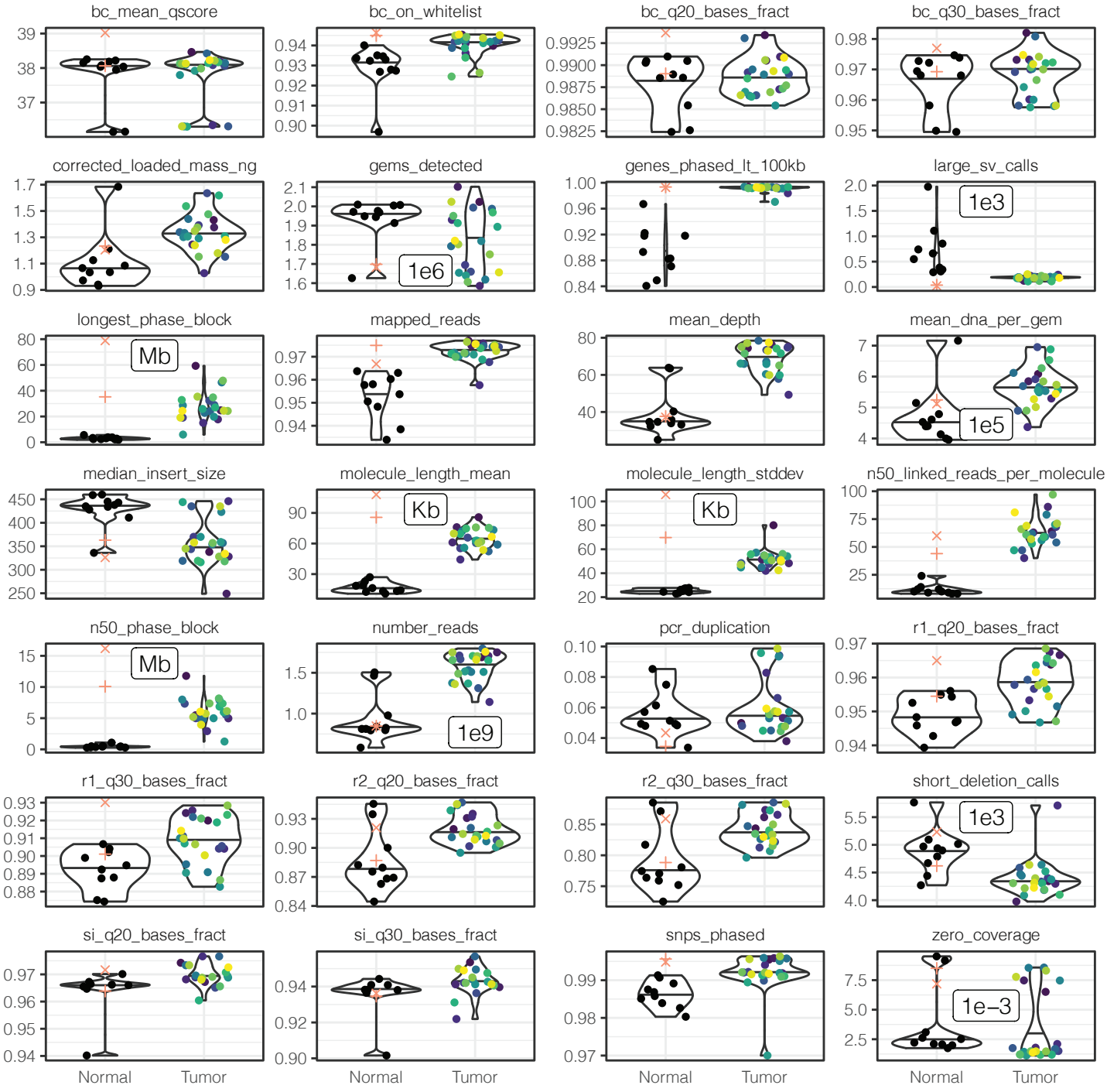
Supplementary Figure 4. Additional information related to somatic mutation phasing. **a.** Precision/recall rates at various cutoffs for the proportion of linked-alleles assigned to one haplotype. **b.** Comparison of phasing results with Long Ranger genotypes.

Supplementary Figure 5. Additional information related to the relationship of pairs of somatic mutation. **a.** Number of barcodes covering each mutation site and those supporting the mutant allele. **b.** Number of overlapping barcodes by distance between somatic mutations less than 100 bp apart.

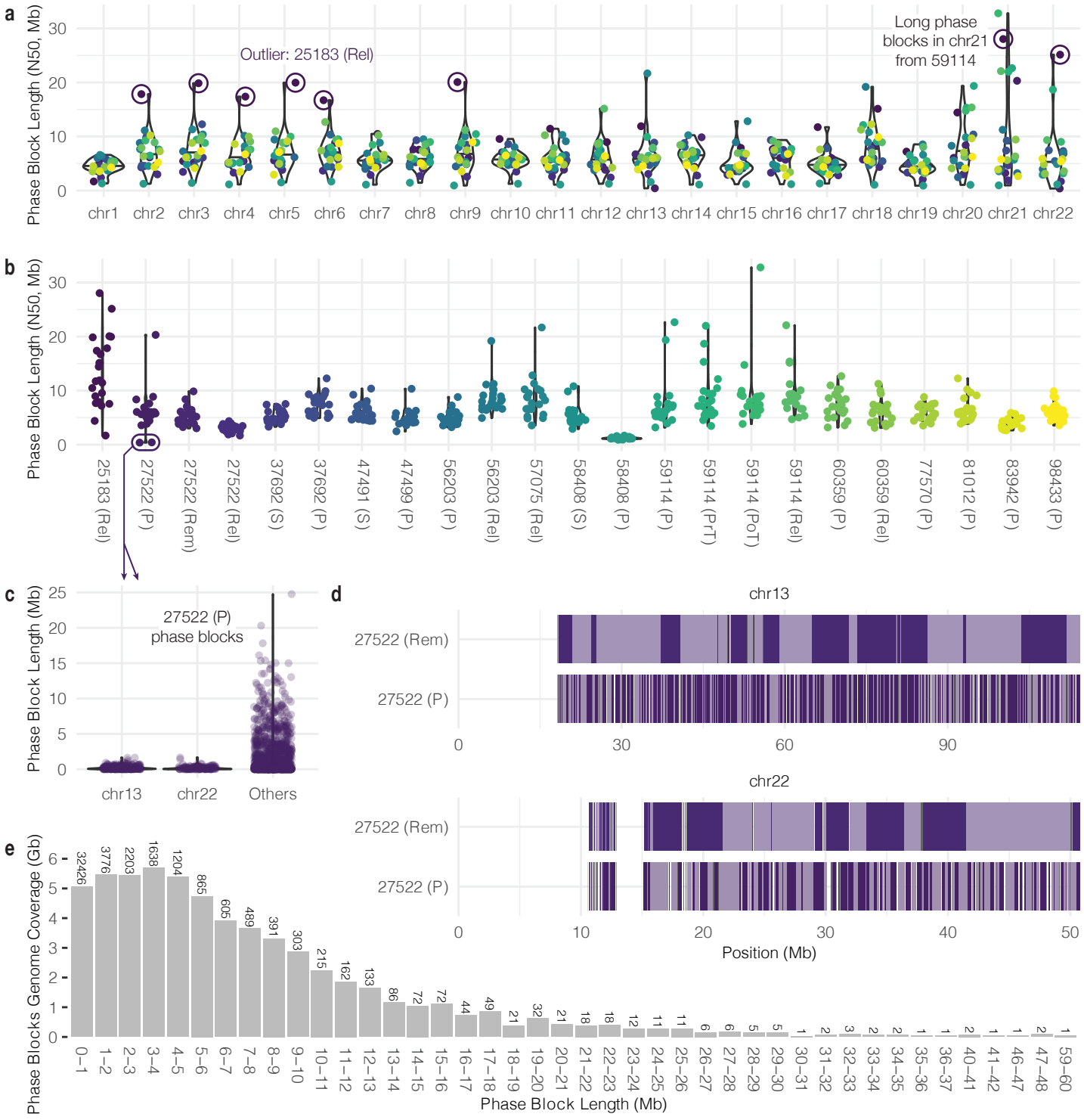
Supplementary Figure 6. Barcodes supporting 27522 (P) NRAS hotspot mutation pair.

Supplementary Figure 7. Common myeloma translocations mapped to haplotypes. a. Overlap of translocations observed in 27522 (P) and (Rel). **b.** Model of t(4;14) translocation. **c.** Barcodes supporting t(4;14) indicate a single haplotype origin. **d.** Translocations observed in 77570 (P). **e.** Model of t(11;14) translocation. **f.** Barcodes supporting t(11;14) indicate a single complex event.

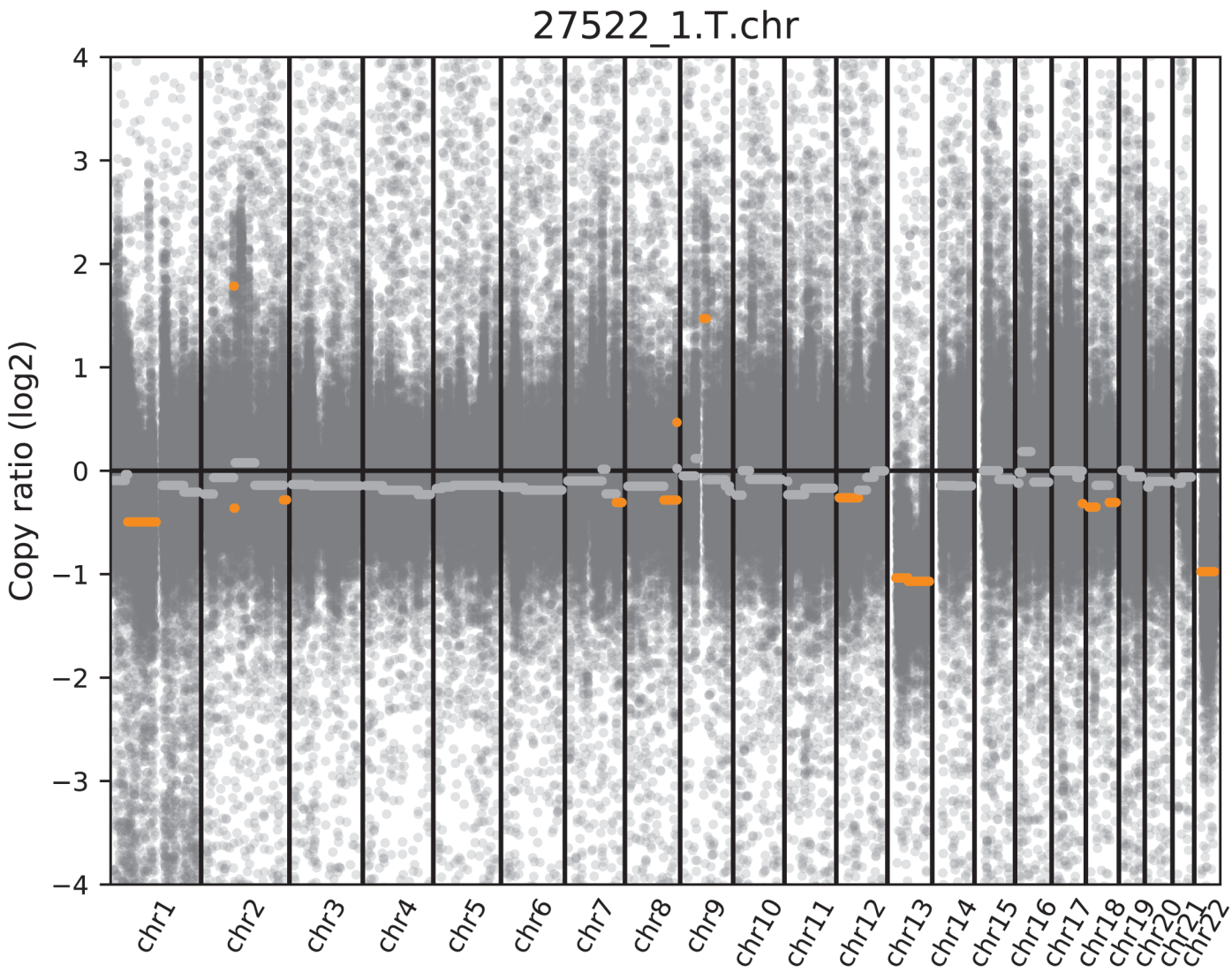
Supplementary Figure 8. Barcode support for common myeloma translocations. **a-b.** 27522 (P) t(4;14). **c-f.** 77570 (P) t(11;14).



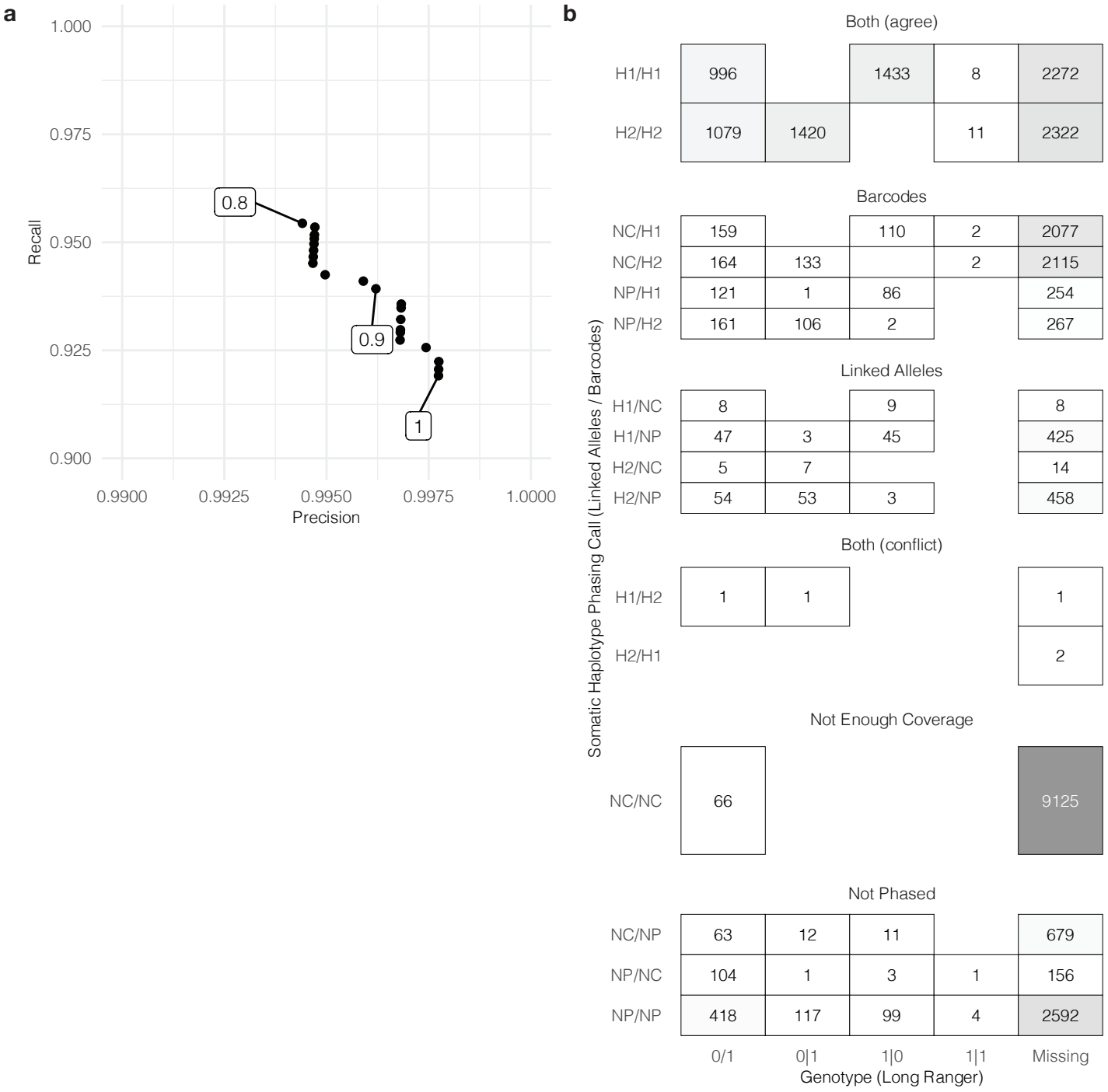
Supplementary Figure 1. Phasing performance quality control summary measures for our tumor and normal samples plus 1000 Genomes samples NA12878 (+) and NA19240 (x). Violin plots defined as: center line, median; violin limits, minimum and maximum values; points, every observation. Definitions of metrics may be found here: https://support.10xgenomics.com/genome-exome/software/pipelines/latest/output/metrics.



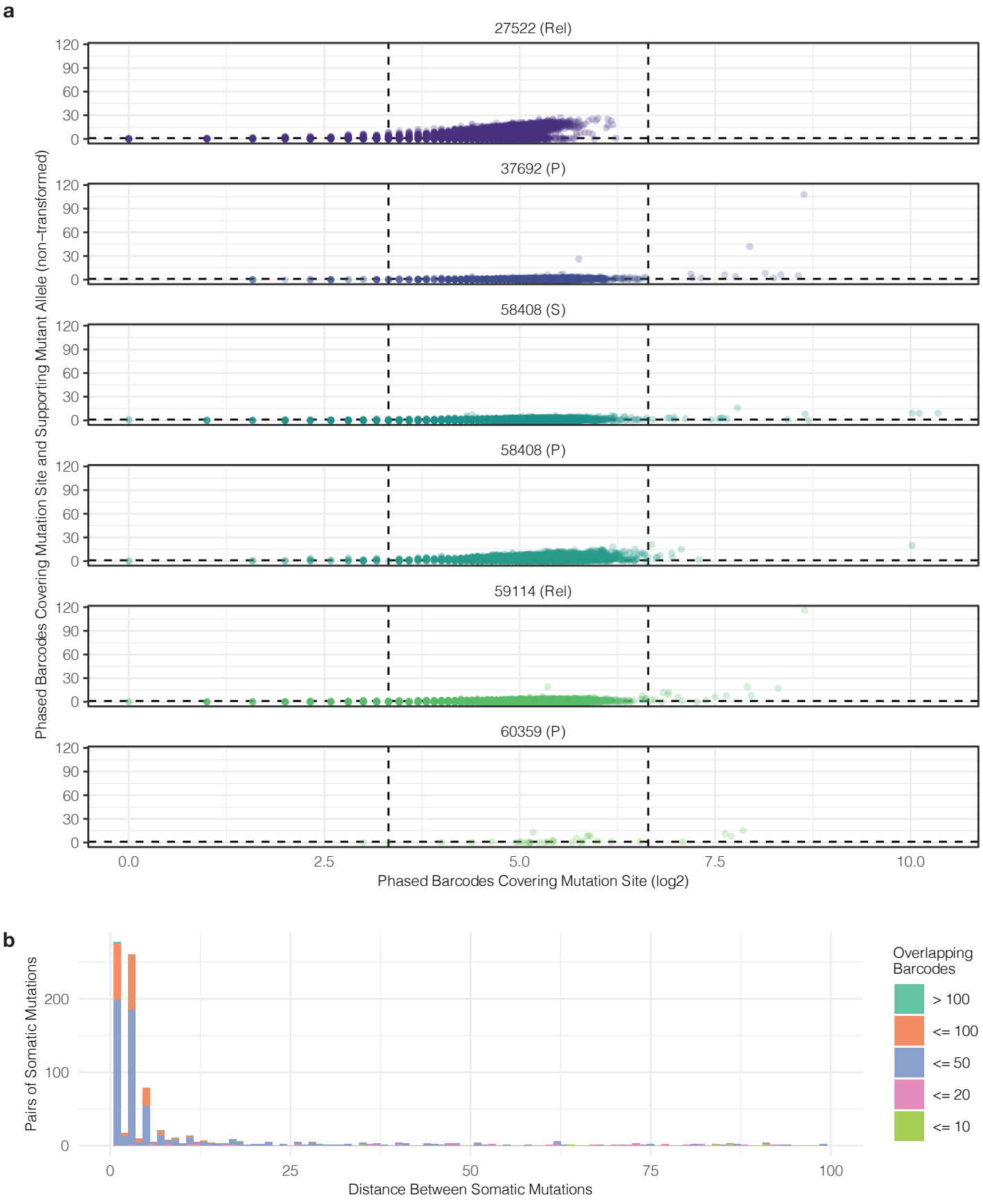
Supplementary Figure 2. Phase block length distribution. a. Phase block length by chromosome across all samples. Outlier phase blocks from sample 25183 (Rel) circled. Violin plots defined as: center line, median; violin limits, minimum and maximum values; points, every observation. b. Phase block length per sample across all chromosomes. c. Phase block lengths of chr13, chr22, and others from 27522 (P). Phase blocks less than 1 kb filtered out for plotting. d. Chr13 and chr22 phase block boundaries from 27522 (P) and 27522 (Rem). Alternating dark and light boxes indicate adjacent phase blocks. e. Total phase block genome coverage from all samples combined, grouped by phase block length.



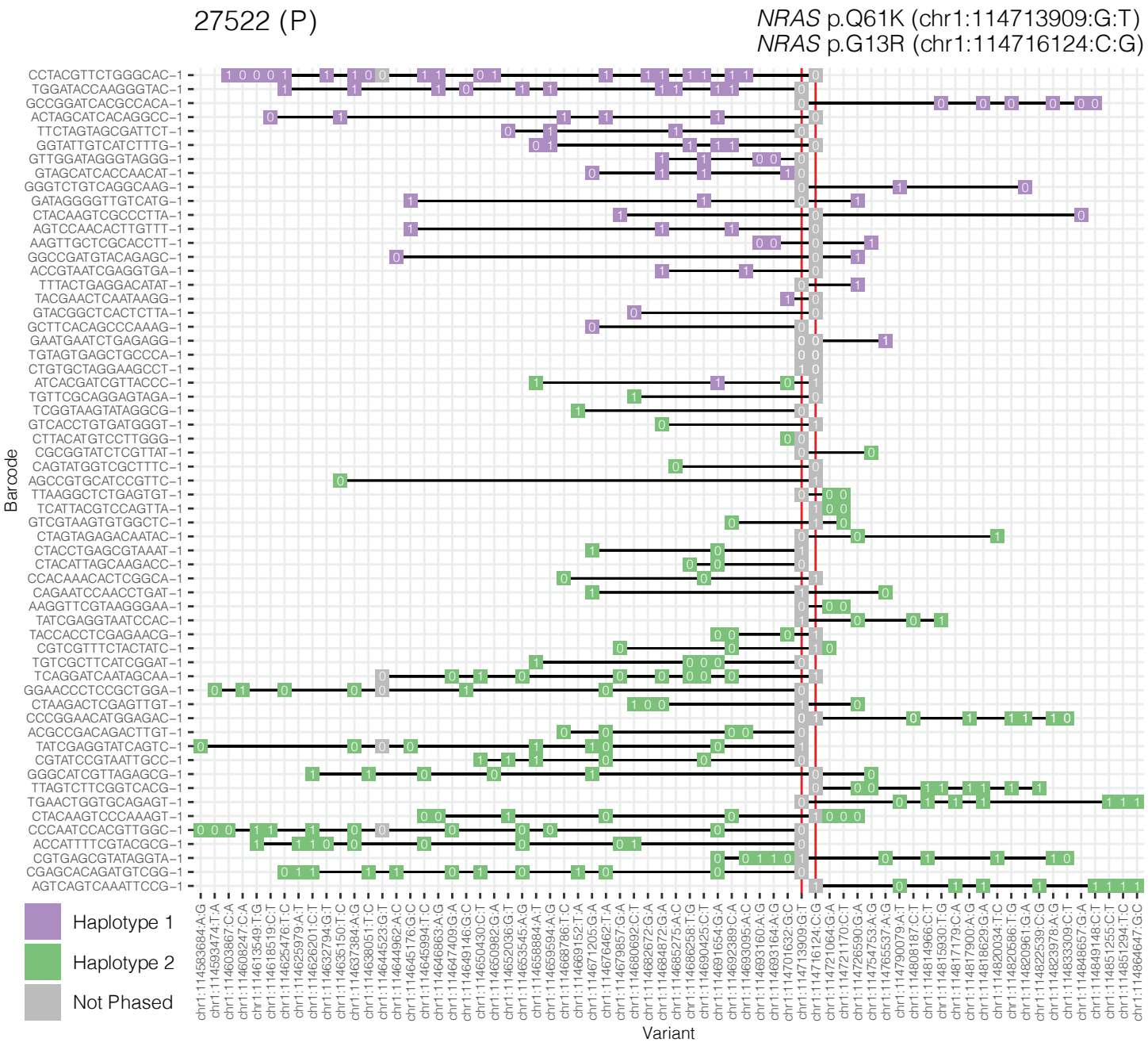
Supplementary Figure 3. Copy number profile of Patient 27522 at the primary disease stage. Y-axis values are copy number ratios on the log2 scale.



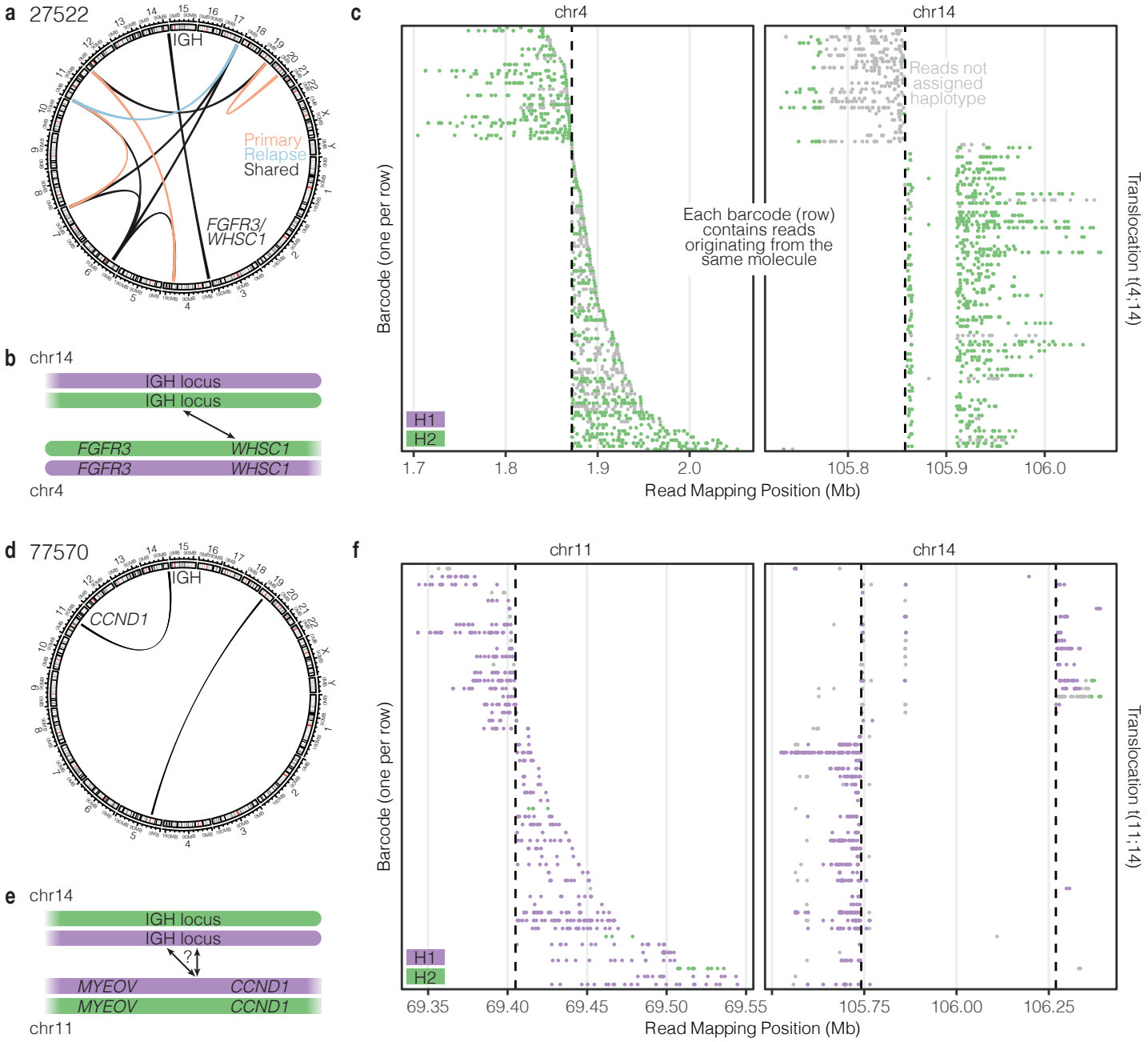
Supplementary Figure 4. Additional information related to somatic mutation phasing. a. Precision/recall rates at various cutoffs for the proportion of linked-alleles assigned to one haplotype. b. Comparison of phasing results with Long Ranger genotypes.



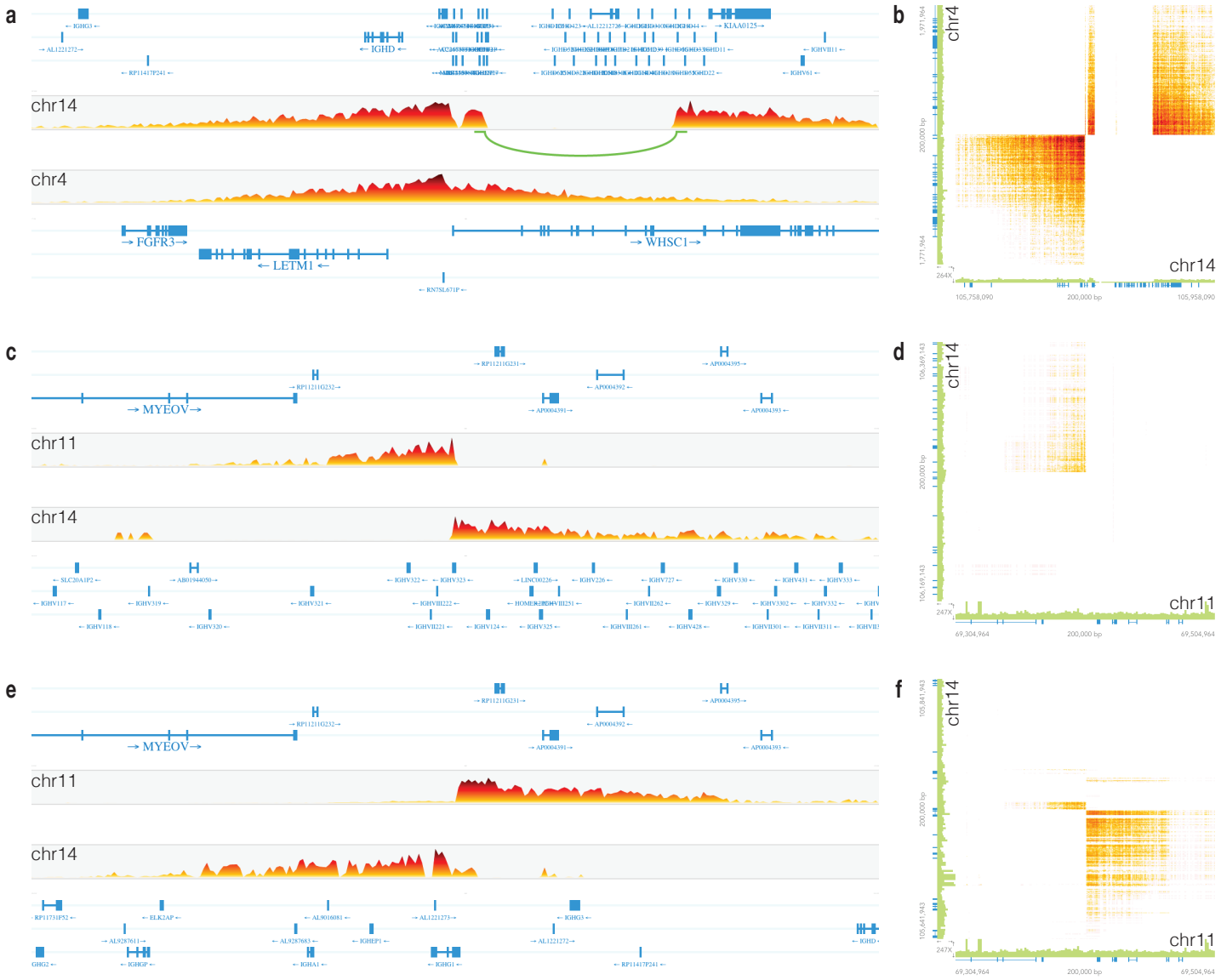
Supplementary Figure 5. Additional information related to the relationship of pairs of somatic mutation. a. Number of barcodes covering each mutation site and those supporting the mutant allele. b. Number of overlapping barcodes by distance between somatic mutations less than 100 bp apart.



Supplementary Figure 6. Barcodes supporting 27522 (P) NRAS hotspot mutation pair.



Supplementary Figure 7. Common myeloma translocations mapped to haplotypes. a. Overlap of translocations observed in 27522 (P) and (Rel). b. Model of t(4;14) translocation. c. Barcodes supporting t(4;14) indicate a single haplotype origin. d. Translocations observed in 77570 (P). e. Model of t(11;14) translocation. f. Barcodes supporting t(11;14) indicate a single complex event.



Supplementary Figure 8. Barcode support for common myeloma translocations. a-b. 27522 (P) t(4;14). c-f. 77570 (P) t(11;14).