Linked-read whole-genome sequencing resolves common and private structural variants in multiple myeloma

Lucía Peña-Pérez,[1,2] Nicolai Frengen,[1,2] Julia Hauenstein,[1,2] Charlotte Gran,[2,3] Charlotte Gustafsson,[1,2] Jesper Eisfeldt,[4,5] Marcin Kierczak,[6] Fanny Taborsak-Lines,[7] Remi-André Olsen,[8] Ann Wallblom,[2,3] Aleksandra Krstic,[9] Philip Ewels,[8] Anna Lindstrand,[4,10] and Robert Månsson[1,2,11]

[1]Department of Laboratory Medicine, [2]Center for Hematology and Regenerative Medicine, [3]Department of Medicine, and [4]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden; [5]Science for Life Laboratory, Karolinska Institutet Science Park, Stockholm, Sweden; [6]Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, Uppsala, Sweden; [7]Division of Gene Technology, Royal Institute of Technology, Stockholm, Sweden; [8]Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden; and [9]Department of Clinical Pathology and Cytology, [10]Department of Clinical Genetics, and [11]Department of Hematology, Karolinska University Hospital, Stockholm, Sweden

Key Points

- Linked-read WGS can be performed without DNA purification and allows for resolution of the diverse structural variants found in MM.

- Linked-read WGS can, as a standalone assay, provide comprehensive genetics in myeloma and other diseases with complex genomes.

Multiple myeloma (MM) is an incurable and aggressive plasma cell malignancy characterized by a complex karyotype with multiple structural variants (SVs) and copy-number variations (CNVs). Linked-read whole-genome sequencing (lrWGS) allows for refined detection and reconstruction of SVs by providing long-range genetic information from standard short-read sequencing. This makes lrWGS an attractive solution for capturing the full genomic complexity of MM. Here we show that high-quality lrWGS data can be generated from low numbers of cells subjected to fluorescence-activated cell sorting (FACS) without DNA purification. Using this protocol, we analyzed MM cells after FACS from 37 patients with MM using lrWGS. We found high concordance between lrWGS and fluorescence in situ hybridization (FISH) for the detection of recurrent translocations and CNVs. Outside of the regions investigated by FISH, we identified >150 additional SVs and CNVs across the cohort. Analysis of the lrWGS data allowed for resolution of the structure of diverse SVs affecting the MYC and t(11;14) loci, causing the duplication of genes and gene regulatory elements. In addition, we identified private SVs causing the dysregulation of genes recurrently involved in translocations with the IGH locus and show that these can alter the molecular classification of MM. Overall, we conclude that lrWGS allows for the detection of aberrations critical for MM prognostics and provides a feasible route for providing comprehensive genetics. Implementing lrWGS could provide more accurate clinical prognostics, facilitate genomic medicine initiatives, and greatly improve the stratification of patients included in clinical trials.

Introduction

Multiple myeloma (MM) is a hematological malignancy affecting terminally differentiated B lineage cells and is characterized by the accumulation of clonal plasma cells in the bone marrow.[1] It has a complex genetic landscape thought to cause the clinical heterogeneity of the disease both in terms

of treatment response and overall outcome.[2] The introduction of novel treatments has significantly improved survival.[3] However, despite these advances, the disease essentially remains incurable, and patients with high-risk aberrations continue to display a poor outcome.[3-5] Therefore, identifying high-risk patients in the context of developing therapy regimens and understanding underlying disease biology to find novel venues for treatment remain critical to improve overall outcome.

On the basis of primary genetic events, MM is largely divided into hyperdiploid (HRD) and non-HRD cases.[2,6] In this division, HRD MM is characterized by multiple trisomies of odd-numbered chromosomes, whereas non-HRD MM is associated with immunoglobulin heavy chain (IGH) translocations. The most common IGH translocations include t(4;14), t(11;14), t(6;14), t(14;16), and t(14;20), which, via colocalization with strong Eμ and 3′ regulatory region (RR) IGH enhancers,[7] cause the dysregulation of MMSET/FGFR3, CCND1, CCND3, MAF, and MAFB, respectively. Traditionally, these primary aberrations, together with common secondary events linked to poor outcome (including deletion of 17p, amplification of 1q21, and *MYC* translocation), have been investigated in clinical routine using fluorescence in situ hybridization (FISH). Despite this seemingly simple dichotomy of initiating events, next-generation sequencing has revealed a complex landscape of genetic aberrations.[5,8-16] This landscape often comprises an array of secondary genetic aberrations, with frequent copy-number variations (CNVs), single-nucleotide variants (SNVs), and structural variants (SVs) resulting from templated insertions, focal amplifications, chromoplexy, chromothripsis, or other complex rearrangements.[12,13,15-17] Collectively, these aberrations affect plasma cell differentiation, cell-cycle regulation, DNA repair, and multiple signaling pathways.[2] Major efforts are being made to exploit targetable aberrations, which, together with high-throughput sequencing–based genomics, create the possibility of using personalized medicine strategies for the treatment of MM.[2,18-20]

Linked-read whole-genome sequencing (lrWGS) is a developing technology that allows for the creation of synthetic long reads from conventional short-read sequencing. Linked-read data are achieved by generating groups of reads or read clouds originating from a single high molecular weight (HMW) DNA molecule, which all carry a common barcode linking them together.[21-24] Mapping the read clouds together and subsequently leveraging the existence of SNVs within the read clouds allow for improved mapping and haplotype reconstruction of large phase blocks.[21] lrWGS data further provide the ability to identify and reconstruct SVs by analyzing read clouds spanning chromosomal breakpoints. Several studies have effectively used lrWGS to identify and describe often-complex SVs in various cancers and other diseases.[25-31] This makes lrWGS an attractive option for providing comprehensive genetics in MM and other diseases with complex genomes.

Here we show that high-quality lrWGS can be generated directly on denatured cells subjected to fluorescence-activated cell sorting (FACS) without the need for prior DNA purification. Applying this to MM in a proof-of-concept study, we found that lrWGS reliably detected translocations and CNVs investigated by FISH. In addition, we show that lrWGS allows for identification and resolution of diverse structural events, including private SVs, with implications for the prognosis of the individual patient.

Materials and methods

MM patient samples and FACS

Bone marrow samples were collected as part of clinical routine from patients with MM at Karolinska University Hospital, Stockholm, Sweden. Collection and use of patient material were approved by the Swedish Ethical Review Authority (2014-526-31/3, 2019-02638, and 2020-00175) and performed in accordance with the Helsinki Declaration. To analyze pure subsets, cells underwent FACS with a FACS ARIA IIu sorter (BD Biosciences). Propidium iodide (PI) was used to exclude dead cells. Using CD319 to compensate for the loss of CD138 on cryopreserved material,[32,33] MM cells undergoing FACS were sorted as PI⁻CD14/CD16/CD11b⁻CD3⁻CD319⁺CD38⁺. In addition, CD138, CD19, CD45, and CD56 were used to define the MM cells depending on expression. Myeloid and T cells undergoing FACS were sorted as PI⁻CD14/CD16/CD11b⁺CD3⁻CD19⁻CD56⁻CD319⁻CD45⁺CD38⁻ and PI⁻CD14/CD16/CD11b⁻CD3⁺CD19⁻CD56⁻CD319⁻CD45⁺CD38⁻, respectively.

FISH

FISH was performed as previously described[34] using FISH probes targeting t(4;14), t(14;16), t(11;14), 1q21, 4p16, 6q21, 8p12, 9p21, 11q13, 13q14, 13q34, 14q32, 15q22, 16q23, 17p13, and 19q13.

lrWGS

To generate libraries, 200 to 240 cells subjected to FACS were sorted into 8 μL of phosphate-buffered saline with 2% fetal calf serum; tubes were briefly pulse spun, frozen on dry ice, and stored at −80°C until use. To denature the cells and make the genomic DNA accessible, 2 μL of 0.25N NaOH was added using narrow-bore tips, and the samples were mixed by carefully flicking the tubes. After 5 minutes of incubation at room temperature, 97.5 μL of sample master mix (89.5 μL of Genome Reagent Mix, 3 μL of Additive A, and 5 μL of Genome Enzyme Mix; Chromium Genome Reagent Kit [v2 Chemistry]) was added to the denatured cells with a narrow-bore tip. Subsequently, the sample was mixed by gentle pipetting using a wide-bore tip. A total of 90 μL of the sample mix (containing ~1.2 ng of DNA) was loaded onto the Chromium Genome Chip, and the remaining steps were performed according to the manufacturer's instructions (Manual CG00043 Rev B). Barcoded libraries were quantified using the Qubit dsDNA HS Kit (Invitrogen), pooled, and paired-end sequenced (2× 150 cycles) using the Illumina platform (Novaseq; Illumina, San Diego, CA).

Data were processed using Long Ranger,[21] and output files were used for downstream analysis. In brief, chromosome numbers were calculated using BarCrawler (https://github.com/J35P312/BarCrawler) in combination with in-house scripts; for selected samples, chromosome numbers were also assessed using CNVkit (v0.9.8),[35] ASCAT (v2.5.2),[36] and the Battenberg approach (v2.2.9)[37]; CNVs were called using FindSV (https://github.com/J35P312/FindSV) and Long Ranger; somatic variant calling was done using the nf-core/Sarek pipeline (v2.6[38]; with Strelka2 [v2.9.10][39] and Mutect2 [GATK v4.1.7.0][40] as variant callers) and a combination of VEP (v99.2)[41] and SnpEff (v4.3t)[42] for variant annotation; SVs were identified using GROC-SVs (v0.2.5)[43] in combination with Long Ranger; and called SVs were visually confirmed in Loupe and IGV.[44] Details are provided in the data supplement.

RNA sequencing

Strand-specific RNA sequencing data were generated from MM cells subjected to FACS as previously described[45] (details provided in data supplement). Reads were mapped to hg38 using STAR (v2.5.2b),[46] reads in exons quantified using HOMER,[47] and transcripts per million (TPMs) calculated using R.

ChIPseq

Chromatin immunoprecipitation sequencing (ChIPseq) was performed on 20 000 MM cells subjected to FACS as previously described[48] using H3K27Ac antibodies (lot A1723-0041D/2; catalog #C15410196; Diagenode) but with minor modifications to the preparation of input controls. Fastq files were mapped to hg38 using bowtie2, and downstream analysis was performed using HOMER,[47] DEseq2,[49] and R. Details are provided in the data supplement.

Results

To investigate if lrWGS could be used to identify recurrent and private genetic aberrations in MM, we subjected a cohort of 37 MM samples to lrWGS (supplemental Table 1). All bone marrow samples were taken for diagnostic purposes (35 of 37 at diagnosis) and analyzed by an extended FISH panel. To ensure that analysis was performed on pure populations, all cells used underwent FACS (Figure 1A). In addition, germ line controls were generated in 32 of 37 cases by performing lrWGS primarily on T cells (supplemental Table 1). Prepared lrWGS libraries were sequenced to average depths of 31.2× and 31.6× for MM and germ line controls, respectively (supplemental Table 1).

High-quality lrWGS libraries can be prepared from low numbers of denatured cells

To fully use the lrWGS technology, libraries must be generated from purified HMW DNA.[21,50] To circumvent the need for HMW preparation, we established a protocol for preparing lrWGS libraries directly on cells denatured with sodium hydroxide. This allowed us to prepare lrWGS libraries on 200 to 240 cells subjected to FACS (containing ~1.2 ng of DNA) on the Chromium Genome platform.[21] We found that this simple protocol allowed for maintaining very large DNA molecules (median DNA fragment length, 216 kbp), which was directly reflected in the size of the reconstructed phase blocks (median N50, 14.8 Mbp; median longest phase block, 60.2 Mbp; Figure 1B-D; supplemental Table 1). The longest assembled phase block was 120.9 Mbp and constituted almost the entire q arm of chromosome 4 (120.9 of 138.4 Mbp). Thus, this phase block nearly reached the theoretical maximum length achievable without crossing centromeric regions. As expected, MM samples typically displayed poor phasing in areas with loss of heterozygosity (supplemental Figure 1), but overall, the N50 phase block size of the MM samples was comparable to that of the normal controls (Figure 1B-C). Comparing our lrWGS data with high-quality data sets generated from purified HMW DNA on the 10X platform,[21,50] we found that our libraries were generated from significantly longer DNA molecules and maintained similar phase block sizes (Figure 1B-D).

Overall, we conclude that performing lrWGS library preparation directly on denatured cells maintains DNA integrity and allows for

the generation of high-quality lrWGS from very limited numbers of cells subjected to FACS.

CNVs identified by lrWGS and FISH are highly concordant

CNVs are a common feature of MM, defining the HRD subgroup and high-risk patients.[3] To identify CNVs, we calculated copy numbers and computationally called CNVs. Comparing the chromosomal setup calculated from the lrWGS data with that from FISH, we found that chromosome numbers were correctly predicted in 35 (95%) of 37 MMs. The 2 cases where copy number was not accurately calculated displayed widespread and complex subclonal copy-number changes. No CNV caller used (including FindSV, CNVkit,[35] ASCAT,[36] and the Battenberg approach[37]) accurately predicted the chromosomal setup in these 2 samples (data not shown). Nevertheless, correcting the copy numbers based on chromosome 4 coverage (the only chromosome known to have a normal copy number from the FISH analysis) allowed for visual identification of all (12 of 12) CNVs found by FISH (Figure 2A). In the 35 patients for whom the chromosomal setup was accurately predicted by lrWGS, 96% (143 of 149) of all CNVs identified by FISH, including all 1q amplifications, were concordantly called in the lrWGS data (Figure 2A-B). These results are similar to those in prior studies comparing FISH with conventional WGS for identification of CNVs.[51,52]

Overall, this shows that lrWGS can reliably identify CNVs. However, our data argue that caution should be taken to verify chromosome numbers predicted by WGS in tumors with potentially widespread and complex subclonal copy-number changes.

Identifying double-hit TP53 inactivation

Double-hit TP53 inactivation in MM is associated with poor outcome, even with novel therapy.[3-5] This prompted us to carefully investigate the existence of such events in our patient cohort. Haplotype-specific chromosome 17p deletions could readily be observed in the phased lrWGS data (Figure 2C). On the basis of the computational calling of CNVs, we identified 7 (87.5%) of 8 patients with CNVs involving the TP53 locus (supplemental Figure 2). The 17p deletion with the lowest clonal involvement identified in the lrWGS data was found in 24% of cells by FISH (supplemental Figure 2). In comparison, the undetected 17p deletion identified by FISH was found in only 14% of the cells. We next identified acquired mutations using Sarek[38] and analyzed sequence changes in genes recurrently mutated in MM[10,12] (supplemental Figure 3A-B). We identified 2 patients with frame-shift mutations deleterious to TP53 function (supplemental Figure 3A-B). The first was subclonal and appeared in a patient with an otherwise normal TP53 locus (P13756_103). In contrast, the second mutation was near clonal (present in ≥80% of reads in patient P11603_109) and cooccurred with a clonal 17p deletion (Figure 2C, right), suggesting that this patient presented with high-risk double-hit TP53 MM at diagnosis.

Primary MM IGH locus translocations are readily identifiable by lrWGS

SVs can be identified in lrWGS data as enrichments of read clouds spanning the areas surrounding the chromosomal breakpoints (Figure 3A). This can be used to identify translocations, templated
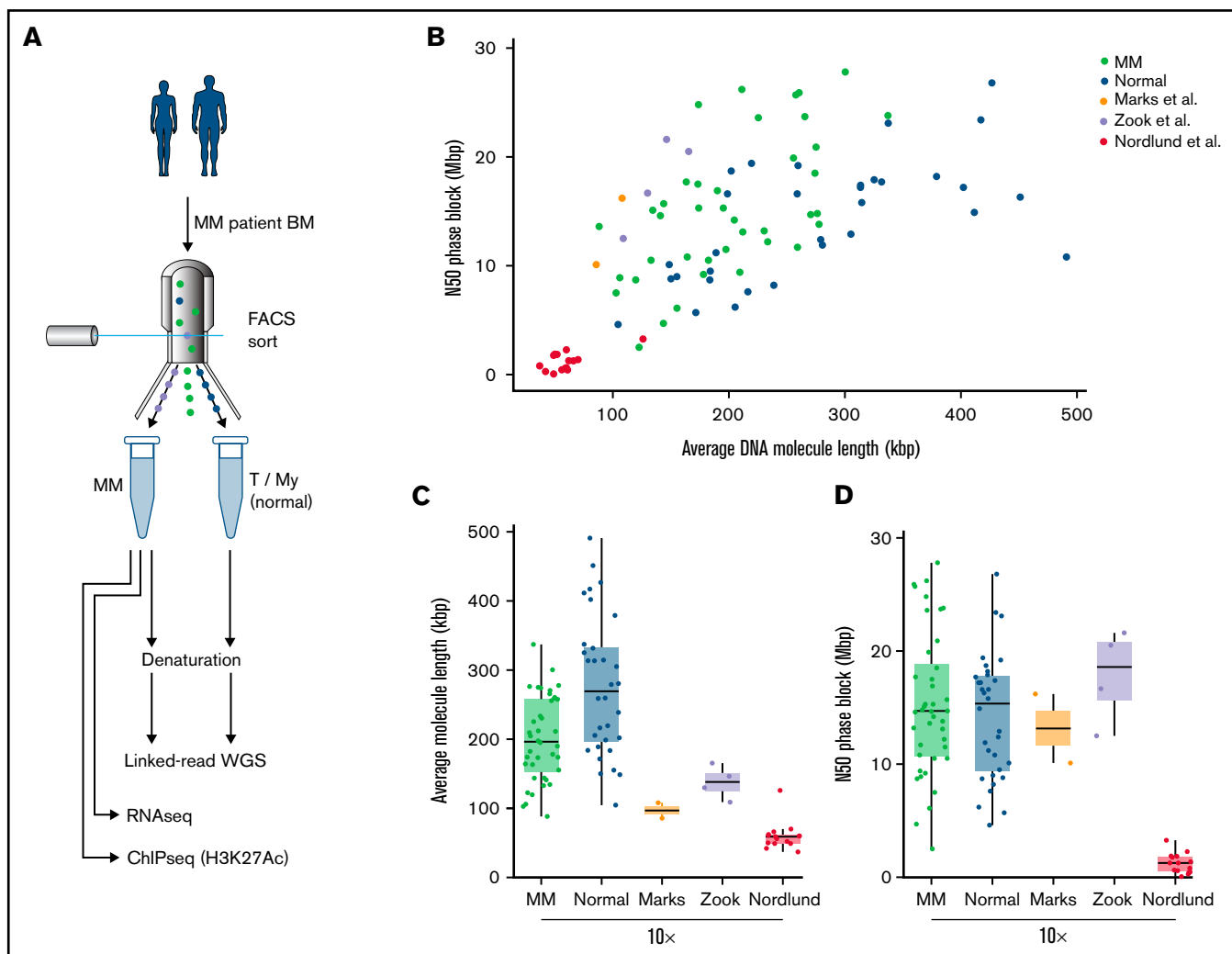
Figure 1. lrWGS can be performed directly on denatured cells subjected to FACS from diagnostic bone marrow (BM) samples. (A) Schematic overview of experimental setup. To perform genetic, transcriptional, and epigenetic characterizations, MM and normal cells (T cells [T] or myeloid cells [My]) were subjected to FACS from patient BM samples. lrWGS libraries were prepared directly on denatured cells without prior DNA purification. (B-D) Scatter plot (B) and box plots (C-D) showing the N50 phase block size and average DNA molecule length of lrWGS libraries prepared using the indicated cell type and method. Quality control data from published lrWGS generated from prepared DNA (HMW or column purified) on the 10X platform are shown for comparison.[21,31,50]

insertions, focal amplifications, and other complex SVs (Figure 3B-C; supplemental Figure 4A-D). We used Long Ranger[21] and GROCS-SVs[43] to call interchromosomal SVs. This approach identified 198 SVs across all samples (Figure 4A; supplemental Table 2). No significant differences in the number of SVs per sample were found between HRD and non-HRD cases, although 3 cases of HRD MM had no detectable SVs (Figure 4B).

To identify IGH translocations, we visualized the common t(4;14), t(11;14), and t(14;16) breakpoint regions using Loupe. We found that the translocations could readily be identified visually through the enrichment of breakpoint-spanning read clouds (Figure 4C; supplemental Figure 5). The enrichment was highly specific to the individual IGH translocation events and clearly distinguishable from background signals (supplemental Figure 5). Among the computationally called SVs (supplemental Table 2), we found 16 MMs with IGH translocations involving the *MMSET*, *CCND1*, or *MAF* locus. Overall, 16 (94%) of the 17 IGH translocations identified by FISH

were found by lrWGS, and no false-positive calls were made. The unidentified t(11;14) was subclonal (28% of cells), and read clouds supporting the existence of the SV were identified upon visual inspection (supplemental Figure 5).

Together, this suggests that lrWGS allows for the accurate identification of SVs, including common IGH translocations.

t(11;14) is associated with amplification of the 3′RR IGH enhancer regions

In line with the notion that IGH translocation events often occur as a result of erroneous class-switch recombination,[2] the IGH translocations were localized on the centromeric side of the IGHJ regions (Figure 4C; supplemental Figure 5). Looking specifically at t(4;14) and t(14;16), we found that they were all reciprocal, with the chromosome 14 breakpoint occurring between the Eμ and 3′RR enhancer regions (flanking the IGHA1/2 regions; supplemental
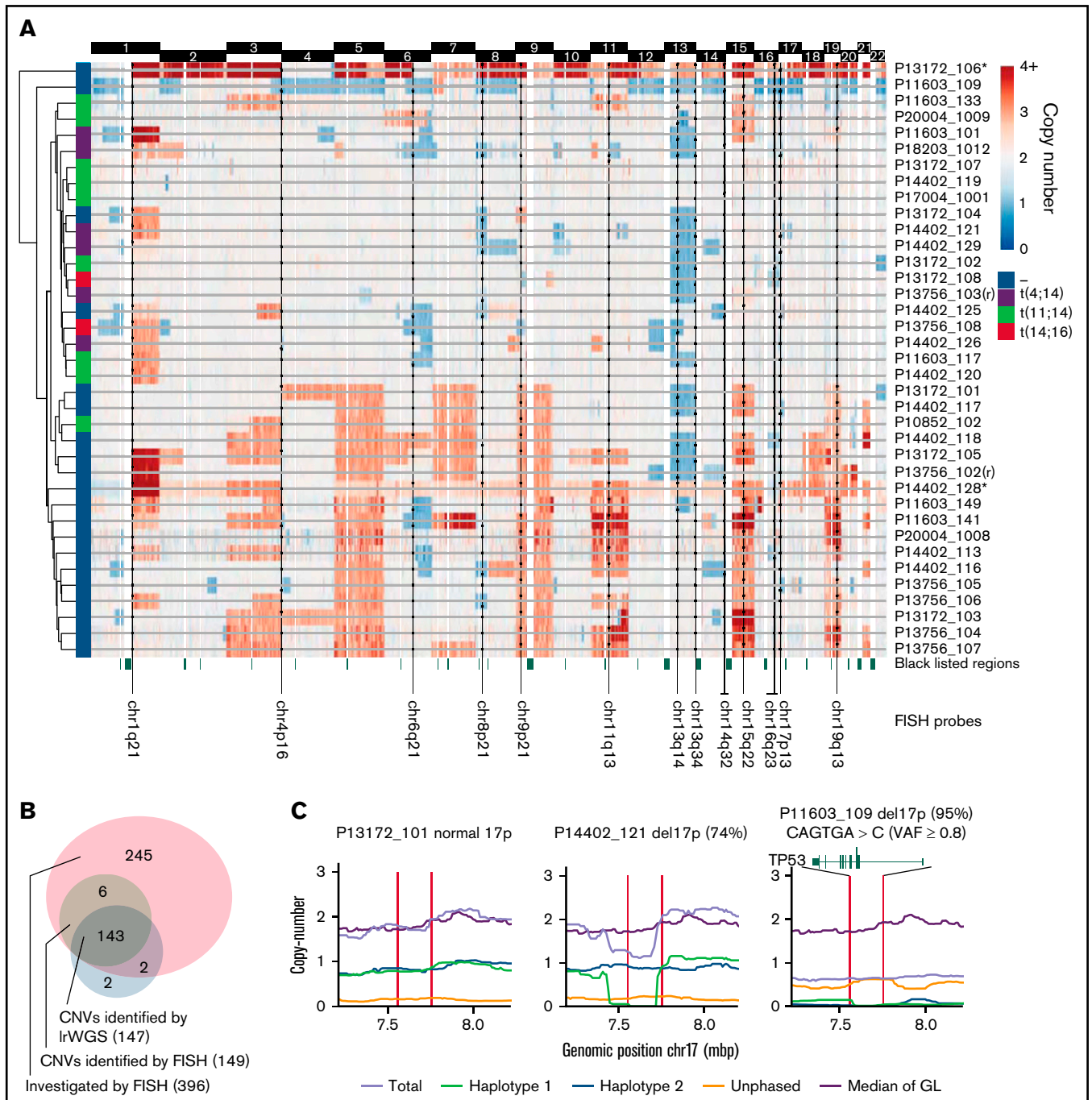
Figure 2. lrWGS allows for reliable detection of CNVs and identification of double-hit MM. (A) Heat map showing the copy numbers across the somatic chromosomes (chr1-22) as calculated from lrWGS coverage. Copy-number was in two cases (indicated by*) corrected to the chr4 copy-number (supplemental Methods). Vertical black lines indicate genomic localization of FISH probes. Regions investigated by FISH are indicated by black dots, with the position of the dot above or below the gray line indicating copy-number gains or losses, respectively. (B) Venn diagram showing the overlap between CNVs called by FISH and by lrWGS among regions included in the FISH panel (excluding the IGH regions). Regions investigated by FISH are indicated in panel A. (C) Copy numbers of total, haplotype-specific, and unphased lrWGS coverage in the area surrounding the TP53 locus. Median copy number in the normal control samples (germ line [GL]) is shown for comparison. Vertical lines indicate the boundaries of the FISH probes used to identify chr17p CNVs. For P11603_109, the position of the acquired TP53 mutation (detected using the lrWGS data) is indicated. VAF, variant allele frequency.
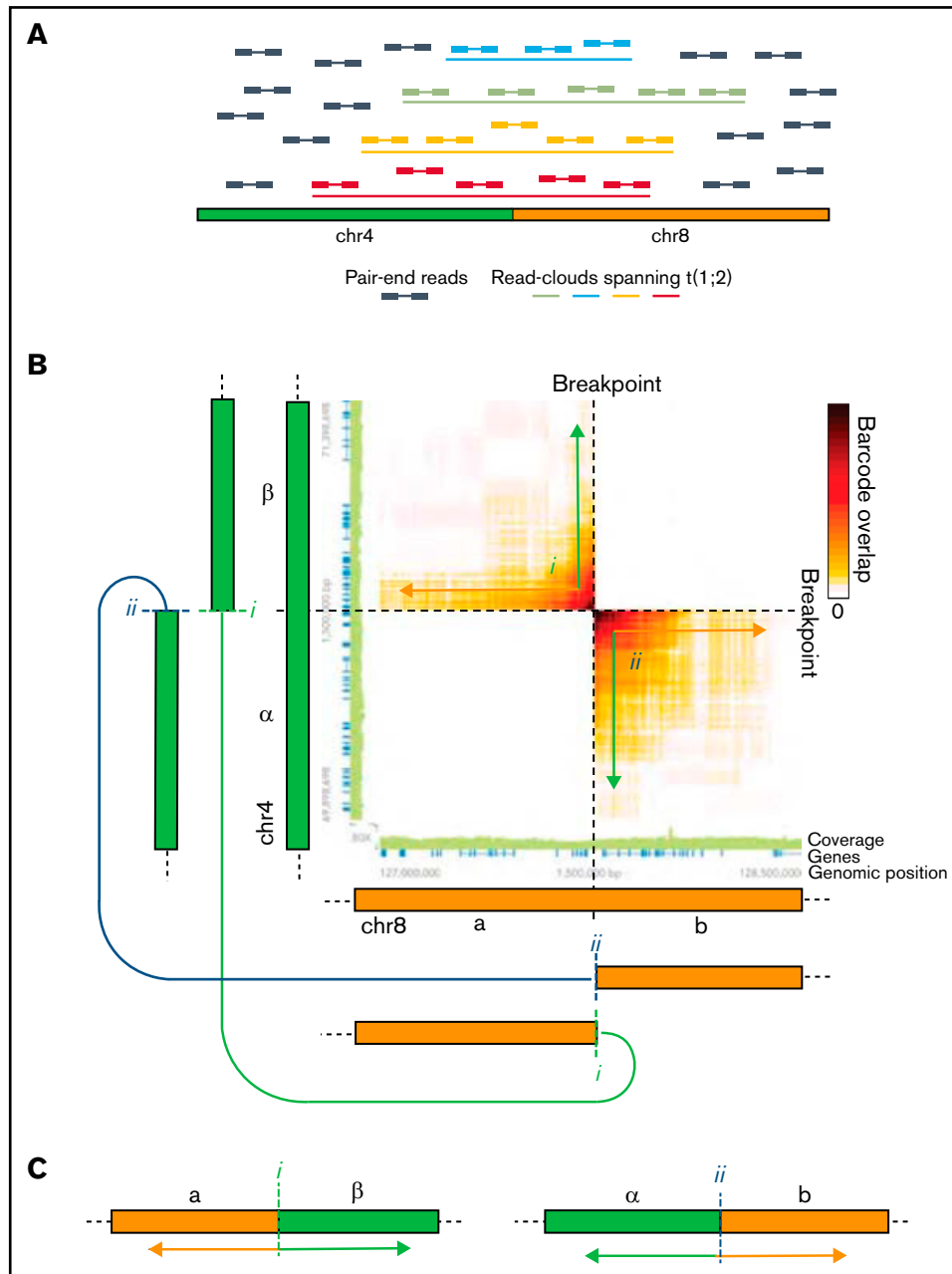
Figure 3. SVs are identified in lrWGS as enrichments of breakpoint-spanning read clouds. (A) Schematic illustration of read clouds spanning a chromosome 4 (chr4) to chr8 translocation breakpoint and the pair-end reads constituting the read clouds. (B) Heat map showing the enrichment of shared read cloud (barcodes) between regions on chr4 and chr8 fused by a reciprocal translocation. As the enrichment of shared read clouds progressively increases closer to the breakpoint, the signals in the heat map resemble 2 arrowheads pointing toward the breakpoint. Effectively, this allows for determining that the reciprocal translocation results in the fusion of the chr4 β-region with the chr8 a-region (i) and the chr4 α-region with the chr8 b-region (ii). The connected green and orange arrows schematically indicate the areas joined by events i and ii. (C) Schematic illustration of the derivate chromosomes generated by the reciprocal translocation. Supplemental Figure 4 provides additional details on the interpretation of read-cloud data and examples of other SVs.

Figures 6 and 7).[7] Similarly, approximately half (5 of 9) of the t(11;14) events were conventional translocations (Figure 4C; supplemental Figure 5) juxtaposing the 3′RRs with the CCND1 locus (supplemental Figure 8). Interestingly, 1 of the reciprocal t(14;16)s and 3 of the t(11;14)s were part of complex SVs with additional breakpoint junctions outside of the IGH locus (supplemental Figure 9).

The remaining t(11;14) MMs (4 of 9) presented with patterns of read-cloud clusters indicative of more complex events affecting the breakpoint region (Figure 4C; supplemental Figure 5). Three of the patients (P14402_119, P17004_1001, and P13172_102) shared a similar pattern with 3 overlapping read-cloud clusters (supplemental Figure 5). Analyzing the read-cloud overlaps of the clonal t(11;14) in P14402_119, we found that the pattern was consistent
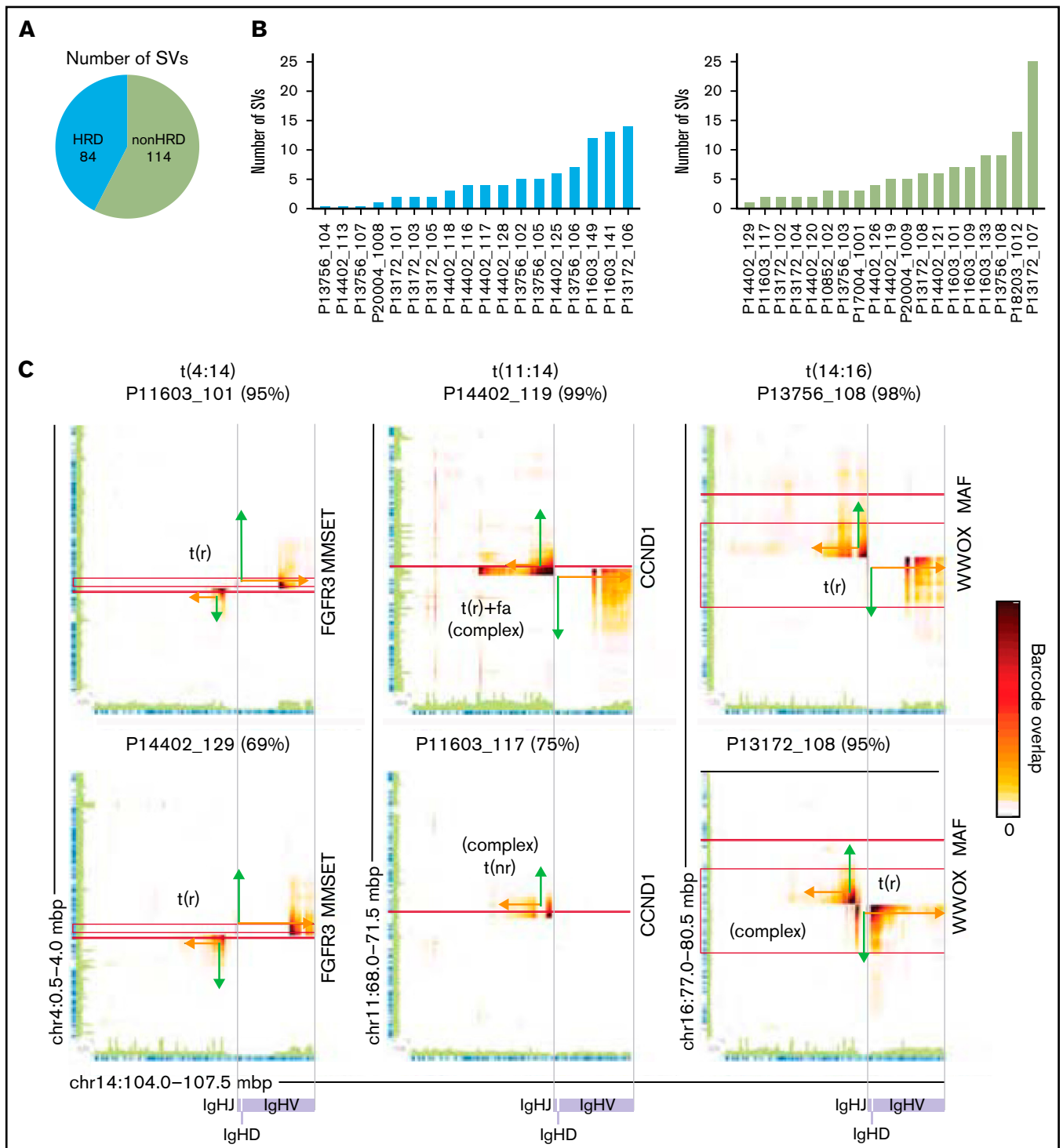
Figure 4. Breakpoint-spanning read clouds allow for identification and reconstruction of IGH translocation events. (A) Total number of SVs involving different chromosomes identified in HRD and non-HRD MM. (B) Number of SVs involving different chromosomes in HRD (left) and non-HRD MM (right). (C) Heat maps displaying the number of read clouds shared between the IGH locus (chromosome 14 [chr14]) and breakpoint regions on chr4 (MMSET), chr11 (CCND1), and chr16 (MAF). The connected green and orange arrows schematically indicate the areas joined by the translocations. The type of SV is indicated next to the read-cloud clusters. Figure 3 and supplemental Figure 4 provide details on the interpretation of the read-cloud patterns. If the SV depicted is part of a more complex SV, this is indicated. The percentage of cells found by FISH to carry the indicated translocations is shown in parentheses above each heat map. Horizontal (red) and vertical (purple gray) lines mark the position of indicated genes and IGH VDJ regions, respectively. (D) Read-cloud overlap on chr11 (top left), from chr11 to chr14 (top right), and on chr14 (bottom right) in P14402_119. Track shows the H3K27Ac signal (median ± standard deviation [SD]) in t(11;14) MM. The connected green and orange lines schematically indicate the areas joined by events i to iv. (E) Schematic representation of the derivate chromosomes in P14402_119 as determined by lrWGS. t(nr), nonreciprocal translocation; t(r), reciprocal translocation; t(r)+fa, reciprocal translocation with a focal amplification of the breakpoint region.
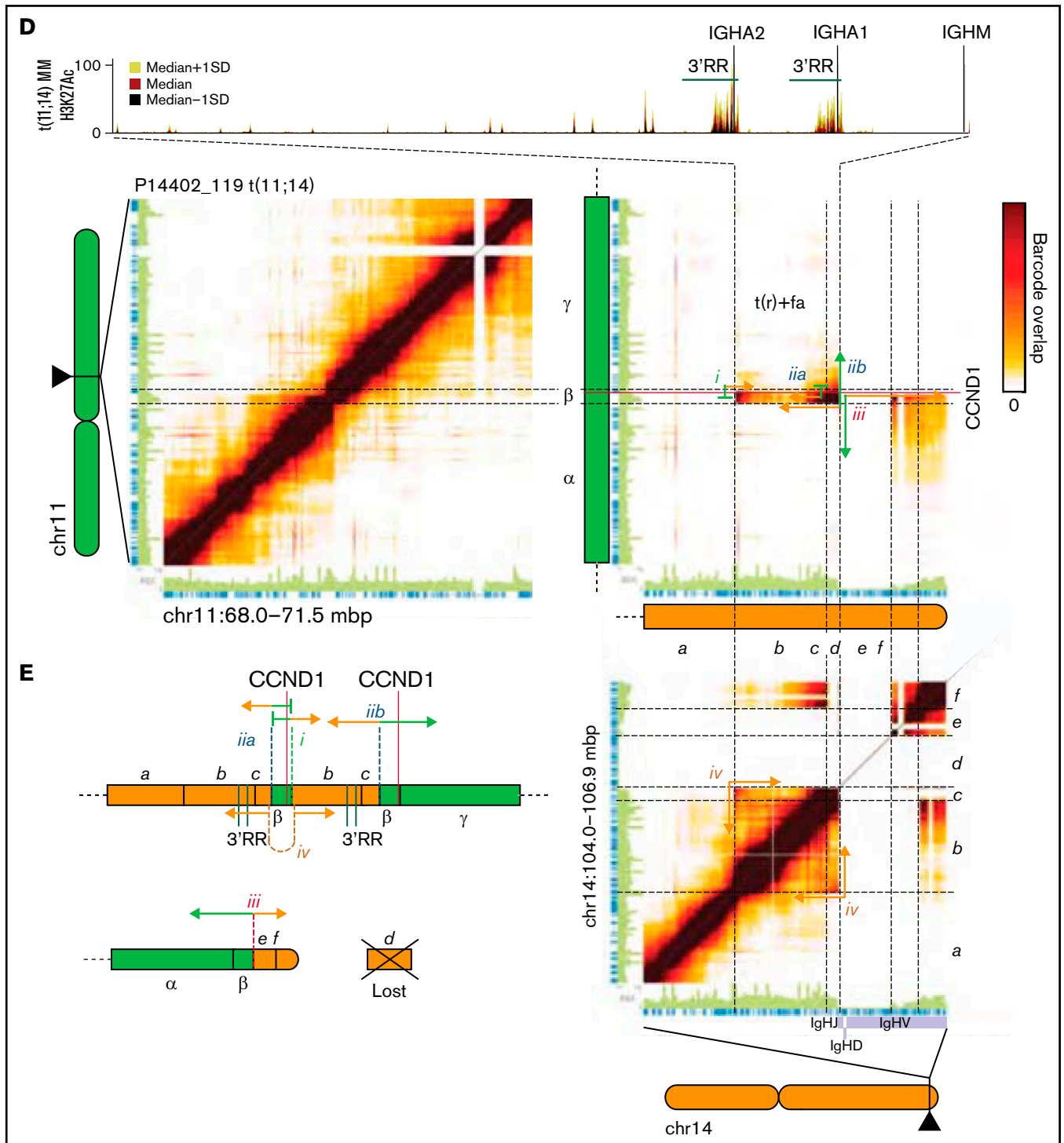
Figure 4. (continued)

with a reciprocal translocation (Figure 4Di-iib) carrying a focal amplification of the translocation breakpoint region on the CCND1-carrying derivative chromosome (Figure 4Dii-iv,E). This could be deduced by the copy-number gains on chromosome 11 (Figure 4D, β region) and chromosome 14 (Figure 4D, b-c region), the shared read clouds between centromeric (5′) and telomeric (3′) ends of

the amplified region on chromosome 14 (Figure 4Div, b-c region), and the read-cloud overlaps in the middle being compatible with 2 separate events, with 1 ending abruptly near the end of the β-region without a gradual decrease in overlapping read clouds (Figure 4Diia) and 1 displaying a conventional pattern with gradually reduced numbers of shared read clouds at increasing distance from
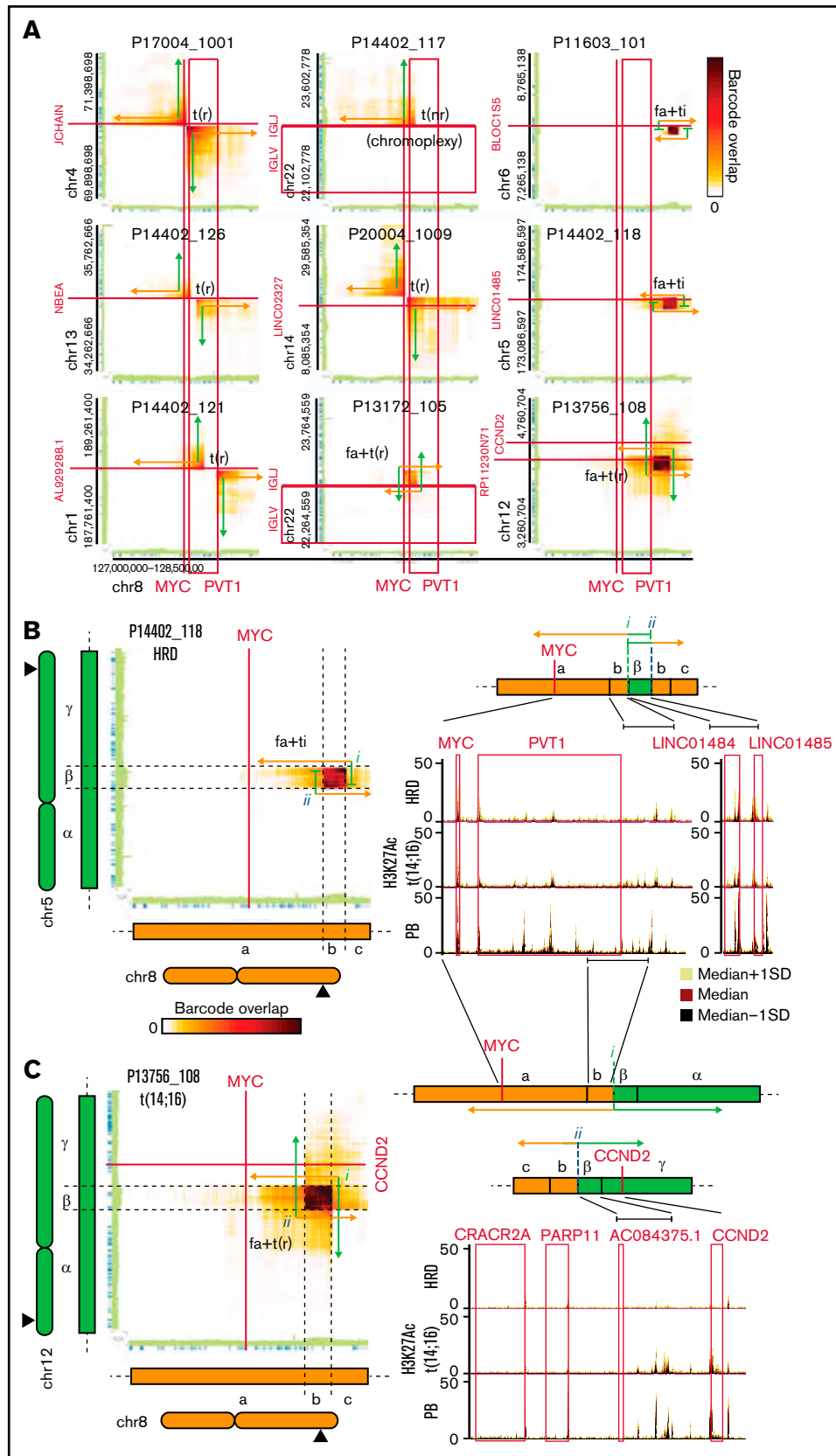
Figure 5.

the breakpoint (Figure 4Diib). The 2 other cases with a similar pattern of read clouds involving overlapping regions (P17004_1001 and P13172_102) could be resolved in the same manner. In addition, the IGH translocation in P14402_119 was part of a complex SV with additional translocations occurring outside of the IGH and CCND1 loci (supplemental Figure 9). Looking at the fourth case (P14402_120), which had a different pattern with 2 (in terms of genomic location) partially overlapping read-cloud clusters, we concluded that this pattern represented a reciprocal translocation occurring within a focal amplification that resulted in duplicated flanking regions on either side of the translocation breakpoint (supplemental Figure 10). Overlapping the involved regions with H3K27Ac (marking active gene regulatory elements) ChIPseq data, we found that the 3′RRs were within the duplicated regions in all 4 cases (Figure 4D; supplemental Figure 11A). As expected, this was associated with very high expression of CCND1 (>1000 TPMs in all 4 patients; supplemental Figure 11B).

Taken together, this suggests that t(11;14) events are frequently associated with amplifications of the CCND1 locus and the 3′RR of the IGH locus. In addition, the IGH translocations were, as expected,[16] frequently (5 of 17 cases) part of complex SVs.

lrWGS allows for detection and resolution of diverse SVs affecting the MYC locus

MYC translocations have been shown to be associated with poor outcome.[9,13] SVs involving the MYC locus are often complex, involving focal amplifications, insertions, and translocations that are thought to dysregulate MYC expression by positioning the gene in proximity to strong enhancer elements.[9,13,15,17] Relying on the computational identification of SVs (supplemental Table 2), we found SVs affecting the MYC locus in 9 (24%) of 37 patients, with the IGL region being the only recurrently involved locus (2 of 9 cases). The identified SVs were supported by distinct read-cloud clusters (Figure 5A) that could easily be distinguished from the background (supplemental Figure 12), and in 1 case (P14402_117), the event was part of a more complex SV (supplemental Figure 9). A majority of the identified MYC SVs were simple translocations (5 of 9; Figure 5A). Looking closer at the remaining cases with overlapping read-cloud clusters, we concluded that these cases represented focal amplifications with templated insertions (Figure 5A-B) or reciprocal translocations (Figure 5A,C). Both kinds of SVs have previously been described for the MYC locus.[13,15,17,53] As expected, identified SVs involved either the region containing the MYC and PVT1 promoters (6 of 9 cases) or the area just telomeric of the PVT1 gene (3 of 9 cases) known to harbor elements influencing the MYC gene.[54] All 3 SVs affecting the PVT1 flanking enhancer region resulted in the juxtaposition of this element with myeloma-active gene regulatory elements from another chromosome through

insertion (P14402_118 and P11603_101; Figure 5B; supplemental Figure 13A) or by reciprocal translocation (P13756_108; Figure 5C). Similarly, the translocations affecting the MYC/PVT1 promoter area in most cases resulted in the MYC locus being juxtaposed with myeloma-active elements from various genomic locations, including the JCHAIN and IGL loci (supplemental Figure 13A-B). Overall, MM with MYC SVs expressed higher levels of MYC than MM cases with a normal MYC locus, but the difference was not statistically significant (supplemental Figure 13C).[9]

Together, this shows that the MYC locus is affected by a diverse spectrum of SVs in MM and that these can be detected and structurally resolved using lrWGS.
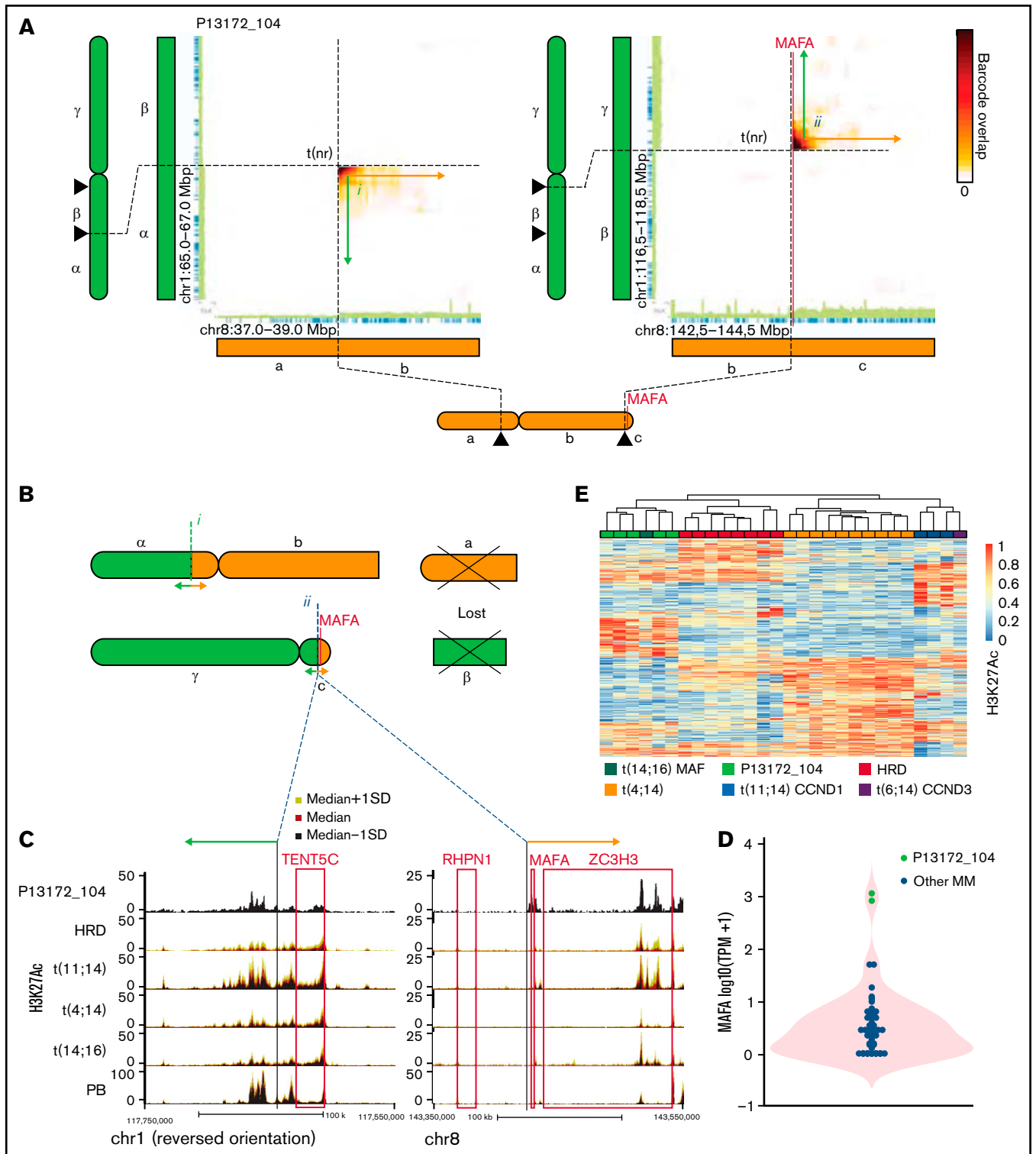
Private SVs cause the overexpression of MAFA and MAP3K14

To determine if we could find additional SVs that may represent potential primary or high-risk variants, we investigated whether we could identify private SVs affecting loci previously found to be translocation partners of the IGH locus.[2,13,55] This identified a t(1;8) in P13172_104 involving the MAFA locus and a t(6;17) in P13172_101 involving the MAP3K14 locus.

Patient P13172_104 had 2 SVs involving chromosomes 1 and 8, with the second juxtaposing the MAFA and TENT5C loci (Figure 6A-C). Analysis of the read clouds in the breakpoint regions suggested that the SV involving MAF was clonal (supplemental Figure 14A-B). Looking closer at the breakpoints, we found that the MAFA gene was juxtaposed with a strong MM-active enhancer region normally flanking the TENT5C gene (Figure 6C). H3K27Ac ChIPseq confirmed that both the TENT5C flanking enhancer and MAFA proximal elements were highly active in MM cells from P13172_104. Therefore, this strongly suggests that the TENT5C enhancer causes activation of the MAFA gene. In line with this, we found very high MAFA expression (~1000 TPMs) specifically in the MM cells from P13172_104 (Figure 6D). Given that MAFA is part of the Maf transcription factor family, we next wanted to assess if the MAFA overexpression resulted in a gene-regulatory landscape similar to that in t(14;16) (MAF-overexpressing) MM. Looking at MM subgroup–specific H3K27Ac signals, hierarchical clustering of all the samples revealed that P13172_104 clustered with the t(14;16) MM and shared H3K27Ac features with the MAF-overexpressing group (Figure 6E). Thus, the private t(1;8) translocation seen in P13172_104 results in the overexpression of MAFA and the establishment of an MAF-type gene regulatory landscape.

The SV affecting the MAP3K14 locus structurally constitutes a focal amplification on chromosome 17 carrying a templated insertion of a small chromosome 6 region (Figure 7A-B). Analysis of read-cloud overlaps indicated that the SV was subclonal (supplemental

---

Figure 5 (continued) Diverse SVs affecting the MYC locus can be identified by lrWGS. (A) Heat maps displaying the number of read clouds shared between the MYC locus and indicated genomic locations. Data from patients (n = 9) with detected SVs affecting the genomic MYC regions are shown. The connected green and orange arrows schematically indicate the areas joined by the SV. The type of SV is indicated next to the read-cloud clusters. Figure 3 and supplemental Figure 4 provide details on the interpretation of the read-cloud patterns. If the SV depicted is part of a more complex SV, this is indicated. (B-C) Read-cloud overlap between the MYC locus on chr8 and chr5 in P14402_118 (B) or chr12 in P13756_108 (C). In panels B-C, features (i-ii) used to resolve the structure of the derivate locus are indicated; schematic representations of the derivate chromosome regions are shown on the right; and tracks show the H3K27Ac signals (median ± standard deviation [SD]) for the involved regions in the indicated MM subtypes and plasmablasts (PBs). fa+ti, templated insertion occurring in a focal amplification; fa+t(r), reciprocal translocation occurring in a focal amplification; t(nr), nonreciprocal translocation; t(r), reciprocal translocation.

Figure 6. **Private SVs cause massive MAFA overexpression and reprogramming to MAF-type MM.** (A) Heat maps displaying the number of read clouds shared between indicated regions on chromosome 1 (chr1) and chr8. The connected green and orange arrows schematically indicate the areas joined by the SV. The type of SV is indicated. (B) Schematic representation of the derivate chromosomes. (C) Tracks showing the H3K27Ac signal (median ± standard deviation [SD]) in the regions involved in t(1;8). (D) Expression of MAFA. (E) Row-normalized hierarchically clustered heat map of H3K27Ac-marked regions showing differential signals between MM subtypes. PB, plasmablast; t(nr), nonreciprocal translocation.
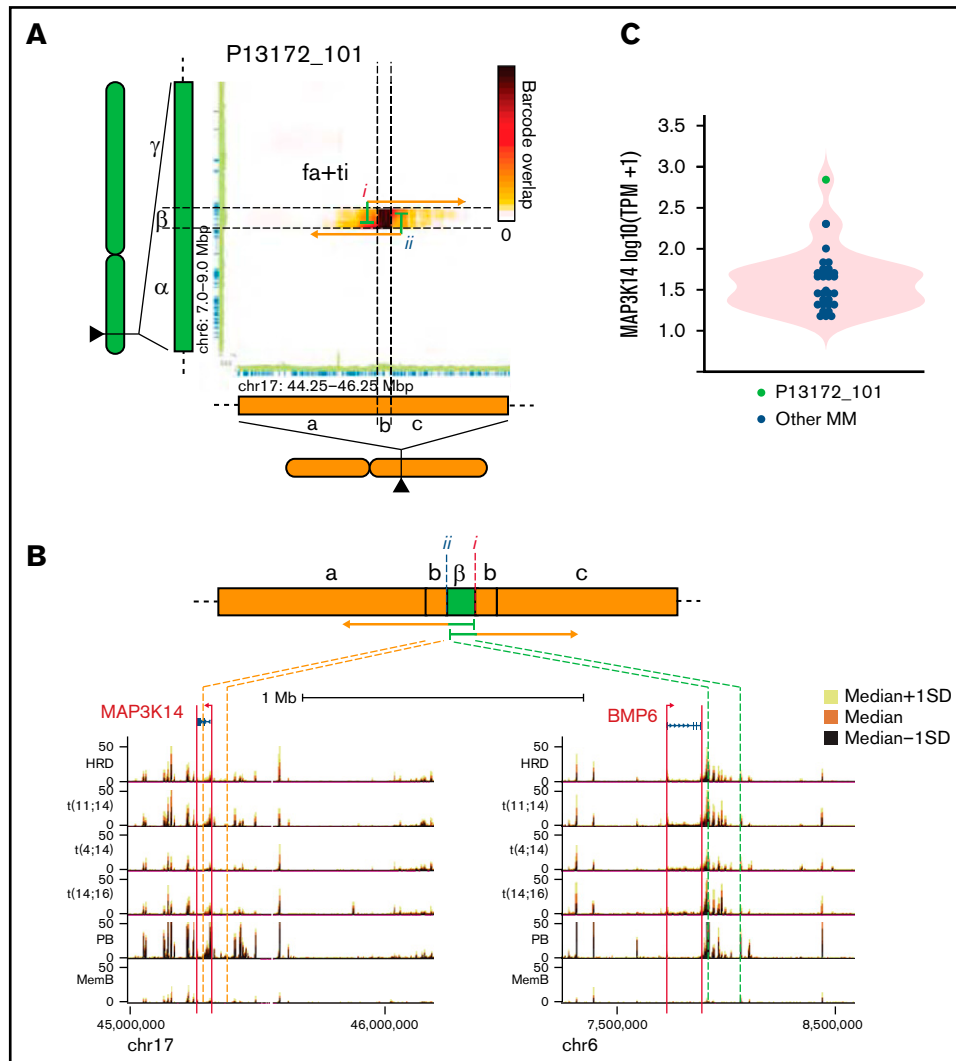
Figure 7. Insertion of the BMP6 flanking enhancer region causes the overexpression of MAP3K14. (A) Heat maps displaying the number of read clouds shared between indicated regions on chromosome 6 (chr6) and chr17. The connected green and orange arrows schematically indicate the areas joined by the SV. The type of SV is indicated. (B) Schematic representation of the derivate chromosome and tracks showing the H3K27Ac signal (median ± standard deviation [SD]) in the regions involved in the SV. (C) Expression of MAP3K14. fa+ti, templated insertion occurring in a focal amplification.

Figure 14C). Analyzing the involved regions, we found that the SV juxtaposed an enhancer region normally flanking the BMP6 gene with the amplified region on chromosome 17 (Figure 7B). The latter contained part of the *MAP3K14* gene but lacked the 3′ most exon, indicating that the amplification does not result in 2 functional copies of the gene. In line with the *BMP6* flanking element dysregulating the MAP3K14 gene, expression was high (>300 TPMs) in P13172_101 and almost 10-fold higher than any other investigated MM (Figure 7C). With *MAP3K14* functioning as an activator of *NF-κB*, this could suggest that the MM subclone carrying this SV could present with elevated or constitutive NF-κB signaling.[56]

Discussion

MM remains an incurable disease with poor outcome for most patients. Therefore, developing strategies to provide comprehensive genetic characterization in clinical routine is of paramount significance to understand the disease at the level of individual patients and establish personalized medicine programs aiming to improve outcome and survival.

Here we performed a proof-of-principle study involving 37 patients with MM and demonstrate that lrWGS can be used to provide comprehensive genetics in MM. Compellingly, lrWGS provided accurate and straightforward identification of the recurrent IGH and IGL translocations through simply visualizing read clouds spanning common breakpoint regions. Given the specificity of the read-cloud clusters, these SVs can be rapidly identified, even without computational analysis (apart from base processing) or germ line controls. In a clinical setting, this is attractive, because it would allow for performance of a quick screen of common SVs before a more time-consuming holistic analysis. As exemplified by the SV affecting both the IGH and *MYC* loci, analysis of the read-cloud clusters can identify and resolve diverse SVs in a simpler and more accurate way than possible with FISH or conventional WGS. Overall, given the high concordance between lrWGS and FISH, this suggests that

lrWGS can function as a standalone assay that can alleviate the need to complement sequencing-based analysis with optical mapping[57,58] or other technologies aimed at mapping SVs.

Taking advantage of the possibility to holistically analyze genetic aberrations in the lrWGS data, we identified both recurrent and private genetic events associated with poor outcome that would have remained undetected in the current clinical routine genetics. Among the recurrent variants, we found cases with t(8;22) MYC-IGL translocations,[5] high-risk double-hit TP53 inactivation, and chromothripsis, which was recently linked to poor outcome.[16] Interestingly, we could also find private structural variants causing the overexpression of genes previously shown to be translocation partners of the IGH locus (*MAP3K14* and *MAFA*). In particular, the t(1;8) involving the MAFA gene is notable, because this SV likely represents an alternative primary event that would be the molecular equivalent of t(14;16) deregulation of *MAF* with a similar patient risk profile and innate resistance to proteasome inhibitors.[59] Furthermore, we found that samples with t(11;14)s often carried duplications of the 3′RR and CCND1 regions. The fact that the t(11;14) MMs analyzed were randomly included suggests that this is a rather common phenomenon, which can be found both in cases with and without involvement of the derivate t(11;14) chromosomes in complex SVs Insert period

Nominally, lrWGS can be performed on very limited amounts of cells (≤1000 cells or 1-5 ng of DNA), but the need for HMW DNA preparations has effectively barred this. Circumventing the need for DNA preparation, as done in this study, to fully take advantage of the low requirements on input material has interesting implications for how to perform routine genetics in MM and other hematological malignancies. By running diagnostic flow cytometry in FACS, lrWGS-based genetics can be performed on any defined tumor or normal cell population. Given the low number of cells needed, this could even be performed on samples where tumor cells are expected to be rare or low in number, including samples from fine-needle biopsies or samples taken for minimal residual disease analysis. This approach would simultaneously allow for performing RNA sequencing and other assays on the pure target cell population in parallel with genetics. The more recent introduction of more automated FACS has the potential to make this kind of analytical workflow feasible to implement in diagnostic flow cytometric laboratories.

Although the removal of Chromium Genome from the market undoubtedly has hampered the wider introduction of lrWGS, several other platforms have emerged to take its place, including single-tube long fragment read,[60] transposase enzyme-linked long-read sequencing,[61] and droplet barcode sequencing.[24] Although larger validation studies must be performed to demonstrate that these effectively can replace the 10X platform, the simple workflow of the transposase enzyme-linked long-read sequencing protocol makes this an interesting candidate for potential clinical implementation.

In summary, lrWGS can with relative ease identify and resolve the structure of diverse SVs, making it an attractive solution for providing comprehensive genetics in MM and other hematological malignancies with complex genomes. Furthermore, using FACS in diagnostics could facilitate lrWGS of pure tumor cells as well as molecular characterization at both the expression and epigenetic levels. This could provide a powerful platform for understanding underlying genetics and disease mechanisms in individual patients.

Authorship

Contribution: R.M. proposed the study and devised strategies for performing lrWGS without DNA preparation; L.P.-P., N.F., C. Gran, and R.M. planned the study; A.W. biobanked patient samples and performed FISH; C. Gran, C. Gustafsson, and F.T.-L. performed proof-of-principle lrWGS experiments; R.-A.O. and P.E. performed initial analysis of proof-of-principle experiments; N.F., J.H., C. Gran, A.K., and R.M. performed FACS of cells; C. Gran, C. Gustafsson, and F.T.-L. prepared 10X lrWGS libraries; L.P.-P. analyzed lrWGS data; J.E., M.K., and A.L. assisted in establishing analysis pipelines; C. Gustafsson performed ChIPseq; J.H. analyzed ChIPseq data; C. Gustafsson and N.F. performed RNA sequencing (RNAseq); L.P.-P. and N.F. analyzed RNAseq data; P.E., A.L., and R.M. supervised the study; L.P.-P. and R.M. integrated data; L.P.-P. and R.M. wrote the manuscript with input from the other authors; and all authors reviewed the manuscript before submission.

Conflict-of-interest disclosure: A.L. has received honoraria from Illumina. The remaining authors declare no competing financial interests.

ORCID profiles: L.P.-P., 0000-0002-5044-7754; N.F., 0000-0003-1834-7638; J.H., 0000-0001-6674-4297; F.T.-L., 0000-0001-7198-5116; R.M., 0000-0003-0738-0328.

Correspondence: Robert Månsson, Center for Hematology and Regenerative Medicine, Karolinska Institutet, Neo Building, floor 7, Blickagången 16, 141 52 Huddinge, Sweden; e-mail: robert.mansson@ki.se.

References

1. Palumbo A, Anderson K. Multiple myeloma. *N Engl J Med.* 2011;364(11):1046-1060.

2. Manier S, Salem KZ, Park J, Landau DA, Getz G, Ghobrial IM. Genomic complexity of multiple myeloma and its clinical implications. *Nat Rev Clin Oncol.* 2017;14(2):100-113.

3. Rajkumar SV. Multiple myeloma: 2020 update on diagnosis, risk-stratification and management [published correction appears in *Am J Hematol*. 2020;95(11):1444]. *Am J Hematol.* 2020;95(5):548-567.

4. Goldschmidt H, Lokhorst HM, Mai EK, et al. Bortezomib before and after high-dose therapy in myeloma: long-term results from the phase III HOVON-65/GMMG-HD4 trial. *Leukemia.* 2018;32(2):383-390.

5. Walker BA, Mavrommatis K, Wardell CP, et al. A high-risk, double-hit, group of newly diagnosed myeloma identified by genomic analysis. *Leukemia.* 2019;33(1):159-170.

6. Morgan GJ, Walker BA, Davies FE. The genetic architecture of multiple myeloma. *Nat Rev Cancer.* 2012;12(5):335-348.

7. Ghazzaui N, Saintamand A, Issaoui H, Vincent-Fabert C, Denizot Y. The IgH 3′ regulatory region and c-myc-induced B-cell lymphomagenesis. *Oncotarget.* 2017;8(4):7059-7067.

8. Lohr JG, Stojanov P, Carter SL, et al; Multiple Myeloma Research Consortium. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer Cell.* 2014;25(1):91-101.

9. Walker BA, Wardell CP, Brioli A, et al. Translocations at 8q24 juxtapose MYC with genes that harbor superenhancers resulting in overexpression and poor prognosis in myeloma patients. *Blood Cancer J.* 2014;4(3):e191.

10. Walker BA, Boyle EM, Wardell CP, et al. Mutational spectrum, copy number changes, and outcome: results of a sequencing study of patients with newly diagnosed myeloma. *J Clin Oncol.* 2015;33(33):3911-3920.

11. Bolli N, Biancon G, Moarii M, et al. Analysis of the genomic landscape of multiple myeloma highlights novel prognostic markers and disease subgroups. *Leukemia.* 2018;32(12):2604-2616.

12. Maura F, Bolli N, Angelopoulos N, et al. Genomic landscape and chronological reconstruction of driver events in multiple myeloma. *Nat Commun.* 2019;10(1):3835.

13. Barwick BG, Neri P, Bahlis NJ, et al. Multiple myeloma immunoglobulin lambda translocations portend poor prognosis. *Nat Commun.* 2019;10(1):1911.

14. Hultcrantz M, Yellapantula V, Rustad EH. Genomic profiling of multiple myeloma: new insights and modern technologies. *Best Pract Res Clin Haematol.* 2020;33(1):101153.

15. Mikulasova A, Ashby C, Tytarenko RG, et al. Microhomology-mediated end joining drives complex rearrangements and overexpression of *MYC* and *PVT1* in multiple myeloma. *Haematologica.* 2020;105(4):1055-1066.

16. Rustad EH, Yellapantula VD, Glodzik D, et al. Revealing the impact of structural variants in multiple myeloma. *Blood Cancer Discov.* 2020;1(3): 258-273.

17. Demchenko Y, Roschke A, Chen W-D, Asmann Y, Bergsagel PL, Kuehl WM. Frequent occurrence of large duplications at reciprocal genomic rearrangement breakpoints in multiple myeloma and other tumors. *Nucleic Acids Res.* 2016;44(17):8189-8198.

18. Cardona-Benavides IJ, de Ramón C, Gutiérrez NC. Genetic abnormalities in multiple myeloma: prognostic and therapeutic implications. *Cells.* 2021;10(2):336.

19. Harding T, Baughn L, Kumar S, Van Ness B. The future of myeloma precision medicine: integrating the compendium of known drug resistance mechanisms with emerging tumor profiling technologies. *Leukemia.* 2019;33(4):863-883.

20. Manni S, Fregnani A, Barilà G, Zambello R, Semenzato G, Piazza F. Actionable strategies to target multiple myeloma plasma cell resistance/resilience to stress: insights from "omics". *Front Oncol.* 2020;10:802.

21. Marks P, Garcia S, Barrio AM, et al. Resolving the full spectrum of human genome variation using linked-reads. *Genome Res.* 2019;29(4):635-645.

22. Zheng GXY, Lau BT, Schnall-Levin M, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol.* 2016;34(3):303-311.

23. Chen Z, Pham L, Wu T-C, et al. Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. *Genome Res.* 2020;30:898-909.

24. Redin D, Frick T, Aghelpasand H, et al. High throughput barcoding method for genome-scale phasing. *Sci Rep.* 2019;9(1):18116.

25. Bell JM, Lau BT, Greer SU, et al. Chromosome-scale mega-haplotypes enable digital karyotyping of cancer aneuploidy. *Nucleic Acids Res.* 2017; 45(19):e162.

26. Xia LC, Bell JM, Wood-Bouwens C, Chen JJ, Zhang NR, Ji HP. Identification of large rearrangements in cancer genomes with barcode linked reads. *Nucleic Acids Res.* 2018;46(4):e19.

27. Eisfeldt J, Pettersson M, Vezzi F, et al. Comprehensive structural variation genome map of individuals carrying complex chromosomal rearrangements. *PLoS Genet.* 2019;15(2):e1007858.

28. Greer SU, Nadauld LD, Lau BT, et al. Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. *Genome Med.* 2017;9(1):57.

29. Viswanathan SR, Ha G, Hoff AM, et al; PCF/SU2C International Prostate Cancer Dream Team. Structural alterations driving castration-resistant prostate cancer revealed by linked-read genome sequencing. *Cell.* 2018;174(2):433-447.e19.

30. Collins RL, Brand H, Redin CE, et al. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.* 2017;18(1):36.

31. Nordlund J, Marincevic-Zuniga Y, Cavelier L, et al. Refined detection and phasing of structural aberrations in pediatric acute lymphoblastic leukemia by linked-read whole-genome sequencing. *Sci Rep.* 2020;10(1):2512.

32. Frigyesi I, Adolfsson J, Ali M, et al. Robust isolation of malignant plasma cells in multiple myeloma. *Blood.* 2014;123(9):1336-1340.

33. Jin Y, Chen K, De Paepe A, et al. Active enhancer and chromatin accessibility landscapes chart the regulatory network of primary multiple myeloma. *Blood.* 2018;131(19):2138-2150.

34. Gran C, Uttervall K, Borg Bruchfeld J, et al. Translocation (11;14) in newly diagnosed multiple myeloma, time to reclassify this standard risk chromosomal aberration? *Eur J Haematol.* 2019;103(6):588-596.

35. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLOS Comput Biol.* 2016;12(4):e1004873.

36. Van Loo P, Nordgard SH, Lingjaerde OC, et al. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci USA.* 2010;107(39): 16910-16915.

37. Nik-Zainal S, Van Loo P, Wedge DC, et al; Breast Cancer Working Group of the International Cancer Genome Consortium. The life history of 21 breast cancers [published correction appears in *Cell.* 2015;162(4):924]. *Cell.* 2012;149(5):994-1007.

38. Garcia M, Juhos S, Larsson M, et al. Sarek: a portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000 Res.* 2020;9:63.

39. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods.* 2018;15(8):591-594.

40. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol.* 2013;31(3):213-219.

41. McLaren W, Gil L, Hunt SE, et al. The ensembl variant effect predictor. *Genome Biol.* 2016;17(1):122.

42. Cingolani P, Platts A, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* 2012;6(2):80-92.

43. Spies N, Weng Z, Bishara A, et al. Genome-wide reconstruction of complex structural variants using read clouds. *Nat Methods.* 2017;14(9): 915-920.

44. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol.* 2011;29(1):24-26.

45. Glaros V, Rauschmeier R, Artemov AV, et al. Limited access to antigen drives generation of early B cell memory while restraining the plasmablast response. *Immunity.* 2021;54(9):2005-2023.e10.

46. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15-21.

47. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell.* 2010;38(4):576-589.

48. Gustafsson C, De Paepe A, Schmidl C, Månsson R. High-throughput ChIPmentation: freely scalable, single day ChIPseq data generation from very low cell-numbers. *BMC Genomics.* 2019;20(1):59.

49. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12): 550.

50. Zook JM, Catoe D, McDaniel J, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data.* 2016;3:160025.

51. Höllein A, Twardziok SO, Walter W, et al. The combination of WGS and RNA-Seq is superior to conventional diagnostic tests in multiple myeloma: ready for prime time? *Cancer Genet.* 2020;242:15-24.

52. Klintman J, Barmpouti K, Knight SJL, et al. Clinical-grade validation of whole genome sequencing reveals robust detection of low-frequency variants and copy number alterations in CLL. *Br J Haematol.* 2018;182(3):412-417.

53. Sharma N, Smadbeck JB, Abdallah N, et al. The prognostic role of MYC structural variants identified by NGS and FISH in multiple myeloma. *Clin Cancer Res.* 2021;27(19):5430-5439.

54. Cho SW, Xu J, Sun R, et al. Promoter of lncRNA gene PVT1 is a tumor-suppressor DNA boundary element. *Cell.* 2018;173(6):1398-1412.e22.

55. Chng WJ, Glebov O, Bergsagel PL, Kuehl WM. Genetic events in the pathogenesis of multiple myeloma. *Best Pract Res Clin Haematol.* 2007; 20(4):571-596.

56. Gardam S, Beyaert R. The kinase NIK as a therapeutic target in multiple myeloma. *Expert Opin Ther Targets.* 2011;15(2):207-218.

57. Neveling K, Mantere T, Vermeulen S, et al. Next-generation cytogenetics: comprehensive assessment of 52 hematological malignancy genomes by optical genome mapping. *Am J Hum Genet.* 2021;108(8):1423-1435.

58. Levy B, Baughn LB, Chartrand S, et al. A national multicenter evaluation of the clinical utility of optical genome mapping for assessment of genomic aberrations in acute myeloid leukemia. *medRxiv.* Preprint posted online 10 November 2020.

59. Qiang Y-W, Ye S, Chen Y, et al. MAF protein mediates innate resistance to proteasome inhibition therapy in multiple myeloma. *Blood.* 2016; 128(25):2919-2930.

60. Wang O, Chin R, Cheng X, et al. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res.* 2019;29(5):798-808.

61. Chen Z, Pham L, Wu T-C, et al. Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information [published correction appears in *Genome Res.* 2021;31(5):934]. *Genome Res.* 2020;30(6):898-909.