Rapid detection of myeloid neoplasm fusions using Single Molecule Long-Read Sequencing

Olga Sala-Torra MD[1,3], Shishir Reddy [3], Ling-Hong Hung[3], Lan Beppu[1], David Wu[2], Jerald Radich, MD[1,2], Ka Yee Yeung[3], Cecilia CS Yeung, MD[1,2]

[1]Fred Hutchinson Cancer Research Center, Seattle, WA

[2]University of Washington, Seattle, WA

[3]School of Engineering and Technology, University of Washington Tacoma, Tacoma, WA

Keywords (6): Long-read sequencing, CRISPR-Cas9, myeloid neoplasm, nanopore, CML, sequencing

Short Title: **Long-Read Sequencing for Myeloid Neoplasm Fusions**

Address correspondence to:

Cecilia Yeung, MD

1100 Fairview Ave N, G7-910

Seattle, WA 98109

cyeung@fredhutch.org

Key points:

1) We describe a CRISPR-Cas9 enrichment Nanopore sequencing assay with streamlined bioinformatics that outperforms other fusion detectors.
2) We successfully detected both fusion genes and specific breakpoints in CML, APL, and AML in under 8 hours in 80% of patients.

Visual Abstract (Figure 1):

1) We successfully detected fusion genes in hematological malignancies with a fast and efficient long-read sequencing workflow in under 8 hours. The method makes the genomic characterization of *BCR-ABL1* DNA breakpoint in patients quick and simple, which potentiates design of patient specific primers for personalized monitoring MRD assays.
2) Our assay is based on a CRISPR-Cas9 non-amplification enrichment library preparation strategy and uses Nanopore sequencing single stranded genomic DNA coupled with streamlined bioinformatic workflow containing a novel fusion detector software which outperforms current fusion detection software.

Abstract

Recurrent gene fusions are common drivers of disease pathophysiology in leukemias. Identification of these structural variants helps stratify disease by risk and assists with therapy choice. Current fusion detection methods require long turnaround time (7-10 days) or advance knowledge of the genes involved in the fusions. To address the need for rapid identification of clinically actionable fusion genes in heme malignancies without *a-priori* knowledge of the

genes, we describe a long-read sequencing DNA assay designed with CRISPR guides to select and enrich for recurrent leukemia fusion genes. By applying rapid sequencing technology based on nanopores, we sequenced long pieces of genomic DNA and successfully detected fusion genes in cell lines and primary specimens (e.g., *BCR-ABL1, PML-RARA, CBFB-MYH11, KMT2A-AF4*) using cloud-based bioinformatics workflows with novel custom fusion finder software. We detected fusion genes in 100% of cell lines with the expected breakpoints and confirmed the presence or absence of a recurrent fusion gene in 12 of 14 patient cases. With our optimized assay and cloud-based bioinformatics workflow, these assays and analyses could be performed in under 8 hours.

Introduction

Current classification of myeloid malignancies is largely based on the molecular and genetic aberrations [1,2]. Recurrent gene rearrangements are present in 30-40% of acute myeloid leukemias (AML), and well described driver fusions that in some cases suffice to diagnose leukemias with *PML-RARA, RUNX1-RUNX1T1,* etc. even when the blast percentage is below 20%[1]. Recurrent fusion genes confer certain clinical and biological characteristics as drivers of leukemogenesis, and their identification assists in prognosis stratification and inform treatment decisions. Identification of driver fusion genes is especially relevant when targeted therapies are available. Examples include tyrosine kinase inhibitors (TKI) in chronic myelogenous leukemia (CML)[3] that bind to the kinase domain in ABL1 deregulated as a consequence of the fusion with BCR, and differentiation therapy in acute promyelocytic leukemia (APL) in which identification of the promyelocytic leukemia- retinoic acid receptor alpha (*PML-RARA*) fusion confirms the

66 diagnosis and indicates that the patient will likely respond to treatment with all-trans retinoic
67 acid (ATRA) therapy, a nontoxic and highly effective treatment[4] that achieves 90% long term
68 survival rates when combined with anthracyclines and/or arsenic trioxide.[5]
69
70 Fusion genes can be detected through several techniques that include cytogenetics techniques
71 such as fluorescence *in situ* hybridization (FISH), and molecular testing such as polymerase
72 chain reaction (PCR) and next generation sequencing (NGS), which are the clinical gold
73 standard. Clinical molecular assays for *BCR-ABL1* and *PML-RARA* primarily target RNA
74 transcripts by RT-PCR because of the greater abundance of fusion gene transcripts compared to
75 the DNA copies per cell, and because of the large variability in the genomic sequence of the
76 fusion that can encompass large intronic region making routine DNA PCR impossible (e.g., *BCR-*
77 *ABL*).
78
79 Typical NGS reads are 150 to 250 bps in length which are not sufficiently long enough to extend
80 over DNA fragments that are adjoined because of a large genomic aberration thus hindering
81 both alignment and detection of large indels and other structural abnormalities. NGS is limited
82 by short read assembly mis-mapping, and amplification strategies instill imperfect quantitation
83 of the variant allele frequencies. The diversity of aberrations in myeloid neoplasms include
84 large genomic aberrations, including insertions and deletions, loss of heterozygosity, single
85 nucleotide polymorphism, mutations in homopolymer rich regions and highly repetitive regions
86 such as internal tandem duplications, that are difficult to detect by NGS and require different
87 assays in molecular and cytogenetics labs additionally to those used to confirm fusion genes.

Hence, currently a multiple assay approach is used to obtain a complete diagnostic molecular picture in myeloid malignancies.

The advancement of long-read sequencing technologies has enabled the sequencing of continuous single DNA or RNA molecules up to tens to hundreds of kilobases (kb) long.[6] Ongoing improvements in Nanopore sequencing accuracy have reduced error rates to less than 5%,[7] but remain higher than those for Illumina and Ion Torrent, which are used frequently in clinical laboratories.[8] This technology has already made an impact on the understanding of the pathobiology of various diseases,[9,10] and its impact will increase as the quality of sequencing improves and becomes more accurate.[7,11] Addition of CRISPR-Cas9 for targeted enrichment concentrates the regions of interests prior to sequencing by Nanopore[12] without requiring amplification steps, thus optimizing sequencing time and efficiency. Additionally, the portability and affordability of the Nanopore sequencer MinION and Flongle, hold great promise to impact the clinical field. However, a major limitation has been the lack of analytical software featuring standardized parameters to aid in translation into clinical diagnostics.

Here we report our success in developing an amplification-free CRISPR-Cas9 targeted enrichment sequencing protocol using Nanopore MinION and Flongles to detect fusions relevant in the diagnosis and classification of CML and AMLs. The ONT Flongle is an adaptor for the MinION that provides cost-effective (~$90 per flow cell), real-time sequencing for smaller assays. Our assays were designed to capture varied breakpoints of CML and APL, as well as fusion genes resulting from inv(16) (*MYH11-CBFB*) and t(4;11)(*KMT2A-AF4*). Simultaneous

interrogation of these targets is a first step to a rapid characterization of AMLs in a single assay combining data that up to now required multiple different techniques and provide relevant information promptly. In addition to this amplification-free CRISPR-Cas9 nanopore assay, we extended our previously developed cloud-based nanopore data analysis pipeline [13] to include fusion detection and develop a custom breakpoint detection tool (see **Figure 1**). Using our optimized assay and our custom breakpoint finder, we showed that we can reliably detect and confirm fusion breakpoints in 80% of our specimens in under 3 hours of sequencing and data analysis.

Methods:

Cells lines and patient samples

Our assay was optimized using 6 cell lines: three with the *BCR-ABL1* fusion (K562, KU812, and KCL22), and NB4, MV4;11 and ME-1 that bear the *PML-RARA, KMT2A-AF4, and MYH11-CBFB* fusions respectively. Residual mononuclear cells from primary specimens (6 specimens from 5 patients with CML, 6 specimens from 5 patients with suspected APL, and 2 acute myeloid leukemia, not acute promyelocytic leukemia) were isolated using Ficoll® reagent (Millipore-Sigma) and banked in liquid nitrogen until the time of the experiment. All specimens had been originally tested in a CLIA certified laboratory according to standard clinical protocols.[14] IRB coverage was obtained for use of residual laboratory samples. Patient samples were de-identified to the nanopore testing lab, and cytogenetic or molecular results were confirmed after nanopore results were rendered. Characteristics and demographics of specimens and patients are listed in **Table 1**.

Library preparation and sequencing assay

For the cell lines and 11/14 patient specimens the DNA was extracted with PureGene (Qiagen, Germantown, MD, USA) following the standard protocol. Special caution, including use of wide bore pipette tips and moderate centrifuge spin velocity was exercised to minimize fragmenting DNA strands. Two DNA specimens were extracted with AllPrep DNA/RNA Kit (Qiagen, Germantown, MD, USA), and one with QiAgen X-tractor with Reagent Pack DX (Qiagen, Germantown, MD, USA). cRNA guides were designed to direct Cas9 to cut on genomic proximity of each of the regions involved in each one of the translocations studied. When the target region was big, guides were tiled across the region to maximize coverage. Guides were designed to capture *PML-RARA*, *BCR-ABL1* p210, *KMT2A-AF4*, and *MYH11-CBFB*, including different fusion isoforms. Guides were designed using Chopchop (https://chopchop.cbu.uib.no/) with the CRISPR-Cas9 and nanopore enrichment settings and previously described [15].

We used 5 micrograms of DNA as input for each cell line and 2 to 5 micrograms for primary specimens. Average DNA integrity number (DIN) was 9.2 (range: 7.5-9.8). **Figure 1** shows a schematic of our workflow and details of the library prep are published[15]. Briefly, enrichment of target regions was obtained using Oxford Nanopore Technologies "Targeted, amplification-free DNA sequencing using CRISPR-Cas9" protocol[12]. The different guides used in the assay were pooled in equimolar amounts of each guide. Through an initial dephosphorylation step, the 5' ends of the DNA becomes inaccessible to adapter ligation. Double stranded DNA breaks that

154 excise the region of interest are generated with the directional, target specific RNA guides
155 complexed with tracrRNA and Cas9 enzyme. The Cas9 complex remains bound to the 5' end of
156 the guide, and the resulting new DNA ends contain a phosphorylated 5'end that is available for
157 dA tailing and adapter ligation.[12] All libraries generated in this manner were run on a MinION
158 version 9.4. (Oxford Nanopore Technologies, Oxford, UK) nanopore sequencer using flow cells
159 or Flongles. Modifications for libraries sequenced on the Flongle were only at the library loading
160 step, in which the amount of Sequencing Buffer and library beads (both SQK-LSK 109, ONT) are
161 reduced to from 35 to 13 and 25.5 to 7.5UL respectively, and 0.5UL of SQT is added. QC
162 parameters tracked for each run are listed in **Table 2**.
163
164 **PCR/Sanger sequencing**
165 Primers specific to *BCR-ABL1* patient breakpoints were designed using Primer3 v. 0.4.0.[16] in 2
166 cases. 100 ng of DNA were amplified, the PCR product was run on a 2% gel and Sanger
167 sequenced to confirm the genomic breakpoint.
168
169 **Biodepot-workflow-builder: Interactive and accessible front end for fusion detection**
170 We present a graphical, reproducible, and cloud-enabled fusion detection workflow consisting
171 of all the steps of the analyses, including base calling, alignment, fusion detection, and
172 visualization (see **Figure 1 screenshot of BwB interface)**. In contrast to the NanoFG workflow
173 from Stangle et al[17], our platform includes the computationally intensive base calling step, an
174 interactive graphical user interface, and can readily leverage GPUs and be deployed on the
175 cloud. Thus, analyses are fast, and our platform is accessible to biomedical and clinical

scientists. Specifically, we extended the Biodepot-workflow-builder (Bwb)[18] platform in which each computational task (or module) is represented by a graphical widget that calls a software container in the back end. Software containers, such as Docker, include all software dependencies and libraries required to execute the code. We have recently developed a Bwb workflow[13] to support the processing of Nanopore data that includes the use of base callers Guppy[19] and Bonito[19], alignment using minimap2 and visualization of results using the Integrative Genomics Viewer (IGV) and GRCh37 hg19 as reference genome QC data was generated with PycoQC[20]. Guppy was used as the base caller. Minimap2[21] was used as the aligner and variant caller. Fusions are visualized on IGV[22] and confirmed on Blast. Most importantly, we extend our previous work by adding support for fusion detection, including LongGF[23] and our own custom software "Biodepot Fusion Finder" (BFF).

Bioinformatics pipeline for fusion detection: LongGF

LongGF[23] is a software tool for fusion detection optimized for the high base calling error rates and alignment errors commonly found in long read sequencing data. LongGF takes as input a BAM file containing alignments (generated by minimap2 in this pipeline) and a GTF file containing the definitions of known genes. The output is a log file with detected gene fusions and their supporting reads. We created a graphical widget for LongGF in the Bwb. **Figure 2** shows a screenshot of the comparative workflows with output for LongGF versus BFF.

Custom fusion detection tool: Biodepot Fusion Finder (BFF)

Reads that span a fusion gene will align to coordinates in both parts of the fusion and provide a specific breakpoint coordinate. We wrote a new software tool, the Biodepot Fusion

Finder, BFF that examines the alternate alignments for each read and identifies reads that map to coordinates spanning a set of known breakpoints. This is similar to the strategy used by LongGF except that we allow for errors in the alignment near the breakpoint. Accordingly, the Breakpoint Finder identifies fusions that are not detected by LongGF which looks for supporting reads where the breakpoints are exactly matched. The user can provide a panel of breakpoint coordinates of interest. If no panel is provided, BFF will return candidate fusions that span nonadjacent genomic regions. Incomplete coordinates for breakpoints in the panel are supported – the user does not need to define the exact coordinates, nor do both breakpoints need to be given. BFF will return all the reads that match the panel of breakpoints in a text file for further inspection by the user if desired. The user can also provide a file with guide coordinates to obtain additional enrichment metrics. A containerized widget was developed that can integrated with the Bwb workflows for processing nanopore data. Using these workflows, we can directly analyze raw nanopore data and obtain lists of candidate fusions.

Benchmarking experiments

We performed empirical experiments to benchmark the sensitivity and runtime required to reliably detect fusion. For each individual sample, sequencing metrics including quality scores and timestamps are obtained from the sequencing summary text file obtained as an output of base calling using Guppy. Detected fusions are then acquired from the Breakpoint Finder as well as LongGF. Specific fusion breakpoints are also provided in the data set from the BFF. All samples are then combined into a dictionary and separated by patient and cell line data. Plots

are constructed for each category of data pertaining to the time to reach 3 fusions as well as the number of reads required to reach 3 fusions. Finally, the total number of fusion reads detected for each sample is compared between the BFF and LongGF as shown in **Figure 2**.

Enrichment assessment

Two enrichment metrics were computed by BFF and tracked for each sample. First is the *fusion-specific enrichment* which is calculated with the following formula [(number of fusion reads) / (mean coverage of the genome)]. Second is the *on-target enrichment* which is calculated with the following formula [(number of reads that originate from a guide RNA cut point that includes the region of the breakpoint)/(mean genome coverage)]. Reads originating from a guide RNA cut are distinguished by the guide sequence being at the beginning of a read. As initial electrical signal data is generated by DNA passing through the nanopores (reads at the beginning of a strand) are error prone, which affects base calling and therefore the alignment of the reads so that the start sequence is often misaligned. Consequently, the BFF considers base pairs of the sequence near the start (default within 50 bp) of a read that aligns near the coordinates of a guide RNA cut site to have originated from a guide RNA cut. Specific reads cut by guides are manually confirmed. The allowed error intervals are customizable. SamTools v1.13 is used to sort and convert BAM files and determine the overall average coverage[24]. Picard CollectHsMetrics is used to generate the unique base pairs mapped metric for each sample[25].

Results

Sample sequencing and enrichment:

Details of the sample sequencing and enrichment metrics are included in Table 2. A range of 0.04 Gb – 5.47 gigabases of sequencing data was generated for each sample for an average mean coverage of the human genome of 0.32-fold (range: 0.01 – 1.66). We adopted standard quality metrics for nanopore workflows and tracked N50, which is a quality metric where half the reads are above this length (range 4.65kb-32.2kb) and median read length for total reads (range 0.87kb- 9.60kb). Percentage of reads aligned ranged from 72 - 96%. Fusion specific enrichment was 135 - 837 fold for cell lines and 6 - 509 fold for patient samples. On target enrichment was 849 - 5830 fold for cell lines and 535 - 3007 fold for patient samples.

Concordance of fusion detection with clinical results:

Concordance with expected results was 100% for cell lines as the expected fusion was detected in 6/6 cell lines (3 *BCR-ABL1*, 1 *PML-RARA*, 1 *MYH11-CBFB*, 1 *KMT2A-AFF1*). Breakpoint sequences detected for *BCR-ABL1* cell lines are the same as previously published [26]. We correctly confirmed the presence or absence of fusions in 11/14 (78.5%) primary specimens including both diagnostic and measurable residual disease (MRD) cases with a minimum of three reads, however one case (APL6) showed only 2 fusion reads and was not counted as confirmed. The 3 missed cases (CML3, APL1, and APL6 in tables 1 and 2) were comprised of low disease burden, ~0%, 1% and <5%. We detected *BCR-ABL1* in 5/6 cases, *PML-RARA* in 2/6 cases, *KMT2A-AF9* and *CBFB-MYH11* in 1/1 case each. *BCR-ABL1* genomic breakpoint was confirmed by Sanger sequencing in CML1 and CML2 after designing patient specific primers.

In 3 specimens from 2 patients with suspected APL, we could not detect *PML-RARA* but observed other findings. Clinical and laboratory details are listed in **Table 2**. For the first patient, two specimens, one bone marrow and one peripheral blood (APL4 and APL 5) yielded no *PML-RARA* fusion reads. This patient presented with an AML morphologically suggestive of APL and an isochromosome 17q without t(15;17) detected by karyotype. While fusion detection software did not detect a fusion, manual inspection showed an insertion in an intronic region of RARA with TTMV viral genome; this case was previously reported[15]. Patient APL6 (presented with 22% blasts on flow in a <5% marrow which on unblinding showed a complex karyotype with t(11;17)(q23;q25) including the *KMT2A* gene. Two reads with *KMT2A-SEPT9* fusion were detected in a suboptimal but acceptable run (N50 < 5000bp; on-target enrichment 650.66 fold), confirming the lack of t(15;17) or *PML-RARA* fusion but the threshold was below the requisite 3 reads to confirm the *KMT2A-SEPT9* fusion.

Comparison of fusion detection tools:

A comparison of the bioinformatic workflows for data analysis using different fusion detection widgets including LongGF vs Biodepot Fusion Finder (BFF) was conducted; the specific workflow is demonstrated in **Figure 2**. Additionally, BFF computes the fusion enrichment and on target enrichment statistics; these are summarized in **Table 2**. In most cell line and primary specimens, LongGF shows a particular challenge in the detection of BCR-ABL1 and does not detect all fusion reads that are identified by BFF. Using the BFF, the average sequencing, and data processing time to 3 reads with fusions in the cell line experiments was 42.75 minutes (range: 18.87 - 77.65 min) and 188 minutes in the primary specimens where 3 reads were detected (range: 32 - 654

min) [see **Figure 3**, and **Supplemental Table 1**]. Cell line experiments took an average of 11,711 reads to identify 3 fusion reads (range: 1,321 – 43,326 reads) and 10,273 reads in primary specimens (range: 1,790 - 24,999 reads) for confirmation of the fusion calls.

Comparison of Flongles and flow cells:

In five of our experiments (2 cell lines, 3 primary specimens), we used Flongles, while in the rest we used Flow Cells. The performance of the affordable Flongles was inferior to the Flow Cells, with lower expected average data output from the Flongles (based on manufacture expectations of ~3GB), the N50 and median read length were smaller in Flongle reads (21,614 vs 19,166 and 6058 vs 5250 respectively), and significantly, the median Phred score for Flongle reads being lower than that for the Flow Cells (9.2 vs 11.88). Despite the worse performance of the Flongles, fusions were detected in 2 of 2 cell lines, and 2 of the 3 experiments with primary specimens with at least 8 and 14 reads confirming the fusion and the specimen without fusion confirmation was CML3 with pancytopenia and low disease burden.

Discussion

We developed a rapid assay to detect fusion genes in blood or marrow samples in less than 8 hours with the fastest time achieving 3 fusion reads was 5 hours. This was accomplished by combining CRISPR-Cas9 enrichment during library preparation, nanopore long-read sequencing, a cloud-based data analytic pipeline, with a novel BFF program developed and optimized for finding fusions reads and breakpoints. RNA guides were designed to target genes involved in recurrent fusions in myeloid malignancies and used to enrich an amplification-free library

preparation over 1600-fold. Our modular and containerized pipeline in Bwb allows users to efficiently process raw FAST5 data on the cloud through an accessible graphical user interface allowing for a very fast analysis step (average ~4.5 minutes for basecalling, alignment, and fusion detection). To improve fusion calling, we developed the custom BFF that allow users to identify fusion reads not detected by LongGF. We successfully confirmed published genomic breakpoints in our series of cell lines and archival patient samples, which includes both diagnostic and follow-up samples to test feasibility in confirming both common and novel fusions over a range of tumor burdens. Our study includes 14 patient specimens and demonstrates the usability of this method in primary specimens with 2microgram of DNA.

An advantage of the CRISRP-Cas9 based enrichment protocol is that it allows for targeting of multiple common leukemia fusion genes by pooling multiple guide RNAs. Fusions are particularly well suited for this method as they have large gene segments that aid in alignment despite sequencing errors and wide variation where translocation breakpoints may occur. Other labs have employed similar but different methods that detected fusions by targeting one partner gene in the fusion [17]. In contrast, our assay targets both partners of the fusion thus our approach allows for an expanded capability to detect known and novel fusions, such as in case AML1 with t(9;11), when guides are designed to target t(4;11). However, in 2/14 patients the assay did not detect fusion genes because both cases had a very low amount of fusion target. In one case (CML3) because the BCR-ABL1 was <0.01%, the other in the context of a very hypocellular sample.

Our work differs from standard RNA-based fusion detection assays, and instead interrogated single molecules of DNA rapidly and accurately to detect specific translocation breakpoints. Long-read sequencing technologies, like nanopore (Oxford Nanopore Technologies, Oxford, U.K.), allow sequencing of unamplified, long unbroken fragments of DNA which are more likely to span a breakpoint. This has potential clinical utility for personalized disease monitoring when CML patients are on TKI therapy and suppressing RNA transcription[27]; targeting DNA as a monitoring target may be more robust and reproducible since DNA is stable and present in constant numbers[28]. However genomic breakpoints in *BCR-ABL1* are unique to individual patients requiring patient-specific breakpoint characterization[28-32] as *ABL1* breakpoints occur over an expansive region of about 150Kb, making this is an arduous endeavor previously involving multiple primer sets and Sanger sequencing[33]. Our method allows a single approach spanning the *BCR* and *ABL1* breakpoint regions without the use of multiple primers and PCR reactions. The sensitivity of DNA-based qPCR once the breakpoint is known can be as low as 10^{-7} [34-35]. Specimens CML1 and CML5 are a BM and PB obtained from the same patient and demonstrate high fidelity in confirming genomic breakpoints and the ability to use patient specific primers for personalized MRD monitoring.

Advantages of long-read sequencing over current clinical diagnostic assays are speed and the relatively low complexity of the assay when compared to cytogenetics and targeted NGS panels. While long-read sequencing results could potentially have a turnaround time (TAT) of less than 24 hours, full karyotype analysis TAT is generally longer with the fastest times at days to a week and most targeted NGS panels require ~7-10 days from start of processing to result

352 report. The nanopore sequencing streams data simultaneously to our GPU-enabled data
353 analytic pipeline in the Bwb interface which resides on the cloud to help interpret and reliably
354 identify fusion reads within 5000 seconds(<2 hours) computational time in most specimens.
355 Building on our experience[14,17], we used 3 reads as a threshold for fusion confirmations. With
356 current simultaneous sequencing and data analysis workflow described here, 3 sequences are
357 detected in an average of 3 hours and 7 min (fastest at 30 min) in the 9 patient specimens
358 where a fusion could be confirmed. This means that a diagnostic result with a precise fusion
359 breakpoint with 3 fusion supporting reads would be possible in the same day.
360
361 Five specimens were sequenced on the less costly Flongle device, which has lower sequencing
362 capabilities (pore count ~60, sequencing life 24h), but 3 fusion reads were reached in 4/5
363 specimens (2 cell lines, 3 patient samples) with less resources. The sequencing quality is
364 significantly lower on the Flongles (median Phred 9.2 in Flongles vs 11.9 in Flow Cells), however
365 fusions were detected in all samples with adequate tumor burden. To achieve a cheaper
366 version of our fast and portable nanopore fusion assay, further challenges will need to be
367 addressed. We predict overall less sequencing can be achieved, even with additional
368 optimizations of the CRISPR library guides as multiplexing of the guide RNA appears to increase
369 efficiency and on-target fusion reads. Nanopores on the Flongles have poorer viability and
370 generate more read errors demonstrable by lower Flongle Phred quality on average and
371 therefore additional alignment challenges and a study of errors specific in Flongle bioinformatic
372 data compared to the flow cells are needed to understand potential compensation mechanisms
373 for data analysis.

374

375 **Conclusion**

376 We demonstrate the feasibility of using single molecule long-range sequencing assay to detect
377 fusion genes in heme malignancy (AML, CML and APL) patients. Inherent characteristics of
378 fusions make this assay a promising cost effective tool for rapid detection of recurrent fusions
379 that 1) does not require previous knowledge of the target, , 2) with a rapid TAT (8 hours in 80%
380 of samples) when multiplexing different assays and used with the specific data analysis and
381 fusion detection tools described in our manuscript, 3) can precisely map translocation genomic
382 breakpoints that allow for development of personalized markers for disease monitoring, and 4
383) can potentially allow discovery of novel/different fusion partners

384

385 **Contributions:** OST and CY designed the study and performed background research and
386 secured IRB protocol. KYY and LHH designed the informatics pipeline and algorithms. JR, DW
387 and CY identified specimens. LHH and SR implemented the software. SR, OST and CY tested the
388 software. OST, LB, SR, LHH, CY performed research and analyzed the data. KYY, JR, CY funded
389 the project and provided supervision. OST, SR, KYY, and CY drafted the manuscript. All authors
390 reviewed and edited the manuscript.

391

392 **Funding**

393 This study was supported in part by NCCN young investigator award and the Hyundai Hope on
394 Wheels Scholars Award, R01 CA175008-06, UG1 CA233338-02, and Adult Leukemia Research
395 Center Grant # P01 CA018029 as funded by the National Cancer Institute, National Institutes of

Health, Bethesda, MD. LHH, SR and KYY are supported by NIH grant R01GM126019. SR is also supported by the Vicky L. Carwein and William B. Andrews Endowments for Graduate Programs.

Disclosures: LHH and KYY also have equity interest in Biodepot LLC, which receives compensation from NCI SBIR contract numbers 75N91020C00009 and 75N91021C00022.

Acknowledgements: The authors want to thank Dr Phillip E. Starshak, Kaiser Permanente Oakland, for procuring specimens.

References:

1. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*. May 19 2016;127(20):2391-405. doi:10.1182/blood-2016-03-643544

2. Quessada J, Cuccuini W, Saultier P, Loosveld M, Harrison CJ, Lafage-Pochitaloff M. Cytogenetics of Pediatric Acute Myeloid Leukemia: A Review of the Current Knowledge. *Genes (Basel)*. Jun 17 2021;12(6)doi:10.3390/genes12060924

3. Hehlmann R. Innovation in hematology. Perspectives: CML 2016. *Haematologica*. Jun 2016;101(6):657-9. doi:10.3324/haematol.2016.142877

4. Lo-Coco F, Cicconi L. History of acute promyelocytic leukemia: a tale of endless revolution. *Mediterr J Hematol Infect Dis*. 2011;3(1):e2011067. doi:10.4084/MJHID.2011.067

5. Long ZJ, Hu Y, Li XD, et al. ATO/ATRA/anthracycline-chemotherapy sequential consolidation achieves long-term efficacy in primary acute promyelocytic leukemia. *PLoS One*. 2014;9(8):e104610. doi:10.1371/journal.pone.0104610

6. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol*. Feb 7 2020;21(1):30. doi:10.1186/s13059-020-1935-5

7. Jain M, Koren S, Miga KH, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*. Apr 2018;36(4):338-345. doi:10.1038/nbt.4060

8. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA. Accuracy of Next Generation Sequencing Platforms. *Next Gener Seq Appl*. 2014;1doi:10.4172/jngsa.1000106

9. Jancuskova T, Plachy R, Zemankova L, et al. Molecular characterization of the rare translocation t(3;10)(q26;q21) in an acute myeloid leukemia patient. *Mol Cytogenet*. 2014;7:47. doi:10.1186/1755-8166-7-47

10. Jeck WR, Lee J, Robinson H, Le LP, Iafrate AJ, Nardi V. A nanopore sequencing-based assay for rapid detection of gene fusions. *J Mol Diagn*. Sep 28 2018;doi:10.1016/j.jmoldx.2018.08.003

11. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. Improved data analysis for the MinION nanopore sequencer. *Nat Methods*. Apr 2015;12(4):351-6. doi:10.1038/nmeth.3290

12. Gilpatrick T, Lee I, Graham JE, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat Biotechnol*. Apr 2020;38(4):433-438. doi:10.1038/s41587-020-0407-5

13. Reddy S, Hung LH, Sala-Torra O, Radich JP, Yeung CC, Yeung KY. A graphical, interactive and GPU-enabled workflow to process long-read sequencing data. *BMC Genomics*. Aug 23 2021;22(1):626. doi:10.1186/s12864-021-07927-1

14. Yeung C, Qu X, Sala-Torra O, Woolston D, Radich J, Fang M. Mutational profiling in acute lymphoblastic leukemia by RNA sequencing and chromosomal genomic array testing. *Cancer Med*. Aug 2021;10(16):5629-5642. doi:10.1002/cam4.4101

15. Sala-Torra O, Beppu LW, Abukar FA, Radich JP, Yeung CC. TTMV-RARA fusion as a recurrent cause of AML with APL characteristics. *Blood Adv*. Apr 14 2022;doi:10.1182/bloodadvances.2022007256

16. *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. . Humana Press;

17. Stangl C, de Blank S, Renkens I, et al. Partner independent fusion gene detection by multiplexed CRISPR-Cas9 enrichment and long read nanopore sequencing. *Nat Commun*. Jun 5 2020;11(1):2861. doi:10.1038/s41467-020-16641-7

18. Hung LH, Hu J, Meiss T, et al. Building Containerized Workflows Using the BioDepot-Workflow-Builder. *Cell Syst*. Nov 27 2019;9(5):508-514 e3. doi:10.1016/j.cels.2019.08.007

19. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol*. Jun 24 2019;20(1):129. doi:10.1186/s13059-019-1727-y

20. Leger AL, T. pycoQC, interactive quality control for Oxford Nanopore Sequencing. *Journal of Open Source Software*. 2019;4(34):1236.

21. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. Sep 15 2018;34(18):3094-3100. doi:10.1093/bioinformatics/bty191

22. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. Jan 2011;29(1):24-6. doi:10.1038/nbt.1754

23. Liu Q, Hu Y, Stucky A, Fang L, Zhong JF, Wang K. LongGF: computational algorithm and software tool for fast and accurate detection of gene fusions by long-read transcriptome sequencing. *BMC Genomics*. Dec 29 2020;21(Suppl 11):793. doi:10.1186/s12864-020-07207-4

24. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. Aug 15 2009;25(16):2078-9. doi:10.1093/bioinformatics/btp352

25. Picard Toolkit. Broad Institute, GitHub repository.

26. Ross DM, Schafranek L, Hughes TP, Nicola M, Branford S, Score J. Genomic translocation breakpoint sequences are conserved in BCR-ABL1 cell lines despite the presence of amplification. *Cancer Genet Cytogenet*. Mar 2009;189(2):138-9. doi:10.1016/j.cancergencyto.2008.10.010

27. Pagani IS, Dang P, Kommers IO, et al. BCR-ABL1 genomic DNA PCR response kinetics during first-line imatinib treatment of chronic myeloid leukemia. *Haematologica*. Dec 2018;103(12):2026-2032. doi:10.3324/haematol.2018.189787

28. Ross DM, Branford S, Seymour JF, et al. Patients with chronic myeloid leukemia who maintain a complete molecular response after stopping imatinib treatment have evidence of persistent leukemia by DNA PCR. *Leukemia*. Oct 2010;24(10):1719-24. doi:10.1038/leu.2010.185

29. Zhang JG, Lin F, Chase A, Goldman JM, Cross NC. Comparison of genomic DNA and cDNA for detection of residual disease after treatment of chronic myeloid leukemia with allogeneic bone marrow transplantation. *Blood*. Mar 15 1996;87(6):2588-93.

30. Mattarucchi E, Spinelli O, Rambaldi A, et al. Molecular monitoring of residual disease in chronic myeloid leukemia by genomic DNA compared with conventional mRNA analysis. *J Mol Diagn*. Sep 2009;11(5):482-7. doi:10.2353/jmoldx.2009.080150

31. Bartley PA, Latham S, Budgen B, et al. A DNA real-time quantitative PCR method suitable for routine monitoring of low levels of minimal residual disease in chronic myeloid leukemia. *J Mol Diagn*. Mar 2015;17(2):185-92. doi:10.1016/j.jmoldx.2014.10.002

32. Bartley PA, Ross DM, Latham S, et al. Sensitive detection and quantification of minimal residual disease in chronic myeloid leukaemia using nested quantitative PCR for BCR-ABL DNA. *Int J Lab Hematol*. Dec 2010;32(6 Pt 1):e222-8. doi:10.1111/j.1751-553X.2010.01236.x

33. Cumbo C, Anelli L, Specchia G, Albano F. Monitoring of Minimal Residual Disease (MRD) in Chronic Myeloid Leukemia: Recent Advances. *Cancer Manag Res*. 2020;12:3175-3189. doi:10.2147/CMAR.S232752

34. Burmeister T, Marschalek R, Schneider B, et al. Monitoring minimal residual disease by quantification of genomic chromosomal breakpoint sequences in acute leukemias with MLL aberrations. *Leukemia*. Mar 2006;20(3):451-7. doi:10.1038/sj.leu.2404082

35. Branford S. Molecular monitoring in chronic myeloid leukemia-how low can you go? *Hematology Am Soc Hematol Educ Program*. Dec 2 2016;2016(1):156-163. doi:10.1182/asheducation-2016.1.156

Figure and Table legends:

Figure 1. Chemistry and bioinformatics workflow of rapid single molecule long-read sequencing.

Genomic DNA may contain a target fusion gene. The CRISPR-Cas9 system binds via specific guideRNA (gRNA) designed to enrich for DNA containing regions of interests. The library preparation does not undergo any amplification, and simply requires dA tailing and adapter ligation and a clean-up prior to being loaded onto a sequencing flow cell. On the flow cell libraries are sequenced by nanopores. Data from the sequencing devices are streamed onto biodepot workflow builder and are analyzed starting from FAST5 files, through an initial quality control, then base calling and alignment. After alignment different fusion finder tools were tested including LongGF and BFF before visualizing on IgV for confirmation and interpretation.

Figure 2. Screenshot of our bioinformatics workflow and output. **Panel A:** Bwb workflow showing the workflow including our custom Biodepot Fusion Finder (BFF) and LongGF widgets. **Panel B:** shows enrichment statistics as fusion enrichment and on target enrichment. **Panel C:** alignment can be viewed on IGV based on a BAM file generated from minimap2. This case shows a t(9;22) *BCR-ABL1* fusion in a primary specimen. Reads spanning the breakpoint are colored with the same color on alignment to both genes.

Figure 3. Time required to obtain 3 fusion supporting reads.

524 Top panel shows the time to 3 fusion supporting reads seen in the cell lines samples. Bottom
525 panel shows time to 3 fusion supporting reads seen in patient samples. Details of specific times
526 are listed in supplementary table 1.

527

528 **Table 1.** Characteristics of the Primary Specimens analyzed

529

530 **Table 2.** Characteristics of the Runs

531

532 **Supplementary Table 1.** Times and total reads before 3 fusion reads could be confirmed in cell
533 lines and patient samples.

Figure 1.

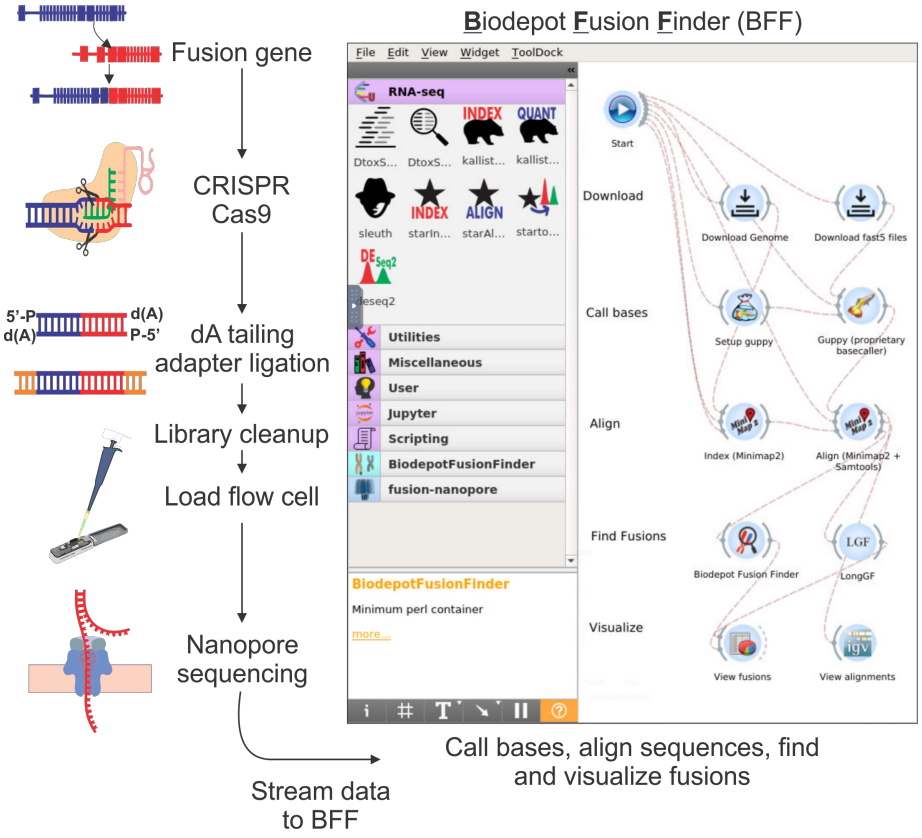
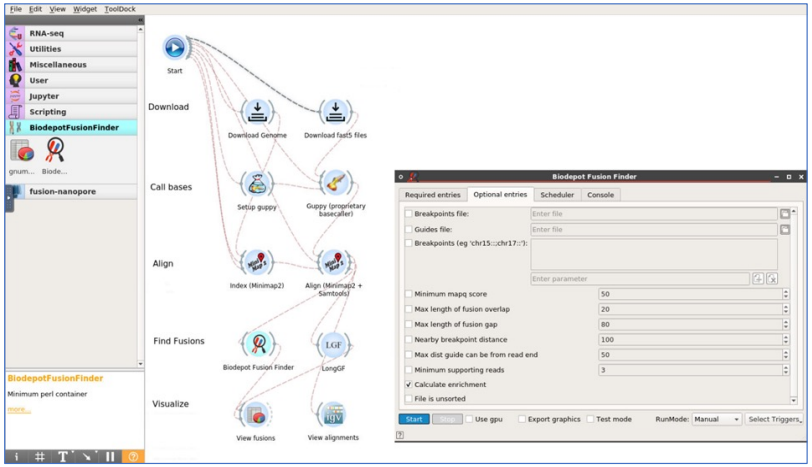

Fusion gene
CRISPR Cas9
5'-P d(A)
d(A) P-5'
dA tailing adapter ligation
Library cleanup
Load flow cell
Nanopore sequencing
Stream data to BFF
Biodepot Fusion Finder (BFF)
File Edit View Widget ToolDock
RNA-seq
Utilities
Miscellaneous
User
Jupyter
Scripting
BiodepotFusionFinder
fusion-nanopore
BiodepotFusionFinder
Minimum perl container
more...
Start
Download
Download Genome
Download fast5 files
Call bases
Setup guppy
Guppy (proprietary basecaller)
Align
Index (Minimap2)
Align (Minimap2 + Samtools)
Find Fusions
Biodepot Fusion Finder
LongGF
Visualize
View fusions
View alignments
Call bases, align sequences, find and visualize fusions

Figure 2.

A: Bwb workflow with LongGF &BFF



B: Enrichment statistics computation output

	A	B	C	D	E	F	G	H	I	J
1	Breakpoint	Gap/Overlap	Count	Nearby-count	ReadThru	NearbyReadThru	Breakpoint enrichment	Breakpoint nearby enrichment	Fraction on target	Fraction on target nearby
2	chr9:133607156;chr22:23632742	2	1	29	30	31	30.19597	875.68318	0.03333	0.93548
3	chr9:133607152;chr22:23632739	4	1	29	29	31	30.19597	875.68318	0.03448	0.93548
4	chr9:133607147;chr22:23632742	5	21	29	28	31	634.11541	875.68318	0.75000	0.93548
5	chr9:133607164;chr22:23632742	5	1	29	31	31	30.19597	875.68318	0.03226	0.93548
6	chr9:133607168;chr22:23632748	5	1	29	30	31	30.19597	875.68318	0.03333	0.93548
7	chr9:133607158;chr22:23632736	8	1	29	31	31	30.19597	875.68318	0.03226	0.93548
8	chr9:133607147;chr22:23632715	16	1	29	31	31	30.19597	875.68318	0.03226	0.93548
9	chr9:133607177;chr22:23632748	17	1	29	30	31	30.19597	875.68318	0.03333	0.93548
10	chr9:133607153;chr22:23632728	20	1	29	31	31	30.19597	875.68318	0.03226	0.93548
11	Coverage	MaxBpFromEnd	GuideReads	Enrichment						
12	0.0331	50	212	6401.546						

C: IGV image confirming fusion

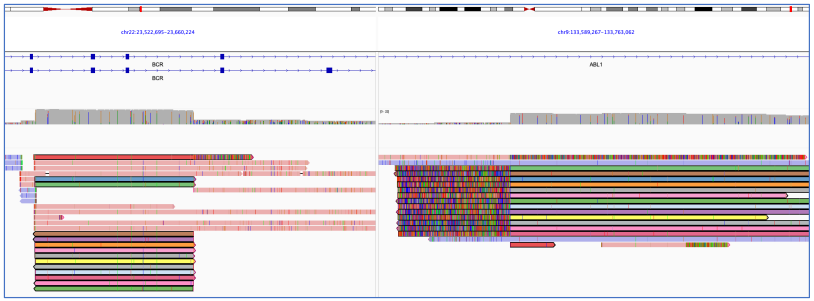


Figure 3.

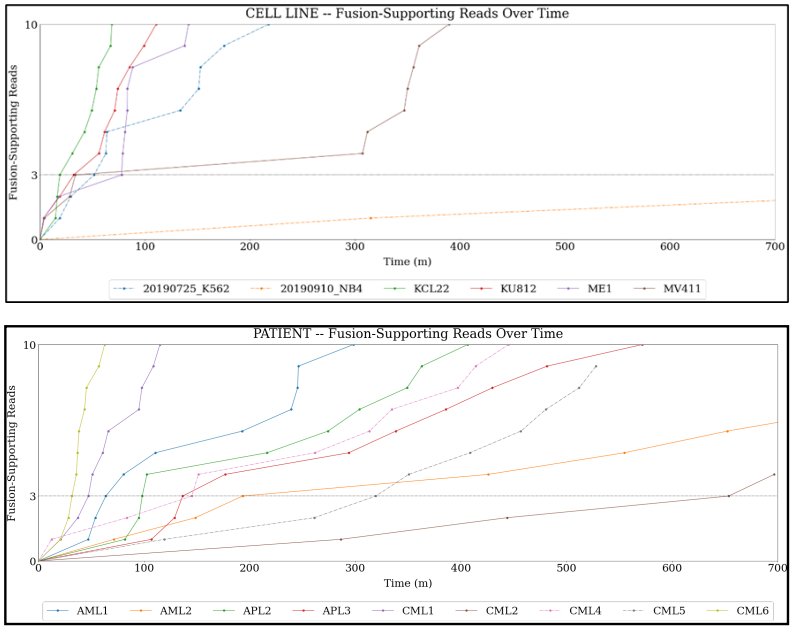

CELL LINE -- Fusion-Supporting Reads Over Time
Fusion-Supporting Reads
Time (m)
20190725_K562
20190910_NB4
KCL22
KU812
ME1
MV411
PATIENT -- Fusion-Supporting Reads Over Time
Fusion-Supporting Reads
Time (m)
AML1
AML2
APL2
APL3
CML1
CML2
CML4
CML5
CML6

Table 1- Characteristics of the Primary Specimens analyzed

Patient Code	Age	Gender	Diagnosis	Specimen Source	Disease Burden	Cytogenetics/Molecular
CML1	61 - 65	Male	CML-AP Rlps	BM	2% blasts	46,XY,t(5;12)(q33;q15),t(9;22)(q34;q11.2)[2]/47,sl,+8[18]
CML2	41 - 45	Female	CML-BC	PB	NA	Outside cytogenetics confirmed t(9;22) but only in 3/21 karyotypes. Full karyotype not available.
CML3	56 - 60	Male	CML-AP	PB	NA	Patient with long standing p190 CML. At this time-point sample presents MDS/MPN with low level BCR-ABL1 by PCR.
CML4	21 - 25	Male	CML-CP	PB	3% blasts	46,XX,t(9;22)(q34;q11.2)[20]
CML5	61 - 65	Male	CML-AP Rlps	PB	3% blasts	46,XY,t(5;12)(q33;q15),t(9;22)(q34;q11.2)[2]/47,sl,+8[18]
CML6	65 - 70	Female	CML-BC	PB	16% blasts	46,XX,t(9;22)(q34:q11.2)[4]; Confirmed by molecular studies.
AML1	36 - 40	Male	AML	PB	90% blasts	46,XY,t(9;11)(q21;q23)[20]
AML2	36 - 40	Female	AML	PB	20% blasts	47,XX+8,inv(16)(p13.1q22)[17]/46,XX[3]
APL1	81 - 85	Male	APL	PB	<1% blasts	Patient with collision Multiple Myeloma and APL in BM with 30% blasts, not available for analysis. No circulating APL in PB.
APL2	31 - 35	Female	APL	BM	80% blasts	46,XX,der(2)t(2;17)(q33;q21)t(15;17)(q22;q21),der(15)t(15;17),der(17)t(2;17)[3]/47,sl+8[3]/48,sdl,+8[12]/46,XX[2]
APL3	21 - 25	Male	APL	PB	75% blasts	NA
APL4	36 - 40	Male	APL	BM	83% blasts	AML with isochromosome 17q
APL5	36 - 40	Male	APL	PB	79% blasts	AML with isochromosome 17q
APL6	41 - 45	Male	APL	BM	22% blasts in a <5% marrow	46, XY,t(1;7)(q21;q21), t(4;10)(q21;q25), add(8)(p23), del(9)(p22), t(11;17)(q23;q25)[4]/46, XY[16} from 3 months prior

CML= Chronic Myeloid Leukemia; AML= Acute Myeloid Leukemia; APL= Acute Promyelocytic Leukemia, AP= Accelerated Phase, BC= Blast Crisis, PB= Peripheral Blood, BM= Bone Marrow; NA= not available

Table 2. 2. Characteristics of the Sequencing Runs

Sample	DIN	Device	Gb of Data	Mean Coverage	All Reads	N50 All Reads	Median Read Length	Median PHRED	% Reads Aligned	Fusion Reads - LongGF	Fusion Reads - BFF	On-target Enrichment	Fusion Enrichment spanning breakpoint	Disease Burden
K562	9.8	Flongle	0.12	0.04	35943	5980	2200	9.516	87.57	0	29	5830.00	797.50	NA
KU812	8.7	Flowcell	0.36	0.11	86579	9690	1820	12.759	96.29	0	51	2108.33	467.50	NA
KCL22	9.7	Flowcell	0.69	0.21	50439	32200	6290	12.259	91.86	5	171	2118.70	817.83	NA
NB4	9.1	Flongle	0.07	0.02	23122	5890	1450	9.133	88.86	3	3	848.57	141.43	NA
MV4;11	8.6	Flowcell	0.28	0.08	40000	27100	1890	11.98	89.34	57	71	2698.93	836.79	NA
ME1	8.2	Flowcell	0.91	0.28	433846	16500	872	11.17	84.69	45	37	1305.49	134.18	NA
CML1	9.8	Flowcell	0.92	0.28	32723	24600	9600	12.309	91.74	0	142	1563.91	509.35	CML-AP Relapse
CML2	9.6	Flowcell	0.04	0.01	4026	22400	6520	11.148	72.08	0	4	2227.50	330.00	CML BC
CML3	9.8	Flongle	0.09	0.03	9857	19900	5860	9.45	85.54	0	0*	3006.67	NA	CML-AP
CML4	9.7	Flongle	0.93	0.28	109648	15200	1800	9.314	85.97	0	14	557.10	49.68	CML-CP
CML5	ND	Flongle	0.04	0.03	3907	22400	8090	8.95	91.06	0	8	1283.33	293.33	CML-AP Relapse
CML6	9.2	Flowcell	5	1.52	381552	30800	7090	11.94	90.73	0	in	1539.12	142.56	CML BC
AML1	9.3	Flowcell	5.47	1.66	639884	15100	6340	9.89	86.76	28	38	535.12	22.93	90% blasts
AML2	9.2	Flowcell	1.59	0.48	161010	19400	6540	12	89.95	0	0	649.62	NA	20% blasts
APL1	9.6	Flowcell	0.17	0.05	13486	24700	7450	12.942	92.49	0	0**	1242.35	NA	0% blasts in PB, APL confirmed by FISH on BM sample
APL2	7.76	Flowcell	1.1	0.33	40919	24200	8770	12.12	92.69	8	10	618.00	30.00	80% blasts
APL3	9.7	Flowcell	0.31	0.09	13625	22500	6910	12.422	93.51	8	10	872.90	106.45	75% blasts
APL4	9.6	Flowcell	1.63	0.49	176665	25200	3740	12.38	93.27	0	0***	1785.64	NA	83% blasts
APL5	9.3	Flowcell	0.22	0.07	31863	24200	2450	11.82	91.49	0	0***	2520.00	NA	79% blasts
APL6	7.5	Flowcell	1.06	0.32	485374	4650	1230	11.7	84.27	0	2****	650.66	6.23	22% blasts on flow in a <5% marrow

* Pt had very low BCR-ABL1 levels that was only detected by qualitative PCR and not quantitative PCR (meaning below 0.01%)

** Pt with confirmed BM disease but PB sample that is used for this assay had <1% blast count.

*** Pt had an unusual fusion between a viral gene and RARA(reference: Sala-Torra, Blood Adv. Apr 14 2022), however a human reference genome was used on BFF to generate this data and thus this fusion could not be detected.

**** Pt was suspected to have APL, but subsequent cytogenetic results showed complex karyotype with an unexpected KMT2A fusion and this was a challenged sample with an N50 below 5000bp.

CML= Chronic Myeloid Leukemia; AML= Acute Myeloid Leukemia; APL= Acute Promyelocytic Leukemia, AP= Accelerated Phase, BC= Blast Crisis, PB= Peripheral Blood, BM= Bone Marrow; NA= not available

FISH = fluorencence in-situ hybridization