SmMIP-tools: a computational toolset for processing and analysis of single-molecule molecular inversion probes derived data

Jessie J. F. Medeiros[1,2,3], Jose-Mario Capo-Chichi[4], Liran I. Shlush[5], John E. Dick[1,3], Andrea Arruda[1], Mark D. Minden[1,6], Sagi Abelson[2,3*]

[1] Princess Margaret Cancer Centre, University Health Network (UHN), Toronto, Ontario, Canada.

[2] Ontario Institute for Cancer Research, Toronto, Ontario, Canada.

[3] Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada.

[4] Genome Diagnostics, Department of Clinical Laboratory Genetics, University Health Network, Toronto, Canada.

[5] Department of Immunology, Weizmann Institute of Science, Rehovot, Israel

[6] Department of Hematology and Medical Oncology, University Health Network, Toronto, ON, Canada.

*Correspondence:sabelson@oicr.on.ca

Keywords: Molecular inversion probes, Sequencing, Mutations calling, Molecular barcodes, Error correction

Abstract

Single-molecule molecular inversion probes (smMIPs) provides a modular and cost-effective platform for high-multiplex targeted next-generation sequencing (NGS). Nevertheless, translating the raw smMIP-derived sequencing data into accurate and meaningful information currently requires proficient computational skills and a large amount of computational work, prohibiting wide-scale adoption of smMIP-based technologies. To enable easy, efficient, and accurate interrogation of smMIP-derived data, we developed SmMIP-tools, a computational toolset that combines the critical analytic steps for smMIP data interpretation into a single computational pipeline. Here, we describe in detail two of the software's major components. The first is a read processing tool that performs quality control steps, generates read-smMIP linkages and retrieves molecular tags. The second is an error-aware variant caller capable of detecting single nucleotide variants (SNVs) and short insertions and deletions (indels). Using a cell-line DNA dilution series and a cohort of blood cancer patients, we benchmarked SmMIP-tools and evaluated its performance against clinical sequencing reports. We anticipate that SmMIP-tools will increase accessibility to smMIP-technology, enabling cost-effective genetic research to push personalized medicine forward.

Introduction

Padlock probes, initially described in 1994 (Nilsson et al. 1994), are linear DNA oligos that contain two target arm segments, physically connected by a linker sequence, designed to complement immediately adjacent target DNA. Upon hybridization of both complementary arms to their target, padlock probes can circularize via ligation, capturing the DNA molecule in a “padlock” conformation to enable localized detection of DNA. The concept of padlock probes was later extended via the introduction of molecular inversion probe (MIP) technology for genotyping of single nucleotide polymorphisms (SNP) (Hardenbol et al. 2003, 2005). Here, probes were designed such that target arms contained a single nucleotide gap flanking a SNP of interest. Upon MIP hybridization, gap-filling by DNA polymerase with the correct deoxynucleoside triphosphate allowed the capture of the SNP locus, later amplified by PCR with standard primers and read-out using DNA microarray. Critically, these studies exhibit unprecedented levels of MIP multiplex capabilities by combining thousands of MIPs in a single reaction while maintaining high levels of specificity. This is in contrast to the notoriously poor multiplex capabilities of standard PCR (Mamanova et al. 2010). Thus, specificity and multiplex capability have positioned MIPs as a vital tool for target enrichment in DNA sequencing. Indeed, one study utilizing MIPs designed with larger gaps between target arms for the multiplex capture and amplification of intervening sequence showed that thousands of exons could be interrogated in a single reaction with subsequent shotgun sequencing (Porreca et al. 2007). However, while individual MIPs were highly specific, uniformity in coverage across MIPs was extremely poor, presenting a major technical challenge. High coverage variability has been addressed with optimized protocols, empirical probe rebalancing, and reduced panel size to enhance uniformity across MIPs (Turner et al. 2009). Critically, these subsequent studies enhanced MIP design elements to permit direct resequencing using NGS platforms such that libraries could be multiplexed to an unprecedented level using barcoded PCR primers for library amplification (O’Roak et al. 2012). High customizability of MIP design has enabled initial groups to generate custom target panels for mutation analysis on large human cohorts, exhibiting the usefulness of MIPs to provide significant biological findings in a high-throughput, cost-effective manner (O’Roak et al. 2012). MIPs continued to evolve and were further enhanced to include unique molecular identifiers (UMIs) into the oligo structure (Hiatt et al. 2013), transforming MIPs into smMIPs and allowing the precise identification of single molecules. UMIs permit the collapse of PCR amplified capture products into single-strand consensus sequences (SSCS), enabling the correction of PCR amplification-induced errors. A second

61 UMI sequence adjacent to a target arm further enhanced smMIP SSCS capabilities (Acuna-Hidalgo et
62 al. 2017).
63 Though these advancements in smMIP technology have largely addressed the technical challenges of
64 smMIP-based sequencing, there is much room for informatic approaches to enhance capabilities
65 further. This is exemplified by MIPgen (Boyle et al. 2014), a machine learning algorithm that has
66 dramatically improved smMIP panel design, increasing the chances of successful target capture and
67 coverage uniformity across smMIPs. The simplification of panel design has driven the recent uptake of
68 smMIP-based technology for customized, targeted sequencing and subsequent SNV, indel, and in
69 some cases, copy number variant mutational analysis (Neveling et al. 2017; Pippucci et al. 2019; Khan
70 et al. 2020). However, while smMIP-sequencing has been used to generate massive amounts of data,
71 most groups have relied on in-house custom pipelines and there remains a paucity of publicly available
72 informatic tools specifically designed for the effective and efficient handling of smMIP-derived
73 sequencing data. For this reason, we developed SmMIP-tools. SmMIP-tools functionalities include
74 quality control metrics of aligned reads, filtering reads which are inadequate for subsequent analysis,
75 generating base call summaries for both raw reads and SSCS, and the calling of both SNV and short
76 indels with high accuracy. To mitigate false-positive calls arising due to the high level of NGS associated
77 errors (Ma et al. 2019), SmMIP-tools generates probabilistic models of allele-specific error rates.
78 SmMIP-tools dramatically enhances the fidelity of error-corrected mutation calling by analyzing data
79 from each read derived from sequencing read pairs, overlapping smMIPs and smMIP library technical
80 replicates when available. We benchmarked the use of SmMIP-tools via a dilution assay of established
81 cell lines with known mutations. Critically, we validated its utility in side-by-side comparative analysis to
82 detect SNVs and indels previously identified by clinical reports in a large clinical cohort. SmMIP-tools
83 represents the next step in the evolution of smMIP-based technologies as a crucial informatic advance
84 to promote widespread, efficient use of smMIPs to reveal significant biologic findings in a high-
85 throughput and cost-effective manner.
86

87 ## Results

88 ### Overview

89 Proper smMIP design (Fig. 1A-C, Methods) is fundamental for the overall success of the sequencing
90 experiment. It is critical to maintain high capture uniformity and specificity and to improve accuracy in

91 detecting rare, subclonal mutations by reducing NGS-associated artefacts. For example, one can
92 improve smMIP panel design by accounting for the sequencing platform intended for use with respect
93 to its output read length. Assuring that each of the paired reads thoroughly interrogates its
94 corresponding target locus will maximize the coverage at each base and support the suppression of
95 errors that preferentially occur on one of the two amplified cDNA strands.
96 Following sequencing, quality control and read processing steps are critical to achieving reliable variant
97 discovery (Koboldt 2020). SmMIP-tools takes as an input a read-alignment BAM file and a smMIP
98 design file generated by MIPgen (Boyle et al. 2014) (Fig. 1D). It uses information concerning each probe
99 and its target sequence to apply a set of filters to discard hard-clipped reads, reads with low mapping
100 quality, paired reads with an unexpected insert size or improper alignment orientations (Fig. 1E). To
101 confirm the proper structure of the remaining reads and to identify corrupted UMI sequences,
102 associations between reads and their precise probe-of-origin (termed here, read-smMIP linkages) are
103 generated using two authentication steps. These include the identification of smMIP targets that overlap
104 with the paired reads' targeted genomic region and validation of a single smMIP among those
105 candidates through local alignment of their extension and ligation arm sequences to the paired reads
106 sequences. The final output contains quality control summary files concerning raw and consensus reads
107 (Fig. 1F) and a BAM file with the remaining high-quality reads. At that point, UMI sequences and smMIP-
108 of-origin are included in each read's header.
109 In the following steps, SmMIP-tools uses the processed BAM file to generate base call summaries of
110 raw reads and SSCS which are subsequently those are refined by the software's error-aware variant
111 detection algorithm (Fig. 1G). To call mutations, binomial models are used to conduct allele-specific
112 frequency comparisons between each sample of interest and either a single control or a cohort of control
113 samples. SmMIP-tools is also capable of precise allele-specific error rates estimation without using
114 dedicated controls by comparing a sample of interest to the remainder of the experimental cohort. Prior
115 knowledge concerning the location of common cancer mutations and germline polymorphisms is used
116 to increase the sensitivity of detecting recurrently mutated alleles (Methods). To improve specificity,
117 non-reference alleles are evaluated separately in each of the paired sequencing reads, in reads derived
118 from different but overlapping smMIPs, and in technical replicates when available. The final output is a
119 comprehensive report that includes the detected mutations, key variant annotations, information
120

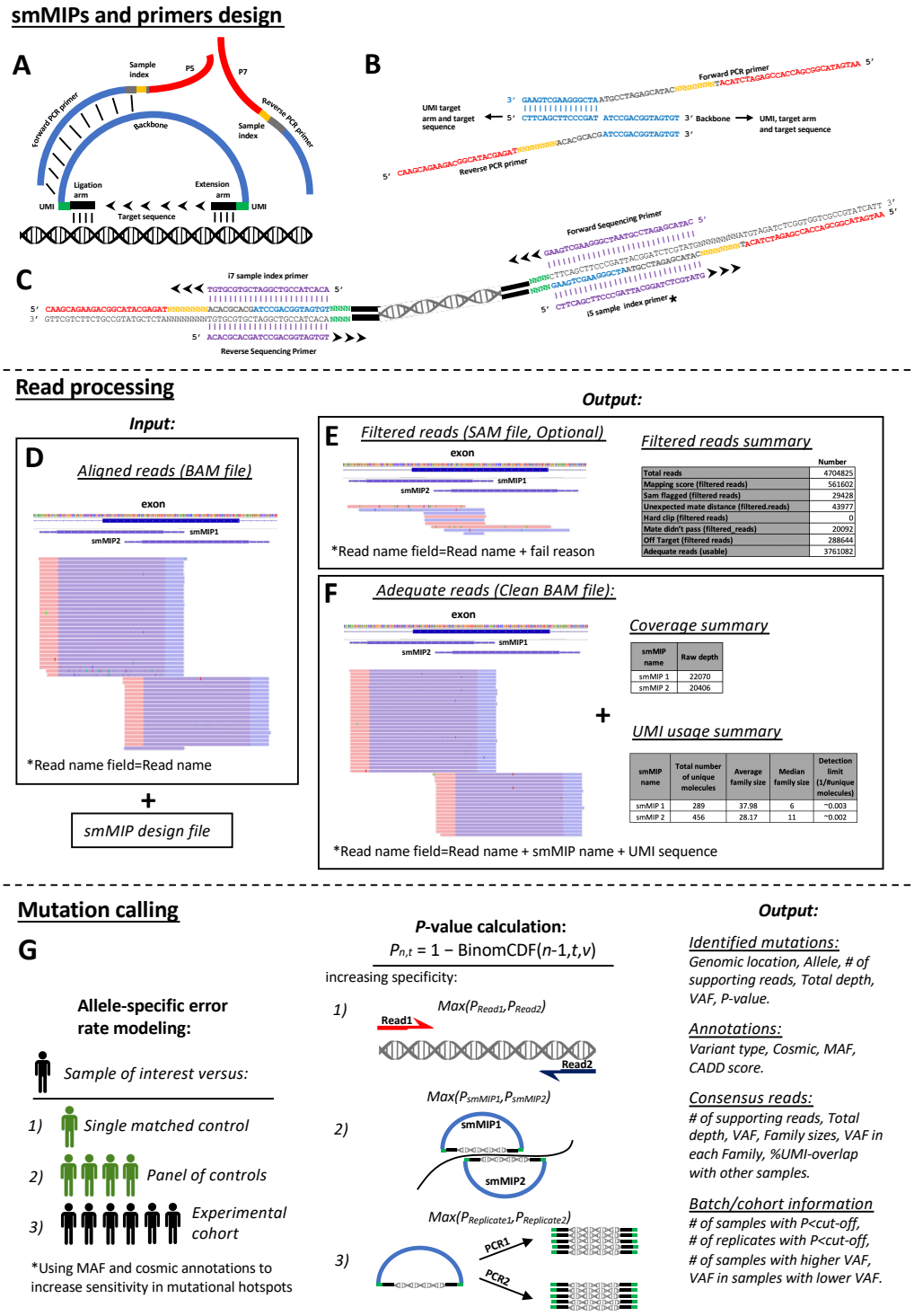


Figure. 1 Overview of smMIP assay design and key steps in data analysis using SmMIP-tools. (A) General layout of a smMIP. (B) Backbone and PCR primers' sequences and primers annealing sites. (C) Library composition and sequencing primers' sequences. *the i5 sample index primer is only required for select sequencing platforms. >>> indicates the direction of extension. (D) SmMIP-tools BAM processing inputs. (E,F) SmMIP-tools BAM processing outputs. (G) Probabilistic models of error rates and SmMIP-tools mutation calling output. Let 'v' be the investigated allele's fraction in the cohort (option 1,2 or 3 in the left panel). Then, the P-value of seeing at least 'n' reads supporting the non-reference allele out of 't' total coverage in the investigated genomic position in the sample of interest is Pn,t = 1 − BinomCDF(n-1,t,v) where BinomCDF denotes the binomial cumulative distribution function.

concerning consensus reads' support, sequencing batch summaries and mutations flags; all of which are valuable for the ranking and prioritization of variants.

A comparison matrix emphasizing differences in strengths and limitations between SmMIP-tools and other popular software is included in Supplemental Table 1.

Feature	smMIP-tools	Mipgen[1]	BWA[2]	FastQC[3]	FastProNGS[4]	SAMtools[5]	GATK HaplotypeCaller[6]	Vardict[7]	iDES[8]	Pindel[9]	UMI-tools[10]	Annovar[11]	Integrative Genomics Viewer[12]
Designed to work on amplicon-based data	Yes, MIPs and smMIPs	Yes, smMIPs	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No
Designed to work on capture-based data	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
Probe design	No	Yes	No	No	No	No	No	No	No	No	No	No	No
Read alignment to the genome	No	No	Yes	No	No	No	No	No	No	For the software's specific purpose only	No	No	No
Data quality assessment :													
1) Sample level	Yes	No	No	Yes	Yes	No	No	No	Yes	No	No	No	No
2) Probe level	Yes	No	No	No	No	No	No	No	No	No	No	No	No
3) Sequencing batch level	Yes	No	No	No	No	No	No	No	No	No	No	No	No
Read preprocessing and filtering :													
1) Chimeric reads	Yes	No	No	No	No	No	No	No	No	No	No	No	No
2) Self-annealing probes	Yes	No	No	No	Possible with the correct user	No	No	No	No	No	No	No	No
3) Other, not smMIP specific (e.g. quality AND alignment flag filters)	Yes	No	No	No	Yes	Yes	Yes	Yes	No	No	No	No	No
4) Clean SAM/BAM output*	Yes	No	No	No	Yes	Yes	No	No	No	No	Yes		
Base call summaries (e.g. pileup) output*	Yes	No	No	No	No	No	No	No	Yes	No	No	No	No
Variant calling :													
1) Germline mutations	Yes	No	No	No	No	No	Yes	Can identify germline background mutations based on presence in public databases	No	Yes	No	No	No
2) Somatic SNVs	Yes	No	No	No	No	No	No	Yes	Yes	No	No	No	No
3) Somatic indels :	Yes	No	No	No	No	No	No	Yes	No	Yes	No	No	No
4) Structural variants	No	No	No	No	No	No	No	Yes	No	Yes	No	No	No
Probabilistic approach to analyze error profiles for rare event detection	Yes. Using a single control sample, a cohort of controls or the studied cohort itself	No	No	No	No	No	No	Yes, tumor-normal calling only	Yes. Designed to work with a cohort of controls	Yes, can compare tumor to normal	No	No	No
Use of prior knowledge concerning recurrent cancer mutations to improve their detection	Yes	No	No	No	No	No	No	No	No	No	No	No	No
Flagging system for variant ranking and prioritization	Yes	No	No	No	No	No	Yes	Yes	No	Minimal **	No	No	No
Unique molecular identifiers (UMI) :													
1) Identify reads with compromised UMI sequences	Yes	No	No	No	No	No	No	No	No	No	Yes	No	No
2) Construction of consensus sequences	Yes	No	No	No	No	No	No	By the same genomic location. Not by UMIs	No	No	Yes	No	No
3) Identify errors due to sample index misassignment	Yes	No	No	No	No	No	No	No	No	No	No	No	No
Variant annotation	Comprehensive annotation for SNVs. Genomic location annotation only for indels	No	No	No	No	No	Can integrate different annotations if were generated externally	Variant type (e.g. SNVs, Insertion)	No	Variant type (e.g. SNVs, Insertion)	Variant type (e.g.inversion, or tandem duplication)	Yes	No
Visualization for review	No	No	No	No	No	No	No	No	No	Yes	No	No	Yes

* To be used with other tools for subsequent analysis if those better fit user's requirements

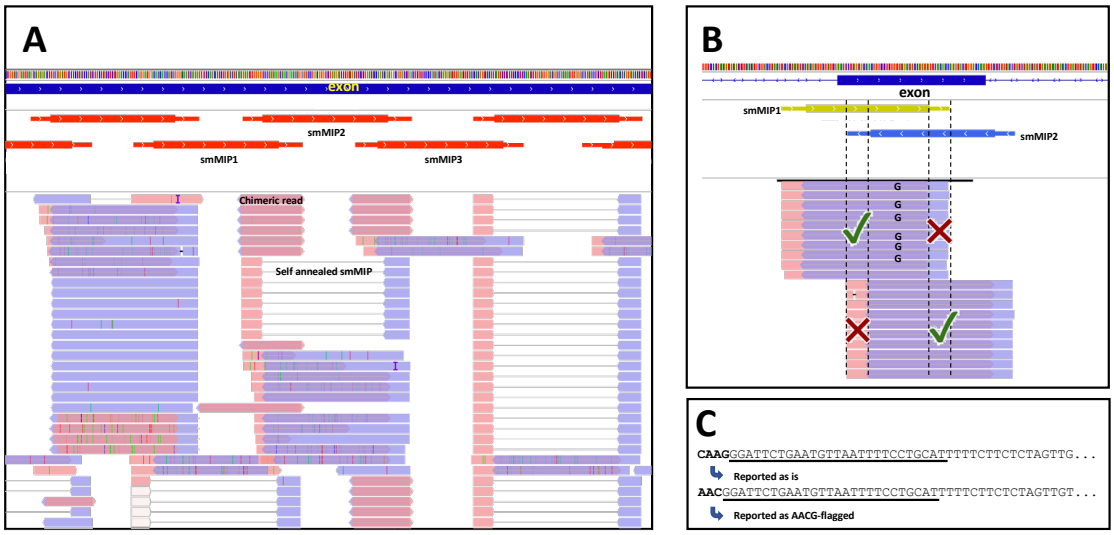
** Include information concerning the appearance of variants in other samples if multiple samples were tested. This information can potentially be used to identify recurrent errors

Software references

1 Boyle, E. A., O'Roak, B. J., Martin, B. K., Kumar, A. & Shendure, J. MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics* **30**, 2670–2672 (2014).
2 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* (2010). doi:10.1093/bioinformatics/btp698
3 Babraham Bioinformatics. FastQC:A quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. .
4 Liu, X. *et al.* FastProNGS: fast preprocessing of next-generation sequencing reads. *BMC Bioinformatics* **20**, 345 (2019).
5 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp352
6 Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2018). doi:10.1101/201178
7 Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).
8 Newman, A. M. *et al.* Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat. Biotechnol.* (2016). doi:10.1038/nbt.3520
9 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
10 Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
11 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).
12 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

Evaluating SmMIP-tools ability to generate authentic read-smMIP linkages

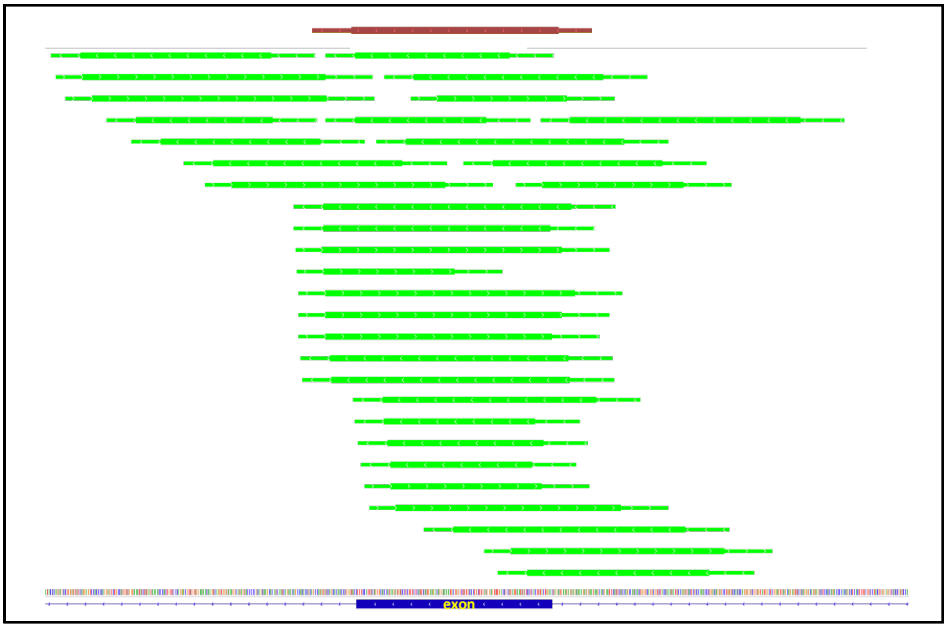
Identifying the correct smMIP arm sequences and validating their expected position in the sequenced read pairs is essential to pinpoint self-annealed probes that lack target sequences and to eliminate chimeric inserts generated by partially overlapping probes (Supplemental Fig. 1A). Furthermore, read-smMIP linkages can help to remove mutations that are called outside the target region of individual smMIPs (Supplemental Fig. 1B), to eliminate errors based on ambiguous calls in regions with overlapping smMIPs and to validate the UMIs' sequence integrity in the expected insert layout (Supplemental Fig. S1C).



Supplemental Figure 1. Addressing issues related to the smMIP technology by linking reads to their probe-of-origin with read-smMIP linkages. (A) Read-smMIP linkages can support the identification and removal of typical undesired sequencing library products. The example shows a representative chimeric read generated by one arm that belongs to smMIP1 and another to smMIP2, as well as a read without a target sequence generated from smMIP2 both of which are are emphasised by the added text. (B) Read-smMIP linkages support the identification of the regions targeted by a given smMIP, thus preventing mutation calling from their arm sequences. (C) Validating the correct position of the smMIP arm's sequences in the expected reads layout helps to identify UMI errors and preserve high quality data for subsequent analysis.

We evaluated the ability of SmMIP-tools to link the correct probe-of-origin to each sequencing read by designing an experiment consisting of real smMIPs that were used for sequencing and simulated smMIPs that were not.

Overall, 22,299 smMIPs were simulated by Mipgen (Boyle et al. 2014), that is, 56-80 simulated smMIPs that overlaps each real smMIP. The simulated probes were designed to include smMIPs with arms and target loci of variable length that can either partially overlap with those of the real smMIPs, are fully included within a real smMIP's genomic insert, or extend beyond the real smMIPs' 3' and/or 5' direction (Supplemental Fig. 2).



Supplemental Figure 2. Simulating smMIPs to assess the performance of read-smMIP linkages generation. Simulated smMIPs (green, n=22,299) were designed in the proximity of the real smMIPs that were used for sequencing (brown, n=284). A simulated smMIP can either overlap with a real smMIP, be smaller and included within the genomic loci belonging to a real smMIP or be larger and flank the entire genomic loci belong to a real smMIP. The closer the start and end genomic positions of simulated smMIPs are to those of real smMIPs the harder it is to accurately determine the reads' origin. Broad and narrow lines represent smMIP's targeted region and arms respectively.

We further controlled for the distance between real and simulated smMIPs' ends to separate the simulated smMIPs into six sets (R0-R5). Those sets ranged from R0, which include simulated smMIPs with an identical start or end genomic position as those of real smMIPs (i.e., all the probes are included). Therefore, R0 represents the set of which the task of separating between real and simulated smMIPs is most challenging. On the other extreme is set R5, which in addition to real smMIPs, only includes simulated smMIPs with ends that are more than five genomic bases away from those of the real smMIP. Therefore, with R5 the tasks of accurately assigning the correct probe-of-origin to each read and differentiate between real and simulating smMIPs is expected to be the least challenging.

A total of 25,353,671 read pairs from 16 sequenced cord blood samples were subjected to read-smMIP linkage performance analysis. On average, 6.1% of the reads could not be linked to any smMIP (real or simulated) and were considered "off-target" due to insufficient overlap with the targeted loci (Fig. 2A). From the other 93.9% "on-target" reads, SmMIPs-tools correctly assigned the authentic probe-of-origin to 89.72% of the reads in R0, the most restrictive simulated smMIP set (Fig. 2B). In the most restrictive set, 8.79% of the reads could equally be associated with more than one smMIP, and 0.04% were falsely

assigned to a simulated smMIP (Fig. 2C). Performance was significantly improved when simulated smMIPs that start or end at an identical genomic base as real smMIPs were omitted (R1).

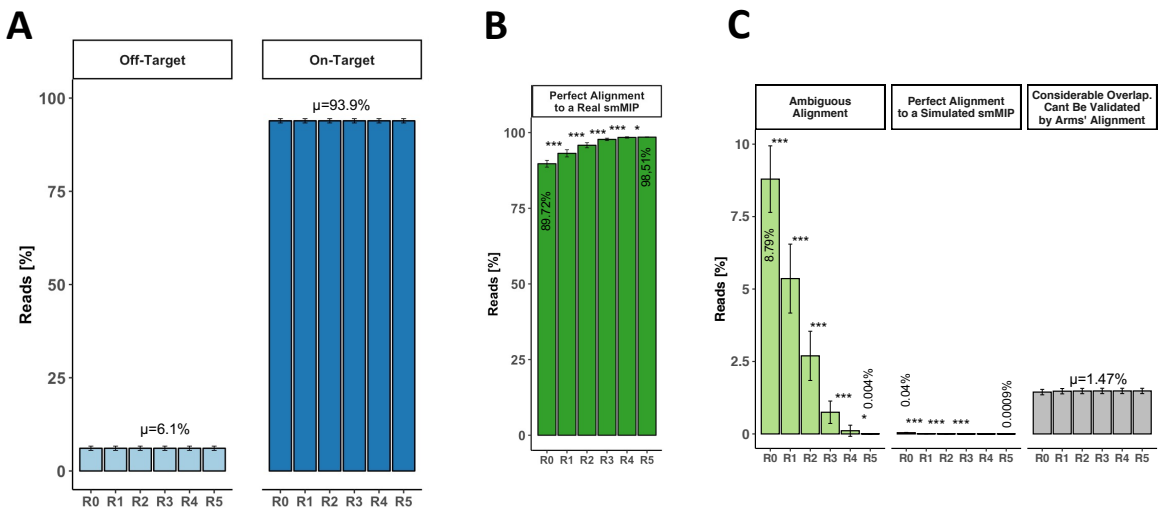
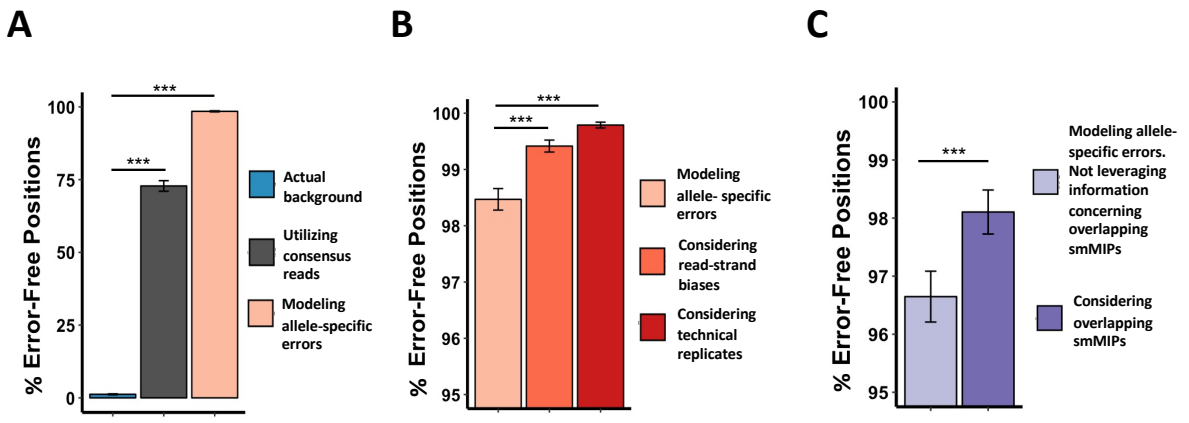


Figure. 2 SmMIP-tools delivers accurate read-smMIP linkages allowing proper downstream analysis of data from complex smMIP panels. **(A)** Average percentage of off-target and on-target reads. **(B)** Average percentage of perfect alignment to a real smMIP. **(C)** Average percentage of imperfect alignments including ambiguous alignment, perfect alignment to a simulated smMIP and alignments that could not be validated by target arms. Error bars representing standard deviations for reads derived from 16 sequenced samples. Paired Sample T-Test: * P<0.05, *** P<0.001. Perfect alignment is defined by first searching for smMIPs that their target overlaps with the read's genome alignment followed by validation of a single smMIP by local alignment of the candidates smMIPs' arms to the paired reads. R0 includes real smMIPs and simulated smMIPs that start or end at the same genomic location as real smMIPs. R3 for example, includes real smMIPs and simulated smMIPs that start or end more than 3 bases apart from real smMIPs start and end sites. R5 for example corresponds to >5 base difference.

In the most permissive simulated smMIP subset (R5), 98.51% of the on-target reads were correctly assigned to a single real smMIP, and a negligible percentage of the reads were falsely assigned to a simulated smMIP (0.0009%). We noticed that sequencing and library amplification errors confound ambiguous or inaccurate assignment of reads to their probe-of-origin. Specifically, due to non-reference bases, on average, 1.47% of the on-target reads showed considerable overlap with the target locus yet failed validation by local alignment of the probes' arms to the reads sequence. SmMIP-tools can still salvage such reads. Nevertheless, since their UMI sequence integrity cannot be validated, such reads are flagged and not included in downstream analyses that consider SSCS (Supplemental Fig. 1C). Our results indicate that SmMIP-tools can generate read-smMIP linkages to accurately filter problematic reads and prepare smMIP-derived data from highly complex target panels, including those containing highly overlapping smMIPs for more efficient downstream analyses.

293 **Benchmarking SmMIP-tools mutations calling algorithm using cell line DNA mixtures**

294 SmMIP-tools incorporates multiple techniques to suppress false positive calls (Fig. 1G). To benchmark
295 their use, we first constructed high confident lists of true and false positive mutations by sequencing 8
296 leukemia cell lines (Supplemental Table 2, Methods). DNA from each cell line was then mixed to
297 generate 6 separate pools containing varying concentrations of each cell line's genomic material. Each
298 mix was sequenced twice to enable the use of information derived from technical replicates to enhance
299 error suppression using SmMIP-tools. In each of the 12 sequenced libraries, we counted the number
300 of error-free positions, defined as positions in the interrogated genomic space, represented exclusively
301 with reference alleles before and after applying error suppression techniques. Both the consensus reads
302 assembly and error modelling techniques (Methods) delivered significant levels of error suppression as
303 indicated by the sole presence of reference alleles in 72.83% and 98.47% of the investigated genomic
304 positions respectively; as compared with an average of 1.22% before error correction (Fig. 3A). Error
305 suppression using the binomial error rate modelling approach was further augmented when information
306 derived from separate strands or from separate technical replicates was incorporated (Fig. 3B).
307 Improved error suppression was also observed when alleles that are covered by different overlapping
308 smMIPs were evaluated against error models derived from each smMIP independently (Fig. 3C).
309

310 

311
312
313
314
315
316
317
318 **Figure. 3.** SmMIP-tools uses multiple approaches to supress errors. (A) The percentage of error-free
319 positions across the targeted genomic space before and following error suppression by SSCS assembly
320 or by probabilistic error-rate modeling. The average error background in the cell line DNA mixes is
321 showed for comparison. (B) Percentage of error-free positions following probabilistic error-rate
322 modeling and consideration of information derived either from paired reads or technical replicates. (C)
323 Percentage of error-free positions following probabilistic error-rate modeling alone versus the additional
324 consideration of overlapping smMIPs. Here, only positions covered by overlapping smMIPs were
325 considered to derive the percentage of error-free positions.
326

Next, we evaluated SmMIP-tools' performance in differentiating real mutations and errors by using all of the error suppression techniques mentioned above. We required at least 1 SSCS in each strand, and for error rate modelling, we accounted for data derived from separate strands, overlapping smMIPs and technical replicates. SmMIP-tools accurately identified the real mutations among the high background of NGS-associated errors, as evident by a near-perfect trade-off between sensitivity and specificity (Fig. 4A). Sensitivity and precision remained high down to a VAF of 0.005, only decreasing to a lower limit of 9.1% sensitivity and 50% precision for mutations detected in the 0.001<VAF<0.005 range (Fig. 4B).

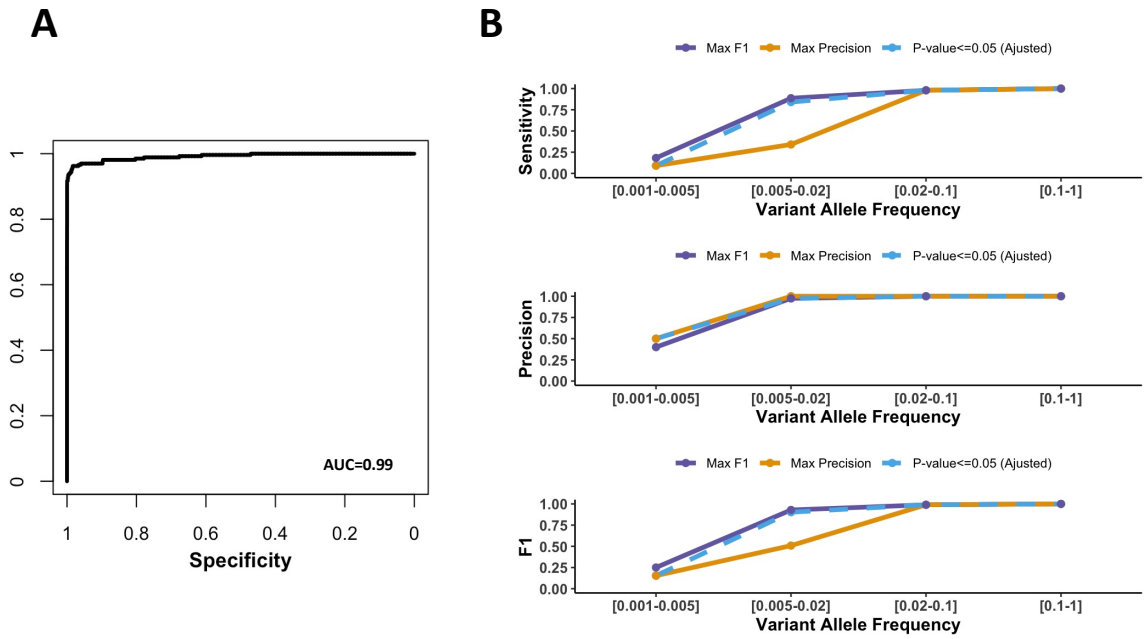


Figure. 4. SmMIP-tools accurately differentiates real mutations from NGS-associated errors. (A) Receiver operating characteristic (ROC) curve indicating high performance of SmMIP-tools to detect real mutations in DNA cell line mixes among the high background of NGS-associated errors. Area under the curve (AUC). (B) Accuracy (precision, sensitivity and F1 score) is shown across different VAF ranges. Colored lines represent the results obtained when the core algorithm was set to achieve maximum F1, maximum precision or when the default P-value cut-off (0.05) was used.

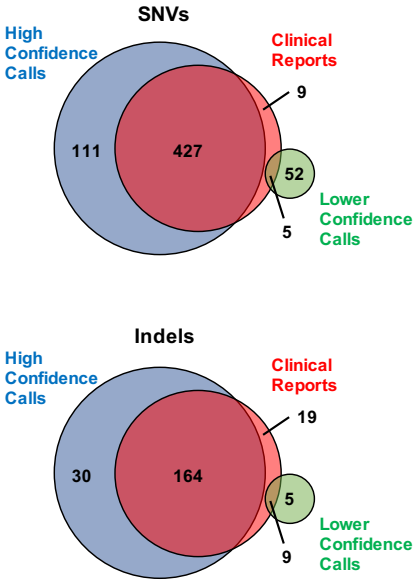
Importantly, SmMIP-tools performance using a P-value threshold of 0.05 showed to be remarkably similar to its performance at maximum F1, suggesting that a P-value of 0.05 is a good default cut-off to use when a sequencing cohort does not include artificially diluted samples with known mutations for P-value optimization. Errors are key confounding factors for sensitive detection of low-frequency variants by deep sequencing (Ma et al. 2019). This analysis illustrates SmMIP-tools error-aware variant calling algorithm's capabilities in accurately detecting mutations with varying VAF ranges among the high number of errors acquired during NGS.

361 **A comparative analysis between SmMIP-tools output and clinical reports obtained from**
362 **leukemia patients**

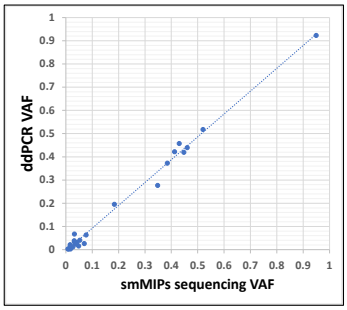
363 As supervised DNA dilutions may be limited in their ability to fully represent clinically relevant mutational
364 processes, we also carried a comparative analysis between SmMIP-tools mutations calling algorithm
365 output and clinical genetic reports generated for patients diagnosed with myeloid malignancies
366 (Methods). Re-sequencing of 168 samples from 162 patients using smMIPs was conducted in technical
367 duplicates. Clinical reports were available for 135 of the patients (Supplemental Table 3). For
368 appropriate comparison with SmMIP-tools, we accounted only for genomic loci covered by both
369 sequencing panels (Supplemental Table 4). Two mutation categories termed "High Confidence" and
370 "Lower confidence" were evaluated based on SmMIP-tools' generated flags (Methods).

371 Overall, 95.6% of the somatic SNVs and indels detected using the clinical pipeline were also detected
372 by SmMIP-tools (Supplemental Table 3). Of these, 97.7% are in the "High Confidence" category.
373 Technical issues were identified to be the primary reason for the missing 4.4% of the variants
374 (Supplemental Table 3). On the other hand, SmMIP-tools detected an additional 111 high confidence
375 SNVs (25.2% increase) and 30 indels (32.6% increase) that were not mentioned in the clinical reports
376 of the patients (Fig. 5A, Supplemental Table 5).

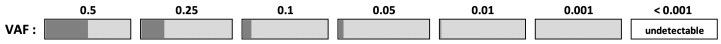
377 A



C



B

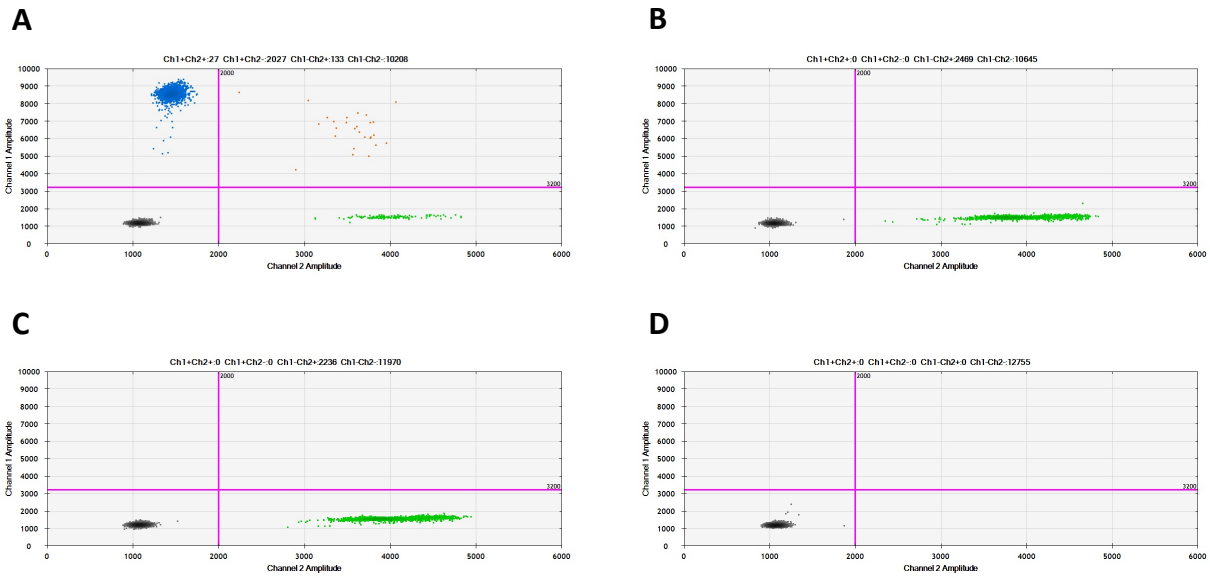


Assay #	Assay	Gene	Sample ID	Sample Type	VAF (%)		
					AGILE	smMIPseq	ddPCR
1	chr1_115258747_C>A	NRAS	SE16-2571	Positive control			
			SE15-0538	Test			
			CB gDNA	Negative control	X		
			water	Negative control	X	X	
2	chr1_115258747_C>T	NRAS	SE17-0221	Positive cntrl			
			SE16-0533	Test			
			SE15-0573	Test			
			SE16-1440	Test			
			SE16-2577	Test			
			SE15-0445	Test			
			SE16-1424	Test			
			CB gDNA	Negative control	X		
			water	Negative control	X	X	
3	chr15_90631934_C>T	IDH2	SE16-0153	Positive control			
			SE15-2817	Test			
			CB gDNA	Negative control	X		
			water	Negative control	X	X	
4	chr2_209113113_G>A	IDH1	SE16-0971	Positive control			
			SE16-0065	Test			
			CB gDNA	Negative control	X		
			water	Negative control	X	X	
5	chr9_5073770_G>T	JAK2	SE16-1043	Positive control			
			SE15-0538	Test			
			CB gDNA	Negative control	X		
			water	Negative control	X	X	
6	chr1_115258747_C>G	NRAS	SE17-0505	Positive control			
			SE15-2084	Test			
			SE17-1169	Test			
			CB gDNA	Negative control	X		
			water	Negative control	X	X	
7	chr12_25398281_C>T	KRAS	SE15-0298	Positive control			
			SE15-2084	Test			
			SE16-1905	Test			
			SE16-1440	Test			
			CB gDNA	Negative control	X		
			water	Negative control	X	X	
8	chr7_148507447_G>C	EZH2	SE15-1986	Positive control			
			SE17-0872	Test			
			CB gDNA	Negative control	X		
			water	Negative control	X	X	
9	chr17_74732959_G>T	SRSF2	SE16-1172	Positive control			
			SE16-0488	Test			
			CB gDNA	Negative control	X		
			water	Negative control	X	X	

Figure. 5. Mutations detected by SmMIP-tools in comparison with patients' clinical reports. (A) Venn diagrams for SNVs and indels illustrate the number of mutations that are shared between SmMIP-tools output and the clinical genetic reports as well as mutations that differ between the two sources. (B) Validation of a subset of high confidence mutations identified in SmMIP-tools output but not clinical reports by ddPCR. White background indicate that the variant was not reported while light grey background indicates the variant was detected above VAF of 0.1%. Dark grey bars represent the VAFs. X - not tested. (C) Scatterplot showing a strong linear correlation (Pearson r=0.996, P-value=9.81×10^{-7}), between VAFs calculated by ddPCR and those derived from the SmMIP-tools output for the 17 high confidence SNVs interrogated in test samples and positive controls.

An average of 29.7 (median=29.6) scaled CADD score (Rentzsch et al. 2019) was calculated for these 111 SNVs, indicating a striking enrichment for deleterious mutations. A subset of high confidence SNVs representing 15.3% of the newly detected SNVs was chosen for validation using digital droplet PCR (ddPCR, Methods). The authenticity of these variants was successfully validated at a 100% rate (Fig. 5B). Furthermore, a high correlation was observed between the VAFs obtained by both smMIP sequencing and ddPCR results (Fig. 5C).

SmMIP-tools was also capable of providing a plausible explanation for the source of some false non-reference alleles, including those detected at relatively high VAF yet not mentioned in the clinical sequencing reports (Supplemental Fig. 3). Information concerning each read's smMIP-of-origin, reads' family size, and UMI sequences suggest that these errors could be generated from index misassignment (also termed as "index-hopping") between multiplexed libraries.



Supplemental Figure 3. ddPCR results for the oncogenic variant JAK2^{V617F}, flagged as "Potential index-hopping". (A) Sample SE16-015 served a positive control. (B) Inability to detect the mutated allele from the test sample, SE16-0218, where the mutation was detected using smMIP sequencing yet was flagged as an error that may arise due to sample-index misassignment. (C) Cord blood DNA was used as a negative control. (D) A water sample was served as "blank". Blue dots correspond to the FAM fluorophores that marks the detection of the mutated allele. Green dots correspond to the HEX fluorophores that marks the wild type allele.

430 Supporting this is the fact that multiple other samples in the cohort presented the same non-reference
431 alleles, some with high VAF (Supplemental Table 5), and shared one sample index barcode with these
432 samples, thus providing an abundant source of reads that could be misassigned. Nine of the 23 variants
433 flagged by SmMIP-tools as "Potential index hopping" were also flagged as "Common SNP" and were
434 detected with atypical VAF far below the expected ~50% for germline mutations. This further supports
435 SmMIP-tools use of UMIs and read-smMIP linkages to flag errors that would otherwise be easily
436 mistaken as true mutations using other pipelines that do not take this information into account
437

438 **Discussion**

439 Molecular inversion probes enable cost-effective multiplex targeted sequencing. However, the immense
440 analytical burden associated with handling the resulting data poses a major obstacle to many users of
441 the technology. Several of the fundamental steps in smMIP-derived data interpretation include initial
442 data quality assessment, read processing, alignment of reads, summarizing the raw base calls, and
443 mutation calling. To address the high amount of analytical work associated with NGS data interpretation,
444 many tools have been developed, yet the critical need to address many fundamental analytical steps
445 under a single software suite, especially as they apply to smMIP-derived data, still remains. The need
446 to join together different programs makes the construction of efficient analytic pipelines difficult and
447 time-consuming. Furthermore, and most importantly, some functions that address critical issues specific
448 to the smMIP technology are entirely missing in existing tools. As an example, one key feature that is
449 typically absent from other NGS data analysis tools is the ability to link the reads to their precise probe-
450 of-origin. We have shown that by generating read-smMIP linkages, one can identify and filter chimeric
451 reads and non-informative reads that originated from self-annealing smMIPs. Furthermore, we have
452 shown that read-smMIP linkage can support the suppression of errors by comparing results derived
453 from overlapping smMIPs. We also designed SmMIP-tools to identify the correct position of the smMIP
454 arms' sequences in the reads' layout. This unique feature is necessary to prevent mutation calling
455 outside of the target regions of any particular smMIP and support the marking of reads with
456 compromised UMIs to improve the overall quality of the data for downstream analysis.
457 Defining the absolute ground truth is a major challenge when reporting mutations from NGS data
458 (Koboldt 2020). Errors that arise during library preparation and sequencing are abundant and can easily
459 obscure real mutations. The introduction of molecular barcode sequences into the probes' sequence

460 (Hiatt et al. 2013) was an essential step in improving the original MIP design towards more accurate
461 mutation calling. In addition to its ability to use UMIs to assemble SSCS, SmMIP-tools provides a
462 versatile error modelling approach to calculate a P-value for every non-reference allele observed in the
463 data to reflect the probability of a false observation. To address different experimental configurations,
464 SmMIP-tools was designed to derive with error rates models using either a single matched control
465 sample, a cohort of controls or with no controls by leveraging data from the experimental cohort itself.
466 Specificity can be improved by modelling all types of base substitutions, insertions and deletions in both
467 the paired reads, overlapping smMIPs, and technical replicates if available.
468 To our knowledge, SmMIP-tools is the only software that uses sequencing batch information to support
469 variants' authenticity. In addition to P-values, users are provided with sample-specific and allele-specific
470 information, including the VAF of the called allele in other samples, the number of samples in which the
471 same allele was observed with a higher VAF and the number of other samples in which the allele was
472 detected above the sequencing noise. Other types of outputs that are unique to SmMIP-tools include
473 the number of reads that were grouped to generate individual consensus sequences for every detected
474 non-reference allele and their intra-family VAFs, all of which are important to rank and prioritize mutation
475 calls effectively. We have shown that batch-related UMI information can also be used to flag false
476 positives potentially arising due to sample-index misassignment.
477 To mitigate index misassignment, specific recommendations, including the use of specific adaptor
478 sequences and sequencing systems have been investigated (Costello et al. 2018). Here we provide a
479 novel *in-silico* approach to identify calls that potentially arose due to index misassignment that could be
480 otherwise mistaken for real mutations using alternative mutation calling algorithms. This functionality of
481 leveraging information other than sample barcode sequences can reduce the cost associated with
482 generating large multiplexed libraries, and thus further increases the smMIP technology's cost-
483 effectiveness. Nevertheless, it is important to note that unlike our *in-silico* approach, using non-
484 combinatorial dual indexes does allow direct identification and removal of swapped reads which induce
485 sample contamination (Costello et al. 2018).
486 Sensitive and cost-effective NGS sequencing opens a myriad of clinical applications, including
487 screening of microbial populations, non-invasive prenatal testing and early cancer detection. It can
488 improve routine diagnostic testing of tumours in molecular diagnostic laboratories and inform cancer
489 treatment decisions by enabling longitudinal sequencing efforts to monitor emerging treatment resistant

490 clones. We anticipate that the public availability of SmMIP-tools will lead to improved smMIP-derived
491 data analysis and increased use of the technology.
492

493 ## Methods

494 **smMIP panel design to maximize error identification and suppression**

495 Proper smMIP design includes all the discrete genomic elements encased within a single DNA
496 oligonucleotide. This includes a standard backbone sequence, UMIs and custom extension and ligation
497 arms (Fig. 1A). The two halves of the backbone sequence are complementary to sequences located on
498 the forward and reverse PCR primers, thus permitting subsequent PCR amplification of captured
499 sequences (Fig. 1B). We used dual sample indices, each 8 bases in length, on each PCR primer to
500 enable the high multiplex sequencing conducted in this study. We also designed dual UMIs, each 4
501 bases in length, at each side of the backbone sequence to allow for PCR amplified target sequences to
502 be collapsed to the single smMIP-captured molecule from which they derived. This UMI design allows
503 the detection of up to 65,536 unique molecules per smMIP targeted loci. All the oligos sequences that
504 are needed for successful sequencing are illustrated in Fig. 1.
505 In this study, we targeted hotspot variants and exons known to be commonly mutated in myeloid
506 malignancies (Supplemental Table 4). smMIPs were designed using MIPgen (Boyle et al. 2014). To
507 maximize the error-correction capabilities of SmMIP-tools we followed two guiding principles. First, the
508 read length of the intended sequencing platform of interest was considered to help determine the
509 optimal target sequence length that will allow every smMIP's target to be sequenced by both paired-
510 end sequencings reads fully. Second, when a relevant genomic region could not be covered entirely by
511 a single smMIP, we designed additional overlapping smMIPs to tile the region. No preference was set
512 to design overlapping smMIP on opposite DNA strands. Based on analysis of simulated data, smMIPs
513 were designed to align >5 bases apart from the start or end site of their overlapping smMIPs to eliminate
514 erroneous read-smMIP linkages. SmMIPs were ordered as separate standard oligos (Thermo Fisher
515 Scientific), pooled 1:1, deployed on cord blood DNA samples and then empirically rebalanced by
516 increasing the concentration of low-performing smMIPs to achieve a more uniform sequencing
517 coverage across the panel.
518
519

520 **smMIP-Library preparation and sequencing**

521 SmMIP libraries were generated as described elsewhere (Cantsilieris et al. 2017; Hiatt et al. 2013).
522 Phosphorylation of the rebalanced smMIP pool was preformed using a smMIP phosphorylation reaction
523 previously described [18]. Once phosphorylated, the smMIP pool was diluted such that 4ul would
524 achieve a 1:1000 DNA:smMIP molecular ratio in subsequent target capture steps when using 5ul of
525 DNA prepared at 20ng/ul for a total input of 100ng. The probe capture reaction, including the probe
526 hybridization phase, and gap-fill and ligation phase as well as exonuclease treatment, were performed
527 as previously described [8]. PCR amplification of exonuclease-treated smMIP captured products were
528 also performed as previously described [8]. Technical duplicates of PCR amplified libraries were
529 generated for each sample, using unique combinations of barcoded PCR primers. Once generated,
530 sample libraries in each 96-well plate were pooled by volume. Each pool was quantified using Qubit
531 (High-Sensitivity dsDNA Assay, Thermo Fisher Scientific) and underwent a PCR clean-up using
532 MinElute columns (Qiagen). Up to 5ug of input was used per column. Elution was made with 15ul of
533 molecular-grade water. Each pool was subjected to quality control steps, including quantification and
534 fragment analysis on the BioAnalyzer 2100 high-sensitivity assay (Agilent). All the different plate-
535 derived pools were then pooled together by concentration taking into consideration the number of
536 samples in each plate. This final pool then underwent size selection using PippinHT (Sage Science) to
537 mitigate the carryover of smaller and larger products suspected to be self-annealed smMIPs without
538 target sequences and complex oligo hybrids, respectively. Lastly, pooled and size-selected multiplexed
539 smMIP libraries were subjected to sequencing. Bulk cell line libraries were sequenced on the NextSeq
540 using the Mid-Output kit (Illumina). Cord blood, cell line DNA mixes and patients' libraries were
541 sequenced using a single NovaSeq SP flow cell (Illumina). 10% PhiX was used to ensure sufficient
542 library complexity.

543

544 **Derivation of mutation lists for SmMIP-tools benchmarking using cell line DNA mixes**

545 To construct lists of high confidence errors and mutated alleles to benchmark SmMIP-tools mutation
546 calling algorithm, we first sequenced the DNA of 8 cell lines, namely ME-1, Kasumi-1, KG-1A, U937,
547 SU-DHL-10, OCI-AML3, OCI-AML-22, and OCI-AML-8227. Each cancer cell line was sequenced in 4
548 replicates, and SmMIP-tools was used to report mutations at each of the 32 samples independently
549 using the experimental cohort itself for allele-specific error rate estimations. That is, allele frequencies

550 in each investigated sample were evaluated against those observed in all the samples of the other cell
551 lines. Altogether, we identified 44 high confidence SNVs and 4 insertions (Supplemental Table 2). True
552 positive status was determined if a mutation: 1) Received a P-value<=0.05, 2) Was found in at least 3
553 out of 4 replicates with VAF>=0.02, and 3) Was supported by at least 1 SSCS from each strand (Read1
554 and Read2). We removed mutations if they were shared by more than 3 different cell lines to prevent
555 the list from being populated with a large number of common germline mutations. Alleles assigned with
556 a P-value<=0.05 in individual samples but did not pass the additional criteria above were eliminated
557 from benchmarking and were assigned with a status "Uncertain" (Supplemental Table 2). Every other
558 non-reference allele in the remaining interrogated genomic positions was considered as false positive
559 if it was seen in the follow-up sequencing of the 6 DNA mixes (Supplemental Table 2).
560 For cell line mixes data analysis, a minimum of one SSCS in each strand was required. For the
561 derivation of probabilistic error rate models, a control cohort of 16 umbilical cord blood samples was
562 used. A p-value threshold of 0.05 was set to suppress errors following Bonferroni correction for multiple
563 tests. When calculating the percentage of error-free positions, the focus is on false positives. Therefore,
564 both real mutations and alleles with the status "Uncertain" were removed for this analysis. Alleles with
565 the status "Uncertain" were not included in sensitivity, precision and F1 calculations.
566
567 **Generating read-smMIP linkages**
568 To determine the probe-of-origin for every sequenced read pair, SmMIP-tools first searches for smMIPs
569 that their targeted genomic loci, including the extension and ligation arms, substantially overlap with the
570 genomic loci determined by the paired-reads' alignment to the genome (default 0.95, user defined
571 parameter). Once smMIP candidates are selected, the algorithm proceeds with local alignment of
572 each smMIP's extension and ligation arms to the reads. The exact probe-of-origin is determined if both
573 of its arms will align the number of UMI bases (here 4nt) from the reads' extremes. The exact location
574 of UMI bases in each read of the pair (i.e, 5' or 3') is determined by the reads' SAM flags and the number
575 of UMI bases in each read which is automatically determined from the user-provided panel design file
576 obtained from MIPgen (Boyle et al. 2014). When the above alignment expectations are not met, the
577 UMI will be considered unreliable and the paired reads will not be used for any further analysis
578 concerning SSCS.
579

580 **Assembly of single strand consensus sequences**

581 SmMIP-tools uses the R-package Rsamtools (Morgan M, Pagès H, Obenchain V, Hayden N .
582 Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF)) to generate base call summaries which
583 are subsequently used to calculate allele frequencies across the target panel. Base call summaries
584 were generated for every UMI-smMIP combination. To minimize information loss, the minimum depth
585 parameter was set to one read, the minimum read mapping quality was set to 50 and the base quality
586 cut-off to 10 (user-defined parameters). At each genomic position, the most frequent allele was
587 determined based on a 70% threshold as previously described (Kennedy et al. 2014) (user-defined
588 parameter). We did not restrict read family size since this information is being captured in the final
589 SmMIP-tools report. The final report allows complete evaluation of the data and is being used to flag
590 and rank reported alleles.

591

592 **Probabilistic modeling of error-rates**

593 SmMIP-tools uses the pbinom R function to calculate for each observed allele in a sample of interest a
594 P-value reflecting the probability of obtaining a number equal to or higher than the observed number of
595 non-references supporting reads for the same allele in a single matched control, a larger control cohort
596 or with no controls. If the latter option is chosen, SmMIP-tools uses the entire cohort except for the
597 sample of interest (and its technical replicate, if available) to estimate error rates. Using the cellbaseR
598 R package (Abdallah 2020), sequenced alleles are annotated, and information concerning recurrent
599 cancer mutations and polymorphisms is leveraged to increase sensitivity at those positions. Samples
600 detected with recurrent alleles with VAF>=0.05 (user-defined parameter) were removed from error rate
601 calculations. The allele frequency in all the other samples was set to their median value. To derive
602 binomial probabilities, allele-specific error rates were determined as the sum of all the non-reference
603 supporting reads in all the controls divided by the total number of reads covering the allele. This value
604 was evaluated against the number of non-reference supporting reads and the allele's coverage in the
605 sample of interest. If any allele was detected by both Read1 and Read2 it received the higher P-value
606 of the two models. This process was repeated for alleles that are covered by overlapping smMIPs and
607 technical replicates. For benchmarking, we modelled allele-specific error rates using a set of 16 control
608 umbilical cord samples where true mutation burden and its effect on real error rates are expected to be
609 low. For comparison with patients' clinical reports, the experimental cohort itself served as control.

610 **Mutation categories**

611 By default, SmMIP-tools generates a report which includes baseline mutations calls defined as those
612 mutations detected with a Bonferroni corrected P-value that is lower than or equal to the user-defined
613 P-value cut-off. In this study, we used a cutoff of 0.05. The number of tests used for Bonferroni
614 correction is equal to the total number of alleles that can potentially be detected using the smMIP target
615 panel. These include "A", "C", "T", "G", "-" (deletion) and "+" (insertion) at every targeted genomic
616 position (Supplemental Table 4). SmMIP-tools uses the R-package cellbaseR (Abdallah 2020) to
617 annotate single base substitutions. To enrich the baseline calls for somatic and pathogenic mutations,
618 we removed common SNPs with a minor allele frequency (MAF)>0.01 as well as mutations assigned
619 with the "potentially germline" flag that was determined based on MAF cut-offs of 0.001<=MAF<0.01
620 (user-defined parameters). SNVs with the "potentially germline" flag were kept only if they have been
621 reported in COSMIC (Tate et al. 2019). We also removed SNVs that were assigned the "likely benign"
622 flag. Those include mutations with one of the following annotations: "synonymous variant", "intron
623 variant", "splice region variant" (rather than "splice acceptor/donor variants"), "non-coding transcript
624 exon variant", "5 prime UTR variant", or "3 prime UTR variant". For Indels, we removed those that fall
625 entirely within intronic regions.

626 The report generated by the SmMIP-tools mutation calling algorithm also includes additional information
627 concerning consensus reads and details about how baseline called alleles are being represented in
628 other samples in the sequenced cohort. We used this information to derive two categories for mutation
629 ranking, namely "High confidence" and "Lower confidence".

630 1) High confidence mutations were determined as those mutations that: a) Received a P-value that is
631 lower than or equal to 0.05. b) Are enriched for somatic and pathogenic mutations as described above.
632 c) Had no flags indicating possible technical issues that may interfere with their detection (i.e., the flags:
633 "Cannot be supported by both Read1 and Read2", "Cannot be supported by all overlapping smMIPs",
634 "Cannot be supported by both technical replicates", "No SSCS support in one of the replicates", "No
635 SSCS support at all", "No SSCS support in either Read1 or Read2 in at least one of the replicates"). d)
636 Had no flags indicative of potential false positives (i.e., the flag "Potential index hopping"). e) Did not
637 have the "VAF Warning" flag. The absence of this flag indicates that the mutation was detected with the
638 highest VAF in the reported sample and that its VAF is at least 2-fold (user-defined parameter) higher
639 as compared with all the other samples in the entire cohort. Of note, by default, the algorithm that

640 generates this flag permits that a single sample with a lower VAF will not reach the 2-fold VAF
641 difference. The algorithm was also designed to work properly with authentic recurrent mutations
642 detected as baseline calls in the cohort.
643 2) Lower confidence mutations were determined as those mutations that received flags indicating
644 potential technical issues that could have interfered with their detection. Those include mutations that
645 were detected only by Read1 or Read2 or only in a single smMIP where overlapping smMIPs were
646 present. Of note, optimal smMIP design where every targeted base is sequenced twice, once with
647 Read1 and once with Read2, can reduce the number of lower confidence calls. Other mutations in this
648 category are mutations that were not detected in both the technical replicates due to low coverage
649 anomalies in one of the replicates. A minimum coverage of 10x from each read (Read1 and Read2)
650 (user-defined parameter) was required to consider the strand-derived information for P-value
651 calculations. For any mutation call, when the coverage does not reach the set threshold, a detailed "Low
652 coverage" flag will be generated. More specifically, we defined and reported lower confidence mutations
653 if: a) They were detected in at least 50% of the available detection options mentioned above (i.e.
654 overlapping sequencing reads, overlapping smMIPs and technical replicates). For example, if a
655 mutation was detected in both Read1 and Read2 but in only one replicate, where overlapping smMIPs
656 do not exist, it will be reported since it was detected in 50% of the possible detection options. If a
657 mutation that is covered by a single smMIP was detected with only Read1 in only one replicate, it would
658 not be reported since it was detected in only 1 of the 4 possible options. b) There were not more than
659 5 other samples in the entire sequenced cohort where those mutations received a P-value<=0.05. c)
660 There were no other samples that were detected with those mutations with a higher VAF (only relevant
661 to mutations that lack a Cosmic ID).
662

663 **Patient cohort**

664 All patients referred to the MPN/leukemia program at the Princess Margaret Cancer Center were
665 approached for targeted sequencing as part of the Advanced Genomics in Leukemia (AGILE)
666 prospective study (Alduaij et al. 2018). All biological samples were collected according to procedures
667 approved by the Research Ethics Board of the University Health Network (01-0573.28) and were viably
668 frozen in the Princess Margaret, Leukemia Tissue Bank. Written informed consent was obtained from
669 all patients in accordance with the declaration of Helsinki.

670 **Clinical reports**

671 DNA samples extracted from peripheral blood or bone marrow were used for NGS. Sequencing was
672 performed using the TruSight Myeloid Sequencing Panel (Illumina) and run on the MiSeq Illumina
673 platform as previously described using a protocol that was validated by the University Health Network
674 Advanced Molecular Diagnostics Laboratory (Thomas et al. 2017) . Briefly, 54 genes implicated in
675 myeloid malignancies were profiled using amplicon-based library preparation (Supplemental Table 4).
676 Data analysis and quality assessment for reporting of SNVs and short indels was performed using
677 NextGene v.2.3.1 (SoftGenetics, State College, PA). Detected variants were then annotated using the
678 WHO classification and diagnostic criteria as previously reported (Barbui et al. 2018) to differentiate
679 between variants of unknown significance and those that are oncogenic. Oncogenic variants detected
680 at coverage >100X, with VAF>5%, were included for subsequent investigation. Known hotspot variants
681 detected below these thresholds were verified using either Sanger sequencing or ddPCR. Detection of
682 recurrent NPM1, insertions or deletions, and FLT3 internal tandem duplications was performed using
683 fluorescent polymerase chain reaction (PCR) followed by fragment analysis. Variants that repeatedly
684 occurred in >10% of cases, excluding mutation in hotspot regions or variants previously reported in the
685 literature, were filtered out.

686

687 **Digital droplet PCR**

688 We used ddPCR to validate a subset of SNVs that were detected as “high confidence” mutations by
689 SmMIP-tools but were not reported in the clinical sequencing reports. Nine ddPCR assays were
690 designed to target oncogenic SNVs spanning a VAF range of 0.008-0.077 as determined by SmMIP-
691 tools in 17 patients (n=1 VAF<0.01, n=5 0.01<VAF<0.02, n=8 0.02<VAF<0.05, n=3 0.05<VAF<0.08).
692 Wild type and mutant specific PrimeTime LNA qPCR probes were conjugated to HEX and FAM
693 fluorophores, respectively (Integrated DNA Technologies). 20ng of DNA, originally used for clinical
694 sequencing and re-sequencing using smMIPs, was subjected to ddPCR assays in a 96-well plate
695 according to the manufacturer’s protocol (Bio-Rad). Droplets were generated and read using the QX200
696 ddPCR system (Bio-Rad) to detect the absolute number of reference (HEX) and non-reference (FAM)
697 alleles. VAF was calculated as the number of FAM captured targets over the total number of targets
698 (FAM+HEX). As previously reported (Shlush et al. 2017), variants were considered positive if >3
699 droplets were captured in the mutant (FAM) channel with a final VAF>0.001. Samples with the same

700 mutations detected by both the clinical pipeline and SmMIP-tools were used as positive controls. Cord
701 blood DNA was used as a negative control and water was used as a blank. ddPCR assay conditions
702 and associated primer and probe sequences are supplemented (Supplemental Table 6).
703

704 **Supplementary information**

705 **Supplemental Figure 1.** Addressing issues related to the smMIP technology by linking reads to their
706 probe-of-origin with read-smMIP linkages
707 **Supplemental Figure 2.** Simulating smMIPs to assess the performance of read-smMIP linkages
708 generation
709 **Supplemental Figure 3.** ddPCR results for the oncogenic variant JAK2^{V617F}, flagged as "Potential
710 index-hopping"
711 **Supplemental Table 1.** Comparison between SmMIP-tools and popular analytical methods for NGS
712 **Supplemental Table 2.** True cell lines mutations and alleles that were removed from the benchmarking
713 process
714 **Supplemental Table 3.** Sequenced samples and clinical genetic results
715 **Supplemental Table 4.** Targeted genomic loci
716 **Supplemental Table 5.** SmMIP-tools mutation calls in the clinical cohort
717 **Supplemental Table 6.** Details concerning ddPCR assays preformed
718 * Supplemental tables are provided as separated files.
719

720 **Acknowledgements**

721 This project was supported by work done at the Advanced Molecular Diagnostic Lab (AMDL) at the
722 Princess Margaret Cancer Centre. The authors also thank the team at the Princess Margaret Genomics
723 Centre for their sequencing services and genome informatics at the Ontario Institute for Cancer
724 Research for providing high-performance computing. We thank Dr. Jean Wang, Dr. Qiang Liu, Dr. Eric
725 Lechman and Dr. Héléna Boutze for providing us with the cell lines used for SmMIP-tools
726 benchmarking.
727
728
729

730 **Authors' contributions**

731 JJFM optimized the design of smMIPs, generated NGS smMIP libraries, preformed ddPCR assays,
732 contributed to data analysis and wrote the manuscript. JJCC managed clinical sequencing and provided
733 the clinical genomic reports. LIS and JED contributed to panel optimization and sequencing. MDM and
734 AA managed patients' selection, and provided the clinical genomic reports and DNA samples from the
735 Princess Margaret, Oncology and Hematology Tissue Bank for smMIP resequencing. SA wrote the
736 code and the manuscript, contributed to data analysis and directed and supervised all aspects of the
737 study. All authors read and approved the final manuscript.

738

739 **Funding**

740 The work conducted in the SA lab is supported by the Investigator Award received from the Ontario
741 Institute for Cancer Research with funds from the province of Ontario and by the University of
742 Toronto's Medicine by Design initiative, with funds from the Canada First Research Excellence Fund
743 (CFREF). JJFM is personally funded by the Canadian Institutes of Health Research Doctoral Award:
744 Frederick Banting and Charles Best Canada Graduate Scholarships (FBD-170928).

745

746 **Competing interests**

747 LIS serves as a consultant for Sequentify LTD and Metasight LTD. All other authors declare that they
748 have no competing interests.

749

750 **Availability of data and materials**

751 The source code for SmMIP-tools, its manual and additional scripts used for data generation and
752 analysis are all available on github (https://github.com/abelson-lab/smMIP-tools). Raw sequencing data
753 generated in this study have been submitted to the European Genome-Phenome Archive (EGA;
754 https://ega-archive.org) under accession number EGAS00001005359. All other information concerning
755 the use of SmMIP-tools, the benchmarking process and the comparison of SmMIP-tools results with
756 the patient's clinical reports are included in the method section and supplemental files.

757

758

759

760 **References**

761 Abdallah M cellbaseR: Q annotation data from the high performance C web. 2020.
762 https://github.com/melsiddieg/cellbaseR.
763 Acuna-Hidalgo R, Sengul H, Steehouwer M, van de Vorst M, Vermeulen SH, Kiemeney LALM,
764 Veltman JA, Gilissen C, Hoischen A. 2017. Ultra-sensitive Sequencing Identifies High
765 Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life. *Am J Hum*
766 *Genet* **101**: 50–64.
767 Alduaij W, McNamara CJ, Schuh A, Arruda A, Sukhai M, Kanwar N, Thomas M, Spiegel J, Kennedy
768 JA, Stockley T, et al. 2018. Clinical Utility of Next-generation Sequencing in the Management of
769 Myeloproliferative Neoplasms: A Single-Center Experience. *HemaSphere* **2**: e44.
770 Barbui T, Thiele J, Gisslinger H, Kvasnicka HM, Vannucchi AM, Guglielmelli P, Orazi A, Tefferi A.
771 2018. The 2016 WHO classification and diagnostic criteria for myeloproliferative neoplasms:
772 document summary and in-depth discussion. *Blood Cancer J* **8**: 15.
773 Boyle EA, O'Roak BJ, Martin BK, Kumar A, Shendure J. 2014. MIPgen: optimized modeling and
774 design of molecular inversion probes for targeted resequencing. *Bioinformatics* **30**: 2670–2672.
775 https://pubmed.ncbi.nlm.nih.gov/24867941.
776 Cantsilieris S, Stessman HA, Shendure J, Eichler EE. 2017. Targeted Capture and High-Throughput
777 Sequencing Using Molecular Inversion Probes (MIPs). *Methods Mol Biol* **1492**: 95–106.
778 Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L, Granger B, Green L, Howd T,
779 Mason T, et al. 2018. Characterization and remediation of sample index swaps by non-
780 redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* **19**: 332.
781 https://doi.org/10.1186/s12864-018-4703-0.
782 Hardenbol P, Banér J, Jain M, Nilsson M, Namsaraev EA, Karlin-Neumann GA, Fakhrai-Rad H,
783 Ronaghi M, Willis TD, Landegren U, et al. 2003. Multiplexed genotyping with sequence-tagged
784 molecular inversion probes. *Nat Biotechnol* **21**: 673–678.
785 Hardenbol P, Yu F, Belmont J, Mackenzie J, Bruckner C, Brundage T, Boudreau A, Chow S, Eberle J,
786 Erbilgin A, et al. 2005. Highly multiplexed molecular inversion probe genotyping: over 10,000
787 targeted SNPs genotyped in a single tube assay. *Genome Res* **15**: 269–275.
788 https://pubmed.ncbi.nlm.nih.gov/15687290.
789 Hiatt JB, Pritchard CC, Salipante SJ, O'Roak BJ, Shendure J. 2013. Single molecule molecular
790 inversion probes for targeted, high accuracy detection of low frequency variation. *Genome Res*
791 **23**: 843–54. http://www.ncbi.nlm.nih.gov/pubmed/23382536.
792 Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, Prindle MJ, Kuong KJ, Shen J-C,
793 Risques R-A, et al. 2014. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat*
794 *Protoc* **9**: 2586–2606. https://www.ncbi.nlm.nih.gov/pubmed/25299156.
795 Khan M, Cornelis SS, Pozo-Valero M Del, Whelan L, Runhart EH, Mishra K, Bults F, AlSwaiti Y,
796 AlTalbishi A, De Baere E, et al. 2020. Resolving the dark matter of ABCA4 for 1054 Stargardt
797 disease probands through integrated genomics and transcriptomics. *Genet Med* **22**: 1235–
798 1246.
799 Koboldt DC. 2020. Best practices for variant calling in clinical sequencing. *Genome Med* **12**: 91.

800 https://doi.org/10.1186/s13073-020-00791-w.
801 Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, Liu Y, Chen X, Newman S, Nakitandwe
802 J, et al. 2019. Analysis of error profiles in deep next-generation sequencing data. *Genome Biol*
803 **20**: 50. https://doi.org/10.1186/s13059-019-1659-6.
804 Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner
805 DJ. 2010. Target-enrichment strategies for next-generation sequencing. *Nat Methods* **7**: 111–
806 118. https://doi.org/10.1038/nmeth.1419.
807 Morgan M, Pagès H, Obenchain V, Hayden N . Rsamtools: Binary alignment (BAM), FASTA, variant
808 call (BCF) and tabix file import. 2020. R package version 2. 6. 0.
809 https://bioconductor.org/packages/Rsamtools.
810 Neveling K, Mensenkamp AR, Derks R, Kwint M, Ouchene H, Steehouwer M, van Lier B, Bosgoed E,
811 Rikken A, Tychon M, et al. 2017. BRCA Testing by Single-Molecule Molecular Inversion Probes.
812 *Clin Chem* **63**: 503–512. https://doi.org/10.1373/clinchem.2016.263897.
813 Nilsson M, Malmgren H, Samiotaki M, Kwiatkowski M, Chowdhary BP, Landegren U. 1994. Padlock
814 probes: circularizing oligonucleotides for localized DNA detection. *Science* **265**: 2085–2088.
815 O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, Carvill G, Kumar A, Lee C,
816 Ankenman K, et al. 2012. Multiplex targeted sequencing identifies recurrently mutated genes in
817 autism spectrum disorders. *Science* **338**: 1619–1622.
818 Pippucci T, Licchetta L, Baldassari S, Marconi C, De Luise M, Myers C, Nardi E, Provini F, Cameli C,
819 Minardi R, et al. 2019. Contribution of ultrarare variants in mTOR pathway genes to sporadic
820 focal epilepsies. *Ann Clin Transl Neurol* **6**: 475–485. https://pubmed.ncbi.nlm.nih.gov/30911571.
821 Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, et
822 al. 2007. Multiplex amplification of large sets of human exons. *Nat Methods* **4**: 931–936.
823 Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. 2019. CADD: predicting the
824 deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**: D886–D894.
825 https://doi.org/10.1093/nar/gky1016.
826 Shlush LI, Mitchell A, Heisler L, Abelson S, Ng SWK, Trotman-Grant A, Medeiros JJF, Rao-Bhatia A,
827 Jaciw-Zurakowsky I, Marke R, et al. 2017. Tracing the origins of relapse in acute myeloid
828 leukaemia to stem cells. *Nature* **547**: 104. http://dx.doi.org/10.1038/nature22993.
829 Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C,
830 Dawson E, et al. 2019. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids*
831 *Res* **47**: D941–D947. https://doi.org/10.1093/nar/gky1015.
832 Thomas M, Sukhai MA, Zhang T, Dolatshahi R, Harbi D, Garg S, Misyura M, Pugh T, Stockley TL,
833 Kamel-Reid S. 2017. Integration of Technical, Bioinformatic, and Variant Assessment
834 Approaches in the Validation of a Targeted Next-Generation Sequencing Panel for Myeloid
835 Malignancies. *Arch Pathol Lab Med* **141**: 759–775.
836 Turner EH, Lee C, Ng SB, Nickerson DA, Shendure J. 2009. Massively parallel exon capture and
837 library-free resequencing across 16 genomes. *Nat Methods* **6**: 315–316.
838 https://pubmed.ncbi.nlm.nih.gov/19349981.
839