EUROPEAN HEMATOLOGY ASSOCIATION

haematologica
Journal of the European Hematology Association

Characterization of gene mutations and copy number changes in acute myeloid leukemia using a rapid target enrichment protocol

by Niccolo' Bolli, Nicla Manes, Thomas McKerrel, Jianxiang Chi, Naomi Park, Gunes Gundem, Michael A. Quail, Vijitha Sathiaseelan, Bram Herman, Charles Crawley, Jenny I.O. Craig, Natalie Conte, Carolyn Grove, Elli Papaemmanuil, Peter J. Campbell, Ignacio Varela, Paul Costeas, and George S. Vassiliou

Haematologica 2014 [Epub ahead of print]

*Citation: Bolli N, Manes N, McKerrel T, Chi J, Park N,
Gundem G, Quail MA, Sathiaseelan V, Herman B, Crawley C, Craig JI, Conte N, Grove C,
Papaemmanuil E, Campbell PJ, Varela I, Costeas P, and Vassiliou GS. Characterization of gene mutations and copy number changes in acute myeloid leukemia using a rapid target enrichment protocol.
Haematologica. 2014; 99:xxx
doi:10.3324/haematol.2014.113381*

Publisher's Disclaimer.0
E-publishing ahead of print is increasingly important for the rapid dissemination of science. Haematologica is, therefore, E-publishing PDF files of an early version of manuscripts that have completed a regular peer review and have been accepted for publication. E-publishing of this PDF file has been approved by the authors. After having E-published Ahead of Print, manuscripts will then undergo technical and English editing, typesetting, proof correction and be presented for the authors' final approval; the final version of the manuscript will then appear in print on a regular issue of the journal. All legal disclaimers that apply to the journal also pertain to this production process.

Characterization of gene mutations and copy number changes in acute myeloid leukemia using a rapid target enrichment protocol

Authors

Niccolò Bolli[1, 2, 3], Nicla Manes[3, 4], Thomas McKerrell[4], Jianxiang Chi[5], Naomi Park[6], Gunes Gundem[1], Michael A. Quail[6], Vijitha Sathiaseelan[1], Bram Herman[7], Charles Crawley[3], Jenny I. O. Craig[3], Natalie Conte[4,8], Carolyn Grove[4], Elli Papaemmanuil[1], Peter J. Campbell[1], Ignacio Varela[9], Paul Costeas[5,10], George S. Vassiliou[4]

Affiliations:

1 Cancer Genome Project, Wellcome Trust Sanger Institute, Cambridge, UK

2 Department of Haematology, University of Cambridge , Cambridge, UK

3 Department of Haematology, Addenbrookes Hospital, Cambridge, UK

4 Haematological Cancer Genetics, Wellcome Trust Sanger Institute, Cambridge, UK

5 The Center for the Study of Haematological Malignancies, Nicosia, Cyprus

6 Sequencing Research and Development, Wellcome Trust Sanger Institute, Cambridge, UK

7 Agilent Technologies, Agilent Technologies LDA UK LTD, Cheadle, UK

8 EMBL-European Bioinformatics Institute, Cambridge, UK

9 Instituto de Biomedicina y Biotecnología de Cantabria (CSIC-UC-Sodercan), Departamento de Biología Molecular, Universidad de Cantabria, Santander, Spain

10 Molecular Haematology and Immunogenetics Center, The Karaiskakio Foundation, Nicosia, Cyprus

Running heads:

HaloPlex in AML diagnosis

Corresponding authors:

Dr Niccolo Bolli
nb8@sanger.ac.uk
Cancer Genome Project
Wellcome Trust Sanger Institute
Wellcome Trust Genome Campus
Cambridge
CB10 1SA
UK

Dr George S Vassiliou
gsv20@sanger.ac.uk
Haematological Cancer Genetics
Wellcome Trust Sanger Institute
Wellcome Trust Genome Campus
Cambridge
CB10 1SA
UK

Word count:

Abstract: 201
Main text: 3851
Tables: 1
Figures: 5
Supplemental Files: 1

Acknowledgments

This project was funded by the Wellcome Trust. NB is a fellow of the European Hematology Association and was supported by the Academy of Medical Sciences. EP is a European Hematology Association Advanced Research Fellow. GV is a Wellcome Trust Senior Fellow in Clinical Science. IV is funded by Spanish Ministerio de Economía y Competitividad subprograma Ramón y Cajal. We thank the Cambridge Blood and Stem Cell Biobank (CBSB,) National Institute of Health Research (NIHR) and the Cambridge Cancer Molecular Diagnosis Laboratory (CMDL) for assistance with sample acquisition and processing.

Abstract

Prognostic stratification is critical for making therapeutic decisions and maximizing survival of patients with acute myeloid leukemia. Advances in the genomics of acute myeloid leukemia have identified several recurrent gene mutations whose prognostic impact is being deciphered. We used HaloPlex target enrichment and Illumina-based next generation sequencing to study 24 recurrently mutated genes in 42 samples of acute myeloid leukemia with a normal karyotype. Read depth varied between and within genes for the same sample, but was predictable and highly consistent across samples. Consequently, we were able to detect copy number changes, such as an interstitial deletion of *BCOR*, three MLL partial tandem duplications, and a novel *KRAS* amplification. With regards to coding mutations, we identified likely oncogenic variants in 41/42 samples. *NPM1* mutations were the most frequent, followed by *FLT3*, *DNMT3A* and *TET2*. *NPM1* and *FLT3* indels were reported with good efficiency. We also showed that *DNMT3A* mutations can persist post-chemotherapy and in 2 cases studied at diagnosis and relapse, we were able to delineate the dynamics of tumor evolution and give insights into order of acquisition of variants. HaloPlex is a quick and reliable target enrichment method that can aid diagnosis and prognostic stratification of acute myeloid leukemia patients.

Introduction

Acute myeloid leukemia (AML) is a heterogeneous group of hematological malignancies characterized by a differentiation block and unrestricted proliferation of myeloid precursors. Historically, AML classification was based on phenotypic criteria of the French-America-British (FAB) co-operative group[1]. More recently, the World Health Organization (WHO), formulated an updated classification based on key genetic lesions underlying distinct clinico-pathological subgroups[2]. With the exception of FAB AML-M3 (acute promyelocytic leukemia), there is limited overlap between subgroups of the FAB and WHO classifications. As recent clinical advances in AML have been driven by better prognostic stratification[3], the WHO classification has rapidly made its way into routine clinical practice in view of its prognostic and therapeutic implications.

However, advances in AML genomics[4,5] have demonstrated that even within WHO classes there exists significant heterogeneity, which can translate into different clinical outcomes[6]. This is particularly true of patients with normal karyotype AML (AML-NK), who could be either over- or under-treated in the absence of prognostic information. In fact, AML-NK is driven by a complex interplay of several diverse leukaemogenic mutations that may confer different prognosis based on their combinatorial patterns of co-occurrence. For example, the good prognostic value of NPM1- or CEBPA-mutations[6-8] is annulled by the presence of *FLT3* internal tandem duplications (*FLT3-ITDs*)[9,10], in the same way as *c-KIT* mutations can negate the good prognostic impact of core binding factor translocations[11]. Similarly, other genes or gene combinations appear to carry prognostic value[5,12], and this is being assessed in large patient cohorts. Additionally, gene mutations may serve as therapeutic targets as shown for example by the clinical efficacy of the tyrosine kinase inhibitor dasatinib for AML with *c-KIT* mutations [13,14], and by therapies targeting *FLT3-ITD*[15].

Next generation sequencing (NGS) technologies introduced rapid sequencing of entire human genomes[16]. AML with normal karyotype was the first cancer whose genome was fully sequenced[17], and the spectrum of its genomic alterations has since been characterized in hundreds of patients[4]. Several technologies are now available that selectively enrich for relevant genes/regions (target enrichment) before NGS is performed. This allows for cheaper multiplexed sequencing of more cases, and moderates the complexity of downstream bioinformatics analyses. Such an approach, employing DNA pulldown with cRNA probes (Sureselect®, Agilent Technologies) was recently described in AML[18] and myelodysplastic syndromes[19,20]. However, this approach suffers from the need for laborious library preparation, long turnaround times and reduced sensitivity for detecting long insertions such as *FLT3-ITDs*[18]. In this study, we employed the HaloPlex® (Agilent Technologies) target enrichment system, which is based on digestion of genomic DNA to produce fragments tiling target regions, followed by sequence-specific annealing to custom-made probes followed by PCR-amplification to produce tagged amplicons for sequencing. This system uses little input DNA and promises a more affordable, quick, and efficient target enrichment that may be more suitable for analysis in diagnostic laboratories[21]. We used HaloPlex to study 24 recurrently mutated genes in 42 AML samples, mostly in the absence of matched normal DNA. Here we report its performance in identifying coding and copy number mutations affecting target genes.

Methods

Samples, DNA target enrichment, sequencing and alignment

DNA was extracted from bone marrow of 40 AML-NK patients with >80% leukemic infiltrate at diagnosis. All patients had either karyotyping or multiplex PCR to rule out recurrent chromosomal translocations (HemaVision®-Screen, DNA Diagnostic A/S). Tumor samples were compared to an unrelated normal DNA sample (human placenta) for variant calling. For 2 patients we collected bone marrow samples at diagnosis and at molecular relapse, identified by increased NPM1/ABL ratio by RT-qPCR. For 5 patients a matched bone marrow sample was also available post-chemotherapy. Informed consent was obtained within our ethics-approved study (IRB 07/MRE05/44) and samples were stored in accordance with the declaration of Helsinki. The 24 genes studied were selected based on their recurrence rate in AML and their relevance to pathogenesis and prognosis (Table 1). The targeting design was generated using an on-line design tool for HaloPlex and target enrichment was performed using HaloPlex standard protocol (version 2.0, November 2011). Briefly, 900 ng of DNA per sample were aliquoted into 8 digestion reactions, each containing 2 restriction enzymes. DNA from the 8 reactions was then pooled, hybridized to HaloPlex probes, and purified using magnetic beads. Fragments were ligated, amplified and barcoded through 19 PCR cycles and two pools of 12 and 35 samples sequenced on one lane each of HiSeq2000 (Illumina), 100 bp paired-end protocol.

Before alignment, 5 bp were trimmed from the start of each read to minimize possible mis-mapping due to restriction site sequence retention. Paired-end sequencing reads were aligned to the human genome (NCBI build 37) using BWA[22]. Unmapped or off-target reads were excluded. Apparent PCR duplicates were not removed as HaloPlex generates fragments of the same start and end positions that cannot be distinguished from each other before or after PCR.

On-target performance and copy-number analysis

To determine the coverage of the target region, we used a BED file encoding the coordinates of the coding sequence of each of the 24 genes and retrieved the number of reads covering each base-pair position using Bedtools v2.15[23]. We then normalized coverage in each sample by dividing the read count at each position by the total number of on-target mapped bases for that sample. Coverage data and plots were produced using open-source software and bespoke R scripts (R v3.0.3)[24]. To identify copy number variants at individual exons, we compared the average coverage of each exon with that of normal samples. Genes with three or more exons showing read depths above or below the standard deviation of normal samples were examined further for amplifications or deletions.

Mutation calling algorithms

Substitutions and insertions/deletions were detected using CaVEMan and Pindel as previously described[19,25,26]. Our main aim was to define driver events and therefore we only reported "likely oncogenic variants", defined as variants already reported as somatic in AML literature, or novel variants clustering with known somatic variant hotspots, or truncating variants in genes implicated in AML through loss of function mutations. Relevant variants and copy number events were validated with orthogonal techniques. More details are provided in the supplementary material.

Results

Patients and sequencing metrics

The target region of 140,811bp did not include UTRs or introns and was sequenced with a mean coverage of 3,655x (total output 39.91 gigabases (Gb)) (Figure 1A). The number of bases mapped on-target per sample was dependent on the degree of multiplexing and ranged from 0.13 to 1.26Gb (Figure 1A), representing an average of 66.33% of the total output. Unsurprisingly, there was a correlation between the depth of sequencing and the percentage of the target region covered at >1000x (p <2.2e-16, Figure 1A) and at >30x, which we consider the minimum depth for reliable analysis (p=0.04, Figure 1A). Coverage of each gene varied between samples depending on total sequencing output (Figure 1B), as did coverage of different genes within the same sample presumably due to factors such as PCR efficiency and GC content. Nevertheless, our study performed well as all genes were covered at >30x for at least 90% of their coding regions with the exception of the GC rich and notoriously hard to target *CEBPA*[19] (Figure 1C).

Factors affecting local coverage

Each fragment/read of HaloPlex target enrichment has a defined start site unlike target enrichment generated using shearing, which produces fragments with different start and end points. We therefore asked whether the position of restriction sites could influence coverage of target regions.

We found significant variability looking at raw coverage across gene loci within each sample, with read depth following a "square wave" pattern. For example, coverage across consecutive bases of the CEBPA locus varied by several fold (Figure 2A), with drops in coverage likely dictated by PCR amplification differences as well as number and size of amplicons. Some reads of our 100 bp paired-end sequencing did not reach the middle portion of the few large amplicons longer than 200bp (Figure 2B) due to positions of restriction sites used in the genome. Therefore, we investigated whether amplicon length correlated with coverage across the entire target region. Coverage of amplicons <100bp was variable, whilst amplicons longer than 200 bp showed a percentage of missed bases that increased proportionally with their length (Figure 2C). Unsurprisingly, we found that coverage at each base-pair position strongly correlated with the number of amplicons covering it (Figure 2D), suggesting that tiling more amplicons over a region rescued coverage gaps in long amplicons. This also explains why not all amplicons longer than 200 bp demonstrate a drop in coverage (Figure 2C), as this phenomenon was mainly limited to regions covered by single amplicons. Finally, we asked if coverage was influenced by length of exons rather than amplicons, and we found that this was not the case (Figure 2E), again suggesting that tiling regions of interest with multiple amplicons can overcome gaps of coverage within long amplicons. Our data show therefore that the regional drops in coverage of HaloPlex target enrichment are predictable based on amplicon length and tiling, and not influenced by the size of the region/feature of interest. These factors should be considered as part of HaloPlex target enrichment designs.

Detection of copy-number changes

We observed that coverage varied significantly between different base positions from the same sample, however coverage patterns appeared consistent between samples. In this context, we asked whether HaloPlex target enrichment data could identify copy number aberrations, as is the case for SureSelect target enrichment[18,19]. We normalized coverage of each sample for on-target mapped bases, and plotted average depth for all genes in our samples (Figure 3A). All samples showed read depths for X- and Y- chromosome genes consistent with patient gender, with females consistently showing a ~2 fold increase in coverage of X-linked genes (*BCOR* and *KDM6A*, also known as *UTX*) and no coverage of the Y-linked gene *UTY* (the Y homolog of *KDM6A*). Interestingly one male sample, PD19747a, showed a *BCOR* depth that was lower than other males in the cohort (black bar in Figure 3A). Coverage of all BCOR exons was significantly lower compared to the average of normal male samples (Figure 3B), suggesting this patient carries a *BCOR* deletion and this was indeed confirmed by quantitative PCR (Figure 3C). As sample PD17940a was previously shown to carry a *MLL* partial tandem duplication (PTD)[18], we checked coverage of MLL exons between 2 and 10 and found that most showed a higher coverage than normal samples (Figure 3D) consistent with a duplication of the region. We found another 2 patients showing the same pattern (PD17948a and PD17957a, Figure 3D), and went on to confirm the presence of MLL-PTDs by long-range PCR (Supplementary Figure S1A). Finally, one patient showed an amplification involving the *KRAS* locus (red bar in Figure 3A), which we confirmed by quantitative PCR (Figure 3E) and by CGH/SNP array (Supplementary Figure 1B).

Given that read depth of gene loci returned a linear estimate of the copy number of the locus, we next looked at the quantitative value of substitution calls, and to this end we analyzed 90 of the most polymorphic SNPs within our target region[27]. 84.6% of the heterozygous SNP calls were confined in a narrow allelic fraction window of 50+/-10% (Supplementary Figure S1).

Therefore, despite HaloPlex target enrichment returning variable coverage of different target regions, this variation is predictable, consistent across samples, and not significantly biased by PCR amplification. Depth of coverage retained quantitative value at the gene- and base-pair level and could identify copy number alterations with pathogenic and prognostic value.

Study controls

We next turned our attention to DNA sequence variants. First, we demonstrated that our algorithm identified likely oncogenic somatic variants and not inherited polymorphisms without the use of matched normal DNA. We did this by comparing the 16 variants called by our unmatched variant detection pipeline to matched post-chemotherapy DNA in 5 patients for whom this was also available (Figure 4A). 13/16 mutations were not present in the post-chemotherapy sample suggesting these were somatic mutations. Of three patients showing persistence of one oncogenic variant each, two were in complete hematological remission and one in partial remission with normal blood counts. Interestingly, the two variants with high allelic frequency in the post-chemotherapy sample were *DNMT3A* R882H substitutions, recently reported to persist in pre-leukemic cells after AML remission[28]. The other, a *TET2* nonsense mutation, showed a marked drop in allelic fraction consistent with incomplete molecular response. This shows that our pipeline can reliably identify somatic oncogenic events in unmatched samples, but underscores the limitation of using post-chemotherapy samples as matched controls in AML NGS studies.

Next, we confirmed that HaloPlex identifies real variants by looking at the 25 mutations found in 8 patients that were previously studied using SureSelect DNA pulldown[18]. These 25 variants included all 23 called by SureSelect[18], including those present at subclonal level (Figure 4B), showing a high reliability of HaloPlex calls. An additional two variants were missed by SureSelect, both FLT3-ITDs, which are notoriously hard to identify by targeted enrichment approaches[18,29] (and Papaemmanuil E., Wellcome Trust Sanger Institute, personal communication, July 2014). Additionally, and notwithstanding the fact that the allelic burden of indels is hard to assess reliably, the correlation between allelic fractions of variants from the two enrichment methods was good, indicating that HaloPlex has similar quantitative properties to SureSelect.

Caveman is a proprietary algorithm and thus we asked whether HaloPlex data would allow for reproducible results with other software. We compared Caveman substitution calls and allelic frequencies to those generated by SureCall (v1.1, Agilent Technologies). SureCall missed 23 of 61 substitutions detected by Caveman, including known oncogenic ones. All missed variants had an allelic burden <15%, suggesting that SureCall performs less well in detecting subclonal variants (Figure 4C), although this may be surmountable by newer versions of the software. Nevertheless, for variants detected by both algorithms, the correlation between allelic frequencies was near perfect (Figure 4C).

Because *NPM1* and *FLT3* indels are frequent variants and key prognostic indicators in AML-NK, we specifically evaluated the performance of the open source software Pindel in detecting these variants as compared to PCR-based genotyping. NGS and PCR were concordant on the *FLT3*-ITD status in 36/40 evaluable samples (Figure 4D). In three cases, the ITD was found by PCR but not by NGS, and these were found to be large ITDs that may have not been amplified or mapped by BWA. In one case, a short ITD was only found by NGS, and we presume that it represented a subclonal event that PCR could not detect/discriminate. Conversely, Pindel only reported *NPM1* C-terminal indels in 7/26 cases shown to carry the mutation by PCR. Looking at *NPM1* exon 12, we found a marked coverage drop of position chr 5:170837554, i.e. few bp away from the insertion site of most *NPM1* indels. The reason for this was that all but one amplicons covering the region were >200bp long, and thus their midpoints were beyond the reach of either 100bp paired-end read (Figure 4E, bottom panel, arrowhead). This design pitfall also caused *NPM1* indels to be close to the end of the reads, and thus discarded by Pindel and under-reported. Since only one amplicon covered the mutation in a position amenable to sequencing (Figure 4E, asterisk), *NPM1* variants were only called in samples where this amplicon was sequenced with enough coverage (p=0.01). Nevertheless, *NPM1* indels from all amplicons were mapped by BWA, and visual inspection of the reads did allow their identification in all mutated cases (Figure 4F). To confirm that a short read length relative to the size of the amplicons covering the mutations was the reason for the poor detection of *NPM1* indels, we re-sequenced HaloPlex libraries for 33 samples using MiSeq (Illumina) with a 150bp paired-end protocol. As expected, coverage of the *NPM1* indel region was much higher (Figure 4E, green line), and all indels were called by Pindel (Figure 4G) with 100% sensitivity and specificity (Figure 4H). The presence of *NPM1* mutations was further validated by capillary sequencing in all but one sample for which we did not have additional DNA (Supplementary Table S2).

Overall, 115 of 119 variants identified by HaloPlex were studied by PCR and/or MiSeq. Of the 103 that passed quality control, 96 were confirmed. Importantly, we could validate both clonal and subclonal variants indicating that HaloPlex can enrich target DNA allowing identification of variants across a range of allelic frequencies. Of the remaining 7 variants, 4 were false positives and 3 were sublclonal indels below the detection threshold of standard PCR (Supplementary Table S2).

Gene mutations

We reported 119 variants in 20 genes in 41 out of 42 samples, with a median of 3 variants per sample (Figure 5A and Supplementary Table S2). The most frequently mutated gene was *NPM1* (62%), followed by *FLT3* (50%), *DNMT3A* and *TET2* (33% and 29%, respectively). As previously described, there was a positive correlation between *NPM1* mutations and *FLT3* (p=0.008, Fisher's exact). We also observed a tendency towards correlation between NPM1 and *DNMT3A*, and towards mutual exclusivity between *TET2* and *IDH1/2* mutations.
Two or more *FLT3-ITD* alleles were identified in 3/14 samples. Allelic frequency couldn't be reliably estimated in these indels making it impossible to determine if they occurred in the same cells (compound heterozygosity), or in different subclones of the tumor (convergent evolution). Similarly, two *TET2* mutated alleles were found in 2/10 patients, reflecting a heterogeneous and evolving mutational pattern. Lastly, we annotated a p.S1018Y missense variant in *UTY*, a paralog of *KDM6A* not implicated in AML before. The variant was previously reported as somatic in a gastrointestinal cancer invoking a possible pathogenic role in AML.

While allelic frequency can be used to assess the subclonal structure of tumors[25], most of our variants were represented by indels and this precluded such analysis. Nevertheless, in two patients from whom paired diagnosis-relapse samples were available, we showed loss of a subclonal *TET2* mutation in PD17932, and loss of a biallelic *FLT3*-ITD and a subclonal *FLT3* N676K substitution in PD17936 at molecular relapse (Figure 5B). This confirms that the subclonal structure of AML can develop through continuous acquisition of subclones with new driver mutations and loss of others, in a pattern consistent with branching evolution and differential sensitivity to chemotherapy as has been shown by others[28,30].

Discussion

Dramatic advances in defining the somatic genome of AML[4] have defined the major mutational drivers of this disease[31]. As a result, the field is ready for targeted follow-up studies aimed at better characterizing the prevalence, prognostic value and pathogenic role of these genetic lesions in large cohorts of patients. Indeed, information on mutated genes is making its way into new prognostic models[5], especially in cases without recurrent karyotype rearrangements[12]. In this paper we describe a rapid, robust and high-throughput approach for the characterization of gene mutations and copy number changes in AML samples using HaloPlex target enrichment followed by NGS and standard bioinformatic analysis.

We showed that amplicon tiling and read length relative to amplicon length are the two most important parameters affecting coverage of target regions. In HaloPlex, the position of restriction sites limits the extent to which sequencing start sites and amplicon lengths can be customized in the target enrichment design. Therefore, depending on tiling and amplicon length, adjacent genomic regions can show variable coverage. While the automated HaloPlex design tool works well in general, if mutational hotspots are anticipated it is advisable that these positions are checked manually to ensure they will be adequately covered. We showed that variability of coverage of HaloPlex data is reproducible and consistent across samples. Normalized coverage of each gene locus correlated with its copy number status, relative to the other samples in the cohort. This enabled us to identify small copy number changes without the need for matched normal DNA, as exemplified by the identification of three cases of MLL-PTDs. Furthermore, we report the novel finding that *KRAS* can be amplified and *BCOR* deleted in AML, reflecting the power of NGS techniques to interrogate tumor genomes in a high-throughput fashion. Clinical follow-up was not available for our patients, and future studies will define the recurrence rate and prognostic role of these events in AML. Compared to genome-wide CGH arrays, we could only infer copy-number of regions targeted in our design. Nevertheless, in the future this property could be harnessed for the capture and study of a large number of polymorphic SNPs evenly spaced across the genome to allow the identification of whole-genome copy-number and loss-of-heterozygosity changes.

Our study had a positive predictive value of 96% for the identification of recurrent mutations in AML. Its ability to report indels, a frequent event in AML, was especially good. Large genomic insertions such as *MLL*-PTDs were identified by copy-number profile of individual exons. While *NPM1* indels were initially under-reported by 100bp reads because of a design flaw, employing longer reads allowed us to achieve 100% accuracy. We also found good efficiency for *FLT3*-ITDs, as we identified 14/17 ITD samples. This was facilitated by targeting both *FLT3* exons and introns around the breakpoints, although the allelic fraction of such events was lower than expected for driver mutations. Therefore, we could only capture and/or map a fraction of the mutated DNA molecules, and our detection sensitivity could have been lower had we not sequenced so deeply. Capture, mapping and quantification of *FLT3*-ITD alleles is a major challenge that will likely require bespoke targeting and bioinformatic approaches, especially for longer ITDs that were missed in our study[29,32]. On the other hand, we suggest that deep sequencing can provide increased sensitivity for short and subclonal ITDs that may be easily missed by conventional PCR, leading to incorrect prognostic characterization of the patient. Indeed, in our study we identified 3 subclonal *NPM1* and *FLT3* indels that could not be confirmed by PCR followed by agarose gel electrophoresis or capillary sequencing. We believe these were true positive results, and the fact that other subclonal variants were validated in our study suggests their veracity. Subclonality in AML is increasingly recognized as a biological event with clinical implications[28,30,33]. HaloPlex target enrichment led to the identification and validation of a number of subclonal variants, and loss/gain of variants at AML relapse. This has the potential to inform on the order of acquisition of such variants during pre-clinical stages of leukemia development and suggests that future, larger studies may be able to inform which variants are associated with better response to chemotherapy and which ones are most likely to confer chemoresistance. For example, our finding that *TET2* mutations can be lost at relapse confirms that mutations in this gene can be late[34] as well as early[35] events in AML. Also, further studies will be required to assess the prognostic value of *DNMT3A* R882H persistence at morphological remission, and whether this variant should be used for assessment of minimal residual disease (MRD).

We anticipate that NGS technologies will soon be used for a combined gene sequencing and copy number analysis of tumors, thus providing a one-stop diagnostic platform that has the potential to enhance current analysis relying on the integration of karyotype, FISH, PCR and RT-PCR data. Future studies with large numbers of patients and longitudinal follow-up will establish the diagnostic and prognostic value of recurrent abnormalities, and in our paper we show that HaloPlex target enrichment can provide a solid platform for this exercise.

Supplementary Information

Supplementary information is available at Haematologica's website.

Authorships and disclosures

NB analysed data and wrote the manuscript. NM, NP, MAQ, VS performed research. TM, GG, BH, NC, CG, EP, PJC, IV analysed data. JC and PC provided samples and performed research. CC and JIOC provided samples. GSV designed research, analysed data and wrote the manuscript.

GSV is a consultant to the pharmaceutical companies KYMAB and Celgene. BH is an employee of Agilent Technologies. PJC holds equity in, and is a paid consultant for, 14M Genomics Limited.
All authors declare no conflict of interest.

References

1. Bennett JM, Catovsky D, Daniel MT, Flandrin G, Galton DA, Gralnick HR, et al. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *British J Haematol*. 1976;33(4):451-8.
2. Jaffe E, Harris N, Stein H, Vardiman J. Pathology and genetics of tumours of hematopoietic and lymphoid tissues. Lyon, France: IARC Press; 2001.
3. Döhner H, Estey EH, Amadori S, Appelbaum FR, Büchner T, Burnett AK, et al. Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood*. 2010;115(3):453-74.
4. Cancer Genome Atlas Research N. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059-74.
5. Patel JP, Gonen M, Figueroa ME, Fernandez H, Sun Z, Racevskis J, et al. Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *N Engl J Med*. 2012;366(12):1079-89.
6. Vardiman JW, Thiele J, Arber DA, Brunning RD, Borowitz MJ, Porwit A, et al. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood*. 2009;114(5):937-51.
7. Falini B, Martelli MP, Bolli N, Sportoletti P, Liso A, Tiacci E, et al. Acute myeloid leukemia with mutated nucleophosmin (NPM1): is it a distinct entity? *Blood*. 2011;117(4):1109-20.
8. Dufour A, Schneider F, Metzeler KH, Hoster E, Schneider S, Zellmeier E, et al. Acute myeloid leukemia with biallelic CEBPA gene mutations and normal karyotype represents a distinct genetic entity associated with a favorable clinical outcome. *J Clin Oncol*. 2010;28(4):570-7.
9. Döhner K, Schlenk RF, Habdank M, Scholl C, Rücker FG, Corbacioglu A, et al. Mutant nucleophosmin (NPM1) predicts favorable prognosis in younger adults with acute myeloid leukemia and normal cytogenetics: interaction with other gene mutations. *Blood*. 2005;106(12):3740-6.
10. Renneville A, Boissel N, Gachard N, Naguib D, Bastard C, de Botton S, et al. The favorable impact of CEBPA mutations in patients with acute myeloid leukemia is only observed in the absence of associated cytogenetic abnormalities and FLT3 internal duplication. *Blood*. 2009;113(21):5090-3.
11. Cairoli R, Beghini A, Grillo G, Nadali G, Elice F, Ripamonti CB, et al. Prognostic impact of c-KIT mutations in core binding factor leukemias: an Italian retrospective study. *Blood*. 2006;107(9):3463-8.
12. Schlenk RF, Döhner K, Krauter J, Fröhling S, Corbacioglu A, Bullinger L, et al. Mutations and treatment outcome in cytogenetically normal acute myeloid leukemia. *N Engl J Med*. 2008;358(18):1909-18.
13. Wang YY, Zhao LJ, Wu CF, Liu P, Shi L, Liang Y, et al. C-KIT mutation cooperates with full-length AML1-ETO to induce acute myeloid leukemia in mice. *Proc Natl Acad Sci U S A*. 2011;108(6):2450-5.
14. Chevalier N, Solari ML, Becker H, Pantic M, Gartner F, Maul-Pavicic A, et al. Robust in vivo differentiation of t(8;21)-positive acute myeloid leukemia blasts

to neutrophilic granulocytes induced by treatment with dasatinib. *Leukemia*. 2010;24(10):1779-81.

15. Leung AY, Man CH, Kwong YL. FLT3 inhibition: a moving and evolving target in acute myeloid leukaemia. *Leukemia*. 2013;27(2):260-8.

16. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456(7218):53-9.

17. Ley TJ, Mardis ER, Ding L, Fulton B, Mclellan MD, Chen K, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*. 2008;456(7218):66-72.

18. Conte N, Varela I, Grove C, Manes N, Yusa K, Moreno T, et al. Detailed molecular characterisation of acute myeloid leukaemia with a normal karyotype using targeted DNA capture. *Leukemia*. 2013;27(9):1820-5.

19. Papaemmanuil E, Gerstung M, Malcovati L, Tauro S, Gundem G, Van Loo P, et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*. 2013;122(22):3616-27; quiz 99.

20. Haferlach T, Nagata Y, Grossmann V, Okuno Y, Bacher U, Nagae G, et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia*. 2014;28(2):241-7.

21. Berglund EC, Lindqvist CM, Hayat S, Overnas E, Henriksson N, Nordlund J, et al. Accurate detection of subclonal single nucleotide variants in whole genome amplified and pooled cancer samples using HaloPlex target enrichment. *BMC Genomics*. 2013;14:856.

22. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26(5):589-95.

23. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-2.

24. R Core T. R: A language and environment for statistical computing: R Foundation for Statistical Computing; 2014.

25. Bolli N, Avet-Loiseau H, Wedge DC, Van Loo P, Alexandrov LB, Martincorena I, et al. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun*. 2014;5:2997.

26. Ye K, Schulz MH, 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*. 2009;25(21):2865-71.

27. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65.

28. Shlush LI, Zandi S, Mitchell A, Chen WC, Brandwein JM, Gupta V, et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature.* 2014;506(7488):328-33.

29. Spencer DH, Abel HJ, Lockwood CM, Payton JE, Szankasi P, Kelley TW, et al. Detection of FLT3 internal tandem duplication in targeted, short-read-length, next-generation sequencing data. *J Mol Diagn*. 2013;15(1):81-93.

30. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*. 2012;481(7382):506-10.

31. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 2014;505(7484):495-501.
32. Luthra R, Patel KP, Reddy NG, Haghshenas V, Routbort MJ, Harmon MA, et al. Next-generation sequencing-based multigene mutational screening for acute myeloid leukemia using MiSeq: applicability for diagnostics and disease monitoring. *Haematologica*. 2014;99(3):465-73.
33. Klco JM, Spencer DH, Miller CA, Griffith M. Functional Heterogeneity of Genetically Defined Subclones in Acute Myeloid Leukemia. *Cancer Cell*. 2014;25(3):379-92.
34. Schaub FX, Looser R, Li S, Hao-Shen H, Lehmann T, Tichelli A, et al. Clonal analysis of TET2 and JAK2 mutations suggests that TET2 can be a late event in the progression of myeloproliferative neoplasms. *Blood*. 2010;115(10):2003-7.
35. Busque L, Patel JP, Figueroa ME, Vasanthakumar A, Provost S, Hamilou Z, et al. Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nat Genet*. 2012;44(11):1179-81.

Table 1

Table 1. Genes and transcripts used for the targetd enrichment study

gene	NCBI RefSeq transcripts used (NCBI RefSeq ID)			
ASXL1	NM_001164603	NM_015338		
BCOR	NM_017745	NM_001123385	NM_001123384	NM_001123383
CBL	NM_005188			
CEBPA	NM_004364			
CSF1R	NM_005211			
DNMT3A	NM_022552	NM_153759	NM_175629	NM_175630
FLT3	NM_004119			
EZH2	NM_152998	NM_004456		
IDH1	NM_005896			
IDH2	NM_002168			
JAK2	NM_004972			
KIT	NM_001093772	NM_000222		
KDM6A	NM_021140			
KRAS	NM_004985	NM_033360		
MLL	NM_005933			
NF1	NM_000267	NM_001042492	NM_001128147	
NPM1	NM_002520	NM_001037738	NM_199185	
NRAS	NM_002524			
PTPN11	NM_002834			
RUNX1	NM_001001890	NM_001754	NM_001122607	
SF3B1	NM_012433	NM_001005526		
TET2	NM_001127208	NM_017628		
UTY	NM_007125	NM_182659	NM_182660	
WT1	NM_024426	NM_000378	NM_024424	NM_024425

Figure Legends

Legend to Figure 1

A) Stacked bar chart showing the total sequence output in gigabases (Gb, left Y axis) per sample: grey, bases unmapped; yellow, bases mapped off target; blue, bases mapped on target. Samples plexed and sequenced in different HiSeq lanes are segregated by the dashed vertical line. Lines indicate the percentage of target covered at >30X (grey) and >1000X (red) – right Y axis. A Pearson's test shows the correlation between sequence output and percentage of target covered at the above percentages.

B) Bar chart displaying the absolute coverage of each gene in the study, calculated as the mean coverage of that gene in all samples. Error bars represent standard deviation.

C) Bar chart showing, for each gene, the average percentage of the coding region covered at >30X in all samples (i.e., the minimum coverage used for variant calling). Error bars represent standard deviation. *UTY*, the Y chromosome homologue of *KDM6A* (*UTX*), was only covered by males in the study.

Legend to Figure 2

A) Line graph showing base-by-base average normalized coverage of the CEBPA gene footprint. The CEBPA coding region is shown by a thick blue bar, and the UTR regions by a thin blue bar. The horizontal red line highlights positions not covered (0 bp coverage). Below, the amplicons from HaloPlex design are shown in green.

B) Boxplot showing the distribution of amplicon size in the design. The central line is the median, and the box includes values between the first and third quartile.

C) Plot showing, for each amplicon in the design (blue dots), the relation between its length (X axis) and the percentage of its bases covered at >30X. Note that the coverage drops in a fraction of amplicons longer than 200 bp (i.e. the combined length of the paired-end sequencing protocol), suggesting that the middle region of such large amplicons wasn't covered by either of the 100bp paired-end reads, and no other amplicons were overlapping on that region.

D) Boxplot showing, for each base in the design, the positive correlation between the number of amplicons covering it (X axis) and its average coverage in all samples (Y axis). The central line is the median, and the box includes values between the first and third quartile.

E) Plot showing that coverage (Y axis) of individual exons (blue dots) in the design is independent of their length (X axis, in log scale).

Legend to Figure 3

A) For each gene in the study, the normalized average depth of coverage (Y axis) is plotted individually for all patients (X axis) on a linear arbitrary scale.

B) PD17947a (black bar) shows a *BCOR* deletion involving all exons, whose coverage is lower than the average coverage of three normal male samples (white bar). The residual signal from all BCOR exons in PD17947a likely reflects that the deletion is subclonal, although a percentage of normal cells admixed in the tumor sample must also be taken into account. In the particular case of BCOR exon 5 the ratio between WT samples and PD17947a is different compared to neighboring exons, but this must be interpreted with caution. This exon showed the lowest coverage of all BCOR exons and a high number of homologous regions that could lead to mismapping and make it insensitive to copy number changes.

C) Quantitative PCR on genomic DNA shows lower levels of BCOR exons 1 and 4 in PD17947a (black, solid and dashed bar respectively) compared to a control male sample (PD17948a, yellow).

D) PD17940a, PD17948a, and PD17957a (green, yellow and blue bars, respectively) show a *MLL* partial tandem duplication as shown by increased coverage of most exons between 1 and 10 compared to the average of 5 normal samples (white bars).

E) PD17946a (red bar) shows a *KRAS* amplification confirmed by qPCR on genomic DNA compared to a control sample in the study.

Legend to Figure 4

A) For 5 patients in the study where a post-chemotherapy sample was available, the somatic status of 16 variants was checked. Y axis represents the raw allelic fraction of the variant, and X axis represents individual variants, clustered by patient, in the tumor (blue bar) and remission sample (yellow). Note that DNMT3A R882H persists at a similar allelic fraction in post-chemotherapy samples, independent of the remission status of the patient. Also, the persistence of a low-level TET2 p.L1119* nonsense mutation suggests that patient had a partial molecular response to treatment.

B) For 25 variants, validation data was available from a previous study performed with SureSelect target enrichment. The plot shows the allelic fraction of variants in the HaloPlex study (X axis) and that of the Sureselect study (Y axis). Variants are represented as solid circles (substitutions) or open triangles (indels), and are blue if shared between the two studies and yellow if only reported by the HaloPlex study. The plot shows good correlation of allelic fraction between the two studies.

C) For the 61 substitutions in the study, two different algorithms were compared (CaVEMan and SureCall). Shared variants are in blue, variants missed by SureCall are in yellow. For the shared variants, the correlation between allelic fractions is near perfect.

D) For 40 samples in the study (X axis), FLT3-ITDs are plotted by length (Y axis, value=0 if no ITD present). Variants confirmed by both Pindel and PCR are blue circles, those only found by PCR are yellow upwards triangles, and those only found by NGS are yellow downwards triangles.

E) Top: the black line shows base-by-base average normalized coverage of the NPM1 gene locus. The NPM1 coding region is shown by a thick blue bar, and the UTR regions by a thin blue bar. Below, predicted amplicons from HaloPlex design are shown in green.

Bottom: zoomed-in view of NPM1 exon 12. Coverage by 100 bp reads is shown as a black line and leaves a 1-bp gap close to the NPM1c+ insertion site. The amplicon closer to NPM1c+ mutations and allowing their identification is highlighted by an asterisk, while the amplicons where such mutations were missed are marked by an arrowhead. When 150 bp reads are employed (green line), coverage of the middle region increases.

F) Aligned reads from a 100 bp sample where the mutation was missed because of short read length leaving a gap in coverage (arrow). Yellow: reverse reads. Blue: forward reads. The presence of the indel must be deducted by the BWA calls (red boxes, arrowheads).

G) Aligned reads from a 150 bp sample where the mutation was identified. Yellow: reverse reads. Blue: forward reads. The insertion is identified by Pindel as green boxes (arrow).

H) Stacked bar chart showing the increase in sensitivity of Pindel for detection of NPM1 indels with a 150bp protocol.

Legend to Figure 5

A) Table highlighting relevant genetic alterations and recurrently mutated genes in the study. Samples are represented in columns. Recurrently mutated genes are color-coded for missense (blue), nonsense (red), splice-site (green) substitutions, and indels (yellow). In case of multiple mutations of the same class in a gene in a patient, a black contour is drawn. If two mutations of different class are present, the box is filled by two triangles. For each gene, the number of patients harboring at least one non-silent mutation is provided in the 'TOTAL' column.

B) For two patients where diagnosis (D) and relapse (R) samples where available, the mutational spectrum is provided to show evolution of the cancer over time.

Figure 1

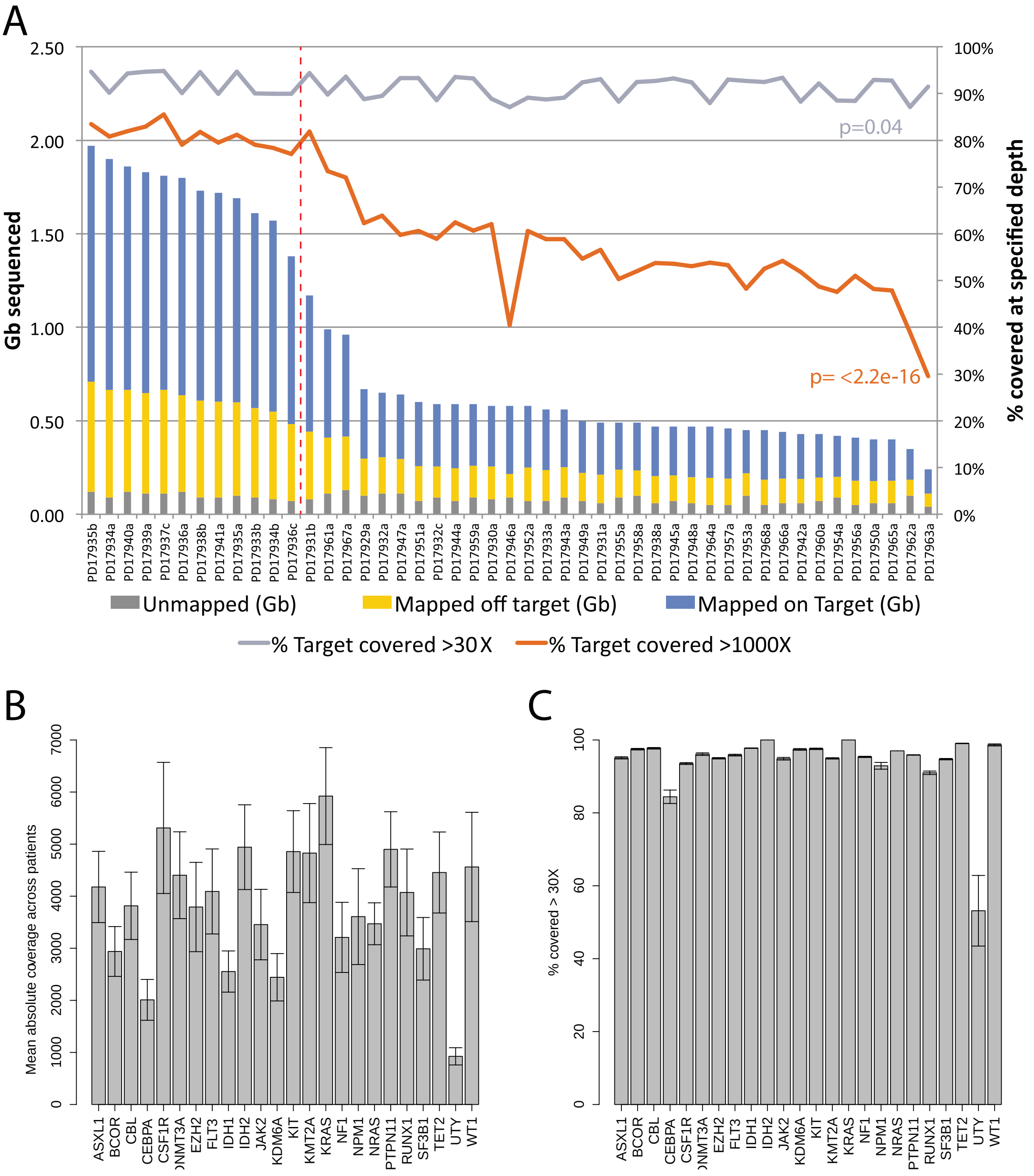

A
Gb sequenced
2.50
2.00
1.50
1.00
0.50
0.00
% covered at specified depth
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
p=0.04
p= <2.2e-16
PD17935b
PD17934a
PD17940a
PD17939a
PD17937c
PD17936a
PD17938b
PD17941a
PD17935a
PD17933b
PD17934b
PD17936c
PD17931b
PD17961a
PD17967a
PD17929a
PD17932a
PD17947a
PD17951a
PD17932c
PD17944a
PD17959a
PD17930a
PD17946a
PD17952a
PD17933a
PD17943a
PD17949a
PD17931a
PD17955a
PD17958a
PD17938a
PD17945a
PD17948a
PD17964a
PD17957a
PD17953a
PD17968a
PD17966a
PD17942a
PD17960a
PD17954a
PD17956a
PD17950a
PD17965a
PD17962a
PD17963a
Unmapped (Gb)
Mapped off target (Gb)
Mapped on Target (Gb)
% Target covered >30X
% Target covered >1000X
B
Mean absolute coverage across patients
0
1000
2000
3000
4000
5000
6000
7000
ASXL1
BCOR
CBL
CEBPA
CSF1R
NMT3A
EZH2
FLT3
IDH1
IDH2
JAK2
KDM6A
KIT
KMT2A
KRAS
NF1
NPM1
NRAS
PTPN11
RUNX1
SF3B1
TET2
UTY
WT1
C
% covered > 30X
0
20
40
60
80
100
ASXL1
BCOR
CBL
CEBPA
CSF1R
NMT3A
EZH2
FLT3
IDH1
IDH2
JAK2
KDM6A
KIT
KMT2A
KRAS
NF1
NPM1
NRAS
PTPN11
RUNX1
SF3B1
TET2
UTY
WT1

Figure 2

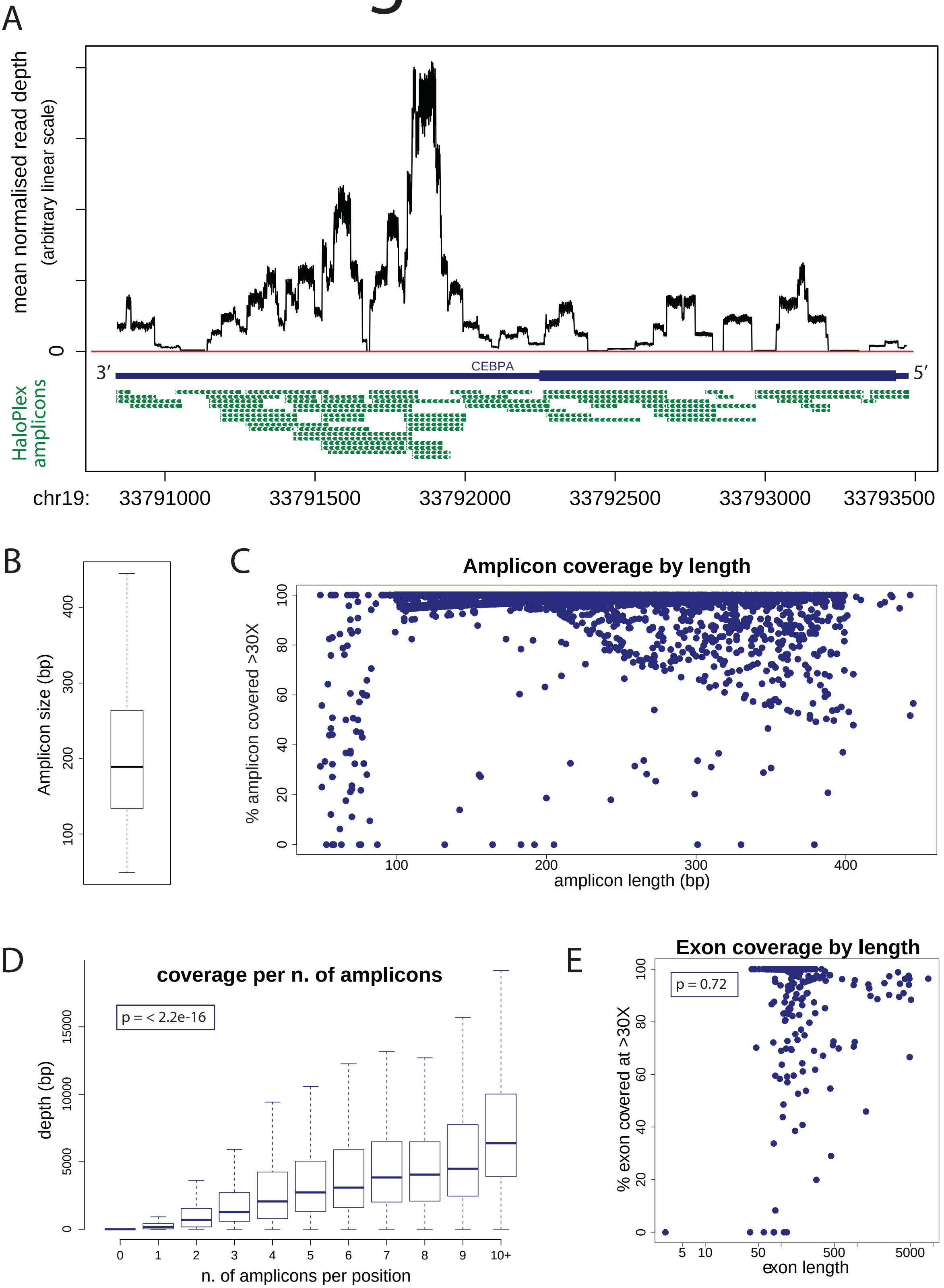
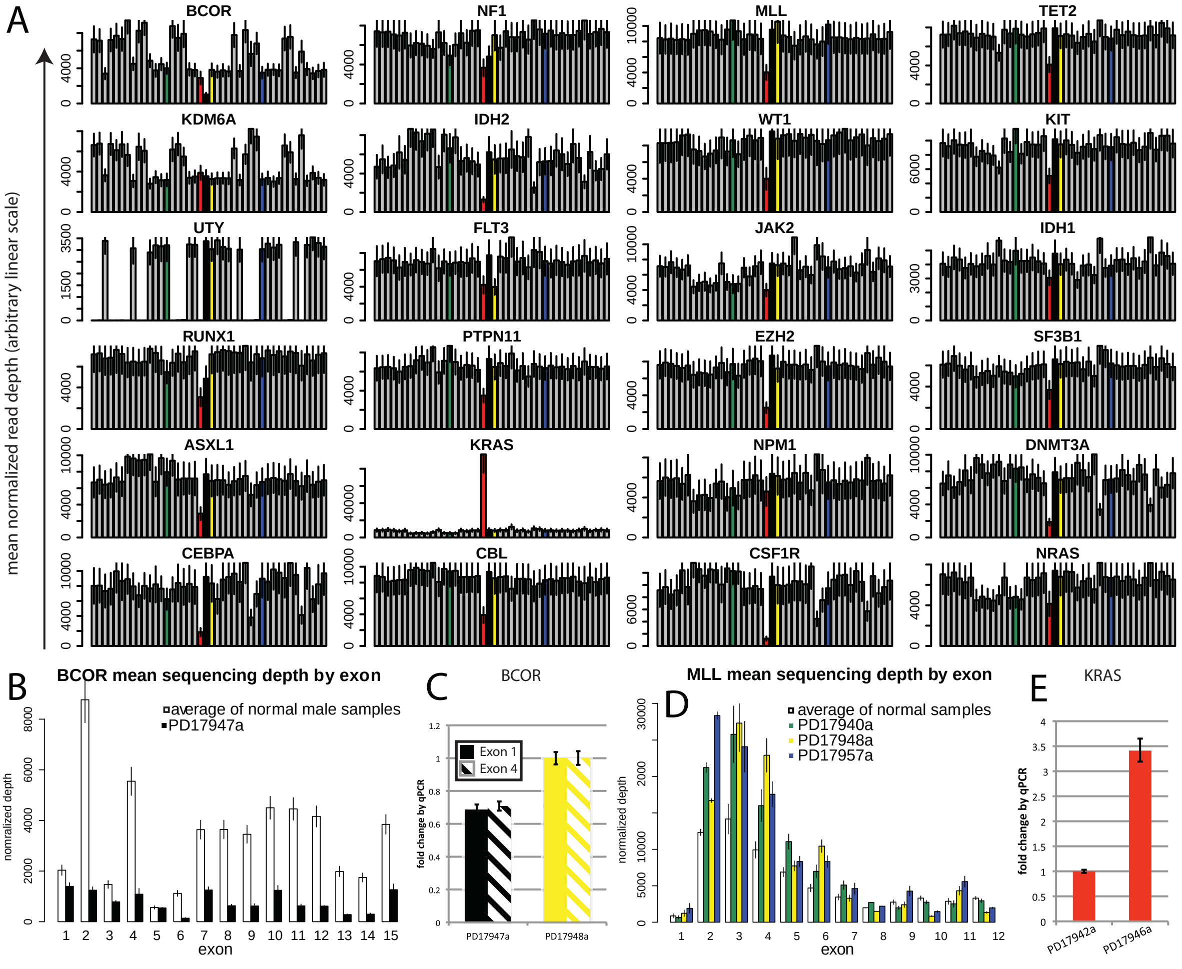

A
mean normalised read depth
(arbitrary linear scale)
0
3′
CEBPA
5′
HaloPlex amplicons
chr19:
33791000
33791500
33792000
33792500
33793000
33793500
B
Amplicon size (bp)
100
200
300
400
C
Amplicon coverage by length
% amplicon covered >30X
0
20
40
60
80
100
amplicon length (bp)
100
200
300
400
D
coverage per n. of amplicons
p = < 2.2e-16
depth (bp)
0
5000
10000
15000
0
1
2
3
4
5
6
7
8
9
10+
n. of amplicons per position
E
Exon coverage by length
p = 0.72
% exon covered at >30X
0
20
40
60
80
100
5
10
50
500
5000
exon length

Figure 3



A
mean normalized read depth (arbitrary linear scale)
BCOR
NF1
MLL
TET2
KDM6A
IDH2
WT1
KIT
UTY
FLT3
JAK2
IDH1
RUNX1
PTPN11
EZH2
SF3B1
ASXL1
KRAS
NPM1
DNMT3A
CEBPA
CBL
CSF1R
NRAS
0
4000
10000
1500
3500
3000
6000
40000
B
BCOR mean sequencing depth by exon
average of normal male samples
PD17947a
nomralized depth
0
2000
4000
6000
8000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
exon
C
BCOR
Exon 1
Exon 4
fold change by qPCR
0
0.2
0.4
0.6
0.8
1
1.2
PD17947a
PD17948a
MLL mean sequencing depth by exon
D
average of normal samples
PD17940a
PD17948a
PD17957a
normalized depth
0
5000
10000
2000
30000
1 2 3 4 5 6 7 8 9 10 11 12
exon
E
KRAS
fold change by qPCR
0
0.5
1
1.5
2
2.5
3
3.5
4
PD17942a
PD17946a

Figure 4

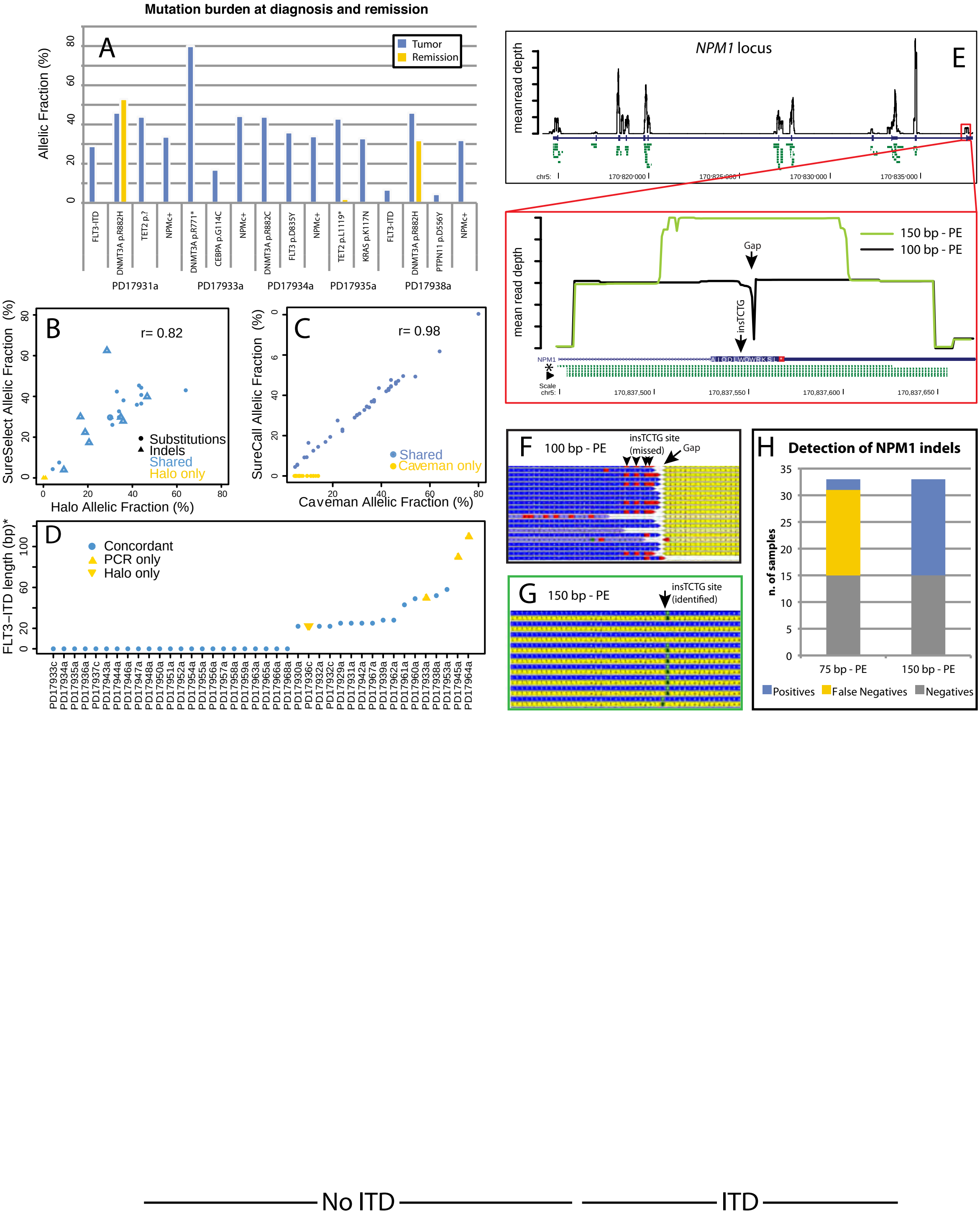
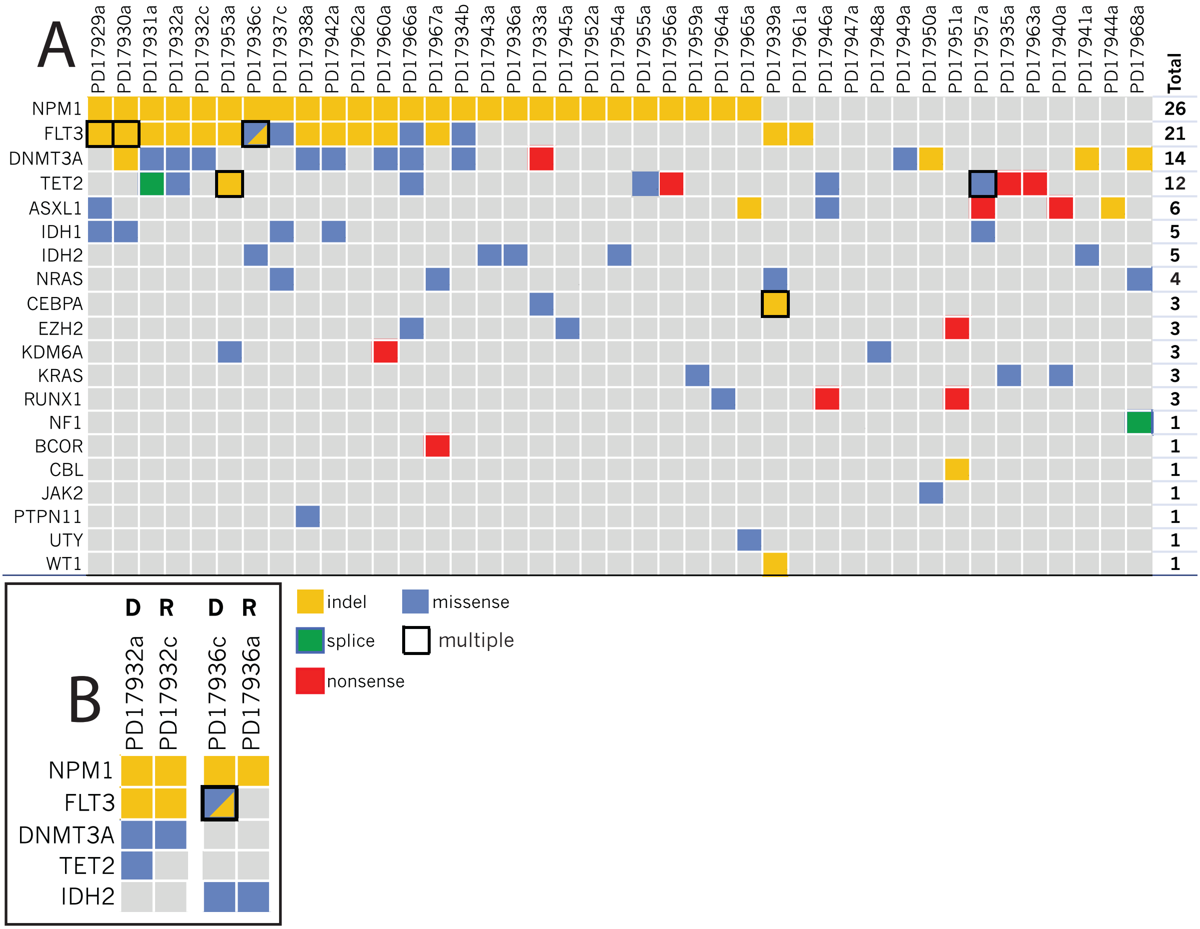

Mutation burden at diagnosis and remission
A
Allelic Fraction (%)
0
20
40
60
80
Tumor
Remission
FLT3-ITD
DNMT3A p.R882H
TET2 p.?
NPMc+
DNMT3A p.R771*
CEBPA p.G114C
NPMc+
DNMT3A p.R882C
FLT3 p.D835Y
NPMc+
TET2 p.L1119*
KRAS p.K117N
FLT3-ITD
DNMT3A p.R882H
PTPN11 p.D556Y
NPMc+
PD17931a
PD17933a
PD17934a
PD17935a
PD17938a
B
r= 0.82
SureSelect Allelic Fraction (%)
Halo Allelic Fraction (%)
Substitutions
Indels
Shared
Halo only
0
20
40
60
80
C
r= 0.98
SureCall Allelic Fraction (%)
Caveman Allelic Fraction (%)
Shared
Caveman only
D
FLT3−ITD length (bp)*
Concordant
PCR only
Halo only
0
20
60
100
PD17933c
PD17934a
PD17935a
PD17936a
PD17937c
PD17943a
PD17944a
PD17946a
PD17947a
PD17948a
PD17950a
PD17951a
PD17952a
PD17954a
PD17955a
PD17956a
PD17957a
PD17958a
PD17959a
PD17963a
PD17965a
PD17966a
PD17968a
PD17930a
PD17936c
PD17932a
PD17932c
PD17929a
PD17931a
PD17942a
PD17967a
PD17939a
PD17962a
PD17961a
PD17960a
PD17933a
PD17938a
PD17953a
PD17945a
PD17964a
No ITD
ITD
E
NPM1 locus
meanread depth
chr5:
170·820·000
170·825·000
170·830·000
170·835·000
mean read depth
150 bp - PE
100 bp - PE
Gap
insTCTG
NPM1
Scale
chr5:
170,837,500
170,837,550
170,837,600
170,837,650
F
100 bp - PE
insTCTG site (missed)
Gap
G
150 bp - PE
insTCTG site (identified)
H
Detection of NPM1 indels
n. of samples
0
5
10
15
20
25
30
35
75 bp - PE
150 bp - PE
Positives
False Negatives
Negatives

Figure 5


A
PD17929a
PD17930a
PD17931a
PD17932a
PD17932c
PD17953a
PD17936c
PD17937c
PD17938a
PD17942a
PD17962a
PD17960a
PD17966a
PD17967a
PD17934b
PD17943a
PD17936a
PD17933a
PD17945a
PD17952a
PD17954a
PD17955a
PD17956a
PD17959a
PD17964a
PD17965a
PD17939a
PD17961a
PD17946a
PD17947a
PD17948a
PD17949a
PD17950a
PD17951a
PD17957a
PD17935a
PD17963a
PD17940a
PD17941a
PD17944a
PD17968a
Total
NPM1 26
FLT3 21
DNMT3A 14
TET2 12
ASXL1 6
IDH1 5
IDH2 5
NRAS 4
CEBPA 3
EZH2 3
KDM6A 3
KRAS 3
RUNX1 3
NF1 1
BCOR 1
CBL 1
JAK2 1
PTPN11 1
UTY 1
WT1 1
B
D
R
D
R
PD17932a
PD17932c
PD17936c
PD17936a
NPM1
FLT3
DNMT3A
TET2
IDH2
indel
missense
splice
multiple
nonsense

Supplementary Methods

Sequence variant detection and filtering criteria

Base substitutions and small insertions or deletions were identified by comparison of 42 MDS samples against unmatched normal samples using established bioinformatics algorithms[1-3]. To account for the absence of matched control a bespoke variant selection pipeline was developed. Each putative variant was annotated using the following resources:

1. Known constitutional polymorphisms using known human variation databases, Ensembl GRCh37.5, 1000 genomes release 2.2.2 and ESP6500[4 5].
2. Known somatic variation in myeloid and other common malignancies as reported in COSMIC v67[6].
3. Exome or whole genome sequencing data derived from 317 constitutional DNA samples analyzed in CGP (CGP normal panel).
4. Sequence context 5' and 3' to the reported sequence change highlighting regions of homopolymer sequences that are prone to PCR slippage and artifacts altering the last base of the homopolymer or inserting the same base as the homopolymer at +1, +2 of the track.
5. Variant specific metrics to include protein annotation, sequence depth and % of reads reporting the variant allele.

To enrich for high-confidence somatic variants that impact on protein function further filtering was conducted using the following criteria:

1. Removal of all variants with a predicted effect of a silent amino acid change on all transcripts corresponding to each gene.
2. Removal of known polymorphisms present in either of the human variation databases at a population frequency > 0.0014 (reflecting the population incidence of myeloid disease and potentially rare variants that could be associated with myeloid malignancies) unless variant is present as confirmed somatic mutation in COSMIC.
3. Removal of known polymorphisms present in human variation databases at a population frequency < 0.0014 and also represented in the extended normal CGP panel, available form in house exome and whole genome sequencing projects.
4. Retention of all variants present in human variation databases at a population frequency < 0.0014 and also present in COSMIC as confirmed somatic in Haematopoietic tissue.
5. Removal of all sequence variants that were represented in at least 2 normal individuals in the CGP normal panel with a minimum variant allele proportion of 10%.
6. Removal of variants present within regions prone to sequence context specific artifacts, including regions of high depth, enriched for reads of low mapping quality that harbor multiple mismatches.
7. Removal of all 1bp insertions or deletions present adjacent to regions of more than 5 homopolymer bases (i.e insA adjacent to AAAAA) and a variant allele proportion of < 12% and evidence of occurrence in CGP normal panel;

Once low confidence or likely polymorphisms were removed from the dataset, likely oncogenic were annotated and selected for the study among the shortlist of high confidence variants in accordance to prior evidence in the literature. To reflect the confidence that one would use these as diagnostic biomarkers in the clinic, variants were annotated conservatively, so that we only reported known oncogenic variants previously reported in the literature, or novel variants that cluster with known somatic variants in cancer driver genes, or truncating variants (nonsense mutations, essential splice mutations or frameshift indels) in genes implicated in myeloid malignancies through acquisition of loss of function mutations.

Validation

Copy number alterations of KRAS and BCOR were validated on genomic DNA with SYBRgreen quantitative PCR using the ACTB gene as endogenous control, and applying the $\Delta\Delta$CT method to perform a relative quantification[7]. Furthermore, the copy number pattern identified by NGS in sample PD17946a was validated using the Agilent SurePrint G3 ISCA CGH+SNP Microarray. MLL-PTDs were validated by long range PCR as described in[8]. FLT3-ITDs were assessed on genomic DNA by PCR followed by either agarose gel electrophoresis or Bioanalyzer using a high sensitivity analysis kit (Agilent Technologies) for 40 samples. NPM1 exon 12 mutations were validated in 33 samples using genomic DNA PCR followed by capillary sequencing. All primer sequences are provided in Supplementary Table 1.

Supplementary References

1. Bolli N, Avet-Loiseau H, Wedge DC, Van Loo P, Alexandrov LB, Martincorena I, et al. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nature Communications*. 2014;5:2997.
2. Papaemmanuil E, Gerstung M, Malcovati L, Tauro S, Gundem G, Van Loo P, et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*. 2013;122(22):3616-27.
3. Ye K, Schulz MH, 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*. 2009;25(21):2865-71.
4. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. *Nucleic Acids Res*. 2014;42(Database issue):D749-55.
5. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65.
6. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res*. 2011;39(Database issue):D945-50.
7. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods*. 2001;25(4):402-8.
8. Schnittger S, Kinkelin U, Schoch C, Heinecke A, Haase D, Haferlach T, et al. Screening for MLL tandem duplication in 387 unselected patients with AML identify a prognostically unfavorable subset of AML. *Leukemia*. 2000;14(5):796-804.

Supplementary Table S1. Primers used for PCR validation

name	Sequence (5'-3')
BCOR_ex1_F	TTTAGCACAGTCCTCCACCCCA
BCOR_ex1_R	CATTCCGTTCAAACCCAGCAGC
BCOR_ex4_F	CGGAAGACAGCGGTTCAAGACA
BCOR_ex4_R	GTATCGCCCAGTCCAATGCCTT
ACTB_ex3_F	GGAAGGAAGGCTGGAAGAGTGC
ACTB_ex3_R	TGTGCTATCCCTGTACGCCTCT
KRAS_ex3_F	CACTACCGATGCAGTCTGGAGC
KRAS_ex3_R	GGACTGGGGAGGGCTTTCTTTG
NPM1_F	ATTGGCCATATGGGTCTCTG
NPM1_R	AACACGGTAGGGAAAGTTCTCA
FLT3-ITD_F	GCAATTTAGGTATGAAAGCCAGC
FLT3-ITD_R	CTTTCAGCATTTTGACGGAACC
MLL-6.1	GTCCAGAGCAGAGCAAACAG
MLL-2.0	CGCACTCTGACTTCTTCATC

Supplementary Table S2. Variants identified in the study and their validation

Algorithm	Sample	CHR	START	END	Gene	Transcript	Protein	Effect	Validation method	Validation outcome
Pindel	PD17929a	13	28608280	28608281	FLT3	CCDS31953.1	p.D600_L601insDFREYEYD	frameshift	PCR + agarose gel	validated
Pindel	PD17929a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Pindel	PD17929a	13	28608290	28608298	FLT3	CCDS31953.1	p.E598_Y599insNEYFYVDFREYE	frameshift	PCR + agarose gel	validated
Caveman	PD17929a	20	31022938	31022938	ASXL1	CCDS13201.1	p.P808H	missense	MiSeq	validated
Caveman	PD17929a	2	209113112	209113112	IDH1	CCDS2381.1	p.R132L	missense	MiSeq	validated
Pindel	PD17930a	2	25463299	25463300	DNMT3A	CCDS33157.1	p.E733fs*1	frameshift	MiSeq	validated
Pindel	PD17930a	13	28608271	28608272	FLT3	CCDS31953.1	p.K602_W603insEYEYDLK	frameshift	PCR + agarose gel	validated
Pindel	PD17930a	13	28608274	28608275	FLT3	CCDS31953.1	p.E608_N609insYEYDLKWEFPRE	frameshift	PCR + agarose gel	validated
Caveman	PD17930a	2	209113113	209113113	IDH1	CCDS2381.1	p.R132S	missense	MiSeq	validated
Pindel	PD17930a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Pindel	PD17931a	13	28608286	28608287	FLT3	CCDS31953.1	p.Y597_E598insDYVDFREY	frameshift	PCR + agarose gel	validated
Caveman	PD17931a	2	25457242	25457242	DNMT3A	CCDS33157.1	p.R882H	missense	MiSeq	validated
Caveman	PD17931a	4	106180928	106180928	TET2	CCDS47120.1	p.?	ess splice	MiSeq	validated
	PD17931a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Pindel	PD17932a	13	28608280	28608281	FLT3	CCDS31953.1	p.D600_L601insFREYEYD	frameshift	PCR + agarose gel	validated
Caveman	PD17932a	2	25467449	25467449	DNMT3A	CCDS33157.1	p.G543C	missense	MiSeq	validated
Caveman	PD17932a	4	106156570	106156570	TET2	CCDS47120.1	p.Q491K	missense	MiSeq	No coverage
	PD17932a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Pindel	PD17932c	13	28608280	28608281	FLT3	CCDS31953.1	p.D600_L601insFREYEYD	frameshift	PCR + agarose gel	validated
Caveman	PD17932c	2	25467449	25467449	DNMT3A	CCDS33157.1	p.G543C	missense	MiSeq	validated
	PD17932c	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	not confirmed
Caveman	PD17933a	2	25463182	25463182	DNMT3A	CCDS33157.1	p.R771*	nonsense	MiSeq	validated
Caveman	PD17933a	19	33792981	33792981	CEBPA	ENST00000498907	p.G114C	missense	MiSeq	validated

	PD17933a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Caveman	PD17934b	2	25457243	25457243	DNMT3A	CCDS33157.1	p.R882C	missense	SureSelect + NGS	validated
Caveman	PD17934b	13	28592642	28592642	FLT3	CCDS31953.1	p.D835Y	missense	SureSelect + NGS	validated
	PD17934b	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Caveman	PD17935a	4	106158455	106158455	TET2	CCDS47120.1	p.L1119*	nonsense	SureSelect + NGS	validated
Caveman	PD17935a	12	25378647	25378647	KRAS	CCDS8703.1	p.K117N	missense	SureSelect + NGS	validated
Caveman	PD17936a	15	90631934	90631934	IDH2	CCDS10359.1	p.R140Q	missense	SureSelect + NGS	validated
Pindel	PD17936a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	PCR failure
Pindel	PD17936c	13	28608280	28608281	FLT3	CCDS31953.1	p.D600_L601insFREYEYD	frameshift	PCR + agarose gel	not confirmed
Caveman	PD17936c	15	90631934	90631934	IDH2	CCDS10359.1	p.R140Q	missense	SureSelect + NGS	validated
Caveman	PD17936c	13	28602340	28602340	FLT3	CCDS31953.1	p.N676K	missense	SureSelect + NGS	validated
Pindel	PD17936c	13	28608288	28608302	FLT3	CCDS31953.1	p.E608_N609insDNEYFYVDFREYEYDLKWEFPRE	frameshift	PCR + agarose gel	not confirmed
Pindel	PD17936c	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Pindel	PD17937c	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Caveman	PD17937c	2	209113112	209113112	IDH1	CCDS2381.1	p.R132H	missense	SureSelect + NGS	validated
Caveman	PD17937c	13	28592642	28592642	FLT3	CCDS31953.1	p.D835Y	missense	SureSelect + NGS	validated
Caveman	PD17937c	1	115258748	115258748	NRAS	CCDS877.1	p.G12S	missense	SureSelect + NGS	validated
Pindel	PD17938a	13	28608308	28608309	FLT3	CCDS31953.1	p.E598_Y599insCRSSDNEYFYVDFREYE	frameshift	PCR + agarose gel	validated
Caveman	PD17938a	2	25457242	25457242	DNMT3A	CCDS33157.1	p.R882H	missense	MiSeq	validated
Caveman	PD17938a	12	112940014	112940014	PTPN11	CCDS9163.1	p.D556Y	missense	MiSeq	No coverage
	PD17938a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Pindel	PD17939a	19	33792393	33792394	CEBPA	ENST00000498907	p.E309_T310insN	inframe	SureSelect + NGS	validated
Pindel	PD17939a	11	32456252	32456254	WT1	CCDS7878.2	p.N214fs*36	frameshift	SureSelect + NGS	validated
Pindel	PD17939a	13	28608288	28608291	FLT3	CCDS31953.1	p.Y597_E598insDFYVDFREY	frameshift	PCR + agarose gel	validated
Caveman	PD17939a	1	115258747	115258747	NRAS	CCDS877.1	p.G12D	missense	SureSelect + NGS	validated
	PD17939a	19	33792982	33792983	CEBPA	ENST00000498907_r69	p.A111fs*56	frameshift	SureSelect + NGS	validated
Caveman	PD17940a	20	31022297	31022297	ASXL1	CCDS13201.1	p.C594*	nonsense	SureSelect + NGS	validated

Caveman	PD17940a	12	25398284	25398284	KRAS	CCDS8703.1	p.G12V	missense	SureSelect + NGS	validated
Pindel	PD17941a	2	25467105	25467109	DNMT3A	CCDS33157.1	p.G590fs*61	frameshift	SureSelect + NGS	validated
Caveman	PD17941a	15	90631838	90631838	IDH2	CCDS10359.1	p.R172K	missense	SureSelect + NGS	validated
Pindel	PD17942a	13	28608286	28608287	FLT3	CCDS31953.1	p.Y597_E598insDYVDFREY	frameshift	PCR + agarose gel	validated
Caveman	PD17942a	2	25457242	25457242	DNMT3A	CCDS33157.1	p.R882H	missense	MiSeq	validated
Caveman	PD17942a	2	209113113	209113113	IDH1	CCDS2381.1	p.R132C	missense	MiSeq	validated
	PD17942a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Caveman	PD17943a	15	90631934	90631934	IDH2	CCDS10359.1	p.R140Q	missense		
	PD17943a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Pindel	PD17944a	20	31022412	31022413	ASXL1	CCDS13201.1	p.H633fs*2	frameshift	MiSeq	validated
Caveman	PD17945a	1	115256536	115256536	NRAS	CCDS877.1	p.A59S	missense	MiSeq	not confirmed
Caveman	PD17945a	7	148508721	148508721	EZH2	CCDS5891.1	p.G648V	missense	MiSeq	No coverage
	PD17945a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*	frameshift	PCR + capillary sequencing	validated
Pindel	PD17946a	4	106156658	106156660	TET2	CCDS47120.1	p.S521fs*1	frameshift		
Caveman	PD17946a	21	36231774	36231774	RUNX1	CCDS13639.1	p.R204*	nonsense	MiSeq	validated
Caveman	PD17946a	20	31022625	31022625	ASXL1	CCDS13201.1	p.G704R	missense	MiSeq	No coverage
Caveman	PD17946a	4	106164773	106164773	TET2	CCDS47120.1	p.R1214Q	missense	MiSeq	not confirmed
Caveman	PD17947a	17	29557890	29557890	NF1	CCDS42292.1	p.W1048C	missense	MiSeq	not confirmed
Caveman	PD17948a	X	44937750	44937750	KDM6A	CCDS14265.1	p.D980Y	missense	MiSeq	validated
Caveman	PD17949a	2	25458661	25458661	DNMT3A	CCDS33157.1	p.N838D	missense	MiSeq	validated
Pindel	PD17950a	2	25463567	25463568	DNMT3A	CCDS33157.1	p.I705fs*8	frameshift		
Caveman	PD17950a	9	5073770	5073770	JAK2	CCDS6457.1	p.V617F	missense		
Pindel	PD17951a	11	119149254	119149274	CBL	CCDS8418.1	p.I423_E427delIKGTE	inframe	MiSeq	validated
Caveman	PD17951a	21	36231774	36231774	RUNX1	CCDS13639.1	p.R204*	nonsense	MiSeq	validated
Caveman	PD17951a	7	148508719	148508719	EZH2	CCDS5891.1	p.E649*	nonsense	MiSeq	No coverage
Pindel	PD17952a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Pindel	PD17953a	4	106157954	106157955	TET2	CCDS47120.1	p.R953fs*19	frameshift	MiSeq	validated

Pindel	PD17953a	4	106156406	106156418	TET2	CCDS47120.1	p.Y437fs*7	frameshift	MiSeq	validated
Pindel	PD17953a	13	28608316	28608317	FLT3	CCDS31953.1	p.E598_Y599insWVTGSSDNEYFYVDFREYE	frameshift	PCR + agarose gel	validated
Caveman	PD17953a	X	44938474	44938474	KDM6A	CCDS14265.1	p.A1008S	missense	MiSeq	validated
	PD17953a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Caveman	PD17954a	15	90631934	90631934	IDH2	CCDS10359.1	p.R140Q	missense	MiSeq	validated
	PD17954a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Pindel	PD17955a	4	106157812	106157817	TET2	CCDS47120.1	p.M906fs*17	frameshift	MiSeq	validated
Caveman	PD17955a	4	106164741	106164741	TET2	CCDS47120.1	p.S1203R	missense	MiSeq	not confirmed
	PD17955a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Caveman	PD17956a	4	106156348	106156348	TET2	CCDS47120.1	p.Q417*	nonsense	MiSeq	validated
	PD17956a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Caveman	PD17957a	20	31025013	31025013	ASXL1	CCDS13201.1	p.E1500*	nonsense	MiSeq	validated
Caveman	PD17957a	2	209113113	209113113	IDH1	CCDS2381.1	p.R132S	missense	MiSeq	validated
Caveman	PD17957a	4	106182965	106182965	TET2	CCDS47120.1	p.P1335Q	missense	MiSeq	No coverage
Caveman	PD17957a	4	106164878	106164878	TET2	CCDS47120.1	p.T1249N	missense	MiSeq	validated
Caveman	PD17959a	12	25398281	25398281	KRAS	CCDS8703.1	p.G13D	missense	MiSeq	validated
	PD17959a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Pindel	PD17960a	13	28608304	28608305	FLT3	CCDS31953.1	p.D600_L601insSDNEYFYVDFREYEYD	frameshift	PCR + agarose gel	validated
Caveman	PD17960a	2	25467449	25467449	DNMT3A	CCDS33157.1	p.G543C	missense	MiSeq	validated
Caveman	PD17960a	X	44938447	44938447	KDM6A	CCDS14265.1	p.E999*	nonsense	MiSeq	validated
	PD17960a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Pindel	PD17961a	13	28608216	28608219	FLT3	CCDS31953.1	p.?	frameshift	PCR + agarose gel	validated
Pindel	PD17962a	13	28608274	28608275	FLT3	CCDS31953.1	p.K602_W603insCREYEYDLK	frameshift	PCR + agarose gel	validated
Pindel	PD17962a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Caveman	PD17963a	4	106157573	106157573	TET2	CCDS47120.1	p.S825*	nonsense	MiSeq	No coverage
Caveman	PD17964a	21	36259280	36259280	RUNX1	CCDS13639.1	p.L71M	missense	MiSeq	No coverage
	PD17964a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	MiSeq	No coverage

Pindel	PD17965a	20	31022545	31022546	ASXL1	CCDS13201.1	p.R678fs*40	frameshift	MiSeq	No coverage
Caveman	PD17965a	Y	15417990	15417990	UTY	CCDS14783.1	p.S1018Y	missense	MiSeq	No coverage
	PD17965a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Caveman	PD17966a	7	148511205	148511205	EZH2	CCDS5891.1	p.R566L	missense	MiSeq	validated
Caveman	PD17966a	2	25469038	25469038	DNMT3A	CCDS33157.1	p.R474S	missense	MiSeq	validated
Caveman	PD17966a	13	28592641	28592641	FLT3	CCDS31953.1	p.D835V	missense	MiSeq	validated
Caveman	PD17966a	4	106157119	106157119	TET2	CCDS47120.1	p.Q674K	missense	MiSeq	No coverage
	PD17966a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	PCR + capillary sequencing	validated
Pindel	PD17967a	13	28608286	28608287	FLT3	CCDS31953.1	p.Y597_E598insDYVDFREY	frameshift	PCR + agarose gel	validated
Caveman	PD17967a	1	115258744	115258744	NRAS	CCDS877.1	p.G13D	missense	MiSeq	validated
Caveman	PD17967a	X	39932971	39932971	BCOR	CCDS48093.1	p.S543*	nonsense	MiSeq	validated
	PD17967a	5	170837547	170837548	NPM1	CCDS4376.1	p.W288fs*12	frameshift	MiSeq	validated
Pindel	PD17968a	2	25463235	25463244	DNMT3A	CCDS33157.1	p.F752delF	inframe	MiSeq	validated
Caveman	PD17968a	1	115251178	115251178	NRAS	CCDS877.1	p.G183V	missense	MiSeq	validated
Caveman	PD17968a	17	29554540	29554540	NF1	CCDS42292.1	p.?	ess splice	MiSeq	validated

Supplementary Figure Legends

Legend to Supplementary Figure 1

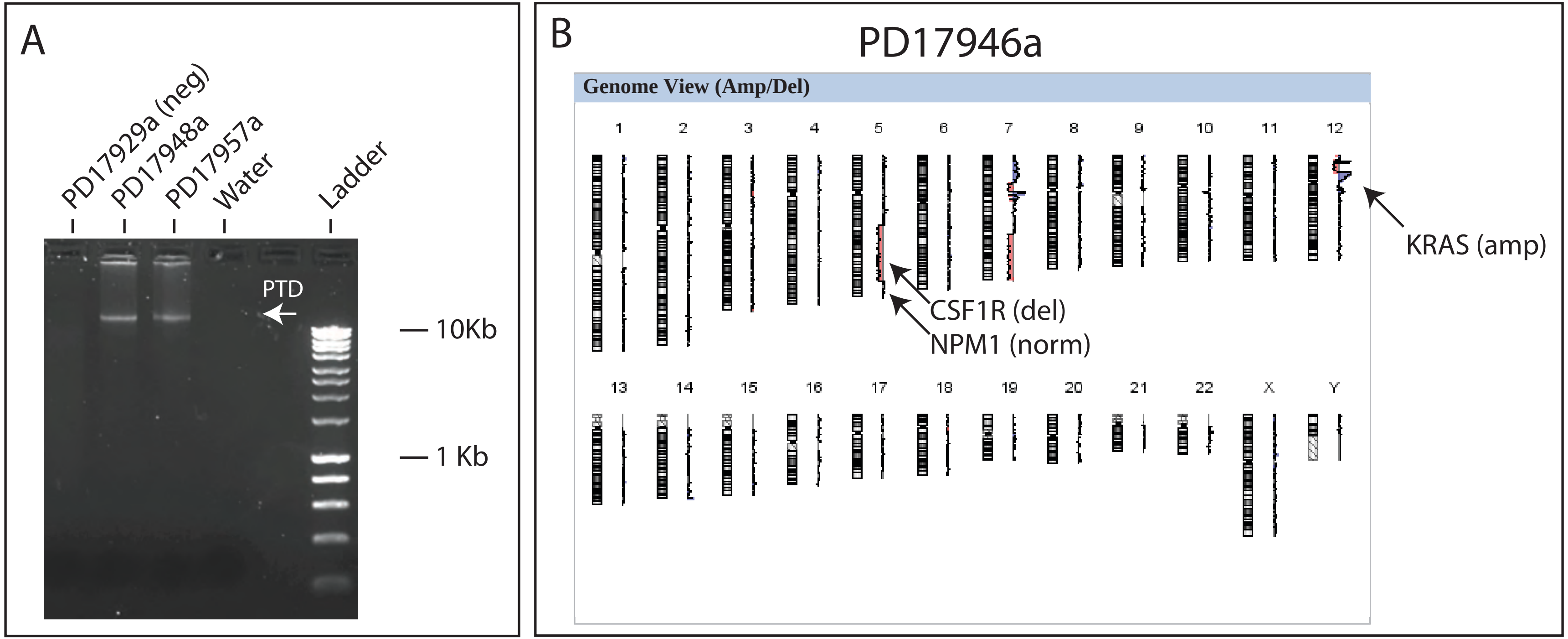
A) Long-range PCR on genomic DNA was performed on samples PD17948a and PD17957a to check for the presence of a MLL-PTD. Sample PD17929a was used as a negative control along with water. The white arrows shows a band at >10 Kb suggestive of an MLL-PTD in the two test samples.

B) Array CHG analysis of sample PD17946a confirms a KRAS amplification in chromosome 12p, and a 5q deletion that involved CSF1R but not NPM1 in keeping with the copy number pattern shown in Figure 3A (red bar).

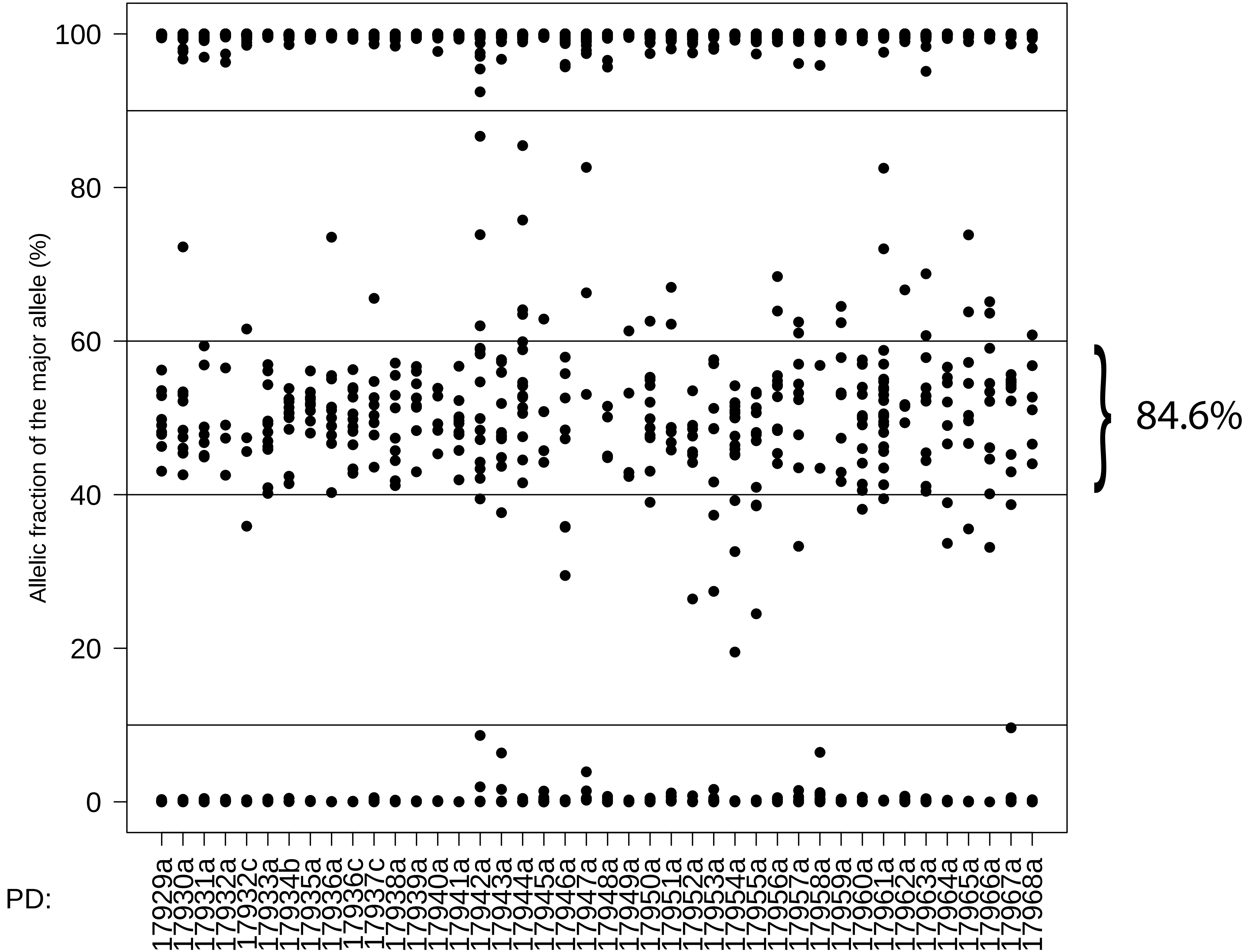
Legend to Supplementary Figure 2

For 90 single nucleotide polymorphisms (SNPs) covered by the study design, the allelic fraction of the major allele (defined as the most prevalent in the general population) is plotted in the Y-axis. Samples are plotted in the X-axis. Note that 84.6% of SNP calls fall close to the 50% mark for heterozygous SNPs, indicating quantitative value of the allelic fraction of single nucleotide substitutions in Haloplex data.

Supplementary Figure 1



Supplementary Figure 2


Allelic fraction of the major allele (%)
PD:
17929a
17930a
17931a
17932a
17932c
17933a
17934b
17935a
17936a
17936c
17937c
17938a
17939a
17940a
17941a
17942a
17943a
17944a
17945a
17946a
17947a
17948a
17949a
17950a
17951a
17952a
17953a
17954a
17955a
17956a
17957a
17958a
17959a
17960a
17961a
17962a
17963a
17964a
17965a
17966a
17967a
17968a