

Mutation accumulation in cancer genes relates to non-optimal outcome in chronic myeloid leukemia

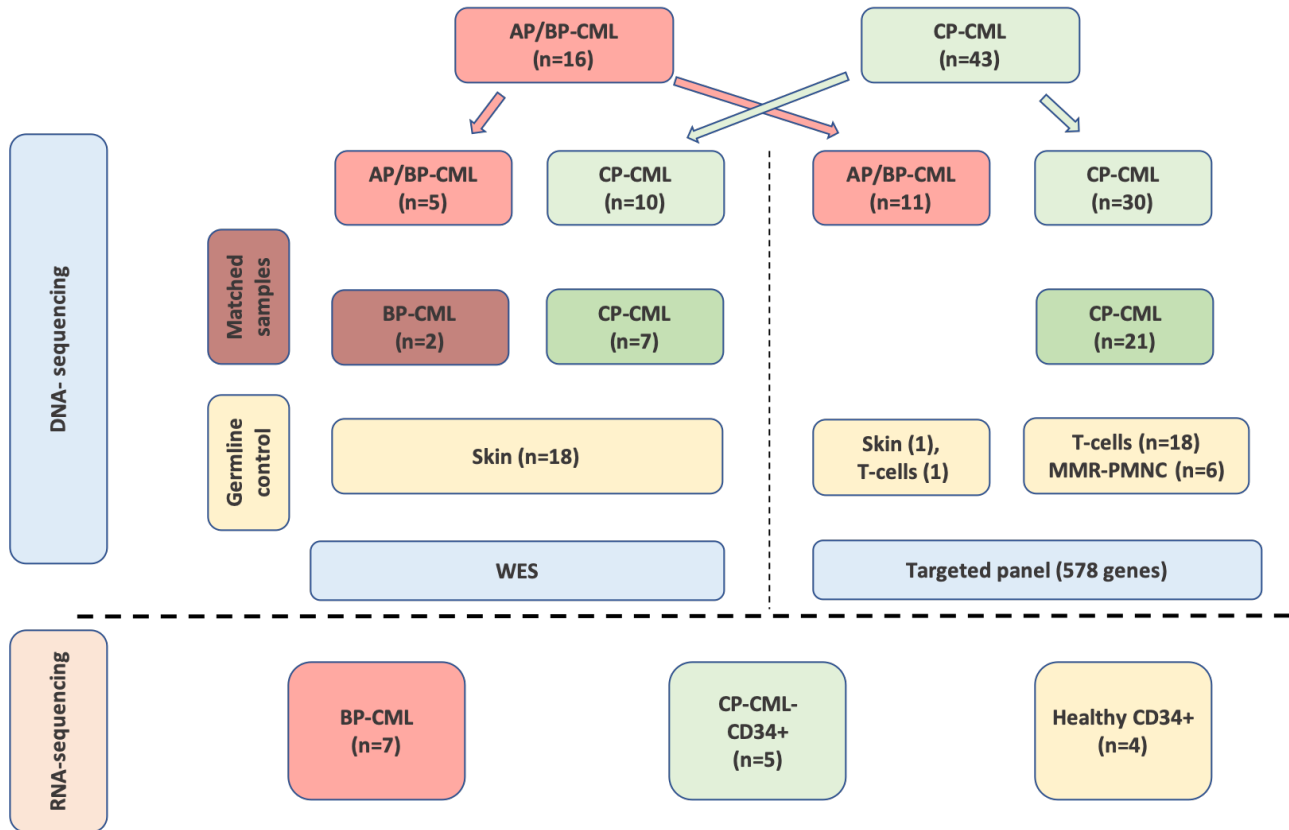
Short title: Genomic landscape of chronic and blast phase CML

Shady Adnan Awad,^{1-3*} Matti Kankainen,^{1,2,4,5*} Teija Ojala,⁶ Perttu Koskenvesa,¹ Samuli Eldfors,⁷ Bishwa Ghimire,⁷ Ashwini Kumar,⁷ Soili Kytölä,⁵ Mahmoud M. Kamel,³ Caroline A. Heckman,⁷ Kimmo Porkka,^{1,4} and Satu Mustjoki^{1,2,4}

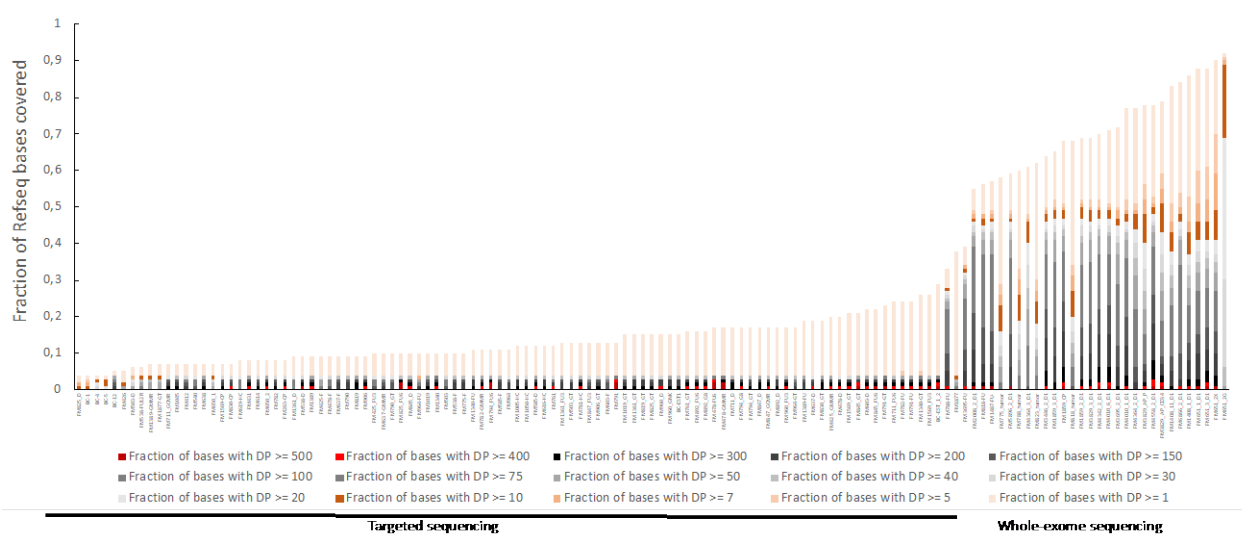
¹Hematology Research Unit Helsinki, Department of Hematology, Helsinki University Hospital Comprehensive Cancer Center, Helsinki, Finland; ²Translational Immunology Research Program and Department of Clinical Chemistry, University of Helsinki, Helsinki, Finland; ³Clinical and Chemical Pathology Department, National Cancer Institute, Cairo University; ⁴iCAN Digital Precision Cancer Medicine Flagship ⁵Medical and Clinical Genetics, University of Helsinki and Helsinki University Hospital, Helsinki, Finland; ⁶Pharmacology, Faculty of Medicine, University of Helsinki, Helsinki, Finland; and ⁷Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; *S.A.A. and M.K. contributed equally to this study.

Corresponding author: Prof. Satu Mustjoki, Hematology Research Unit Helsinki, University of Helsinki and Helsinki University Hospital Comprehensive Cancer Center, Haartmaninkatu 8, P.O. Box 700, FIN-00290 Helsinki, Finland. Tel: +358 9 471 71898, Fax: +358 9 471 71897, e-mail: satu.mustjoki@helsinki.fi

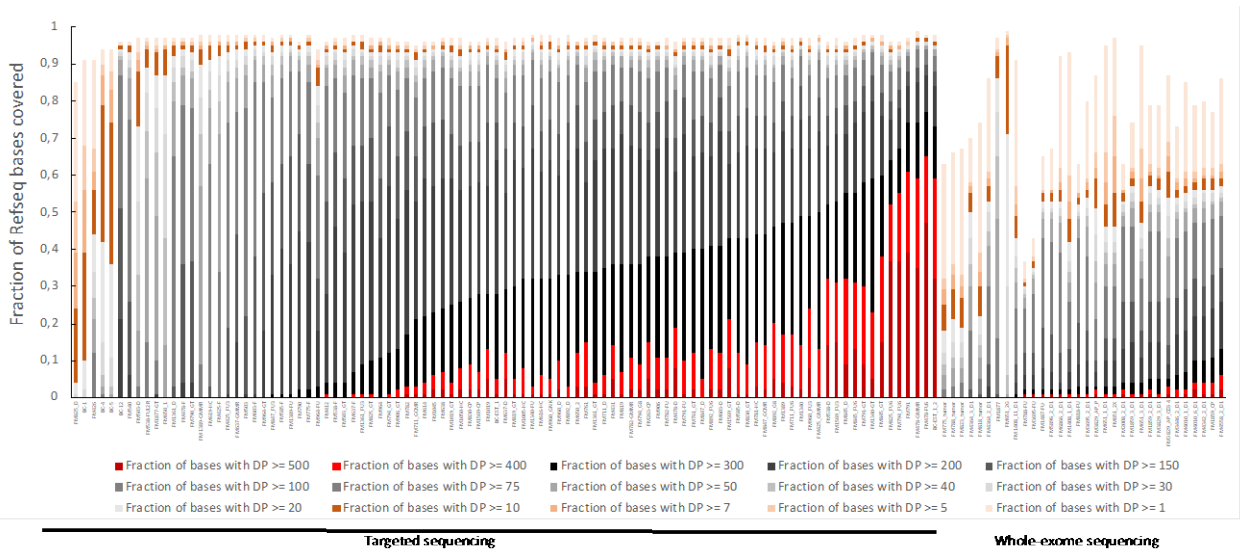
Supplemental Figures



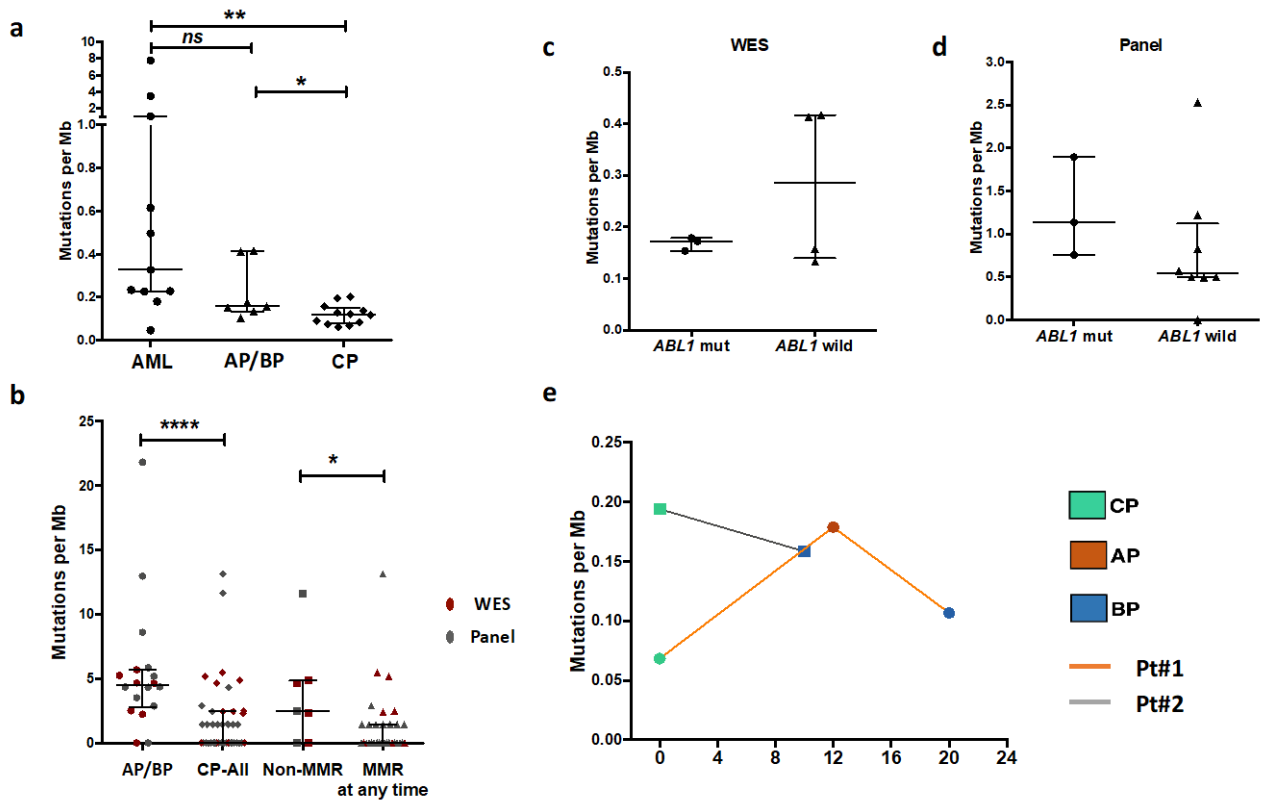
Supplemental Figure 1. Experimental workflow



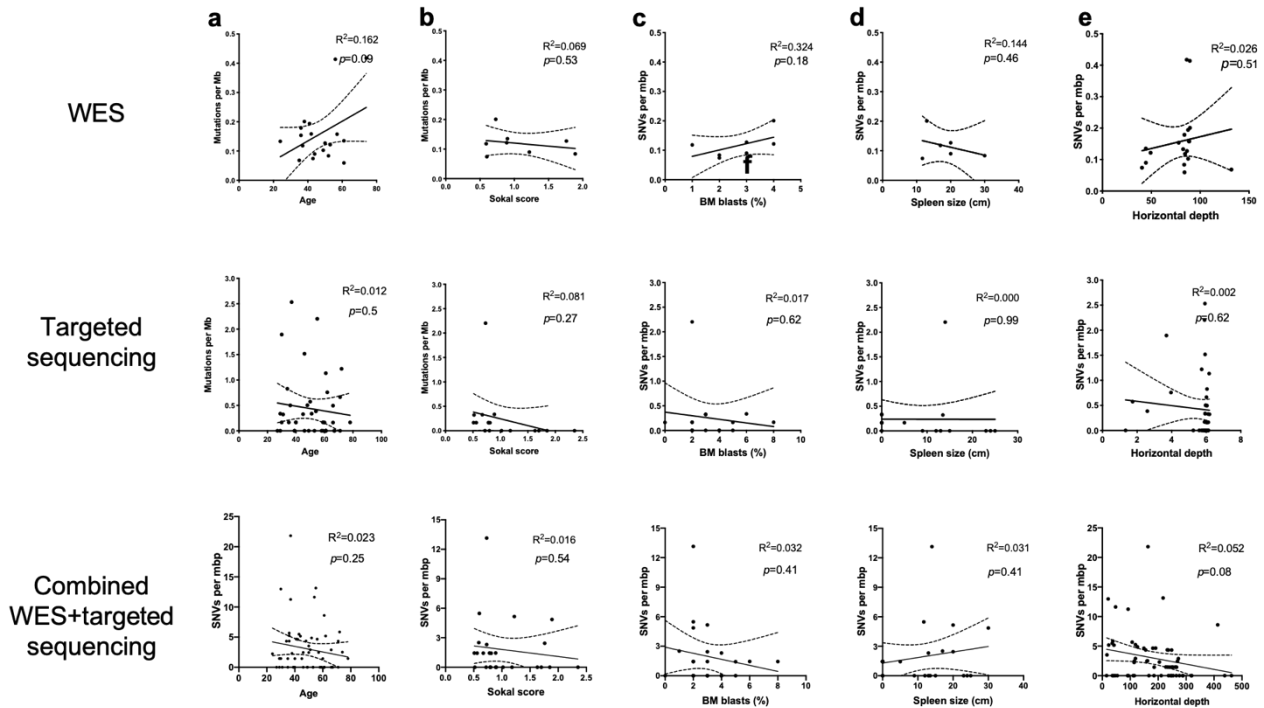
Supplemental Figure 2. Sequencing depth and coverage of each case analyzed. The figure shows fraction of exonic bases ± 5 flanks covered at depth of ≥ 100 , ≥ 75 , ≥ 50 , ≥ 40 , ≥ 30 , ≥ 20 , ≥ 10 , ≥ 5 and ≥ 1 . Annovar RefGene exons were used to define exonic bases that were padded by ± 5 bases. The average sequence depth of whole-exome sequencing (WES) samples was $84 \times (\pm 55)$ and of targeted sequencing samples $191 \times (\pm 95)$.



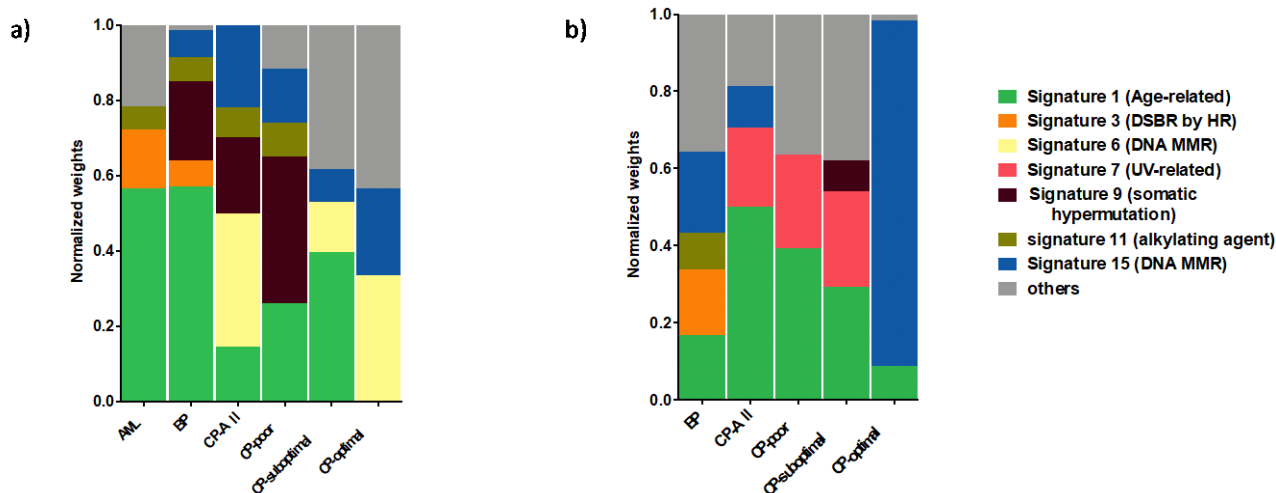
Supplemental Figure 3. Sequencing depth and coverage of each case analyzed over panel targets. The figure shows fraction of exonic bases ± 5 flanks covered at depth of ≥ 100 , ≥ 75 , ≥ 50 , ≥ 40 , ≥ 30 , ≥ 20 , ≥ 10 , ≥ 5 and ≥ 1 . Annovar RefGene exons of genes part of the panel were used to define exonic bases that were padded by ± 5 bases. The average sequence depth of whole-exome sequencing (WES) samples was $124\times (\pm 47)$ and of targeted sequencing samples $271\times (\pm 100)$.



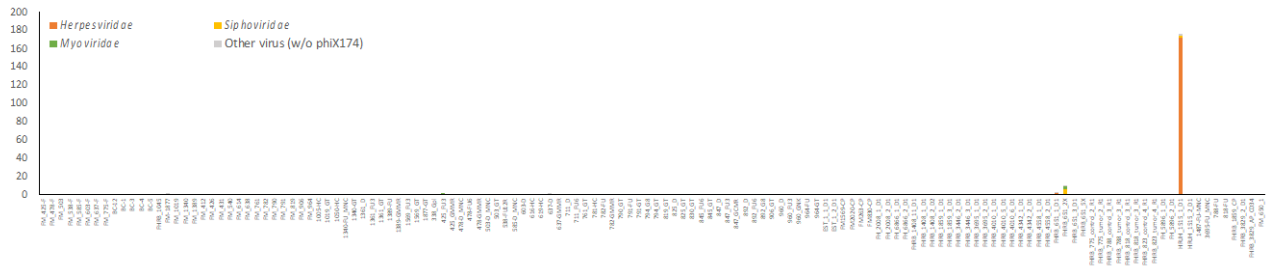
Supplementary figure 4. Numbers of single nucleotide variants (SNVs) per million bp in individual samples in (a) whole-exome sequencing (WES) highlighting the differences between AML (n=11), AP/BP (n=7), and CP (n=12) cases. (b) Combined WES and panel samples with restriction of SNVs to exons covered by the targeted panel sequencing highlights differences in SNV loads between BP/AP (7 WES, 12 panel) and CP cases (12 WES, 31 panel), and also differences between CP cases who fail to achieve MMR (4 WES, 3 panel) compared to cases who achieve MMR during TKI treatment (7 WES, 26 panel). WES cases are marked with dark red and panel cases with grey. SNVs per million bp in individual AP/BP samples from cases with and without ABL1 resistance mutations in samples analyzed using (c) WES or (d) panel. (e) SNVs per million bp in individual samples from cases with matched diagnostic CP and AP/BP samples. In the figure, * indicates $p < 0.05$, ** indicates $p < 0.01$, and **** indicates $p < 0.0001$.



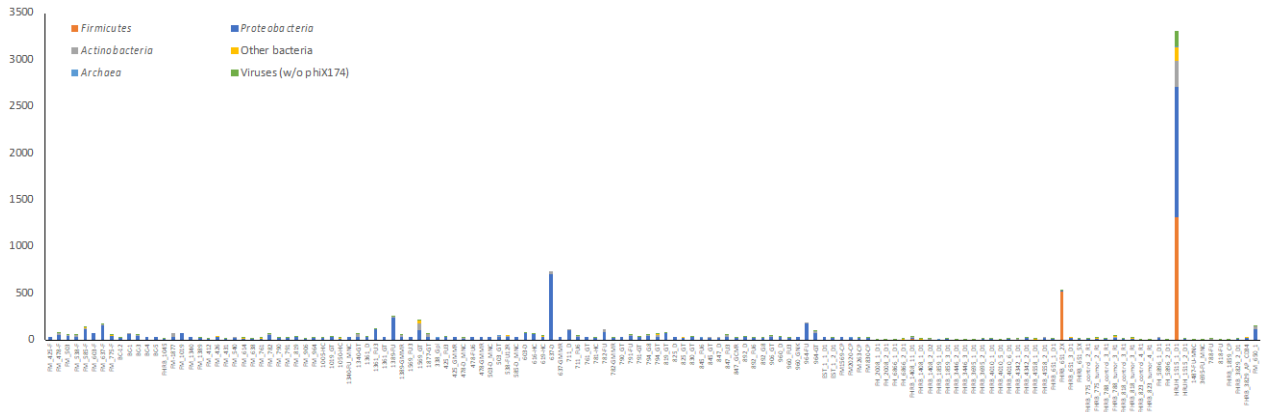
Supplemental Figure 5. Mutation load and clinical characters. Scatter plots comparing mutation load per million base pairs and clinical characters. Data from whole exome sequencing (WES) is presented in the upper row, from targeted panel in middle row and data from combined experiments in the left row. Clinical correlates are presented in columns: a) age, b) Sokal score, c) Bone marrow blast percentage, d) spleen size (in centimeter measured by ultrasound), e) sequencing horizontal depth. Correlations between variables were assessed using Pearson's correlation Coefficient. Only weak correlation was observed between mutation load and clinical criteria.



Supplemental Figure 6. Mutational signatures in CML. Normalized weights of trinucleotide signatures identified **(a)** in 11 AML, 10 AP/BP, and 10 CP cases and in CP cases with optimal (n=3), suboptimal (n=3), or poor respond (n=4) by WES. **(b)** Normalized weights of 11 AP/BP and 31 CP cases and of CP cases with optimal (n=16), suboptimal (n=10) or poor (n=3) respond by targeted panel sequencing. Weights of most frequent signatures in each cancer type are shown across cancers as separate signatures and others.

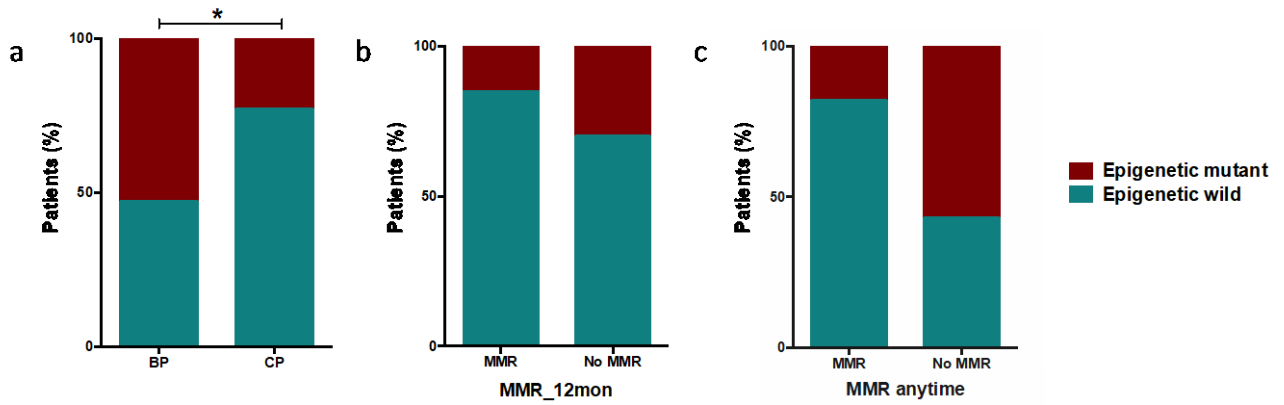


Supplemental Figure 7. Virus classification of each case analyzed. Number of viral reads in each sample analyzed. Viral read counts are expressed as counts per million mapped reads (CPMs). CPMs of three most abundant taxas in any sample are shown across all samples. Assignments of reads to other viral taxas excluding those to phiX174 (a common Illumina sequencing control) are summarised under the category other virus.

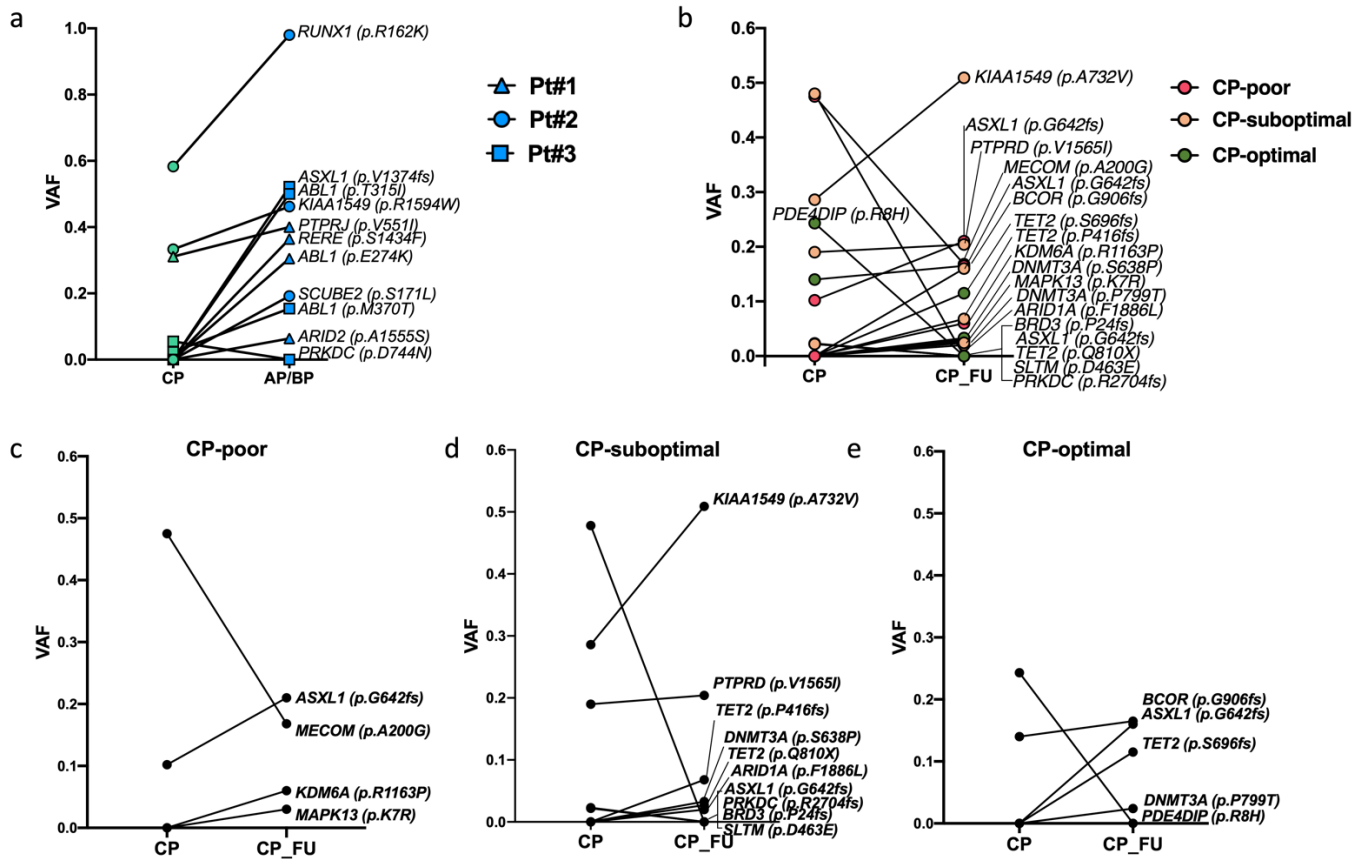


Supplemental Figure 8. *Bacteria, Archaea, and virus classification of each case analyzed.*

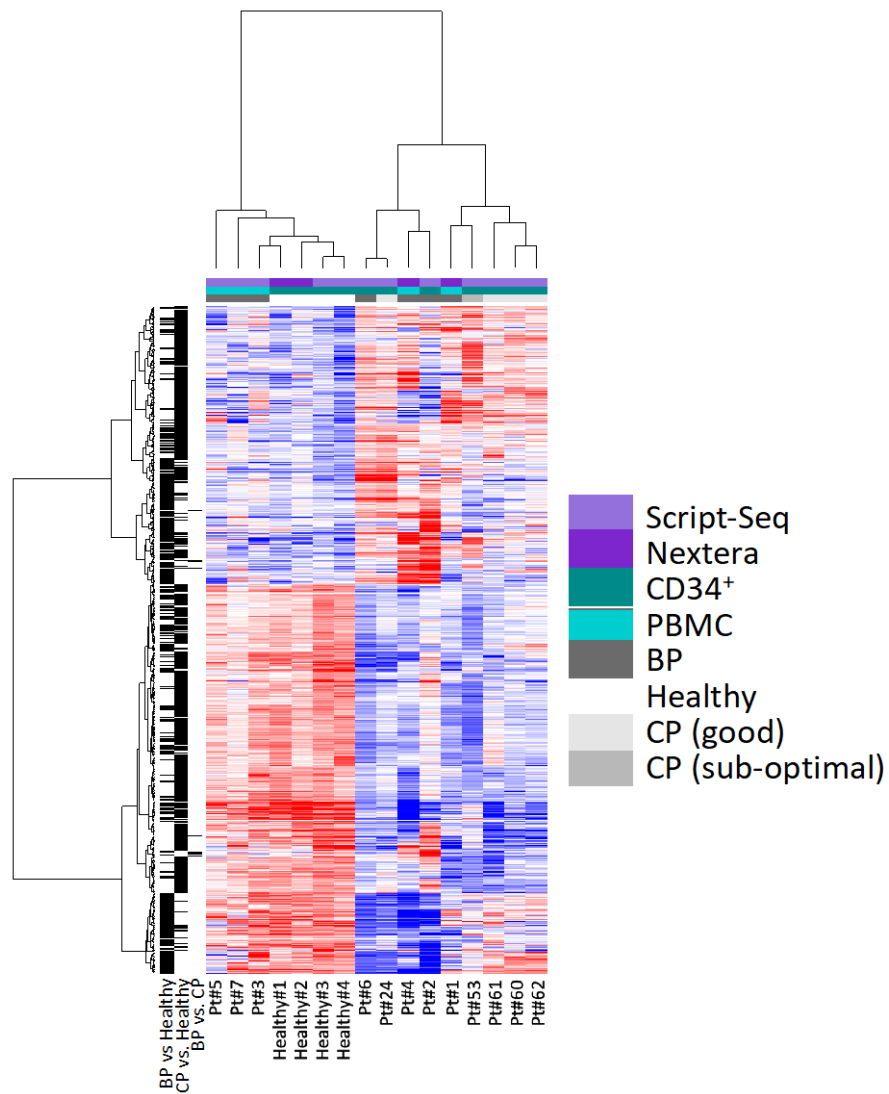
Number of bacterial, archaeal, and viral reads in each sample analyzed. Read counts are expressed as counts per million mapped reads (CPMs). CPMs of total archaeal, total viral assignments excluding those to phiX174 (a common Illumina sequencing control), and three most abundant bacterial taxa across any sample are shown. Assignments of reads to other bacterial taxa are summarised under the category other bacteria.



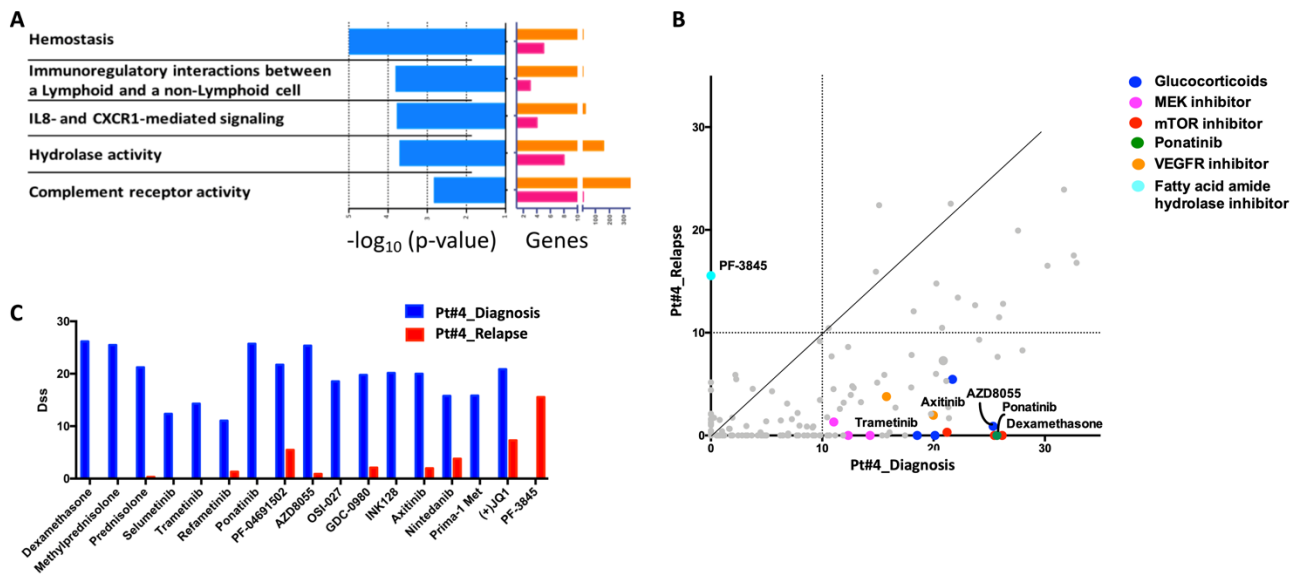
Supplemental Figure 9. Somatic mutations in CML. Stacked columns comparing the prevalence of mutations in epigenetic genes. **(a)** The difference between AP/BP and CP patients. **(b)** The difference between patients with or without optimal response. **(c)** The difference between patients with or without optimal or sub-optimal response. The differences between groups were analyzed with t-test. In the figure, * marks $p < 0.05$.



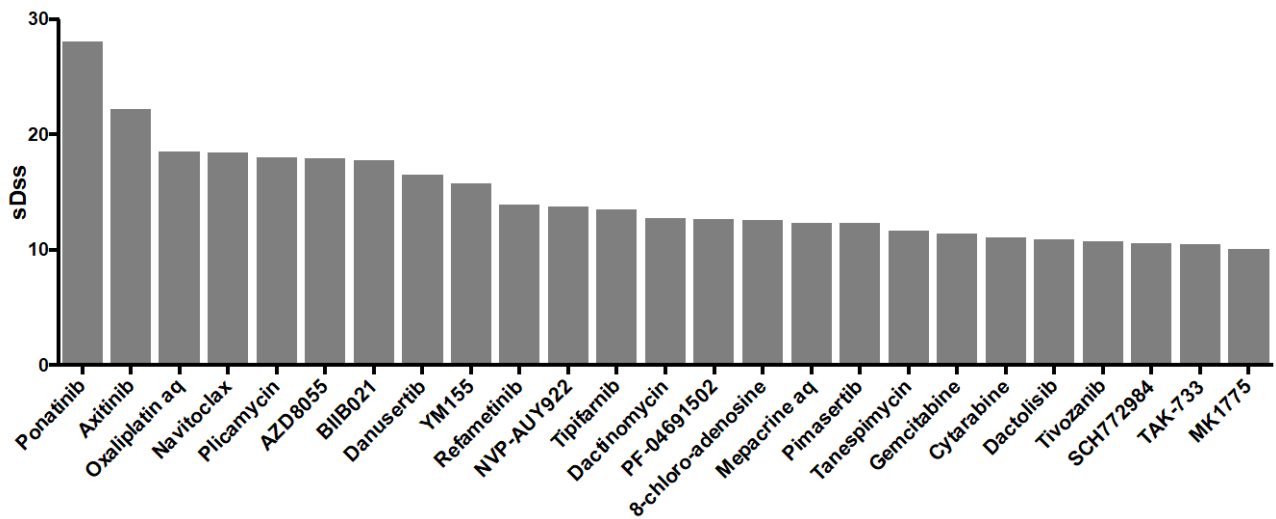
Supplemental figure 10. Relevant variants VAF in longitudinal samples. Plots showing change in relevant variants VAF between **(a)** CP and BP samples in patients with matched CP_AP/BP samples (n=3) **(b-e)** initial diagnosis and follow up time points in CP patients with serial samples (n=25) collectively **(b)**, CP-poor patients (n=2) **(c)**, CP-suboptimal patients (n=11) **(d)**, CP-optimal patients (n=13) **(e)**



Supplemental Figure 11. Heatmap of statistically differentially expressed genes between AP/BP and CP, AP/BP and healthy, and CP and healthy. Fading blue colours indicate down-regulation of the gene in the sample and red its up-regulation relative to the mean expression of the genes across all samples. Explanatory tracks from top to bottom show disease phase, sequencing protocol, and sorting status of the sample. Clustering was performed with both genes and samples using a Euclidean distance and Ward linkage method. Panel on the left shows in which comparison the gene was found as statistically differentially expressed with Q-value ≤ 0.05 .



Supplemental Figure 12. Comparison of drug sensitivity profile diagnosis and relapse samples from first index case. **A)** Depiction of molecular pathways with altered expression between timepoints. The analysis highlighted notable reprogramming of expression of genes associated with pathways X and Y. **B).** Scatter plot comparing DSS of diagnosis (X-axis) to relapse (y-axis). Note that TKI other than ponatinib has DSS<5 in both samples. **C)** Waterfall plot highlighting the most potent cancer-selective drugs for the primary cells extracted from patient. The drug sensitivity screen suggested administration of ponatinib and axitinib to the patient. At the relapse, sensitivity to these drugs were lost in accordance with the loss of the original subclone.



Supplemental Figure 13. Drug sensitivity profile from the second index case. Waterfall plot highlighting the most potent cancer-selective drugs for the primary cells extracted from the patient.

Supplemental Datasets:

Supplemental dataset 1. Clinicopathological features of CML cohort.

Supplemental dataset 2. Identified mutations in all samples.

Supplemental dataset 3. Correlation of mutational signature profiles in CML patients' subsets.

Supplemental dataset 4. Results from pathogen screening analysis.

Supplemental dataset 5. Somatic mutations identified in 59 CML samples after manual curation.

Supplemental dataset 6. Somatic variants dynamics in 28 CML patients with serial samples.

Supplemental dataset 7. Fusion genes in CML patients.

Supplemental dataset 8. Differential expression analysis of RNA-sequencing data.

Supplemental dataset 9. Pathway enrichment analysis.

Supplemental dataset 10. Somatic mutations identified in the index case at diagnosis and relapse.

Supplemental dataset 11. Pathway enrichment analysis of the index case between diagnosis and relapse.

Supplemental dataset 12. Drug sensitivity profiling data of the index case at diagnosis and relapse.

Supplemental Methods

RNA sequencing

Total RNA was extracted from two AP/BP, all five CP, and all four control samples following CD34+ enrichment to minimize signal related to mature granulocytes and other cell types not present in samples with a high blast count. In the case of five AP/BP samples without CD34+ enrichment. The miRNeasy Mini Kit (Qiagen) was used in the RNA extraction. The RNA integrity was measured by Agilent Bioanalyzer RNApico chip (Agilent) and Qubit RNA kit (Life Technologies) was used to quantitate RNA in samples. RNA sequencing libraries were then prepared from 1.5 µg of total RNA using ribo-depletion-based approaches. In the case of two AP/BP and two controls cases, RNA sequencing libraries were prepared using the ScriptSeq v2™ Complete kit for human/mouse/rat (Illumina). In the remaining cases, preparation of RNA sequencing libraries involved the use of Illumina compatible Nextera™ Technology (Epicentre). In each case, RNA-sequencing libraries were purified using SPRI beads (Agencourt AMPure XP, Beckman Coulter). High Sensitivity chips by Agilent Bioanalyzer (Agilent) was used to evaluate the library quality. All libraries were sequenced on Illumina HiSeq instruments (HiSeq 2000, Illumina) with paired-end 100-bp (2 × 100) reads.

Analysis of RNA sequencing data

Analysis of RNA-sequencing data was performed mainly as previously described¹. Briefly, RNA-sequencing data were pre-processed similar to DNA-sequencing data. Filtered paired-end reads were aligned to human reference genome build 38 (Ensembl v82) using STAR² with the guidance of Ensembl v82 gene models. Analysis was done using default 2-pass per-sample mapping settings, except that the overhang of the splice junctions was set to 99. Reads were sorted by coordinate using the SortSAM, PCR duplicates were marked with the MarkDuplicate module of the Picard toolkit,

feature counts were generated using SubRead³, feature counts were converted to expression estimates using Trimmed Mean of M-values (TMM) normalisation⁴, and lowly expressed genomic features with a CPM value ≤ 1.00 in less than half of samples removed. Differential expression testing was then performed using the edgeR⁵ software. In the statistical testing, comparisons between subject groups included factors for cell-sorting status and sequencing kit. The resulting P-values were adjusted Storey's Q-value approach⁶. Genomic features with Q-value ≤ 0.05 were considered differentially expressed. In data visualisations and pathway analyses, we used CPM data that was corrected for cell-sorting and library preparation kit effects. Batches were corrected using the removeBatchEffect function from the package limma⁷. Clustering of gene expression profiles was performed with both genes and samples using a Euclidean distance and Ward linkage method. Clustering revealed majority of samples from the same phase to cluster together (Supplemental Figure 10).

Drug sensitivity and resistance testing

The oncology compound library consisted of 125 FDA/EMA anti-cancer approved drugs and 127 investigational and preclinical compounds (Supplemental Table 11). All compounds were purchased from commercial chemical vendors and dissolved in either 100% dimethyl sulfoxide (DMSO) or water. Drug sensitivity and resistance testing was performed as previously described⁸. Briefly, mononuclear cells were suspended in Mononuclear Cell Medium (MCM; PromoCell) supplemented with 0.5 $\mu\text{g ml}^{-1}$ gentamicin and 2.5 $\mu\text{g ml}^{-1}$ amphotericin B. Each compound was tested covering a 10,000-fold concentration range in five different concentrations and pre-printed on 384-microwell plates (Corning) with an acoustic liquid handling device (Echo 550, Labcyte Inc.). Five μl culture medium per well was added to dissolve compounds and plates were shaken for 10 min. Freshly isolated cells in a single-cell suspension (10,000 cells in 20 μl per well) were dispensed using Multi-

Drop Combi peristaltic dispenser (Thermo Scientific). Plates were then incubated for 72 h at 37 °C and 5% CO₂. CellTiter-Glo 2.0 reagent (Promega) was used to measure cell viability according to the manufacturer's instructions using a Pherastar FS plate reader. Cell viability luminescence data were normalised to DMSO-only wells (negative control) and 100 mM benzethonium chloride-containing wells (positive control). The drug sensitivity and resistance testing data were quantified using the drug sensitivity score⁹.

Variant analysis

Analysis of DNA read data was mainly performed as previously described¹⁰. Briefly, sequence data was pre-processed for low quality, adapter sequences, and short read length using the Trimmomatic software¹¹. Paired-end reads passing filters were then aligned to human reference genome build 38 (Ensembl v82) using BWA-MEM¹², alignments were sorted by coordinate using the SortSAM, and PCR duplicates were marked with the MarkDuplicate module of the Picard toolkit (Broad Institute). Default parameters were used. Calling of variants employed Genome Analysis Toolkit (GATK) toolset¹³ and was based on the GATK somatic short variant best practice (version 3.5), supplemented with the estimation of the cross-sample contamination level and filtering of the 8-oxoguanine and deamination artefacts with GATK4 CalculateContamination, CollectSequencingArtifactMetrics, and FilterByOrientationBias tools. In the case of WES, calling of variants employed tumor-normal variant calling strategy. These variants were filtered against a panel of normals consisting of variants detected in two or more exomes of 24 healthy unrelated Finnish individuals. In the case of targeted sequencing, tumor-only variant calling strategy was used to enable comparison of cases with and without controls. Variants were filtered against the same panel of normals that was used in exome analyses as well as a panel of normals from five unrelated age-matched controls (seen in at least one individual) sequenced using the same sequencing protocol and a panel of normals from patient-matched control samples (seen in three or more samples) that had been sequenced using the same sequencing protocol.

The sensitivity and positive predictive value of the tumor-only variant calls in comparison to the matched tumor-control calls were estimated to be (after variant annotation and filtering) 0.77 and 0.62, respectively. GATK resources used in the variant calling process were converted from GRCh37 to GRCh38 using CrossMap¹⁴ and chain files downloaded from EnsEMBL.

To enable differentiation of variants with a low variant allele frequency from technical or biological artefacts, datasets were filtered after variant calling for vector contamination, RNA or pseudogene associated reads. In this process, reads from the final GATK alignment files were re-mapped to human reference genome build 38 (EnsEMBL v94) using STAR² with the guidance of EnsEMBL v94 gene models. Alignments were sorted by coordinate using the SortSAM, PCR duplicates were marked with the MarkDuplicate, indels were left-aligned using the GATK toolkit, and duplicate pairs, unmapped pairs, and secondary alignment were removed. Read pairs with an internal gap ≥ 10 bp and insert size less than 50 kb or with an insert size of between 1 and 50 kb were then classified as discordant. The fraction of discordant read pairs relative to undiscordant spanning exon-intron boundaries were then assessed per gene and exon. Variants with a variant allele frequency \leq either contamination fraction + 2% were removed with the exception of variants supported by approximately same fractions of discordant and undiscordant reads (*i.e.* variant allele frequency in discordant read pairs $\times 0.8 \leq$ variant allele frequency in undiscordant read pairs \leq variant allele frequency in discordant read pairs $\times 1.2$) at gene and exon level and variants residing in genes and exons without any discordant read pair.

Variants were annotated and filtered using the Annovar tool¹⁵ against the RefGene database. At first, all variant calls were normalised using bcftools¹⁶. Variants other than those passing all MuTect2 filters with a TLOD ≥ 6.3 or a TLOD ≥ 5.0 and supported by five or more independent COSMIC¹⁷ samples were filtered. Variant data were then filtered for false-positives by removing variants in

intronic and intergenic regions, with a total coverage ≤ 10 , and not supported by at least one read in both directions as well as variants with variant quality value ≤ 40 , variant allele frequency $\leq 2\%$, strand odd ratio for SNVs ≥ 3.00 , and strand odd ratio for indels ≥ 11.00 , minor allele frequency $\geq 1\%$ in the 1KG database, minor allele frequency $\geq 3\%$ in the EPS database, minor allele frequency $\geq 1\%$ in general, African, Finnish, Latino, East Asian, and Non-European ExAC, gnomAD exome, or gnomAD genome databases, PHRED-like CADD score ≤ 3.00 , and likelihood ratios score ≤ 2.00 . Variants with a variant allele frequency $\geq 35\%$ were accepted, if supported by five or more COSMIC¹⁷ samples. For functional analyses, the previous variant call set was filtered further by removing synonymous mutations and non-frameshift variants. Finally, cancer associated mutations were picked by removing those without COSMIC identifier and those seen in genes that were mutated in less than two patients. Variants were manually curated, missed known cancer variants checked and rescued, and variants inspected using Integrative Genomics Viewer 2.3.66 (Broad Institute).

Identification of mutational signatures was done using the deconstructSigs¹⁸ software with default parameters and using cancer profiles downloaded from the COSMIC web site on September 2017. In the analysis, function mapSeqlevels from the package GenomeInfoDb was used to convert EnSEMBL chromosome nomenclature to UCSC nomenclature. Sequencing coverage was computed from read regions overlapping Annovar RefGene exons with a 5 bp padding on each side (i.e. regions tested in variant annotation and filtering) using samtools depth, revealing a mean coverage of 89 \times and 187 \times for WES and targeted sequencing (Supplemental Figure 2).

Pathogen discovery

Classification of the DNA sequencing reads into microbial taxa was performed mainly as previously described¹⁰. Briefly, Trimmomatic¹¹ was used in adapter trimming, quality filtering, and removal of

short (<36bp) reads. Pre-processed read-pairs were then mapped against rRNA sequences from RFAM v12.3¹⁹ by using the Burrows-Wheeler Aligner (BWA)²⁰ with default settings, and read pairs matching rRNAs were filtered by using samtools²¹. Surviving paired-end reads were classified into different taxa by using Centrifuge²² and an index made of 27,127 known complete bacterial, archaeal, and viral genome assemblies, the human genome, and 10,615 technical artefact sequences that were available in the RefSeq²³ database on February 2018. Default parameters were used in the classification, with the exception of reporting only one taxonomical assignment (*i.e.* the lowest common ancestor) for read-pairs with multiple primary assignments. Classification results are available in supplemental Table 3. In data visualisations and statistical analyses, read counts were scaled by the total number of reads in the root times one million to obtain CPMs. Total number of viral and total number of bacterial, archaeal, and viral reads in each case analyzed as expressed in CPMs is provided in supplemental Figures 7 and 8, respectively. The statistical significance of the difference in microbial counts was examined using two-tailed Student's t-test with unequal variance.

Pathway enrichment analysis

Pathway enrichment analysis was done using GSEA²⁴ software (Broad Institute) and Enrichr^{25,26}. In the GSEA analysis, a pre-ranked gene list was prepared by sorting genes by their log-fold change in the batch corrected CPM data. GSEA analysis was then performed using default values. Q-value <0.05 was used as a threshold to interpret analysis output. The Enrichr was applied to protein coding genes significantly differentially expressed between AP/BP and healthy, CP and healthy, intersection of the previous two gene lists, as well as AP/BP and CP groups.

Calling of fusion genes

Hybrid genes were identified using FusionCatcher²⁷ with default parameters. FusionCatcher was applied to raw, un-processed RNA sequencing data. Results obtained from the tool were further filtered by excluding fusion calls located in introns, present in healthy control samples, and supported by ≤ 3 spanning unique reads. Fusion calls supported by fewer unique reads were selected if supported by evidence from spanning pairs or reciprocal reads and if clinically relevant (previously described in cancer).

Validation of hybrid genes

To validate the hybrid genes, single-step RT-PCR was performed on cDNA from samples in which hybrid genes were called. The cDNA was synthesized from total RNA using QuantiNova Reverse Transcription kit (QIAGEN). Primers were designed for *CBFB-MYH11*, *RUNX1-DYRK1A*, *TMEM236-MRC1* hybrid genes (supplemental Table 6). The cDNA was amplified with Phusion® High-Fidelity DNA Polymerase (NEB) and using the Applied Biosystem 2720 thermal cycler (ThermoFisher). 10 ul of PCR products were stained with Gel Loading Dye, Purple (6X) (NEB), run on a 2% agarose gel, and visualized on a standard UV trans illuminator. 5ul of PCR products were then cleaned up using ExoSAP-IT (ThermoFisher) and used for Sanger sequencing using both forward and reverse primers of the relevant hybrid gene and standard sequencing protocols. *In situ* hybridization (FISH) and karyotyping data were available for two hybrids (*BCR-ABL1* and *CBFB-MYH11*).

References

1. Kumar A, Kankainen M, Parsons A, et al. The impact of RNA sequence library construction protocols on transcriptomic profiling of leukemia. *BMC Genomics*. 2017;18:.
2. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* 2013;29(1):15–21.
3. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* 2013;41(10):e108.
4. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 2010;11:R25.
5. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* 2010;26(1):139–140.
6. Storey JD. A Direct Approach to False Discovery Rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 2002;64(3):479–498.
7. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43(7):e47.
8. Pemovska T, Kontro M, Yadav B, et al. Individualized systems medicine strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. *Cancer Discov.* 2013;3(12):1416–1429.
9. Yadav B, Pemovska T, Szwajda A, et al. Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies. *Sci. Rep.* 2014;4:5193.
10. Dufva O, Kankainen M, Kelkka T, et al. Aggressive natural killer-cell leukemia mutational landscape and drug profiling highlight JAK-STAT signaling as therapeutic target. *Nat. Commun.* 2018;9(1):1567.
11. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* 2014;30(15):2114–2120.

12. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio*. 2013;
13. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–1303.
14. Zhao H, Sun Z, Wang J, et al. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinforma. Oxf. Engl*. 2014;30(7):1006–1007.
15. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.
16. Li H. BFC: correcting Illumina sequencing errors. *Bioinformatics*. 2015;31(17):2885–2887.
17. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*. 2017;45(D1):D777–D783.
18. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol*. 2016;17:31.
19. Nawrocki EP, Burge SW, Bateman A, et al. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res*. 2015;43(Database issue):D130–D137.
20. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl*. 2009;25(14):1754–1760.
21. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl*. 2009;25(16):2078–2079.
22. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res*. 2016;
23. O’Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1):D733-745.

24. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 2005;102(43):15545–15550.
25. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 2016;44(W1):W90-97.
26. Chen EY, Tan CM, Kou Y, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics.* 2013;14:128.
27. Nicorici D, Satalan M, Edgren H, et al. FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv.* 2014;011650.