Recapitulation of human germline coding variation in an ultra-mutated infant leukemia

Alexander M Gout[1], Rishi S Kotecha[2,3,4], Parwinder Kaur[5,6], Ana Abad[1], Bree Foley[7], Kim W Carter[8], Catherine H Cole[3], Charles S Bond[9], Ursula R Kees[2], Jason Waithman[7,*], Mark N Cruickshank[1,*]

[1]Cancer Genomics and Epigenetics Laboratory, Telethon Kids Cancer Centre, Telethon Kids Institute, University of Western Australia, Perth, Australia.

[2]Leukaemia and Cancer Genetics Program, Telethon Kids Cancer Centre, Telethon Kids Institute, Perth, Australia.

[3]Department of Haematology and Oncology, Princess Margaret Hospital for Children, Perth, Australia.

[4]School of Medicine, University of Western Australia, Perth, Australia.

[5]Personalised Medicine Centre for Children, Telethon Kids Institute, Australia.

[6]Centre for Plant Genetics & Breeding, UWA School of Agriculture & Environment.

[7]Cancer Immunology Unit, Telethon Kids Institute, Perth, Australia.

[8]McCusker Charitable Foundation Bioinformatics Centre, Telethon Kids Institute, Perth, Australia.

[9]School of Molecular Sciences, The University of Western Australia, Perth, Australia.

* These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.W. (Jason.Waithman@telethonkids.org.au) or to M.N.C. (Mark.Cruickshank@telethonkids.org.au). Telethon Kids Institute, 100 Roberts Road, Subiaco, Western Australia, 6008. Phone: +61 8 9489 7777. Fax: +61 8 9489 7700.

Running title: Ultra-mutation targets germline alleles in an infant leukemia.

Total number of Figures/Tables: 5 figures and 2 tables.

Word count: 5,312.

Conflicts of interest: The authors declare no conflicts of interest.

Abstract: *Mixed lineage leukemia/Histone-lysine N-methyltransferase 2A* gene rearrangements occur in 80% of infant acute lymphoblastic leukemia, but the role of cooperating events is unknown. Infant leukemias typically carry few somatic lesions, however we identified a case with over 100 somatic point mutations per megabase. The patient presented at 82 days of age, among the earliest manifestations of cancer hypermutation recorded. The transcriptional profile showed global similarities to canonical cases. Coding lesions were predominantly clonal and almost entirely targeting alleles reported in human genetic variation databases with a notable exception in the mismatch repair gene, *MSH2*. The patient's diagnostic leukemia transcriptome was depleted of rare and low-frequency germline alleles due to loss-of-heterozygosity, while somatic point mutations targeted low-frequency and common human alleles in proportions that offset this discrepancy. Somatic signatures of ultra-mutations were highly correlated with germline single nucleotide polymorphic sites indicating a common role for 5-methylcytosine deamination, DNA mismatch repair and DNA adducts. These data suggest similar molecular processes shaping population-scale human genome variation also underlies the rapid evolution of an infant ultra-mutated leukemia.

Introduction:

Histone-lysine N-methyltransferase 2A (*KMT2A*) (also known as *Mixed lineage leukemia*) gene rearrangements (*KMT2A*-R) at 11q23 occur in infant, pediatric, adult and therapy-induced acute leukemias and are associated with a poor prognosis. *KMT2A*-R occurs in around 80% of infant acute lymphoblastic leukemia (ALL) and 35-50% of infant acute myeloid leukemia (AML) involving more than 60 distinct partner genes [1]. *KMT2A*-R infant ALL (iALL) genomes have an exceedingly low somatic mutation burden, with recurring mutations in *RAS-PI3K* complex genes. Compared to wild-type KMT2A, the fusion oncoprotein has altered histone-methyltransferase activity, causing epigenetic and transcriptional deregulation. While KMT2A-oncoproteins are strong drivers of leukemogenesis, several lines of evidence suggest that *KMT2A*-R alone is insufficient for tumor initiation. For example, there is significant variability in latency of disease onset in experimental models of *KMT2A*-fusions [2, 3] and in patients with *KMT2A*-R detected in neonatal blood spots [4]. Furthermore, *KMT2A*-R acute leukemias [5] in children are enriched for mutations affecting epigenetic regulatory genes and on average harbor more somatic lesions than infants. These observations have prompted speculation that iALL is a distinct developmental and genetic entity and pre-leukemic transformation occurs *in utero*, targeting a cell of origin with a genetic/epigenetic permissive predisposition [6].

As a step towards defining cooperating events we performed next-generation sequencing on a *KMT2A*-R iALL patient cohort and integrated the data with a larger set of acute leukemias. We investigated the combined germline and somatic mutation profiles, identifying a highly mutated specimen from a patient diagnosed at 82 days of age. Here we report sequence analysis of this unique leukemia in relation to its canonical *KMT2A*-R iALL counterparts, revealing common mechanisms driving leukemia ultra-mutation and population-scale human genome variation.

Methods:

Patient specimens, external sequencing data and controls

The study cohort consisted of ten infants diagnosed at Princess Margaret Hospital for Children (Perth), of which we have previously reported RNA-seq data for five patients [7]. Additional RNA-seq data was downloaded for 39 *KMT2A*-R iALL cases at diagnosis, six cases at relapse, six cases lacking the *KMT2A*-R (*KMT2A*-germline) and for pediatric *KMT2A*-R patients at diagnosis, comprising 10 AML cases, seven ALL cases and one undifferentiated case as reported previously by Andersson *et al*. [5] (accession EGAS00001000246). RNA-seq data comprising 28 *KMT2A*-R and 52 *KMT2A*-germline adult diagnosistic AML cases downloaded from the Short Read Archive (BioProject Umbrella Accession PRJNA278766[8]; Leucegene: AML sequencing). Normal human bone marrow transcriptomes were downloaded from the Short Read Archive (BioProject PRJNA284950) comprising 20 pooled sorted human hematopoietic progenitor populations [9]. The full list of samples analysed in this study including those downloaded from external sources is summarised in Supplementary Table S1.

Sample preparation and next-generation sequencing

RNA was isolated with TRIzol™ Reagent (Invitrogen Life Science Technologies, Carlsbad, CA, USA) and purified with RNeasy Purification Kits (Qiagen, Hilden, Germany). Genomic DNA was isolated with QIAamp DNA Mini Kit (Qiagen). Nucleic acid integrity and purity was assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). RNA-seq was performed at the Australian Genome Research Facility, Melbourne with 1 µg RNA as previously reported [7]. Whole exome sequencing (WES) was performed by the Beijing Genome Institute using SureSelect XT exon enrichment (Agilent Technologies) with 200 ng of remission DNA (v4 low input kit) or 1 µg of diagnostic DNA (v5). Diagnostic RNA and

remission genomic DNA sequencing was performed on a HiSeq 2000 (Illumina, Inc., San Diego, CA, USA) instrument with 100bp paired end reads; diagnostic genomic DNA sequencing was performed on a HiSeq 2500 with 150bp paired end reads. Sequencing libraries were constructed using Illumina TruSeq kits (Illumina, Inc.).

Bioinformatics

For gene-expression analysis, RNA-seq reads were mapped to hg19 with HISAT and gene-counts generated with HTseq [10]. Data was normalized with RUV-seq [11] implementing factor analysis of negative control genes identified by fitting a linear model with cell type (hematopoietic sub-set or leukemia sub-type) as a covariate. voom! [12] was used to adjust for library size and generate counts per million (cpm) values. *KMT2A*-fusion split-reads were identified using RNA-seq data with *FusionFinder* [13] using default settings.

Single nucleotide variants (SNVs) were identified from RNA-seq data using the intersection of two methodologies: (1) the "Genome Analysis Toolkit (GATK) Best Practices workflow for SNP and indel calling on RNAseq data" and (2) *SNPiR* [14]. WES data was processed based on the GATK Best Practices workflow. Variants were characterized using custom scripts that extract the coverage of sequenced bases from VCF and bam files, determining the overlap among samples, and calculating variant allele fractions (VAFs).

High confidence somatic mutations were identified using *VarScan2* (v2.3.9) [15] setting cut-offs of *P*-value<0.005; coverage > 50X; VAF > 0.1 or VAF > 0.3 to enrich for clonal mutations. Variants were annotated with ANNOVAR [16] and with metadata from additional sources including seven functional/conservation algorithms (PolyPhen2 [17], Sift [18], MutationTaster [19], likelihood ratio test [20], GERP [21], PhyloP [22], and CADD [23]) and databases of human genetic variation (Exome Aggregation Consortium [ExAC] [24], 1000 genomes, HapMap, dbSNP, Catalog of Somatic

Mutations in Cancer (COSMIC)[25] and ClinVar). Predicted functional alleles were assigned for variants called as damaged/deleterious by at least 4/7 algorithms.

Somatic mutational spectra were investigated using the R package, *DeconstructSigs* (v1.8.0) [26] calculating the fraction of somatic substitutions within the possible 96-trinucleotide contexts relative to the exome background. Copy number analysis was performed on WES data using *VarScan2* to calculate log2 depth ratio and GC-content and *Sequenza* (v2.1.2) [27] to identify chromosomal segments with copy loss. Gene ontology enrichment was performed using GREAT (v3.0.0) [28] bioinformatics tools. Mutations were inspected by mapping the genomic position to crystal structures in the Protein DataBank[29] and analyzed with Pymol (v1.8.6.2).

Data and code availability

Scripts used to execute variant detection software and analysis are available at https://github.com/cruicks/ultramutated-case under an MIT license and are being developed on Github (https://github.com/cruicks/vcfbamCompare). The sequencing data has been deposited in the European Genome Phenome Archive (ega-box-909).

Results:

Characteristics of KMT2A-R iALL cohort and RNA-seq variant detection pipeline

We reported previously *KMT2A*-R fusion transcripts for 5/10 cases [7], and of the remaining five *KMT2A*-R iALL we detected *KMT2A-MLLT3* (P401, P438), *KMT2A-AFF1* (P848), *KMT2A-MLLT1* (P706) and *KMT2A-EPS15* (P809). *KMT2A*-fusion transcripts were therefore detected in all samples and reciprocal fusion sequences detected in 5/10 samples (Table 1). To determine the validity of our RNA-seq based pipeline for detecting rare and common alleles, we examined the concordance of single nucleotide variants (SNVs) called from RNA-seq and WES data. Across seven samples with matched WES and RNA-seq data, we identified a total of 54,792 point mutations from RNA-seq reads of which 166 were absent from matched WES (coverage>20 reads), representing a discordance rate of 0.3% (Figure 1a; Supplementary Table S2). In one sample, P706 there were two events called with RNA-seq data that were not identified in matched WES but with evidence extracted using the Integrated Genome Browser at previously reported somatic mutation sites in *FLT3* (encoding FLT3 p.V491L [30]; 200/1,719 alternate RNA-seq reads; 15/214 alternate WES reads) and *KRAS* (encoding KRAS p.A146T [31]; 24/159 alternate RNA-seq reads; 6/51 alternate WES reads). Furthermore, analysis of 117 discordant events with coverage > 50 reads, revealed 76 recurrences in two or more samples (64%) at 26 unique sites, however none of these sites have been validated as somatic mutations in cancer to date in the COSMIC[25] or The Pediatric Cancer Genome Project (PCGP)[32] databases. These results indicate a high rate of reproducibility of the RNA-seq variant detection pipeline and suggest that RNA-DNA differences mostly comprise artifacts and/or novel RNA-editing events [33] and a minority of putative somatic alterations exclusively detected by RNA-seq.

Identification of a highly mutated infant leukemia

Expressed somatic mutations were defined as variants detected in both diagnostic leukemia RNA-seq and WES, but absent in remission WES at sites with >20 reads (Figure 1b-d). Most patients (5/6) showed the expected "silent genomic landscape" (Supplementary Table S3) except for patient P337 (Figure 1b-d; Supplementary Table S4). Three patients expressed RAS-family hot-spot somatic mutations (P848 and P706: KRAS p.A146T and P438: NRAS.p.G12S). We detected no expressed somatic missense mutations in patients P399 and P401. The pair of monozygotic twins (P809 and P810) lacked private expressed mutations but we detected two shared expressed missense mutations clustered in the *PER3* gene (encoding PER3.p.M1006R and PER3.p.K1007E). These results are consistent with recurring RAS-pathway mutations and low mutation rate reported previously [5, 34]. However, in patient P337 RNA-seq reads we detected 1,420 expressed somatic mutations (Supplementary Table S4). The overlap of both rare and common alleles across the matched datasets verifies the common origin of RNA-seq and diagnostic/remission WES data ruling out sample contamination or mislabeling (Figure 1a-c). We previously reported a complex *KMT2A* translocation involving 2q37, 19p13.3 and 11q23[35] in the highly mutated leukemia sample and established two cell lines from the diagnostic sample [7].

Characterisation and clonal analysis of somatic lesions

Somatic lesions were further evaluated in six cases by comparing diagnostic and remission WES. Two different variant allele fraction (VAF)-cut-offs, either VAF>0.1 or VAF>0.3, were chosen to enrich for sub-clonal and clonal events respectively. There were on average ~82.2 (range 69-95) sub-clonal and ~3.4 clonal (range 3-4) somatic point mutations in canonical *KMT2A*-R iALL exomes, confirming previous studies of a low somatic mutation rate in the dominant clone (Table 1). In contrast, there were 198 sub-clonal and 5,054 clonal somatic point mutations in the highly-mutated sample, equating to a rate of 139 mut/Mb, classifying this sample as ultra-mutated.

There were also ~40-fold increased number of indels in the dominant clone of the ultra-mutated sample (n=260) compared to typical cases (average = 6.6; range: 3-9) including 13 predicted to cause a frame-shift protein-coding truncation (Table 1; Supplementary Table S5). There were 2,420 exonic substitutions encoding 1,109 missense alleles and 12 stop-gain alleles (Supplementary Table S5). We also noted that the ultra-mutated case carried an excess of loss-of-heterozygosity (LOH) positions, with 6,555 high confidence sites; vastly more than typical cases which had a mean of 57.2 LOH positions (range 28-100) (Figure 2a-b; Table 1).

WES was performed on two independent cell lines derived from the ultra-mutated diagnostic specimen, PER-784A and PER-826A, to characterize clonal somatic events. (Figure 2c-d). Most somatic and LOH-events detected in the primary ultra-mutated sample were also called in both, PER-784A and PER-826A cell lines using P337 remission WES for all comparisons (Figure 2d). These data suggest that the somatic lesions are largely clonal in agreement with the VAF-distributions (Figure 2b, d). In total, 4,754/5,054 (94%) somatic events (Supplementary Table S6) and 6,150/6,575 (93%) LOH-events (Supplementary Table S7) were called in the derived cell lines. Moreover, there were relatively few discordant sites when directly comparing cell lines with the primary diagnostic sample (PER-784A: 34 discordant sites; PER-826A: 44 discordant sites; Supplementary Table S6)). We further investigated evidence of clonal somatic lesions in cell lines using an alternative approach, by directly extracting quality filtered WES reads mapping to coding somatic point mutations (n=2,411; Q>50 in diagnosis WES; Figure 2e) and LOH-sites (n=3,117; Q>50 in remission WES; Figure 2f; Supplementary Table S8). All coding clonal somatic mutations were supported by reads in both cell lines of which 121 (5%) were homozygous (VAF>0.95 in cell line and VAF>0.9 in primary sample). Most LOH-sites were evidenced by reference alleles (VAF<10%; 2,179/3,117) in cell lines,

with ~30% of LOH-sites evidenced by alternate alleles (VAF> 90%; 926/3,117) and only 12 sites (0.3%) with both reference and alternate alleles (VAF range 10-90%).

We performed copy number analysis of the ultra-mutated sample to map LOH-sites to chromosomal segments with evidence of deletions. Most genomic regions were called as diploid with limited chromosomal copy loss detected (Supplementary Figure S1a-b). There was a total of 382/6,555 (5.8%) LOH-sites mapping to homozygous deletions, suggesting only a minority of the identified LOH-sites likely arose through chromosome segment deletion. Infant *KMT2A*-R ALL typically carry a small number of copy-number alterations which may coincide with LOH-sites. We determined the distance of LOH-sites, somatic mutations and germline variants (called by *VarScan2*) to the nearest homozygous SNV in each leukemia diagnostic sample finding that in canonical cases LOH-sites tend to be located closer to homozygous sites than either germline or somatic variants (Supplementary Figure S1c; left panel) and show a unique bimodal distribution (Supplementary Figure S1c; right panel). In contrast the average distance of LOH-sites to homozygous SNVs in the ultra-mutated sample was similar to that of germline variants and somatic mutations with each displaying a similar distribution (Supplementary Figure S1c). Thus, most LOH-sites in the ultra-mutated specimen do not show evidence of localising within deleted or copy-neutral loss-of-heterozygosity chromosomal segments.

In summary, these data demonstrate that somatic lesions in the primary sample were predominantly clonal and LOH-sites appear to comprise a large proportion of somatic substitutions rather than chromosomal alterations. We also note relatively few coding sequence changes in derived cell lines compared to the dominant diagnostic leukemic clone, both which are capable of long-term propagation. These results could reflect loss of the hypermutator phenotype during *in vitro* culture or in diagnostic leukemia cells.

Signatures of somatic mutations and LOH-sites correlate with profiles of germline alleles

We next investigated the sequence features of somatic and LOH substitutions in the ultra-mutated sample (Figure 3). We hypothesised that LOH-sites and somatic mutations may show common feature biases depending on the context of alteration. Somatic events were evaluated separately that induce a heterozygous mutation (n=4,773) or homozygous mutation (n=281), and LOH causing reference→alternative changes (n=2,057) were evaluated separately to alternative→reference changes (n=4,453). Since we noted an overlap in the somatic mutations of the ultra-mutated sample with variants in the ExAC database (Figure 1a-c), we first examined the distribution of these sites with respect to the derived allele frequency in ExAC. We observe that each of these categories of sequence alterations in the ultra-mutated sample were distributed across the range of population allele-frequencies (Figure 3a-b). While the distributions of LOH and somatic sites were similar, somatic heterozygous changes were relatively more frequent at higher frequency alleles as were LOH reference→alternate alleles.

Somatic signatures of ultra-mutations were investigated in comparison to heterozygous and homozygous single nucleotide polymorphisms (SNPs) in patient remission samples. Heterozygous positions were defined by VAF<0.9 yielding 51,347 heterozygous and 31,237 homozygous SNPs in canonical remission samples and 10,681 heterozygous and 6 617 homozygous SNPs in the ultra-mutated remission sample. Analysis of WES remission variants from canonical patients and the ultra-mutated sample revealed similar ExAC allele frequency distributions with distinctions between heterozygous and homozygous alleles (Figure 3c-d). The overall enrichment of each category of somatic substitution and LOH across the 96 possible trinucleotide contexts resembled patient remission SNPs with elevated proportions of C>T transitions at NpCpG motifs (Figure 3e).

Pre-defined signatures were used to infer mutational profiles revealing Signature 1A (associated with spontaneous deamination of 5-methylcytosine), Signature 12 (associated with unknown process with strand-bias for T→C substitutions) and Signature 20 (associated with defective DNA mismatch repair, MMR) as enriched in each of the mutation/variant sets, however with distinctive proportions (Figure 3f). Heterozygous somatic mutations and LOH alternative→reference sites were additionally enriched for Signature 1B (variant of Signature 1A) and showed highly correlated profiles (Pearson's correlation 0.919). LOH reference→alternative sites were additionally enriched for Signature 5 (associated with unknown process with strand-bias for T→C substitutions) and were highly correlated with homozygous somatic mutation profiles (0.912) and to profiles of patient heterozygous positions (P337: 0.935; canonical: 0.926). Homozygous somatic mutations also displayed similarities to heterozygous patient SNPs (P337: 0.981; canonical: 0.985). Overall the somatic mutational profiles indicate highly related patterns of both somatic ultra-mutations and LOH-sites. Notably LOH alternative→reference sites showed unique and overlapping features to somatic heterozygous mutations. We also observe underlying similarities in the profiles of somatic substitutions and germline alleles.

Ultra-mutation distribution recapitulates common and rare germline variation

We noted a similar spectrum of rare and common alleles in the ultra-mutated leukemia transcriptome compared to typical *KMT2A*-R iALL cases; but there were comparatively fewer rare/low-frequency alleles shared with the remission sample (Figure 1a-c; ExAC<0.5%). We further explored this trend to disentangle the contribution of somatic acquired alleles and LOH-sites. Analysis of the remission and diagnostic samples from this patient in isolation, revealed a similar proportion of rare/low-frequency coding alleles detected by WES and both were proportionally within the same range as canonical counterparts (Figure 4a). However, when directly

comparing the remission and diagnostic WES, we noted a modest depletion of germline rare/low-frequency coding SNV alleles defined by *VarScan2* (Figure 4a; 2.9% rare alleles in P337 compared to 3.5±0.2% in canonical cases). This apparent depletion of germline alleles in the ultra-mutated leukemic sample was coincident with an increased proportion of somatic acquired rare/low-frequency coding SNV alleles compared to LOH-sites (Figure 4a). Therefore, when considering the total pool of SNVs in the leukemic coding genome, the somatic gains and losses of rare/low-frequency coding SNV alleles were in proportions that resulted in a similar distribution as canonical cases (Figure 4a). We extended this observation by investigating the rare and common allele abundance of expressed germline-encoded variants and somatic mutations in the ultra-mutated leukemic sample compared with a larger cohort of *KMT2A*-R iALL transcriptomes (Figure 4b; n=32 St. Jude patients and n=8 Perth patients). Most typical *KMT2A*-R iALL diagnostic specimens expressed similar numbers of rare and low-frequency expressed SNVs, except for 4/32 cases sequenced by St Jude investigators, which had an elevated number of expressed rare and low-frequency ExAC alleles (Figure 4b). Excluding these samples, typical *KMT2A*-R iALL patients expressed on average 42.3 ±10.4 (standard deviation) variants that were absent from ExAC and 212 ± 32.5 variants with an allele frequency of <0.5% in ExAC. The total number of expressed variants in the ultra-mutated sample was slightly lower than typical cases (38 absent from ExAC and 174 AF <0.5% in ExAC), and were approximately comprised of half germline (17/38 and 88/174 respectively) and half somatic acquired alleles. Overall, the ultra-mutated transcriptome had the lowest prevalence of germline-encoded rare and low-frequency alleles among the larger sample size (Table 2). Furthermore, the somatic mutation abundance within each allele-frequency bin was proportional to their abundance in prototypical silent cases. Consequently, the low numbers of rare germline alleles in the ultra-mutated transcriptome were off-set by somatic mutations such that the total numbers were within the same range as typical cases (Figure 4b-

c). We conclude that despite carrying thousands of somatic mutations, the coding-genome of ultra-mutated patient P337 shows similarities to canonical *KMT2A*-R iALL cases, vis-à-vis rare allele prevalence and representation of common human genetic variation.

RAS-pathway and DNA-replication and repair gene mutations

Gene ontology analysis of the expressed non-synonymous somatic mutations (n=535) revealed over-representation of RAS-pathway genes (7/23 genes; 7-fold enriched; *P*=0.049; *RALGDS*, *RAC1*, *CHUK*, *PIK3R1*, *NFKB1*, *PLD1* and *PIK3CA*) and "*MLL* signature 1 genes" (33/369 genes; 2-fold enriched; *P*=0.046). Of the 535 non-synonymous expressed somatic mutations, 45 were predicted deleterious/damaging, including in the MMR gene *MSH2* (p.E489K). The *MSH2* somatic mutation localizes to a hot-spot, within the MMR ATPase domain in a region at the periphery of the DNA binding site, approximately 20 Å from the DNA molecule in the complex structure of full-length MSH2, a fragment of MSH6, ADP and DNA containing a G:T mispair [36]. The mutation encodes a substitution of a negatively charged glutamate with a positively charged lysine (Figure 4d). There was a non-synonymous somatic mutation in another MMR gene, *MUTYH*; however, this site is a common human allele with a derived AF of 0.297 and was not called as damaged/deleterious by our prediction pipeline. There were no somatic mutations, LOH-sites nor rare/deleterious coding germline alleles detected in *POLE* or *POLD1* which are frequently co-mutated with *MSH2* in ultra-mutated tumours [37, 38] and in sporadic early-onset colorectal cancer [39]. Among the 45 expressed and predicted damaging somatic mutations, there were two rare substitutions in DNA polymerase encoding genes: *POLG2* (p.D308V; ExAC AF 0.0000165), which encodes a mitochondrial DNA polymerase and *POLK* (p.R298H; ExAC AF 0.0003), which encodes an error prone replicative polymerase that catalyzes translesion DNA synthesis. The *POLK* mutation maps to a conserved impB/mucB/samB family domain

(Pfam: PF00817) involved in UV-protection. The encoded substitution of arginine to histidine is located approximately 20 Å from the DNA molecule in the structure of a 526 amino acid N-terminal Pol K fragment complexed with DNA containing a bulky deoxyguanosine adduct induced by the carcinogen benzo[a]pyrene (Figure 4e) [40]. Both wild-type and mutant residues are positively charged, however arginine is strongly basic (pI 11) while histidine is weakly basic (pI 7) and titratable at physiological pH [41]. We additionally observed an expressed and predicted functional germline rare allele in *ATR*, a master regulator of the DNA damage response (p.L2076V; ExAC AF 0.0002).

Global transcriptional similarities of the ultra-mutated and canonical infant leukemias

We next examined the transcriptional profile of the highly mutated *KMT2A*-R iALL specimen in comparison to a collection of patient leukemia samples and purified blood progenitor populations (Figure 5). As expected cell type was the major source of variation distinguishing normal hematopoietic progenitor, lymphoblastic leukemias and myeloid leukemias. We also observe clustering according to *KMT2A* translocation status and partner genes as reported previously [42]. Replicate samples of the highly mutated *KMT2A*-R iALL specimen cluster with canonical counterparts demonstrating global transcriptional similarities (Figure 5).

Discussion:

We report sequence analysis of an ultra-mutated acute lymphoblastic leukemia diagnosed in an infant at 82 days of age which, to our knowledge, represents the earliest onset of a hypermutant tumor recorded. The mutation rate observed was at the upper limit of cases reported previously in childhood cancers [37, 38]. By integrating analysis of somatic mutations and germline alleles with a larger cohort of *KMT2A*-R iALL patients we found that the high somatic mutation rate does not generate large numbers of rare alleles compared to typical cases. Rather, we observe that the ultra-mutated *KMT2A*-R leukemic coding genome closely resembles the rare allele landscape of canonical cases. This apparent paradox is explained by the net effect of somatic gains and losses, both of which were targeted to rare and common alleles in proportions that off-set the depleted pool of germline alleles observed in the leukemic coding sequences. Our data indicate both the rare allele burden and transcriptional profile of the ultra-mutated leukemia were comparable to its canonical counterparts. Therefore, despite its radically different genetic route to leukemogenesis, the molecular profile of the ultra-mutated specimen seems to have converged towards that of canonical cases.

Most ultra-mutated childhood cancers reported previously have a combined MMR and proof reading polymerase deficiency [37, 38] - we detect evidence for the former (*MSH2* point mutation), but not the latter. However, we identify somatic mutations in two genes encoding DNA polymerases (*POLK* and *POLG2*) and an additional MMR gene (*MUTYH*). Notably, the *POLK* mutation is a rare allele, predicted functional and encodes pol κ, a replicative error-prone translesion DNA polymerase that plays a role in bypassing unrepaired bases that otherwise block DNA replication [43, 44]. For example, pol κ preferentially incorporates dAMP at positions complementary to 7,8-dihydro-8-oxo-2′-deoxyguanosine adducts [45]. As a corollary to these observations, *POLK* mutations have been found in prostate cancer

with elevated C→T transition [46]. Further studies are required to determine if any of these mutations cooperate to specify the unique mutational pattern observed in the ultra-mutated *KMT2A*-R iALL genome.

The largest fraction of ultra-mutations was attributed to Signature 1, associated with 5-methylcytosine deamination inducing C→T transitions, which accumulates in normal somatic cells with age over the human lifespan [47]. This signature is found ubiquitously across tumour sub-types [48], and prominently in hypermutated early onset AML patients with germline *MBD4* mutations [49]. LOH-sites (reference→alternative) were enriched for Signature 5 which is also correlated with human aging. We observe enrichment of Signature 20 in LOH-sites and somatic alterations, consistent with an association with a MMR-defect [48]. This signature has also been reported as enriched in tumours with germline-MMR and secondary mutation in the *POLD1* proof-reading gene [48]. We also find enrichment of Signature 12 which arises from a poorly characterized underlying molecular process; however, the dominant base transitions, A→G and T→C have been observed in association with DNA damage (e.g. by adducts) in normal and cancer tissues [50, 51]. These observations suggest common mutagenic and repair deficits driving age-related mutation in humans were operative in the cell of origin during pre-natal evolution of the ultra-mutated infant leukemia genome. *MSH2* was likely the primary MMR-hit in cooperation with an additional replicative repair lesion, however, we find no evidence for an involvement of *POLE* and *POLD1*.

The preponderance of ultra-mutated sites to overlap germline alleles in proportion to their population allele frequencies demonstrates that hypermutation can recapitulate aspects of population scale human genetic variation in a cancer genome. Aproximately half of the somatic substitutions appear to target heterozygous SNPs resulting in loss-of-heterozygosity. We cannot rule out a role for focal chromosomal alterations undetectable by our analyses. However collectively our data suggest a

major fraction of the LOH-sites could have arisen by point mutagenesis of heterozygous positions. They more frequently target the alternate allele inducing reversion to the reference base, with around 30% of events resulting in a homozygous non-reference allele in the diagnostic sample. Most ultra-mutations are therefore likely passengers coinciding with common human alleles, consistent with findings that hypermutated tumours tend to have relatively few drivers proportional to their mutation burden [52]; however, as proposed previously by Valentine and colleagues[53], a minority of rare alleles may contribute to infant leukemogenesis. While our results do not resolve the temporal order of the key oncogenic events - namely *MSH2* point mutation, genomic hypermutation and *KMT2A* translocation – due to the clonal nature of these events they appear to have occurred in rapid succession. This could be explained by a sentinel DNA replicative repair defect inducing hypermutation followed by positive selection of cooperating lesions resulting in clonal expansion and fixation of passengers [54]. An alternative model is also possible in which the *MSH2*-mutation impairs DNA double-strand-break repair [55, 56] promoting *KMT2A*-rearrangement, but with neutral selection of ultra-mutations. Such a scenario appears inconsistent with the observed predominance of clonal mutations, but could be reconciled if the hypermutator phenotype was lost rapidly after the *KMT2A*-rearrangement and prior to clonal expansion. Reports of pervasive positive selection in cancer and non-cancerous tissues [52, 57], lend support to the former model, however since we note loss of hypermutator phenotype in patient-derived cell lines, neutral selection of ultra-mutations is not ruled out by our data analysis. Altogether our results illuminate a link between mutagenesis during human aging, germline variation and somatic ultra-mutation in cancer. Further studies are required to determine if combinations of somatic mutations and/or germline alleles promote *KMT2A*-R iALL leukemogenesis/hypermutation and if somatic mutated sites significantly overlap human germline alleles in other malignancies.

Financial support: Children's Leukaemia and Cancer Research Foundation (Inc.) Fellowship (MNC); NHMRC Career Development Fellowship 1066869 (JW); Raine Medical Research Foundation Clinician Research Fellowship 2016-2018 (RSK); Children's Leukaemia and Cancer Research Foundation (Inc.) project grants (URK, MNC, RSK); Ian Potter Foundation (AG); Healy Foundation (AG). Telethon Kids Institute Blue Sky Research Grant (MNC, URK, RSK, JW). Telethon Kids Institute Research Focus Area Precision Medicine Working Group Funding (JW, MNC).

Acknowledgments: We thank Daniel MacArthur and Fengmei Zhao for assisting with exome sequence analysis; and researchers at The Pediatric Cancer Genome Project at St. Jude Children's Research Hospital for provision of data. We thank all the patients, families and hospital staff for providing samples.

Authorship contributions: AG wrote computer code, performed bioinformatics, conceived analysis strategy, interpreted and analyzed data and edited the manuscript. AA and BF performed experiments. KWC, PK and CB performed additional data analysis and interpretation. RSK provided patient samples and clinical data, assisted with data interpretation and edited the manuscript. CHC, URK and JW acquired funding support, assisted with data interpretation and edited the manuscript. MNC performed bioinformatics, formulated research, conceived analysis strategy, acquired funding support, interpreted data, designed experiments and drafted the manuscript.

References:

1. Meyer C, Burmeister T, Groger D, Tsaur G, Fechina L, Renneville A*, et al.* The MLL recombinome of acute leukemias in 2017. *Leukemia* 2017 Jul 13.

2. Bergerson RJ, Collier LS, Sarver AL, Been RA, Lugthart S, Diers MD*, et al.* An insertional mutagenesis screen identifies genes that cooperate with Mll-AF9 in a murine leukemogenesis model. *Blood* 2012 May 10; **119**(19)**:** 4512-4523.

3. Chen W, O'Sullivan MG, Hudson W, Kersey J. Modeling human infant MLL leukemia in mice: leukemia from fetal liver differs from that originating in postnatal marrow. *Blood* 2011 Mar 24; **117**(12)**:** 3474-3475.

4. Gale KB, Ford AM, Repp R, Borkhardt A, Keller C, Eden OB*, et al.* Backtracking leukemia to birth: identification of clonotypic gene fusion sequences in neonatal blood spots. *Proc Natl Acad Sci U S A* 1997 Dec 9; **94**(25)**:** 13950-13954.

5. Andersson AK, Ma J, Wang J, Chen X, Gedman AL, Dang J*, et al.* The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. *Nat Genet* 2015 Apr; **47**(4)**:** 330-337.

6. Sanjuan-Pla A, Bueno C, Prieto C, Acha P, Stam RW, Marschalek R*, et al.* Revisiting the biology of infant t(4;11)/MLL-AF4+ B-cell acute lymphoblastic leukemia. *Blood* 2015 Dec 17; **126**(25)**:** 2676-2685.

7. Cruickshank MN, Ford J, Cheung LC, Heng J, Singh S, Wells J*, et al.* Systematic chemical and molecular profiling of MLL-rearranged infant acute lymphoblastic leukemia reveals efficacy of romidepsin. *Leukemia* 2016 Jul 22.

8. Lavallee VP, Baccelli I, Krosl J, Wilhelm B, Barabe F, Gendron P*, et al.* The transcriptomic landscape and directed chemical interrogation of MLL-rearranged acute myeloid leukemias. *Nat Genet* 2015 Sep; **47**(9)**:** 1030-1037.

9. Casero D, Sandoval S, Seet CS, Scholes J, Zhu Y, Ha VL*, et al.* Long non-coding RNA profiling of human lymphoid progenitor cells reveals transcriptional divergence of B cell and T cell lineages. *Nat Immunol* 2015 Dec; **16**(12)**:** 1282-1291.

10. Anders L, Guenther MG, Qi J, Fan ZP, Marineau JJ, Rahl PB*, et al.* Genome-wide localization of small molecules. *Nat Biotechnol* 2014 Jan; **32**(1)**:** 92-96.

11. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* 2014 Sep; **32**(9)**:** 896-902.

12. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014; **15**(2)**:** R29.

13. Francis RW, Thompson-Wicking K, Carter KW, Anderson D, Kees UR, Beesley AH. FusionFinder: a software tool to identify expressed gene fusion candidates from RNA-Seq data. *PLoS One* 2012; **7**(6)**:** e39987.

14. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. *Am J Hum Genet* 2013 Oct 3; **93**(4)**:** 641-651.

15. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L*, et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012 Mar; **22**(3)**:** 568-576.

16. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010 Sep; **38**(16)**:** e164.

17. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* 2013 Jan; **Chapter 7:** Unit7 20.

18. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009; **4**(7)**:** 1073-1081.

19. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010 Aug; **7**(8)**:** 575-576.

20. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res* 2009 Sep; **19**(9)**:** 1553-1561.

21. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 2005 Jul; **15**(7)**:** 901-913.

22. Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J*, et al.* Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods* 2010 Apr; **7**(4)**:** 250-251.

23. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014 Mar; **46**(3)**:** 310-315.

24. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T*, et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016 Aug 18; **536**(7616)**:** 285-291.

25. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J*, et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 2017 Jan 4; **45**(D1)**:** D777-D783.

26. 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 Feb 22; **17:** 31.

27. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q*, et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* 2015 Jan; **26**(1)**:** 64-70.

28. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB*, et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 2010 May; **28**(5)**:** 495-501.

29. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H*, et al.* The Protein Data Bank. *Nucleic Acids Res* 2000 Jan 01; **28**(1)**:** 235-242.

30. He J, Abdel-Wahab O, Nahas MK, Wang K, Rampal RK, Intlekofer AM*, et al.* Integrated genomic DNA/RNA profiling of hematologic malignancies in the clinical setting. *Blood* 2016 Jun 16; **127**(24)**:** 3004-3014.

31. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C*, et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat Biotechnol* 2016 Feb; **34**(2)**:** 155-163.

32. Downing JR, Wilson RK, Zhang J, Mardis ER, Pui CH, Ding L*, et al.* The Pediatric Cancer Genome Project. *Nat Genet* 2012 May 29; **44**(6)**:** 619-622.

33. Pickrell JK, Gilad Y, Pritchard JK. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* 2012 Mar 16; **335**(6074)**:** 1302; author reply 1302.

34. Chang VY, Basso G, Sakamoto KM, Nelson SF. Identification of somatic and germline mutations using whole exome sequencing of congenital acute lymphoblastic leukemia. *BMC Cancer* 2013; **13:** 55.

35. Henderson MJ, Choi S, Beesley AH, Baker DL, Wright D, Papa RA*, et al.* A xenograft model of infant leukaemia reveals a complex MLL translocation. *Br J Haematol* 2008 Mar; **140**(6)**:** 716-719.

36. Warren JJ, Pohlhaus TJ, Changela A, Iyer RR, Modrich PL, Beese LS. Structure of the human MutSalpha DNA lesion recognition complex. *Mol Cell* 2007 May 25; **26**(4)**:** 579-592.

37. Campbell BB, Light N, Fabrizio D, Zatzman M, Fuligni F, de Borja R*, et al.* Comprehensive Analysis of Hypermutation in Human Cancer. *Cell* 2017 Nov 16; **171**(5)**:** 1042-1056 e1010.

38. Shlien A, Campbell BB, de Borja R, Alexandrov LB, Merico D, Wedge D*, et al.* Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. *Nat Genet* 2015 Mar; **47**(3)**:** 257-262.

39. Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P*, et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet* 2013 Feb; **45**(2)**:** 136-144.

40. Jha V, Bian C, Xing G, Ling H. Structure and mechanism of error-free replication past the major benzo[a]pyrene adduct by human DNA polymerase kappa. *Nucleic Acids Res* 2016 Jun 2; **44**(10)**:** 4957-4967.

41. Szpiech ZA, Strauli NB, White KA, Ruiz DG, Jacobson MP, Barber DL*, et al.* Prominent features of the amino acid mutation landscape in cancer. *PLoS One* 2017; **12**(8)**:** e0183273.

42. Stam RW, Schneider P, Hagelstein JA, van der Linden MH, Stumpel DJ, de Menezes RX*, et al.* Gene expression profiling-based dissection of MLL translocated and MLL germline acute lymphoblastic leukemia in infants. *Blood* 2010 Apr 8; **115**(14)**:** 2835-2844.

43. Kanemaru Y, Suzuki T, Sassa A, Matsumoto K, Adachi N, Honma M*, et al.* DNA polymerase kappa protects human cells against MMC-induced genotoxicity through error-free translesion DNA synthesis. *Genes Environ* 2017; **39:** 6.

44. Maddukuri L, Ketkar A, Eddy S, Zafar MK, Eoff RL. The Werner syndrome protein limits the error-prone 8-oxo-dG lesion bypass activity of human DNA polymerase kappa. *Nucleic Acids Res* 2014 Oct 29; **42**(19)**:** 12027-12040.

45. Irimia A, Eoff RL, Guengerich FP, Egli M. Structural and functional elucidation of the mechanism promoting error-prone synthesis by human DNA polymerase kappa opposite the 7,8-dihydro-8-oxo-2'-deoxyguanosine adduct. *J Biol Chem* 2009 Aug 14; **284**(33)**:** 22467-22480.

46. Yadav S, Mukhopadhyay S, Anbalagan M, Makridakis N. Somatic Mutations in Catalytic Core of POLK Reported in Prostate Cancer Alter Translesion DNA Synthesis. *Hum Mutat* 2015 Sep; **36**(9)**:** 873-880.

47. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S*, et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* 2015 Dec; **47**(12)**:** 1402-1407.

48. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV*, et al.* Signatures of mutational processes in human cancer. *Nature* 2013 Aug 22; **500**(7463)**:** 415-421.

49. Mathijs A Sanders EC, Christoffer Flensburg, Annelieke Zeilemaker, Sarah E Miller, Adil al Hinai, Ashish Bajel, Bram Luiken, Melissa Rijken, Tamara Mclennan, Remco M Hoogenboezem, François G Kavelaars, Marnie E Blewitt, Eric M Bindels, Warren S Alexander, Bob Löwenberg, Andrew W Roberts, Peter J M Valk, Ian Majewski. Germline loss of MBD4 predisposes to leukaemia due to a mutagenic cascade driven by 5mC. *bioRxiv* 2018.

50. Chen L, Liu P, Evans TC, Jr., Ettwiller LM. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* 2017 Feb 17; **355**(6326)**:** 752-756.

51. Lin J, Shi T. Error-prone DNA polymerase and oxidative stress increase the incidences of A to G mutations in tumors. *Oncotarget* 2017 Jul 11; **8**(28)**:** 45154-45163.

52. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P*, et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* 2017 Nov 16; **171**(5)**:** 1029-1041 e1021.

53. Valentine MC, Linabery AM, Chasnoff S, Hughes AE, Mallaney C, Sanchez N*, et al.* Excess congenital non-synonymous variation in leukemia-associated genes in MLL- infant leukemia: a Children's Oncology Group report. *Leukemia* 2013 Dec 4.

54. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol* 2017 Nov 8.

55. van Oers JM, Edwards Y, Chahwan R, Zhang W, Smith C, Pechuan X*, et al.* The MutSbeta complex is a modulator of p53-driven tumorigenesis through its functions in both DNA double-strand break repair and mismatch repair. *Oncogene* 2014 Jul 24; **33**(30)**:** 3939-3946.

56. Burdova K, Mihaljevic B, Sturzenegger A, Chappidi N, Janscak P. The Mismatch-Binding Factor MutSbeta Can Mediate ATR Activation in Response to DNA Double-Strand Breaks. *Mol Cell* 2015 Aug 20; **59**(4)**:** 603-614.

57. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S*, et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* 2015 May 22; **348**(6237)**:** 880-886.

Figure Legends

Figure 1: Characterization of germline and somatic point mutations identifies an ultra-mutated infant leukemia case. (a) Overlap of variants identified from diagnostic leukemia specimen by RNA-seq and whole exome sequencing (WES). (b) Overlap of variants called in diagnostic RNA-seq and remission bone marrow WES. (c) Overlap of variants called in diagnostic and remission WES. Barplots colored showing overlap binned by derived allele frequency as follows: present with AF<0.5% in ExAC (green), present with AF 0.5-1% in ExAC (yellow), present with AF>1% in ExAC (purple), absent from the ExAC (orange), non-overlapping variants are colored to show sites with coverage <20 reads (grey) or coverage >20 reads (black) in the comparator. (d) Scatterplots of diagnostic leukemia (x-axis) versus remission (y-axis) variant allele fraction (VAF) from WES for coding variants detected in diagnostic leukemia RNA-seq.

Figure 2: Clonal analysis of somatic ultra-mutations. Histograms showing the VAF density distribution of somatic mutations (left, x-axis: dark shaded bars) or LOH-sites (right; x-axis: light shaded bars) in diagnostic leukemia WES from canonical cases (a) and the ultra-mutated case (b). (c) Illustration of cell lines derived from ultra-mutated specimen and histograms displaying the VAF density distribution of somatic mutations (left, x-axis: dark shaded bars) or LOH-sites (right; x-axis: light shaded bars) in cell lines (PER-784A: red; PER-826A: green). (d) Venn diagram displaying the overlap in somatic mutations (left) and LOH-sites (right) identified by *VarScan2* comparing the primary diagnostic specimen and derived cell lines (PER-784A: red; PER-826A: green) to the primary remission sample. (e) Somatic mutation and (f) LOH coding sites showing total read coverage (left y-axis) and VAF (right y-axis) from primary specimens (left panels) and cell lines (right panels) ordered by decreasing VAF in the diagnostic sample (x-axis).

Figure 3: Germline-like somatic signatures of ultra-mutations. (a-d) Histograms showing the number (a-b) or frequency (c-d) of events overlapping the spectrum of ExAC derived allele frequency bins with 5% increments. (a) Somatic mutations called by *VarScan2* grouping homozygous and heterozygous events separately. (b) Loss-of-heterozygosity sites called by *VarScan2* grouping reference→alternative and alterantive→reference sites separately. (c-d) All quality filtered coding variants from each of the canonical patients or P337 ultra-mutated specimen called by GATK grouped by single nucleotide polymorphisms (SNPs) that were homozygous and heterozygous events from (c) remission sample or (d) diagnostic sample. (e) Normalized fraction of variants within unique trinucleotide context (ordered alphabetically on the x-axis) for *VarScan2* calls (left panel; as shown for panels a-b), or remission variants (right panel; as shown for panel c). (f) Somatic signatures identified by *deconstructSigs* for categories of mutations/SNPs (displayed in panel e) showing their relative proportions (bar plot) and pair-wise Pearson's correlation coefficients (heatmap), ordered by unsupervised hierarchical clustering.

Figure 4: Ultra-mutations target rare and common human alleles including a novel *MSH2* allele and a rare *POLK* allele. (a) Proportion of coding single nucleotide variants in diagnostic (D) and remission (R) whole exome sequencing (WES) data binned by derived allele frequency as follows: present with AF<0.5% in ExAC (green), present with AF 0.5-1% in ExAC (yellow), present with AF>1% in ExAC (purple), absent from the ExAC (orange). Grey shading denotes data from the ultra-mutated specimen. The distribution of variants called as germline, somatic or LOH by *VarScan2* are displayed for comparison. (b) Number of variants (y-axis) called from RNA-seq reads from St Jude and Perth cohorts (open circles) are plotted with the total variants detected in the ultra-mutated sample (shaded circles), germline-encoded variants (shaded triangle pointing up) and somatic point mutations

(shaded triangle pointing down) binned according to the allele count in ExAC database (red: not in ExAC; green < 0.5%; yellow: 0.5-1% and blue > 0.5%). (c) Bar plot displaying fraction of variants within rare, low-frequency and common allele frequency bins (as described for panel a) ordered by increasing proportion of common alleles (i.e. derived AF>1%) and excluding outliers (SJINF018, SJINF021, SJINF010 and SJINF008). Proportion of RNA-seq germline and somatic variants identified by overlap with remission/diagnostic WES data are displayed for comparison. (d) Crystal structure of the MutS lesion recognition complex (PDB CODE 2O8D[36]) MSH2:MSH6 mismatch repair proteins (MSH2: green, MSH6: blue) bound to ADP and G:T mismatch DNA (orange) showing the location of the amino acid change (NM_000251:exon9:c.G1465A:p.E489K). (e) Crystal structure of Pol k (green and blue representing different molecules) in complex with a bulky DNA adduct (orange) (PDB CODE 4U6P[40]) showing locations of the amino acid change (NM_016218:exon7:c.G893A:p.R298H).

Figure 5: Global transcriptional concordance of ultra-mutated infant leukemia. Multi-dimensional scaling of RNA-seq count data from acute leukemias and normal bone marrow progenitor populations displaying translocation partner gene by color and blood/leukemia type by symbol shape according to the legend. Labels show replicate measurements from the ultra-mutated diagnostic specimen, P337.

Table 1: Characterisation of somatic lesions in *KMT2A*-R patient samples.

Patient ID	KMT2A-fusion	Reciprocal KMT2A-fusion	Diagnostic leukemia DNA	Remission DNA	Point mutations 0.1 > VAF[1] < 0.3	Point mutations VAF > 0.3	Indels VAF > 0.3	LOH[2]-sites	Point mutations or hotspots with RNA-SEQ reads
P337	*KMT2A-MLLT1*	Not detected	Exome	Exome	198	5,054	260	6,555	> 1,000 expressed somatic mutations
P401	*KMT2A- MLLT 3*	Not detected	Exome	Exome	78	3	9	45	No non-silent somatic mutations
P399	*KMT2A-AFF1*	*AFF1-KMT2A*	Exome	Exome	95	3	7	67	No non-silent somatic mutations
P438	*KMT2A- MLLT3*	Not detected	Exome	Exome	74	4	9	100	NRAS.p.G12S
P810	*KMT2A-EPS15*	*EPS15-KMT2A*	Exome	Exome	95	4	5	46	PER3.p.M1006R, PER3.p.K1007E
P809	*KMT2A-EPS15*	*EPS15-KMT2A*	Exome	Exome	69	3	3	28	PER3 M1006R, PER3.p.K1007E
P848	*KMT2A-AFF1*	*AFF1-KMT2A*	no	Exome	NA	NA	NA	NA	KRAS.p.A146T
P287	*KMT2A-AFF1*	*AFF1-KMT2A*	no	no	NA	NA	NA	NA	No COSMIC[3] hotspots
P706	*KMT2A- MLLT1*	Not detected	no	Exome	NA	NA	NA	NA	KRAS.p.A146T
P272	*KMT2A-AFF1*	Not detected	no	no	NA	NA	NA	NA	No COSMIC[3] hotspots

[1]VAF: Variant allele fraction; [2]LOH: Loss-of-heterozygosity; [3]COSMIC: Catalogue of Somatic Mutations in Cancer

Table 2: Mean and standard deviation of expressed rare (Not in ExAC or <0.5% allele frequency), low-frequency (0.5-1% allele frequency) and common (>1% allele frequency) alleles in 39 *KMT2A*-R patient samples compared to ultra-mutated (P337) sample and its constituent expressed germline alleles and somatic mutations.

	Not in ExAC	**<0.5% AF[1] in ExAC**	**0.5-1% AF in ExAC**	**>1% AF in ExAC**
KMT2A-R iALL	42.3±10.4	212.1±32.6	76.8±14.7	7722.3±266.0
P337 expressed	38	174	89	7590
P337 Germline	17	88	47	6235
P337 Somatic	21	86	42	1355

[1]AF: allele frequency; [2]ExAC: Exome aggregation consortium

FIGURE 1

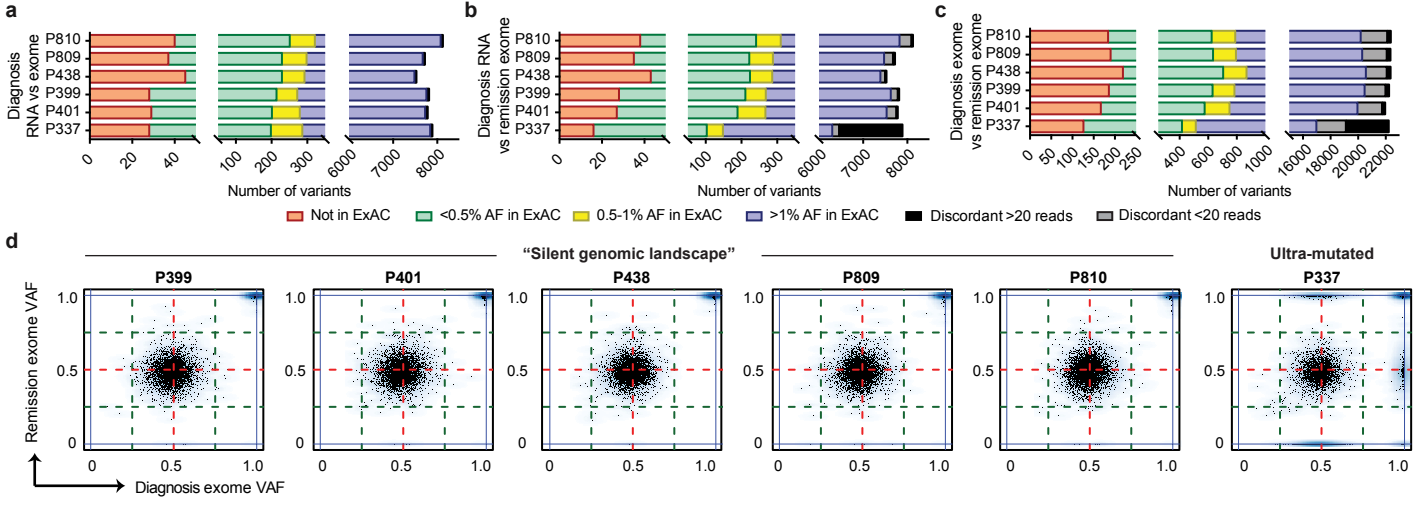

a
Diagnosis RNA vs exome
P810
P809
P438
P399
P401
P337
0
20
40
100
200
300
6000
7000
8000
Number of variants
b
Diagnosis RNA vs remission exome
P810
P809
P438
P399
P401
P337
0
20
40
100
200
300
6000
7000
8000
Number of variants
c
Diagnosis exome vs remission exome
P810
P809
P438
P399
P401
P337
0
50
100
150
200
250
400
600
800
1000
16000
18000
20000
22000
Number of variants
Not in ExAC
<0.5% AF in ExAC
0.5-1% AF in ExAC
>1% AF in ExAC
Discordant >20 reads
Discordant <20 reads
d
"Silent genomic landscape"
Ultra-mutated
P399
P401
P438
P809
P810
P337
Remission exome VAF
Diagnosis exome VAF
0
0.5
1.0

FIGURE 2

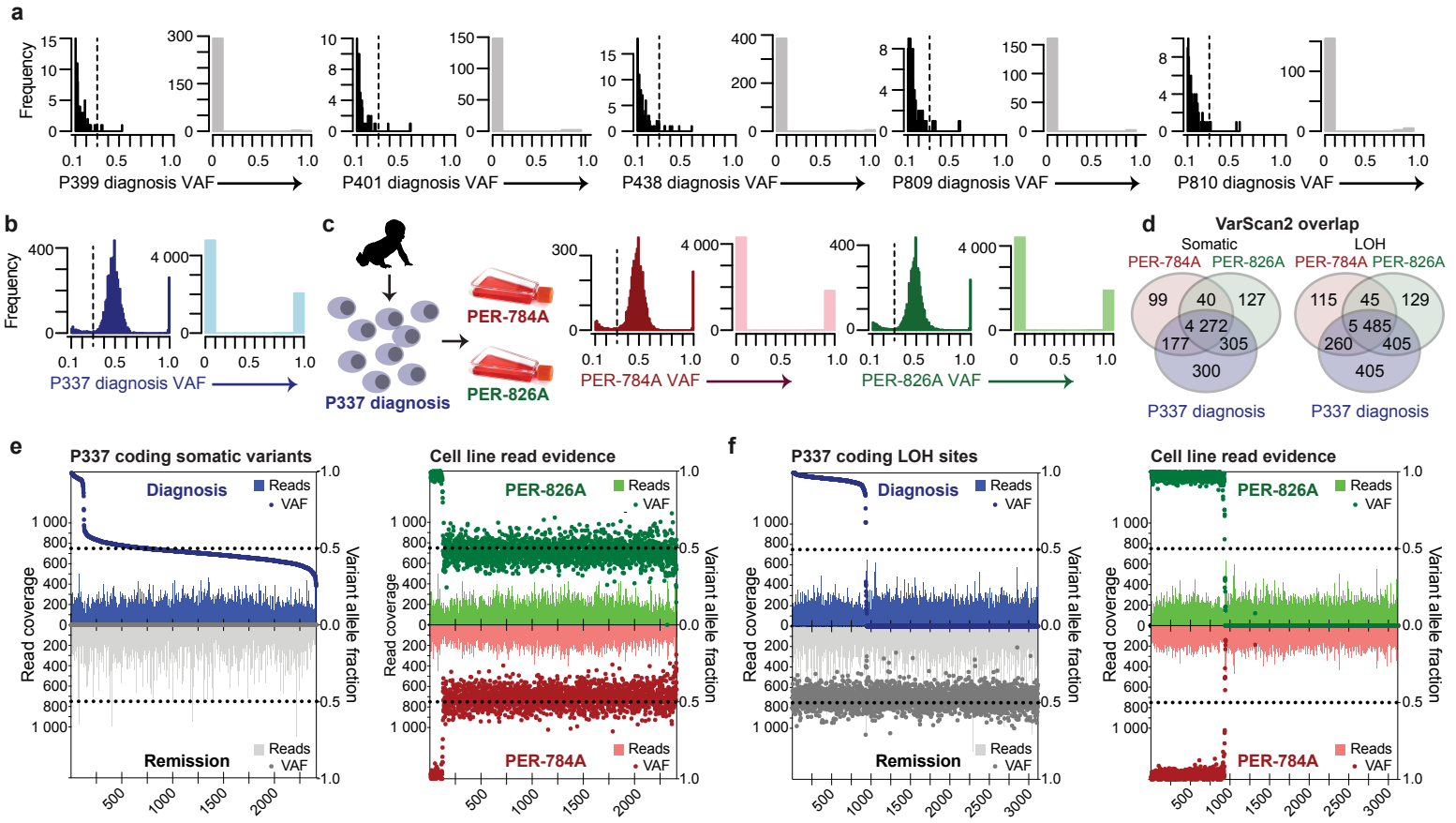

a
Frequency
P399 diagnosis VAF
P401 diagnosis VAF
P438 diagnosis VAF
P809 diagnosis VAF
P810 diagnosis VAF
b
Frequency
P337 diagnosis VAF
c
P337 diagnosis
PER-784A
PER-826A
PER-784A VAF
PER-826A VAF
d
VarScan2 overlap
Somatic
LOH
PER-784A
PER-826A
99
40
127
4 272
177
305
300
115
45
129
5 485
260
405
405
P337 diagnosis
P337 diagnosis
e
P337 coding somatic variants
Diagnosis
Reads
VAF
Remission
Read coverage
Variant allele fraction
Cell line read evidence
PER-826A
PER-784A
f
P337 coding LOH sites
Diagnosis
Remission
Cell line read evidence
PER-826A
PER-784A

FIGURE 3

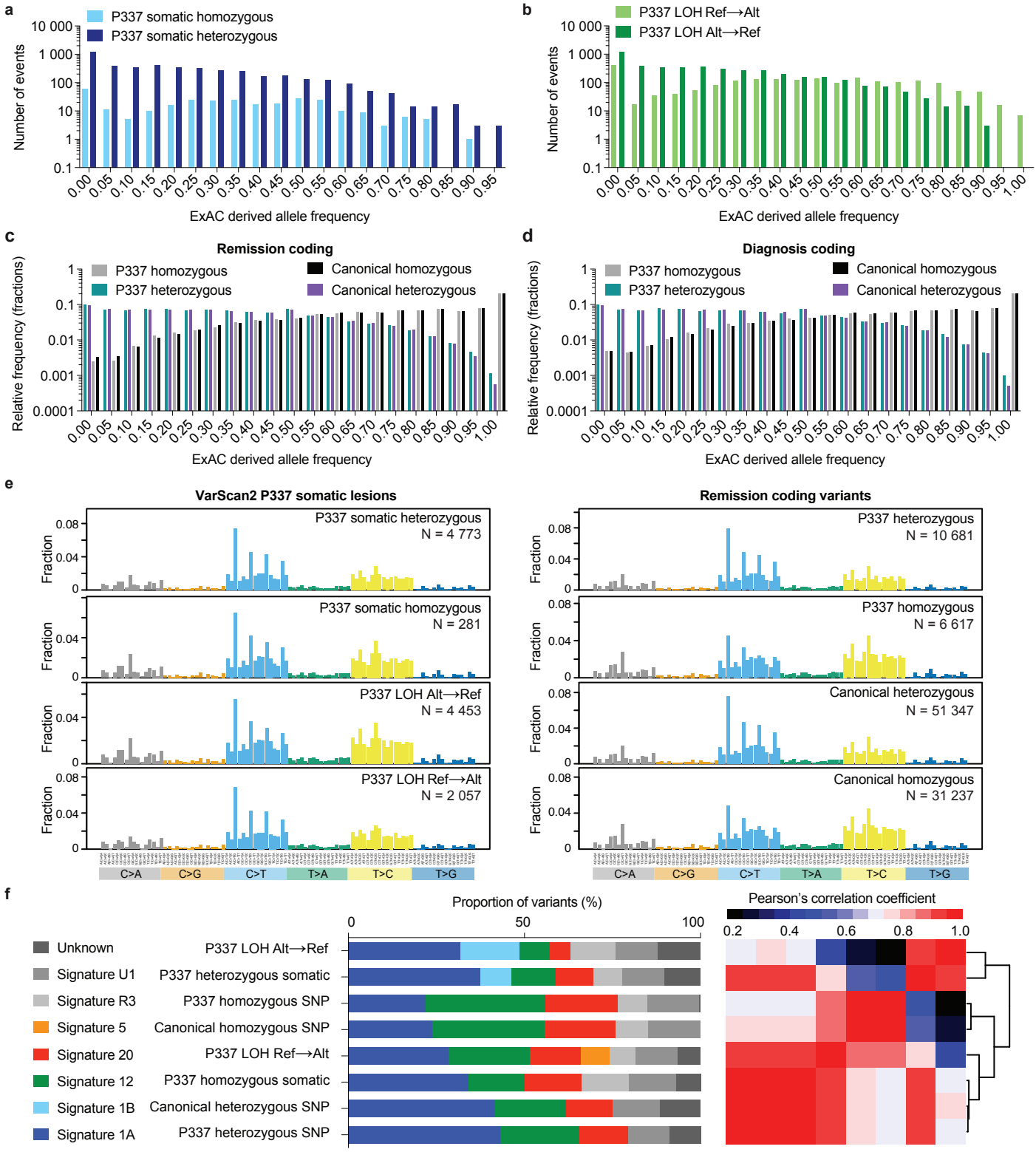

a
P337 somatic homozygous
P337 somatic heterozygous
Number of events
ExAC derived allele frequency
b
P337 LOH Ref→Alt
P337 LOH Alt→Ref
Number of events
ExAC derived allele frequency
c
Remission coding
P337 homozygous
P337 heterozygous
Canonical homozygous
Canonical heterozygous
Relative frequency (fractions)
ExAC derived allele frequency
d
Diagnosis coding
P337 homozygous
P337 heterozygous
Canonical homozygous
Canonical heterozygous
Relative frequency (fractions)
ExAC derived allele frequency
e
VarScan2 P337 somatic lesions
P337 somatic heterozygous
N = 4 773
P337 somatic homozygous
N = 281
P337 LOH Alt→Ref
N = 4 453
P337 LOH Ref→Alt
N = 2 057
Fraction
C>A
C>G
C>T
T>A
T>C
T>G
Remission coding variants
P337 heterozygous
N = 10 681
P337 homozygous
N = 6 617
Canonical heterozygous
N = 51 347
Canonical homozygous
N = 31 237
f
Proportion of variants (%)
Unknown
Signature U1
Signature R3
Signature 5
Signature 20
Signature 12
Signature 1B
Signature 1A
P337 LOH Alt→Ref
P337 heterozygous somatic
P337 homozygous SNP
Canonical homozygous SNP
P337 LOH Ref→Alt
P337 homozygous somatic
Canonical heterozygous SNP
P337 heterozygous SNP
Pearson's correlation coefficient

FIGURE 4

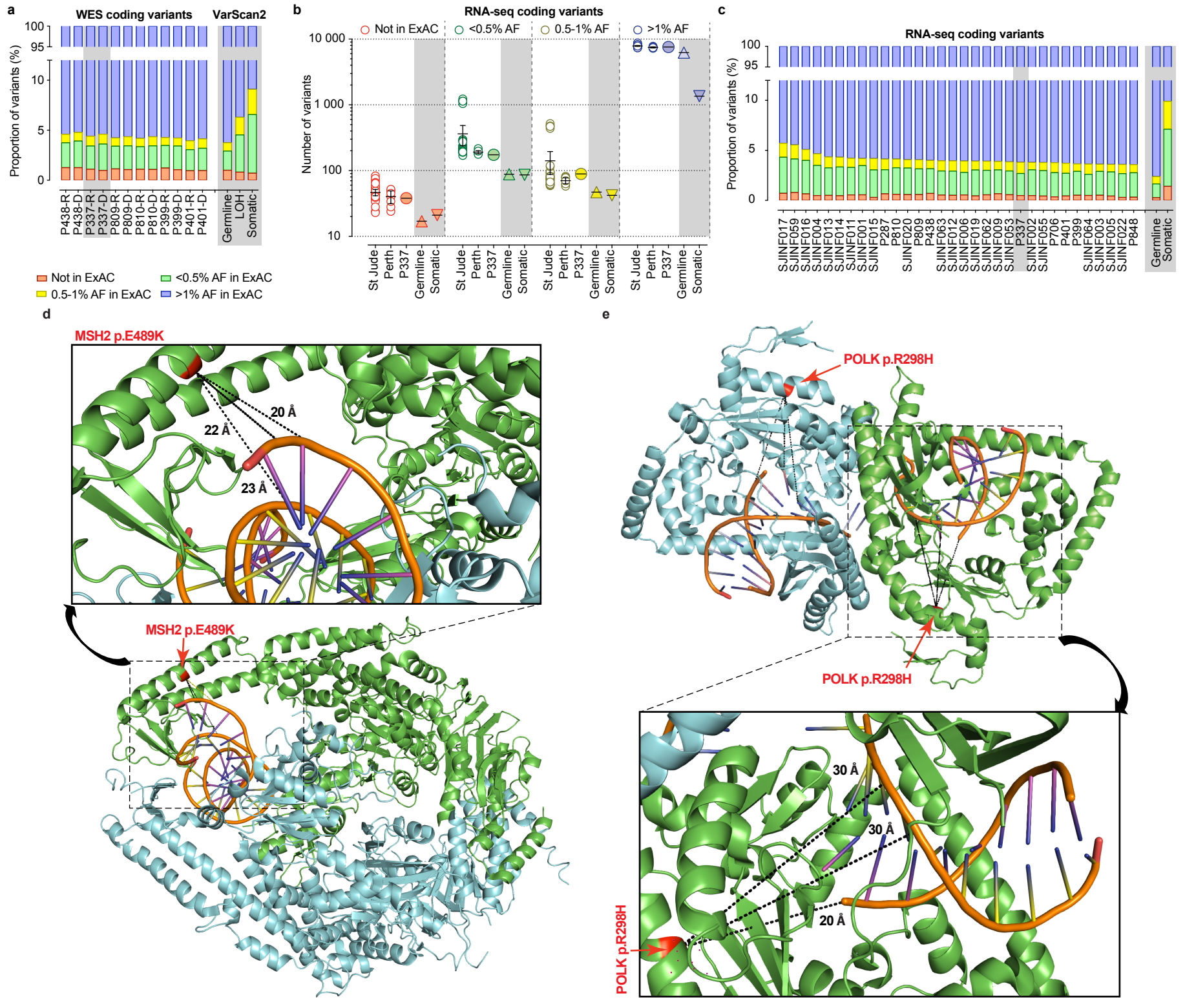
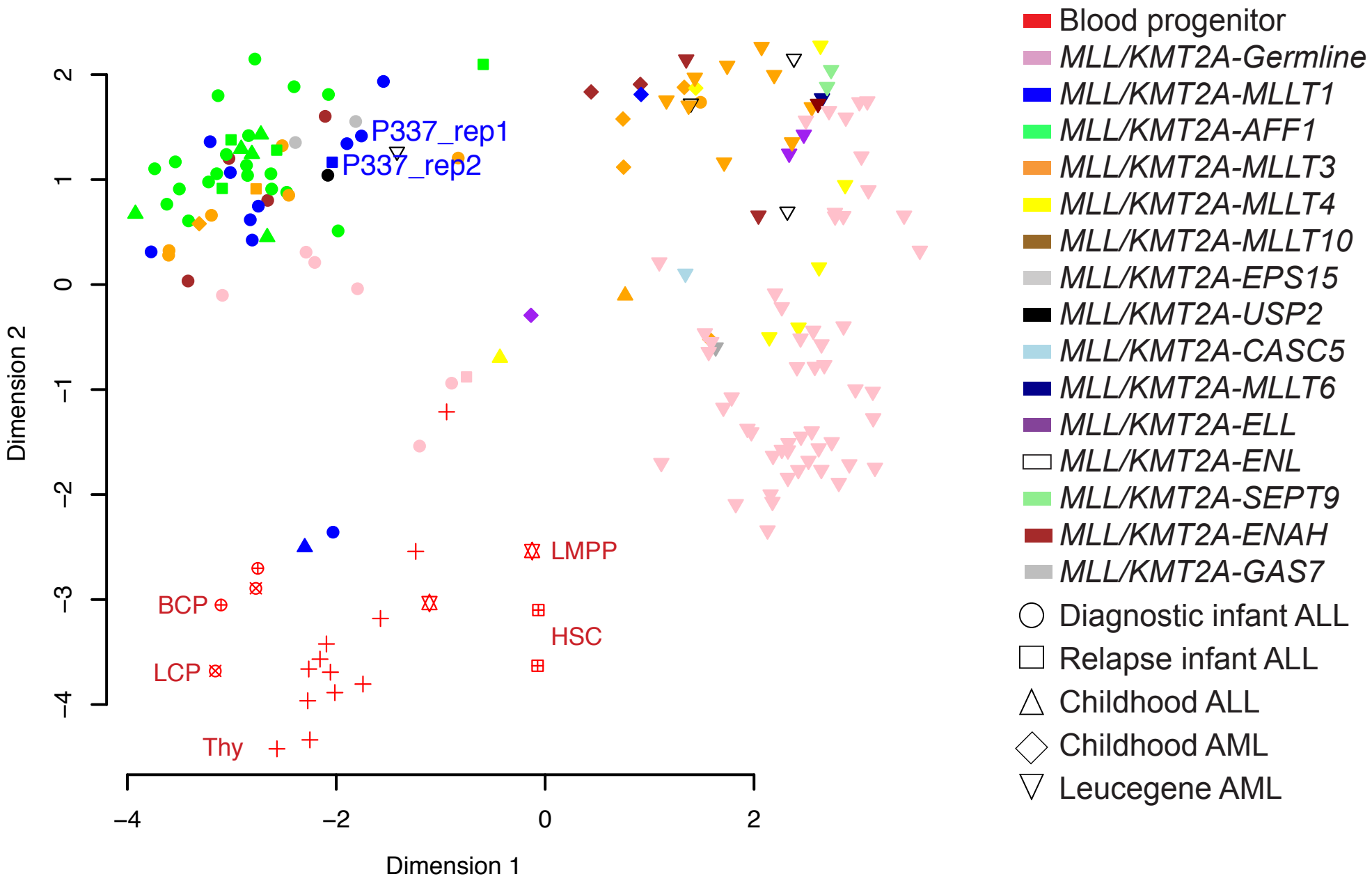

a
WES coding variants
VarScan2
Proportion of variants (%)
P438-R
P438-D
P337-R
P337-D
P809-R
P809-D
P810-R
P810-D
P399-R
P399-D
P401-R
P401-D
Germline
LOH
Somatic
Not in ExAC
<0.5% AF in ExAC
0.5-1% AF in ExAC
>1% AF in ExAC
b
RNA-seq coding variants
Not in ExAC
<0.5% AF
0.5-1% AF
>1% AF
Number of variants
St Jude
Perth
P337
Germline
Somatic
c
RNA-seq coding variants
Proportion of variants (%)
SJINF017
SJINF059
SJINF016
SJINF004
SJINF013
SJINF014
SJINF011
SJINF001
SJINF015
P287
P810
SJINF020
P809
P438
SJINF063
SJINF012
SJINF006
SJINF019
SJINF062
SJINF009
SJINF053
P337
SJINF002
SJINF055
P706
P401
P399
SJINF064
SJINF003
SJINF005
SJINF022
P848
Germline
Somatic
d
MSH2 p.E489K
20 Å
22 Å
23 Å
MSH2 p.E489K
e
POLK p.R298H
POLK p.R298H
30 Å
30 Å
20 Å
POLK p.R298H

FIGURE 5


P337_rep1
P337_rep2
LMPP
HSC
BCP
LCP
Thy
Dimension 1
Dimension 2
−4
−2
0
2
−4
−3
−2
−1
0
1
2
Blood progenitor
MLL/KMT2A-Germline
MLL/KMT2A-MLLT1
MLL/KMT2A-AFF1
MLL/KMT2A-MLLT3
MLL/KMT2A-MLLT4
MLL/KMT2A-MLLT10
MLL/KMT2A-EPS15
MLL/KMT2A-USP2
MLL/KMT2A-CASC5
MLL/KMT2A-MLLT6
MLL/KMT2A-ELL
MLL/KMT2A-ENL
MLL/KMT2A-SEPT9
MLL/KMT2A-ENAH
MLL/KMT2A-GAS7
Diagnostic infant ALL
Relapse infant ALL
Childhood ALL
Childhood AML
Leucegene AML