Cancer drivers and clonal dynamics in acute lymphoblastic leukaemia subtypes

James Studd (james.studd@icr.ac.uk)
Institute of Cancer Research https://orcid.org/0000-0002-7157-754X
Alex Cornish
Institute of Cancer Research https://orcid.org/0000-0002-3966-3501
Phuc Hoang
The Institute of Cancer Research https://orcid.org/0000-0002-9265-8525
Philip Law
Institute of Cancer Research https://orcid.org/0000-0001-9663-4611
Richard Houlston
The Institute of Cancer Research https://orcid.org/0000-0002-5268-0242

Article

Keywords: Acute lymphoblastic leukaemia, paediatric, whole genome sequencing, driver mutations, non-coding variation, clonality

Posted Date: April 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-366981/v1

Abstract

To obtain a comprehensive picture of composite genetic drivers events and clonal dynamics in subtypes of paediatric acute lymphoblastic leukaemia (ALL) we analysed tumour-normal whole genome sequencing and expression data from 361 newly diagnosed patients. We report the identification of both novel coding and structural drivers as well as recurrent non-coding variation in promoters and *cis*-regulatory regions. The transcriptional profile of histone gene cluster 1 and *CTCF* altered tumours shared hallmarks of hyperdiploid ALL suggesting a 'hyperdiploid like' subtype. ALL subtypes are driven by distinct mutational processes with AID mutagenesis being confined to *ETV6-RUNX1* tumours. Subclonality is a ubiquitous feature of ALL, consistent with Darwinian evolution driving selection and expansion of tumours. Driver mutations in B-cell developmental genes (*IKZF1, PAX5, ZEB2*) tend to be clonal and RAS/RTK mutations subclonal. In addition to identifying new avenues for therapeutic exploitation this analysis highlights that targeted therapies should take into account composite mutational profile and clonality.

Introduction

Acute lymphoblastic leukaemia (ALL) is the most common childhood cancer, with around 80% of ALL cases derived from B-cell precursors (BCP-ALL)[1]. The disease is characterised by initiating genetic lesions resulting in characteristic patterns of whole chromosome gain (hyperdiploidy) or loss (hypodiploidy) and the formation of specific fusion genes. Recurrent fusion genes include t(12;21) *ETV6-RUNX1*, t(1;19) *TCF3-PBX1* and t(9;22) *BCR-ABL1*[1]. In addition to these events, copy number changes in *RUNX1* as a consequence of intra-chromosomal amplification (iAMP21), and *ERG* deletion, have more recently been recognised as initiating events[2]. The biological differences between these subtypes is reflected in their clinical behaviour[3–5].

Current first line therapy for ALL is dominated by use of chemotherapeutic and steroidal agents. While their use has driven 5-year survival to > 90%[6] this is at the expense of significant morbidity. Despite these improvements, survival for relapsed ALL is still only 21%-39%[7,8]. Strategies for developing novel therapies for ALL have largely focused on monoclonal antibodies or CAR-T cells. Such therapies are expensive; for example when licenced the anti-cd19 monoclonal blinatumomab was the most expensive cancer therapy ever brought to market[9]. It is therefore desirable to develop additional targeted small molecule therapies to reduce treatment associated morbidity and relapse associated treatment cost. Such developments are likely to require more precise molecular characterisation and risk stratification; both informed by our understanding of ALL genomics.

Precancerous lesions harbouring initiating events can be undetected for years usually requiring the acquisition of additional genetic lesions for symptomatic disease. Most commonly secondary lesions impact genes regulating the cell cycle (*CDKN2A*, *RB1*), B-cell development (*PAX5*, *IKZF1, EBF1*) and the

RAS/RTK pathway (*NRAS, KRAS, FTL1*)[1]. However, the full complement of molecular lesions sufficient to cause ALL, and explain its diversity is unknown.

To obtain a more comprehensive picture of the composite genetic events acting in concert in each of the BCP-ALL subtypes, we performed a genomic analysis of diagnostic samples from 361 ALL patients (**Supplementary Fig. 1**). As well as novel coding, we identify non-coding and copy number drivers. Our analysis also reveals differences in the mutational and biological pathways influencing the initiation and progression of disease subtypes.

Methods

Cases, data and sequencing

Matched tumour-normal whole genome sequencing (WGS) data from 361 treatment-naïve cases of paediatric (< 18 years old) BCP-ALL were obtained from St. Jude Research Hospital (https://www.stjude.cloud/). Data were accessed and analysed through the DNAnexus cloud computing platform. Ethical permission was not required as all data were in the public domain.

WGS data were generated using 100bp paired-end libraries sequenced to an average read depth of 45× and 62× for normal and tumour samples respectively, using Illumina (San Diego, USA) HiSeq2000 technology. Raw sequencing data were aligned with BWAmem v0.7.17[10] to GRCh38. Alignments were performed by Google Genomics. Cross contamination was assessed using GATK v4.0.0.1; no sample having > 2.6%. Tumour RNA sequencing (RNA-seq) on 222 (post quality control [QC]) of the 361 cases was performed on 125bp paired-end libraries using Illumina HiSeq technology to an average number of 55 x 10^6 reads. RNA-seq fastq files were analysed using FastQC and aligned to GRCh38 using STAR v2.6.1[11], discarding samples with < 20% of reads aligning to the genome. Transcript abundance was calculated in transcripts per million (TPM) using RSEM v1.3.0[12] based on GENCODEv30 annotation and was adjusted for batch effects using ComBat-seq[13].

Fusion genes were identified from RNA-seq data using STAR-Fusion v1.5.018 and FusionInspector[14]. Candidate fusions were retained when fusion genes were separated by > 1MB and fusions were absent from normal blood or bone marrow (1000 Genomes [n = 465] and HPA [n = 200] RNA sequencing projects).

The transcriptional impact of histone gene cluster 1 (chr6:26122685–26239852) deletion and CTCF alteration (deletion or mutation) were assessed using DESeq2 v1.329[15], with default settings. Differentially expressed genes were identified after removing tumours with: (1) > 2 copies of the corresponding interval; (2) Alteration of both *CTCF* and the histone gene cluster 1 (n = 1). This resulted in 9 tumours with *CTCF* alterations and 8 tumours with histone cluster deletion.

Variant calling

Somatic single nucleotide variants (SNVs) and indels were called using Strelka v2.8.4[16] adopting default parameters. QC filtering of somatic variants comprised: (1) Retaining only variants marked as 'PASS'; (2) Excluding variants seen in panel of 160 matched germline samples; (3) Excluding variants in repetitive regions (extracted from UCSC) or in homopolymer runs of > 7 nucleotides; (4) Excluding variants with a POPMAX allele frequency > 0.001 in GnomAD v3. *In-silico* predictions of variant effects on protein function were made using Polyphen2 HumDif scores as per[17]. Driver mutation plot generated using Maftools[18].

Mutational signatures

De novo extraction of signatures was performed using SigProfilerExtractor v1.0.18[19]. Extracted signatures were assigned to reference signatures from Catalogue of Somatic Mutations in Cancer (COSMIC) v3.1 using a cosine similarity threshold of 0.9. Mutation enrichment testing was performed on tumours with > 30% of mutations from either SBS7a or SBS2 and SBS13 and accounted for subtype applying Benjamini-Hochberg correction.

Tumour subtyping

Tumour chromosomal ploidy was based on copy number data. Tumours with a total chromosome number > 50 were classified as hyperdiploid and those with < 45, hypodiploid. Near haploid tumours (n = 11) chromosome number 24–30 are included the hypodiploid subtype unless otherwise stated. iAMP21 status was called as *per* Harrison *et al* [20] on the basis of chromosome 21 ploidy and *RUNX1* copy number. Subtypes defined by driver fusion events (*e.g. ETV6-RUNX1, BCR-ABL1*) were assigned on the basis of a clonal SVs consistent with fusion gene expression and corresponding RNA-seq identified fusions. Cases without an established initiating driver event were designated as unclassified/other. The subtype composition of the cohort is detailed in **Supplementary Fig. 2**.

Identification of cancer drivers and Pathways

Identification of SNV/indel drivers in coding regions was based on a consensus-based approach. Per gene *P*-values were calculated combining the output of MutSigCV v1.3.01[21], dndsCV v0.0.1[22] and OncodriveFML[23] using Harmonic means[24] and Benjamini–Hochberg correction. Variants were classified as non-silent using variant effect annotator (VEP)[25] annotations (**Supplementary Table 3**).

Driver CNVs were called using GISTIC2 v2.0.23[26] run in focal mode (excluding arm level events) with default parameters imposing a *Q*-value cut-off of 0.01. Genome regions were excluded if they: (1) overlapped a immunoglobulin locus, (2) contained no protein coding genes, (3) contained no genes expressed in the corresponding RNA-seq data (excluding deleted cases) or (4) the region was both significantly amplified and deleted.

To identify driver mutations in enhancer regions we adopted the strategy of Orlando *et al*[27]. *Cis*-regulatory elements (CREs) were identified from promoter capture chromatin confirmation (PHi-C) contacts (CHiCAGO score > 5) from naïve B-cells[28]. CRE-specific mutation probabilities for each tumour were generated by fitting a logistic regression model using the glm R package, accounting for base

composition, mutation rate, replication timing, and coverage. Mean replication timing was extracted for the lymphoblastoid cell lines (GM12878, GM12813, GM12812, GM12801, GM06990). The Poibin R package was used for approximation of Poisson binomial to derive empiric *P*-values regions as *per* Melton *et al*[29].

Mutational clustering within CREs was tested to infer functionality. Regulatory regions harbouring > 5 mutations were permuted 10,000 times and tested assuming uniform mutation distribution, deriving empiric *P*-values. Frequency and clustering *P*-values were combined using Fisher's method and adjusted for multiple testing using Benjamini–Hochberg correction. Genome regions with a *Q*-value < 0.1 were examined for transcriptional effects. Expression of genes whose promoter were captured by an interaction were compared between mutated and non-mutated samples using Benjamin-Hochberg corrected Wilcox rank-sum test. Tumours with CNVs at either the target gene or CRE were not considered.

Mutation burden in promoters and UTR regions was assessed using OncodriveFML[23]. Promoters (defined from the transcription start site − 2000bp) were extracted from GENCODE v30 GRCh38.p12. Where genes had multiple transcription start sites all promoter sequences were evaluated jointly. Promoters were filtered for any overlapping coding or UTR sequence.

Driver genes were manually assigned to biological pathways. Gene – pathway assignments: RAS/RTK; *NRAS, KRAS, PTPN11, FLT3, NF1, ABL1*. B-cell development; *PAX5, IKZF1, ETV6, ZEB2, RUNX1, TCF3, RAG1, RAG2, EBF1*. Chromatin regulation; *SETD2, HDAC7, NSD2, CTCF, KMT2A, STAG2*, histone gene cluster 1. Cytokine signalling; *JAK2, IL7R, CRLF2*. Gene regulation; *CREBBP, MLLT1, MLLT3, AFF1, BTG1, ERG, TCF4, NCOA6*. Signal transduction; *TBL1XR1, TBL1X, PBX1, PAG1*. Cell cycle regulation; *CDKN2A, CDKN2B, RB1*. Immune regulation; *BTLA, HLA-DRB5*.

Identification of copy number and structural variants

Somatic copy number variation (CNV) was called using CNVkit v0.9.5.3[30]. Tumour WGS data were called against a pooled reference, generated from 45 depth representative matched germline samples (23 male, 22 **female).** CNVkit segment specific coverage log2 ratios were adjusted for tumour cell purity, estimated by cpgBattenberg[31]. Contiguous segments with the **same** copy number were merged. CNVs were annotated as 'arm' level when a copy state occupied > 80% of the mappable length of a chromosome arm differed from chromosome copy number. Other variant regions defined as 'focal'.

Structural variants (SVs) were called using Manta v1.5[32], Lumpy v0.2.13[33] and Delly2 v0.8.1[34]. Manta and Delly2 were run using default parameters. Lumpy was run using the wrapper Smoove v0.2.3. Variants were excluded if they were located in centromeric, telomeric or heterochromatic regions, had a variant allele frequency (VAF) < 0.1, or were seen in a panel of matched normals, generated using the corresponding method. Remaining variants were merged as per Li *et al*[35], retaining only those called by 2 or more methods.

SV cancer cell fractions were estimated using SVclone[36] and SV clustering examined using ClusterSV[35]. Regions of chromothripsis were identified using ShatterSeek v0.4[37], based on thresholds of > 3 adjacent segments of oscillating copy number involving > 5 interleaved SVs. Candidate chromothripsis events were manually reviewed.

SV breakpoint motif enrichment was performed using HOMER v4.10.4[38], by extracting two 100bp sequences (± 50bp) from each breakpoint, excluding SVs where both break points mapped to immunoglobulin regions (**Supplementary Table 1**). HOMER extracted motif sequences are annotated using the most similar sequence, based on Pearson correlation coefficient, from the JASPAR[39] database. Annotated HOMER motifs were further processed with reference to motifs of candidate mutagenic drivers of ALL (**Supplementary Table 2**). Where the Pearson correlation between a HOMER motif and a candidate mutagenic motif exceeded the most similar JASPAR annotation they were substituted. Motifs with a correlation of < 0.85 were excluded from analysis.

To jointly analyse CNV and SV, regions called by GISTIC were additionally filtered, retaining only those with an enrichment of overlapping SVs. For each variant region only simple SVs (not part of a complex rearrangement called by SVClust) of the corresponding type (deletion/amplification) were used. For each GISTIC region chromosome-arm-specific background SV rates were estimated by permuting (n = 1000) arm-specific SVs. *P*-values were then computed as the proportion of permutations where the simulated SVs overlapping the locus was greater than or equal to the number of observed SVs overlapping the locus. Variant regions significantly enriched ($P < 0.01$) for overlapping SVs were retained. Additional copy number variant regions were identified using HMMcopy v1.32 calculating GC and mappability normalised tumour/normal log2 coverage ratios.

Clonality and tumour evolution

Tumour ploidy and SNV cancer cell fractions (CCF) were estimated using cpgBattenberg v3.5.0[31], adopting default parameters with the exception of minimum ploidy, which was thresholded at 1.1. Single nucleotide polymorphisms alleles from the 1000 Genomes Project (v3, GRCh38) were counted in tumour and normal samples, and genotypes phased using impute2[40]. Purity-corrected copy number segments were used to compute SNV/indel CCF estimates and subclones identified by DPClust v2.28[31]. Variants were assigned to most likely clusters. For each tumour the cluster with the highest CCF > 0.9 and < 1.1 was considered clonal, others considered subclonal. Samples were excluded based on the following criteria: (1) a variant cluster with CCF > 1.1; (2) no clonal variant cluster (CCF 0.9–1.1); (3) copy number state-specific SNVs which failed to cluster at predicted VAFs; (4) copy number solutions with homozygous deletions > 3Mb. In the first instance samples were analysed using Battenberg derived purity estimates, when resulting copy number solutions failed QC CCube v1.0[41] estimates were used. 280 samples satisfying QC criteria were retained. Heterogeneity was estimated using the Simpson Index (probability that two individuals/cells, selected from a population/tumour, are from the same species/clone), calculated using VEGAN[42]. Evidence to support neutral evolution was sought using MOBSTER v0.1.1[43], as per authors recommendations (retaining only SNVs and indels in diploid regions).

MOBSTER identifies variants with a VAF distribution consistent with neutral processes, termed a “neutral tail”. Variants belonging to a neutral tail, subclonal or clonal clusters were also analysed using dNdSCV and by calculating mutation rate (number of non-synonymous variants/all non-synonymous sites/total number of mutations) for ALL drivers (**Supplementary Table 8**).

Results

As previously documented, the burden of SNVs and indels was low (median 0.6 Mb^{-1}, range 0.04–5.31) when compared to the majority of solid cancers. Mutation burdens differed significantly across subtypes ($P_{Kruskal-Wallis}$=2.2x10^{-16}), with iAMP21 and KTM2D (MML1) positive tumours having the highest and lowest mutational burdens respectively (Fig. 1A). The most common chromosome-arm level aberrations were loss of 9p (containing *CDKN2A/ CDKN2B*) and gain of 21q (containing *RUNX1*), both occurring in 8% of cases (**Supplementary Fig. 3**). 9p loss preferentially occurred in *TCF3-PBX1* translocated tumours (P_{Fisher}= 0.043), and 21q gain in hypodiploid tumours (P_{Fisher}= 0.012) (**Supplementary Fig. 4A and 4B**).

The median number of SVs was eight per tumour (2x10^{-3} Mb^{-1}), with iAMP21 tumours possessing the highest number (Fig. 1B). The rate of SVs on chromosome 21 (0.027 Mb^{-1}) was 10-fold higher than other chromosomes, largely accounted for by iAMP21 tumours (Fig. 1C). Since iAMP21 tumours are defined by *RUNX1* copy number, we examined the distribution of SVs on chromosome 21, finding no clustering evident (**Supplementary Fig. 5**). Chromothripsis did not account for elevated SV rates in iAMP21 tumours as no events were observed on chromosome 21.

Identification of driver genes

We searched for drivers of ALL by first considering the following classes of somatic coding alterations; single nucleotide variants (SNVs)/indels, copy number variants (CNVs), structural variants (SVs) and loss of heterozygosity (LOH). In addition to established drivers, we identified a number of novel ALL drivers, including *HLA-DRB5*, the histone gene cluster 1, *ZEB2*, *CTCF* and *MAP1B*.

Consistent with previous reports[1,44], the most frequently altered genes included *CDKN2A/B, PAX5, ETV6, ERG, RUNX1, NRAS, KRAS* and *IKZF1* (Fig. 2). By combining CNV and SV data we identified two novel regions of recurrent alteration. Firstly, a 120 kb region of *HLA* (6p21; 32,442,465 – 32,554,750 bps) was deleted in 17% of tumours (Fig. 3A). Within this region only *HLA – DRB5* was expressed and deletion was associated with significantly reduced gene expression (*P* = 3.7x10^{-4}). We further evaluated read depth data in tumours with an HLA SV but no CNV using an additional copy segmentation algorithm[45], finding evidence of a corresponding change number change within 2,000bp an SV breakpoints in every tumour (**Supplementary Fig. 6**). Secondly, a 117 kb region of 6p22.2 overlapping histone gene cluster 1 (26,122,685 – 26,239,852 bps) was deleted in 10% of tumours (Fig. 3B), within which deletion was associated with reduced expression of HIST1H4E (*P* = 0.034) and HIST1H2AE (*P* = 0.023). The cancer cell fraction (CCF) of SVs in the region suggested the majority of these variants are clonal.

Non-silent SNVs or indels in *ZEB2*, *CTCF* and *MAP1B* were seen in 2.2%, 1.7%, and 1.4% of tumours respectively (**Supplementary Table 4 and Supplementary Fig. 7**). *ZEB2* missense mutations were clustered at three base positions, consistent with oncogenic activation (**Supplementary Fig. 8**). A further eight tumours had focal *ZEB2* amplifications. In addition to truncating and damaging mutations in *CTCF*, an additional 20 tumours had *CTCF* deletions, consistent with gene inactivation. None of the *MAP1B* mutations were recurrent and all were predicted to be damaging.

Next we sought to identify non-coding driver mutations. We observed a significant excess of promoter mutations for *BTLA* (4.2%, $Q = 0.002$) and *CHID1* (2.2%, $Q = 0.049$). *BTLA* promoter mutations were clustered within a 27 bp region, and were associated with 5-fold reduced BTLA expression ($P_{\text{Mann-Whitney}}$=0.056), the small number of tumours with corresponding expression data presumably preventing this relationship from attaining significance (Fig. 4A). Mutations were predicted to alter the affinity of various transcription factors (TFs) with evidence of binding from ChIP-sEq. Each *BTLA* promoter mutant possessed a variant predicted to disrupt the binding of an interacting TF, most frequently RUNX1/3, GATA3 and MYB (**Supplementary Table 5**). Of 14 *CHID1* promoter mutations 12 clustered within a 12bp region 1kb upstream of the transcription start site within an AGO1 binding site, corresponding RNA-seq was consistent with mutation conferring reduced CHID1 expression (Fig. 4B).

To search for significantly mutated *cis*-regulatory elements (CREs) we restricted our analysis to sequences interacting with promoters through chromatin looping. A CRE interacting with the *USP22* promoter was mutated in 9.1% of tumours and these were associated with reduced *USP22* expression ($Q_{\text{Mann-Whitney}}$=0.009) (Fig. 4C). Mutations were not uniformly distributed and occupied a number of TF binding sites. A CRE interacting with *XRCC2* was mutated in 4% of tumours, mutations were associated with elevated *XRCC2* expression ($Q_{\text{Mann-Whitney}}$=0.046) (Fig. 4D).

We found no evidence of recurrent mutations within UTRs or non-coding RNAs when imposing a threshold of at least five effected tumours.

Mutated pathways

In addition to documented enrichment of *NRAS* and *KRAS* mutations in hyperdiploid ALL and *TP53* mutations in hypodiploid/near haploid ALL, we identified a number of additional associations (**Supplementary Table 6**). Notably, *TBL1XR1* and *ZEB2* mutations were enriched in *ERG*-deleted ALL (present in 21% and 14% of tumours respectively). iAMP21 tumours were characterised by an excess of *RB1* deletions (40%) and *IL7R* mutations (20%). *NF1* mutations were largely confined to near haploid tumours occurring in 45%. *ETV6-RUNX1* positive tumours were associated with enrichment for the deletion of *TBLXR1* and RAG1/RAG2. Finally undefined tumours (included in other) showed an excess of *IKZF1* deletions.

Given the identification of alterations in both *CTCF* and the histone gene cluster 1 we explored their transcriptional impacts, performing differential expression analysis. We identified five differentially expressed genes in both sets of mutated tumours ($P_{\text{binomial}} = 1.5\text{x}10^{-8}$), including *CLIC5* and *IGF2BP1*

(**Supplementary Table 7 and Supplementary Fig. 9**). Whilst *CLIC5* and *IGF2BP1* are markers of hyperdiploid ALL[46], none of the tumours harbouring these mutations were of this subtype. In total 60 tumours (17%) harboured alterations (deletions or mutations) in either *CTCF* or the histone gene cluster 1.

To produce a composite picture of somatic events we clustered drivers by biological pathways (Fig. 5). The most frequently altered pathway featured B-cell developmental genes, altered in 70% of tumours. This analysis confirmed the importance of RAS/RTK alterations in hyperdiploid biology and highlighted a number of other key pathways, including secondary alterations affecting cytokine signalling in iAMP21, where 37% of tumours possessed a secondary hit in either *IL7R*, *JAK2* or *CRLF2* (including 3/5 cases of P2RY8-CRLF2 translocation). BCR-ABL tumours were characterised by recurrent alteration of genes regulating the cell cycle, whilst hypodiploid tumours were typified by disruption of transcriptional (gene) regulation. As well as disruption of B-cell development genes, driven by loss of the second *ETV6* alleles, 56% of *ETV6-RUNX1* tumours were effected by alterations in chromatin and histone modification genes.

We assessed the clonality of driver gene mutations, finding most occur both clonally and sub-clonally (Fig. 6A **and Supplementary Fig. 10).** Exceptions to this included *ZEB2* mutations which were always clonal, moreover mutations of B-cell development and haematopoiesis genes (*IKZF1*, *PAX5* and *ZEB2*) tended to be clonal. Conversely the majority of RAS/RTK gene mutations were subclonal (65%; P_{Fisher}=0.001). This was especially true of ERG-deleted tumours where 44% possessed a subclonal RAS/RTK variant (accounting for 89% of these mutations in the subtype) compared to only 8% with a clonal variant. Conversely RAS/RTK mutations in hyperdiploid tumours were usually clonal (60%), occurring in 44% of tumours compared to 20% with only a subclonal variant.

To assess the molecular mechanisms promoting tumorigenesis we used non-negative matrix factorization (NMF) to extract COSMIC single base signatures (SBS). Ten signatures were seen contributing at least 1% of mutations (**Supplementary Fig. 11**). SBS5 (aetiology unknown but clock-like) accounted for the most mutations (41%) and was seen in all tumours **(Supplementary Figs. 12 and 15**). SBS2 and SBS13 (AID/APOBEC) were almost exclusively confined to *ETV-RUNX1* tumours ($Q_{Man-Whitney}$=2.3x10^{-33} and $Q_{Man-Whitney}$=1.1x10^{-36} respectively), whilst SBS7a (UV exposure) was highly enriched in iAMP21 tumours ($Q_{Man-Whitney}$=5.3x10^{-12}) (**Supplementary Figs. 13 and 15**). SBS7a was associated with the highest mutation rate, 10-fold higher than SBS1 (**Supplementary Fig. 16**) and was largely responsible for the increased mutation rate in iAMP21 tumours (**Supplementary Fig. 17**).

It has been reported that SVs in *ETV6-RUNX1* positive tumours bear the hallmarks for RAG1 and RAG2 activity[47]. We searched for recurrent DNA motifs at SV breakpoints, firstly agnostically using motif enrichment performed using HOMER, and secondly by assessing the similarity of discovered motifs to those of candidate mutagenic drivers (**Supplementary Table 2**). Overall the most enriched motifs were the RAG heptamer (P< 1x10^{-200}), RAG nonamer (P< 1x10^{-200}) and PRDM9 (P= 1x10^{-121}), found at 8.8%, 7.2% and 1.5% of breakpoints respectively. With the exception of *ETV6-RUNX1* positive tumours the most frequent enriched motifs were the RAG hepamer and RAG nonamer, however in *ETV6-RUNX1* the most

common motif was PRDM9 contained in 28% of breakpoints ($P = 1x10^{-162}$) (Fig. 6B). Overall RAG heptamers were observed at both breakpoints of 3% of SVs.

We also sought evidence of activation induced deaminase (AID) activity at SV breakpoints. Due to the degenerate nature of AID motifs we used the number of repeats of core AID recognition sequences (**Supplementary Table 2**) as a proxy of activity. After comparing SVs in immunoglobulin regions we established a cut-off of >10 repeats as suggestive of AID activity (**Supplementary Fig. 18**). AID signatures were detected in the breakpoints of 2% of all SVs, but 17% of SVs in *TCF3-PBX1* positive tumours ($P_{Fisher}=8x10^{-9}$) (**Supplementary Fig. 18**).

Clonal architecture

The presence of subclonal populations in tumours was almost universal (observed in 98% of tumours; Fig. 7A). Most commonly tumours possessed two subclones, however ERG-deleted tumours tended to have a higher number of subclones ($Q_{Mann-Whitney}$ =0.008) and KMT2A translocated lower ($Q_{Mann-Whitney}$=0.038) (**Supplementary Fig. 20**). The distribution of subclone CCF was similar across subtypes, with the exception of hyperdiploid tumours whose subclones tended to have higher CCFs ($Q_{Mann-Whitney}$=0.004), 50% having a subclone with a CCF between 0.7 and 0.8, compared to 9% of other tumours (**Supplementary Fig. 21**).

The diversity of cell populations (*i.e.* heterogeneity) varied across subtypes, hypodiploid and ERG-deleted tumours were the most heterogeneous (median Simpson index = 0.61 and 0.62; $Q_{Mann-Whitney}=1.7x10^{-3}$, $1.13x10^{-3}$), while hyperdiploid tumours exhibited lower heterogeneity (Simpson index = 0.45; $Q_{Mann-Whitney}=2.8x10^{-7}$).

Accounting for mutational frequency, we found subclones were enriched for driver mutations ($P_{binomal}=1.8x10^{-5}$) relative to clonal populations. To examine the processes influencing tumour evolution we compared the frequency of driver gene alteration in subclones. ERG-deleted subclones were the most likely to possess a mutation in an ALL driver gene (35%; $Q_{binomal}$=0.0016), whereas *BCR-ABL1* positive tumour subclones contained the lowest frequency of driver alterations (4%; $Q_{binomal}$= 0.052) (**Supplementary Fig. 22**).

To explore the possible contribution of neutral evolution to tumour heterogeneity we used MOBSTER[43], which models variant distribution under neutrality. MOBSTER called neutral tails in the majority of tumours, fitting a median of 12% (SNVs) and 16% (SNVs and indels) of variants (**Supplementary Fig. 23**). Evidence of positive selection was sought using dNdSCV, revealing that tail compartments were enriched in *NRAS* ($Q = 3.4x10^{-8}$) and *KRAS* ($Q = 1.9x10^{-3}$) mutations. Additionally rates of non-synonymous substitution in *NRAS*, *KRAS*, *FLT3, NSD2* were higher in tail compartments than clonal groups (**Supplementary Fig. 24**).

Discussion

By analysing whole genome sequencing and transcriptome data from a large series of ALL patients we provide for an enhanced understanding of ALL subtype genetics identifying novel candidate coding, non-coding and copy number drivers. Our analysis reveals differences in the mutational and biological pathways processes influencing the initiation and progression of disease. We also provide evidence of ongoing selection of subclonal mutations as an important feature of ALL evolution.

Around half of *ETV6-RUNX1* and iAMP21 tumours are characterised both by an increased mutation rate and enrichment for specific COSMIC single base signatures. AID/APOBEC related signatures, SBS2 and SBS13, were largely confined to *ETV6-RUNX1* ALL, while UV associated SBS7a was highly enriched in iAMP21 positive tumours. SBS7a has previously been reported to occur in ALL tumours at a similar rate[19]. Moreover SBS7a it occurs at similar rates in a number of tumours types lacking this exposure[19]. These observations provide evidence for an additional mechanistic origin of SBS7a suggesting either an unknown germline genetic influence or environmental exposure promoting this subtype.

As well as identifying novel coding drivers, we also provide the first evidence of non-coding mutations influencing ALL leukaemogenesis. Both hotspot mutation and amplification impacted *ZEB2*, consistent with oncogenic activation. *ZEB2* is a **z**inc-finger, **E**-**b**ox–binding transcriptional repressor and is highly expressed in haematopoietic stem cells and splenic B-cells[48] where it regulates haematopoiesis. *ZEB2* deficient mice have decreased B-cells populations[49] and overexpression of the gene promotes proliferation[50] in AML. Collectively this suggests *ZEB*2 drives tumourigenesis by altering normal haematopoiesis and promoting B-cell proliferation, further *ZEB2* variants were clonal implying that, in those tumours, it was required for transformation.

Deletions of the gene encoding B and T-lymphocyte attenuator (*BTLA*) have previously been reported in ALL[51], which typically overlap *CD200*, however the functional mediator has yet to be elucidated. The identification of *BTLA* promoter variants suggest loss of this genes as opposed *CD200* confers selective advantage. Mutated CREs on chromosomes 17 and 7 were associated with reduced *USP22* and increased *XRCC2* expression respectively. *USP22* has been implicated in T-cell activation however its activity as a deubiquitinase of core histones suggests a wider role in transcriptional regulation[52]. *XRCC2* is a key regulator of double strand break repair. How *XRCC2* promotes ALL is unknown, although the importance of double stand breaks inducing enzymes (RAG, AID) in leukaemogenesis may offer an explanation.

We additionally report recurrent copy and structural variation impacting *HLA-DRB5*. Although the selective basis of these lesions is unclear, genome-wide association studies in chronic lymphocytic leukaemia[53] and lymphoma[54] have identified germline variants in *HLA-DRB5* influencing disease risk.

Around 10% of tumours possessed a deletion overlapping the histone gene cluster 1, which contains 16 different histone isoforms, including at least two of each core histone. Recurrent histone H1 mutations have also been reported in around 30%-50% of lymphomas causing alterations in 3D chromatin architecture inducing stem cell like transcriptional profiles[55]. Further functional characterisation will be required in order to determine the functional gene(s) within these lesions. We also show that tumours

harbouring histone gene cluster 1 deletions and *CTCF* alterations share a common transcriptional profile, with down-regulation of *IGF2BP1* and *CLIC5* occurring in both groups. Interestingly both genes were recently identified alongside *CTCF* as markers hyperdiploid tumours[46]. No hyperdiploid tumours were included in this analysis precluding a co-variant effect driving this relationship. Mutations these genes were however enriched in tumours with no assigned subtype. Alteration of *CTCF*/histone gene cluster 1 was common, occurring in 17% of tumours. Collectively these data raise the possibility of a 'hyperdiploid-like' subtype of ALL.

While there is commonality in disruption of pathways between ALL subtypes there are clear distinctions not only in the particular biological pathways harbouring mutations but also the clonal distribution of these mutations. These differences have implications both for choice of potential targeted therapies and determining which patients will benefit most from their use. As targeting activated oncogenes is generally more tractable than tumour suppressors the biological pathway from most relevance for ALL are RAS/RTK and IL7 signalling. Importantly, RAS/RTK mutations in hyperdiploid tumours were typically clonal whereas in ERG deleted ALL mutations were almost exclusively subclonal, suggesting the efficacy of RAS/RTK inhibitors will differ between subtypes. Alteration of IL7 signalling was common in iAMP21 tumours suggesting that JAK2 inhibitors may have utility in this group.

Variants identified as neutrally occurring by MOBSTER were enriched in ALL drivers, indicating that neutral evolution is not a major contributor to genetic heterogeneity in ALL, this may be reflective of the low mutation rate of the disease comparative to most solid cancers. We show that subclonality in ALL is common suggesting Darwinian evolution drives the selection and expansion of mutations and subclones. Consequently the use novel targeted therapies should take account of the clonality and heterogeneity of tumours.

Declarations

DATA AVAILABILITY

Data available from DNAnexus (DNAnexus.com) subject to application from St Jude hospital.

WEB RESOURCES

Repetitive genomic loci used for variant filtering were downloaded from hgdownload.cse.ucsc.edu/goldenpath/hg38/database/simpleRepeat.txt.gz.

Smooth the wrapper for structural variant caller Lumpy is available from github.com/brentp/smoove.

Structural variant positional filtering was based on https://github.com/dellytools/delly/blob/master/excludeTemplates/human.hg38.excl.tsv.

FusionInspector the RNAseq fusion gene detection software is available from github.com/FusionInspector.

Control RNAseq data for GETex and HPA were downloaded from ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/samples/ and ftp://ftp.sra.ebi.ac.uk/vol1/run/ERR315.

Replication timing was downloaded from 2.replicationdomain.com/.

Promoters were defined using genode v30 downloaded from ftp://ftp.ebi.ac.uk/pub/databases /gencode/Gencode_human/release_30/ gencode.v30.annotation.gtf.gz.

VEGAN package for calculating population diversity is available from github.com/vegandevs/vegan.

HMMcopy is available from http://www.bioconductor.org/packages/release/bioc/manuals/HMMcopy/man/HMMcopy.pdf

ACKNOWLEDGMENTS

This work was supported by Cancer Research UK (C1298/A8362) and Blood Cancer UK.

The authors declare no competing financial interests.

AUTHOR CONTRIBUTIONS

JS, AC, PL performed bioinformatic and statistical analyses. PH and AC provided additional bioinformatics support. JS and RH drafted the manuscript.

References

1. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. Lancet. 2013;381(9881):1943–55.
2. Harrison CJ. Blood Spotlight on iAMP21 acute lymphoblastic leukemia (ALL), a high-risk pediatric disease. Blood. 2015 Feb 26;125(9):1383–6.
3. Carroll WL. Safety in numbers: Hyperdiploidy and prognosis [Internet]. Blood. 2013. p. 2374–6.
4. Moorman A V., Richards SM, Robinson HM, Strefford JC, Gibson BES, Kinsey SE, et al. Brief report: Prognosis of children with acute lymphoblastic leukemia (ALL) and intrachromosomal amplification of chromosome 21 (iAMP21). Blood. 2007 Mar 15;109(6):2327–30.
5. Steeghs EMP, Boer JM, Hoogkamer AQ, Boeree A, de Haas V, de Groot-Kruseman HA, et al. Copy number alterations in B-cell development genes, drug resistance, and clinical outcome in pediatric B-cell precursor acute lymphoblastic leukemia. Sci Rep. 2019 Dec 1;9(1):1–11.
6. Pui CH, Evans WE. A 50-year journey to cure childhood acute lymphoblastic leukemia. Semin Hematol. 2013 Jul 1;50(3):185–96.
7. Freyer DR, Devidas M, La M, Carroll WL, Gaynon PS, Hunger SP, et al. Postrelapse survival in childhood acute lymphoblastic leukemia is independent of initial treatment intensity: A report from the Children's Oncology Group. Blood. 2011 Mar 17;117(11):3010–5.

8. Nguyen K, Devidas M, Cheng SC, La M, Raetz EA, Carroll WL, et al. Factors influencing survival after relapse from acute lymphoblastic leukemia: A Children's Oncology Group study. Leukemia. 2008;22(12):2142–50.
9. Amgen slaps record-breaking $178K price on rare leukemia drug Blincyto | FiercePharma [Internet].
10. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
11. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.
12. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011 Aug 4;12(1):323.
13. Zhang Y, Parmigiani G, Johnson WE. ComBat-Seq: batch effect adjustment for RNA-Seq count data. bioRxiv. 2020;2020.01.13.904730.
14. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol. 2019 Oct 21;20(1):213.
15. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
16. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15(8):591–4.
17. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
18. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: Efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018 Nov 1;28(11):1747–56.
19. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020 Feb 6;578(7793):94–101.
20. Harrison CJ, Haas O, Harbott J, Biondi A, Stanulla M, Trka J, et al. Detection of prognostically relevant genetic abnormalities in childhood B-cell precursor acute lymphoblastic leukaemia: Recommendations from the Biology and Diagnosis Committee of the International Berlin-Frankfürt-Münster study group. British Journal of Haematology. 2010. p. 132–42.
21. 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 Jan 5;505(7484):495–501.
22. 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.e21.
23. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol. 2016 Dec 16;17(1):128.

24. Wilson DJ. The harmonic mean p-value for combining dependent tests. Proc Natl Acad Sci U S A. 2019 Jan 22;116(4):1195–200.
25. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016 Jun 6;17(1):122.
26. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011 Apr 28;12(4):R41.
27. Orlando G, Law PJ, Cornish AJ, Dobbins SE, Chubb D, Broderick P, et al. Promoter capture Hi-C-based identification of recurrent noncoding mutations in colorectal cancer [Internet]. Nature Genetics. 2018. p. 1375–80.
28. Javierre BM, Sewitz S, Cairns J, Wingett SW, Várnai C, Thiecke MJ, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016;167(5):1369-1384.e19.
29. Melton C, Reuter JA, Spacek D V., Snyder M. Recurrent somatic mutations in regulatory regions of human cancer genomes. Nat Genet. 2015 Jun 26;47(7):710–6.
30. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol. 2016 Apr 21;12(4):e1004873.
31. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. The Life History of 21 Breast Cancers. Cell. 2015;162(4):924.
32. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016 Apr 15;32(8):1220–2.
33. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: A probabilistic framework for structural variant discovery. Genome Biol. 2014;15(6).
34. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012 Sep;28(18).
35. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020 Feb 6;578(7793):112–21.
36. Cmero M, Ong CS, Yuan K, Schröder J, Mo K, Group PE and HW, et al. SVclone: inferring structural variant cancer cell fraction. bioRxiv. 2017 Aug 4;172486.
37. Cortés-Ciriano I, Lee JJK, Xi R, Jain D, Jung YL, Yang L, et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nat Genet. 2020 Mar 1;52(3):331–41.
38. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell. 2010;38(4):576–89.

39. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: An open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004 Jan 1;32(DATABASE ISS.).
40. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009 Jun;5(6):e1000529.
41. Yuan K, Macintyre G, Liu W, Markowetz F. Ccube: A fast and robust method for estimating cancer cell fractions. bioRxiv. 2018;484402.
42. Dixon P. VEGAN, a package of R functions for community ecology [Internet]. Journal of Vegetation Science. 2003. p. 927–30.
43. Caravagna G, Heide T, Williams MJ, Zapata L, Nichol D, Chkhaidze K, et al. Subclonal reconstruction of tumors by using machine learning and population genetics. Nat Genet. 2020 Sep 1;52(9):898–907.
44. Tran TH, Hunger SP. The genomic landscape of pediatric acute lymphoblastic leukemia and precision medicine opportunities. Seminars in Cancer Biology. 2020.
45. Lai Daniel ,Ha Gavin SS. HMMcopy: Copy number prediction with correction for GC and mappability bias for HTS data. 2020.
46. Yang M, Vesterlund M, Siavelis I, Moura-Castro LH, Castor A, Fioretos T, et al. Proteogenomics and Hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia. Nat Commun. 2019 Dec 1;10(1).
47. Papaemmanuil E, Rapado I, Li Y, Potter NE, Wedge DC, Tubio J, et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. Nat Genet. 2014;46(2):116–25.
48. Postigo AA, Dean DC. Differential expression and function of members of the zfh-1 family of zinc finger/homeodomain repressors. Proc Natl Acad Sci U S A. 2000 Jun 6;97(12):6391–6.
49. Li J, Riedt T, Goossens S, García CC, Szczepanski S, Brandes M, et al. The EMT transcription factor Zeb2 controls adult murine hematopoietic differentiation by regulating cytokine signaling. Blood. 2017;129(4):460–72.
50. Li H, Mar BG, Zhang H, Puram R V., Vazquez F, Weir BA, et al. The EMT regulator ZEB2 is a novel dependency of human and murine acute myeloid leukemia. Blood. 2017 Jan 26;129(4):497–508.
51. Ghazavi F, Clappier E, Lammens T, Suciu S, Caye A, Zegrari S, et al. CD200/BTLA deletions in pediatric precursor B-cell acute lymphoblastic leukemia treated according to the EORTC-CLG 58951 protocol. Haematologica. 2015 Oct 2;100(10):1311–9.
52. Zhang XY, Varthi M, Sykes SM, Phillips C, Warzecha C, Zhu W, et al. The Putative Cancer Stem Cell Marker USP22 Is a Subunit of the Human SAGA Complex Required for Activated Transcription and Cell-Cycle Progression. Mol Cell. 2008 Jan 18;29(1):102–11.
53. Slager SL, Rabe KG, Achenbach SJ, Vachon CM, Goldin LR, Strom SS, et al. Genome-wide association study identifies a novel susceptibility locus at 6p21.3 among familial CLL. Blood. 2011 Feb 10;117(6):1911–6.

54. Conde L, Halperin E, Akers NK, Brown KM, Smedby KE, Rothman N, et al. Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. Nat Genet. 2010 Aug 18;42(8):661–4.
55. Yusufova N, Kloetgen A, Teater M, Osunsade A, Camarillo JM, Chin CR, et al. Histone H1 loss drives lymphoma by disrupting 3D chromatin architecture. Nature. 2021 Jan 14;589(7841):299–305.

Figures

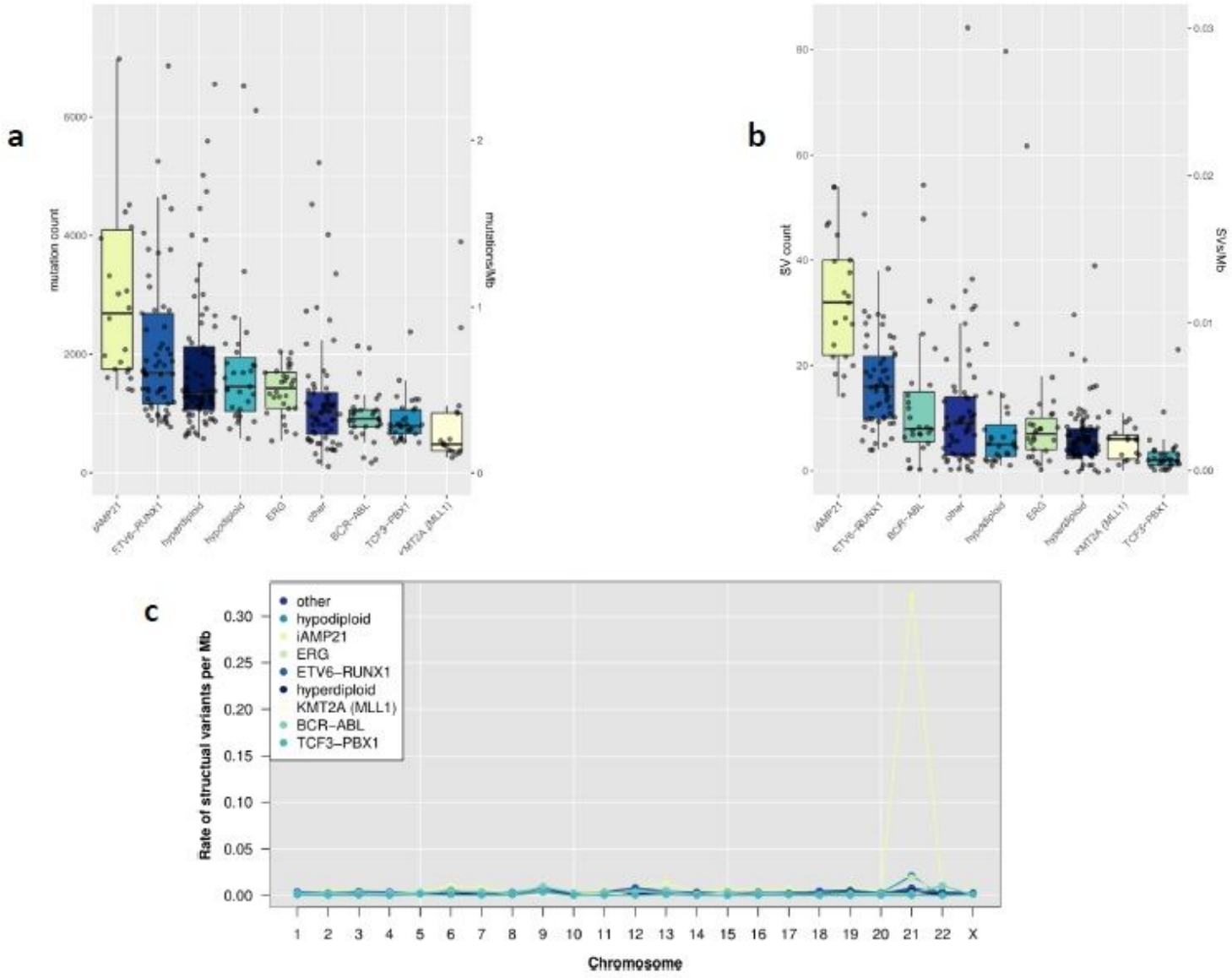

Figure 1

Mutation burden by subtype. Short somatic variants (SNVs and indels) were called in 361 matched normal/tumour whole genome sequencing samples. (a) Burden of SNVs and indels. Box and whiskers plot of mutation count per tumour. (b) Burden of structural variants (SVs). Box and whiskers plot of SV count, dots represent individual tumours. (c) Plot of the SV rate per chromosome retaining only intrachromsomal variants outside immunoglobulin loci. Y-axis; mean SVs rate per Mb. X- axis; chromosome.

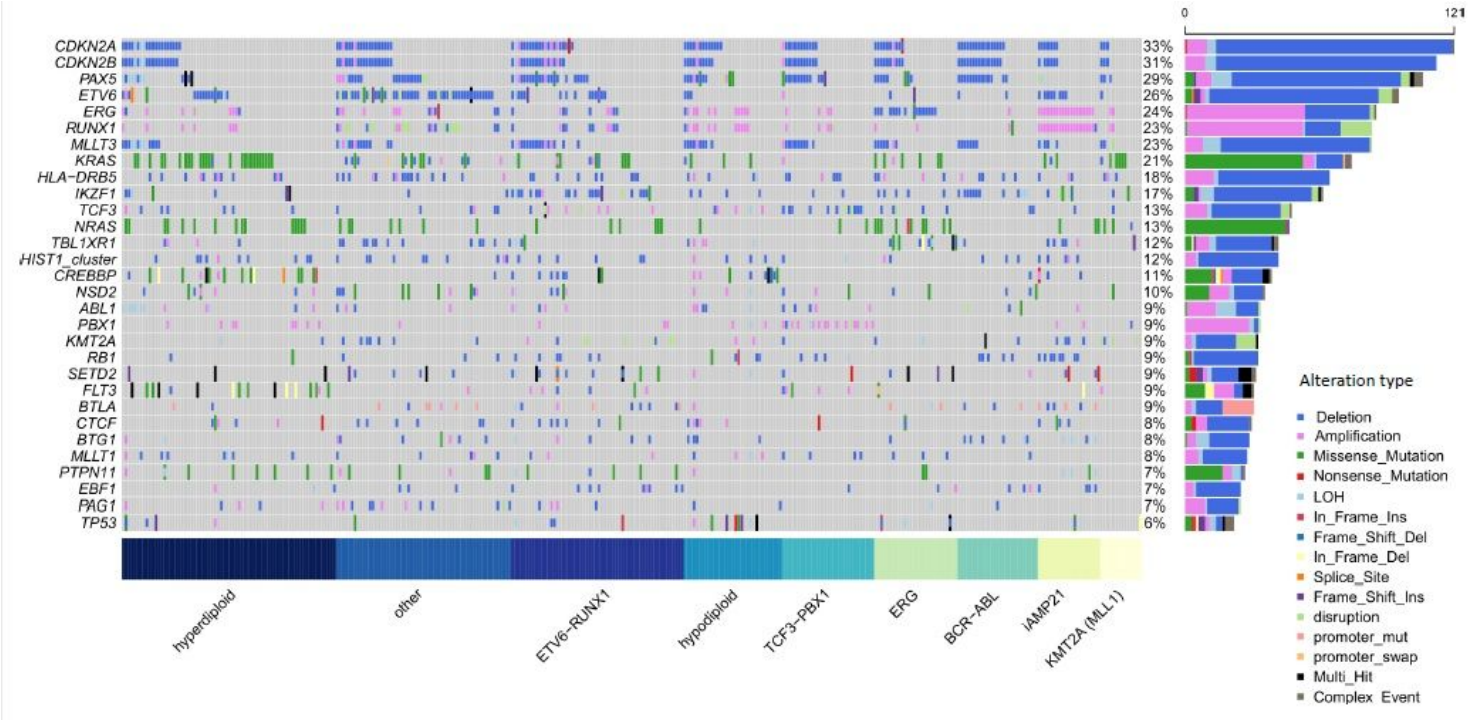

Figure 2

Driver gene analysis. Oncoplot of somatic alterations for selected ALL driver genes (Supplementary Table 8), genes altered in > 20 tumours, compiled from SNV, indel, CNV (only focal events), SV, RNAseq, and LOH. Vertical lines represent one tumour. Coloured sections in grey grid denote alteration type, described in key - “Alteration type”. Short nucleotide variants span entire row, other alterations span half row width. Right bar plot shows the frequency of driver gene alteration, colour denotes alteration type, as prior. Deletions and amplifications derived from CNVs and SVs; disruption from RNAseq and SVs. Alterations are shown non-redundantly; tumours with multiple alterations in the same gene only counted once. Plot generated using Maftools6 after the exclusion of genes in the psuedoautosomal regions.

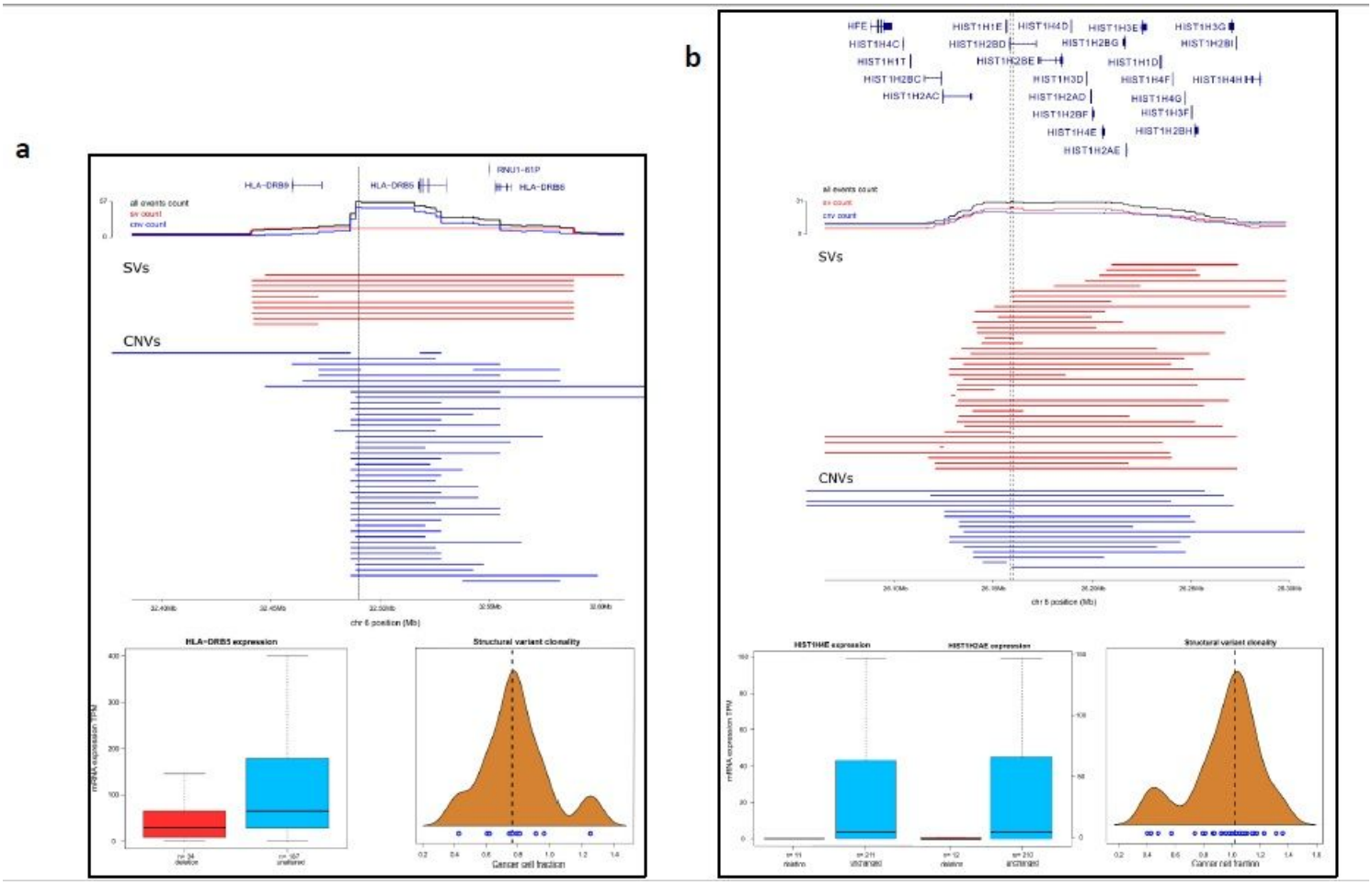

Figure 3

Recurrent copy number and structural variants. Significantly amplified or deleted regions in CNV data, were filtered retaining only those with an enrichment of structural variants, based on a permutation test. Regional genetic plots showing recurrent deletions mapping to (a) HLA-DRB5 and (b) histone gene cluster 1. Upper panes shows gene position. Line plots show number of tumours with an overlapping variant; blue – CNVs, red – SVs, black – total count (tumours with both SVs and CNVs counted once). Central pane shows the individual variants. For convenience only variants starting or ending in the field of view are plotted. Vertical black lines denote region with highest deletion frequency. Lower left pane, box plots of gene expression split by mutational status. Lower right pane, density plots of structural variant clonality; blue circles individual SVs. Genomic coordinates from GRCh38.

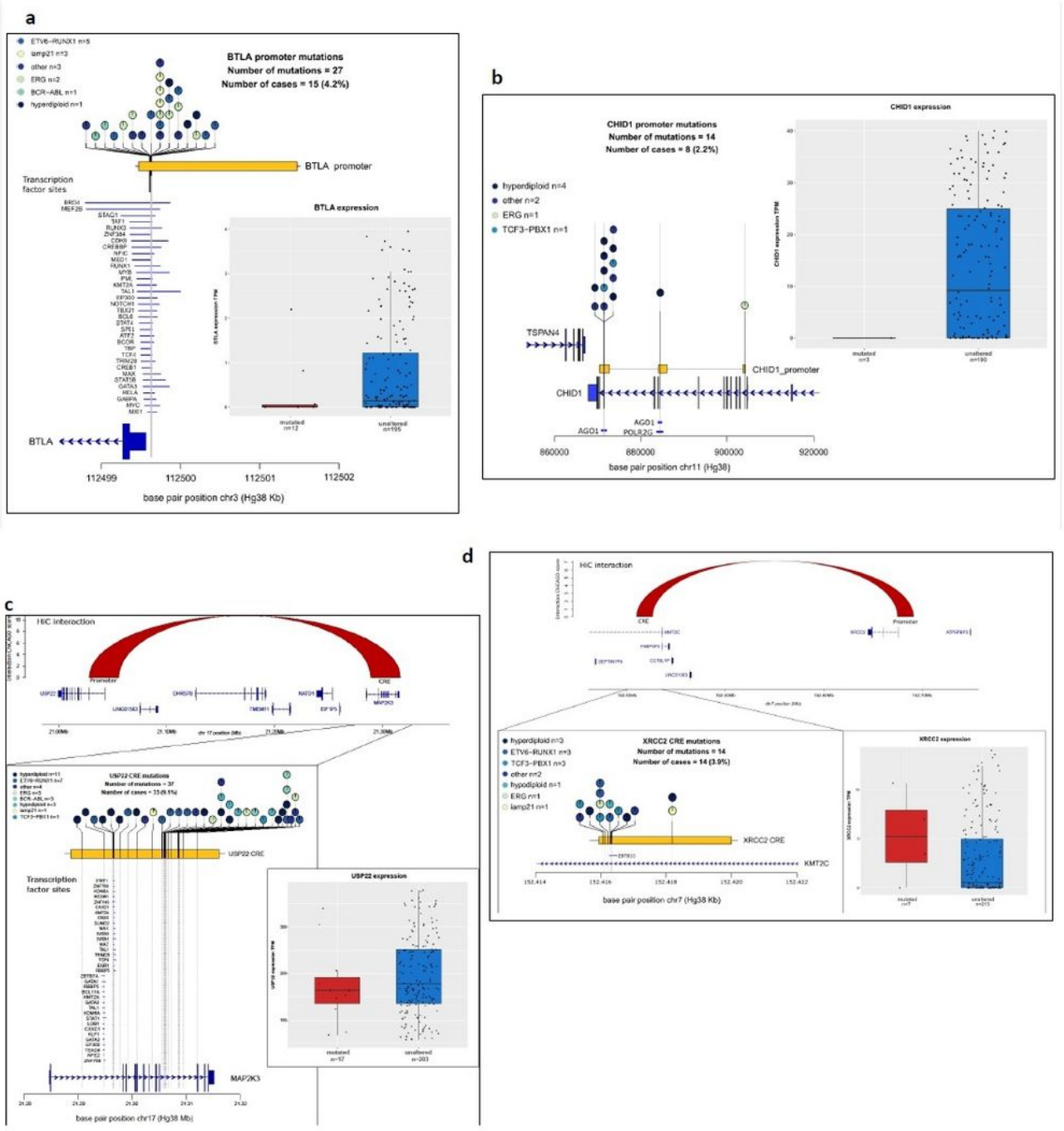

Figure 4

Non-coding driver mutations. Mutation burden within promoters and cis-regulatory elements (CREs) and their transcriptional impact. Promoter mutations of (a) BTLA and (b) CHID1. Regional plot of mutations (coloured circles) relative to coding sequence (dark blue boxes) and promoter (yellow horizontal bar). Transcription factor binding sites (light blue horizontal line) overlapping mutations were extracted from Encode and ChIP atlas. Grey boxes correspond to transcriptional impact on respective gene. Box and

whiskers plot, tumours are split by mutational status, dots represent individual tumours. CRE mutations of (c) USP22 and (d) XRCC2 and associated transcriptional impact. Regional genetic plots as per (a) and (b). Upper pane shows a wide genetic view of each CRE and linked promoter connected by chromatin conformation contact in naïve B-cells (red arch). Lower pane, shows zoomed in view of the mutated CRE. Genomic coordinates from GRCh38.

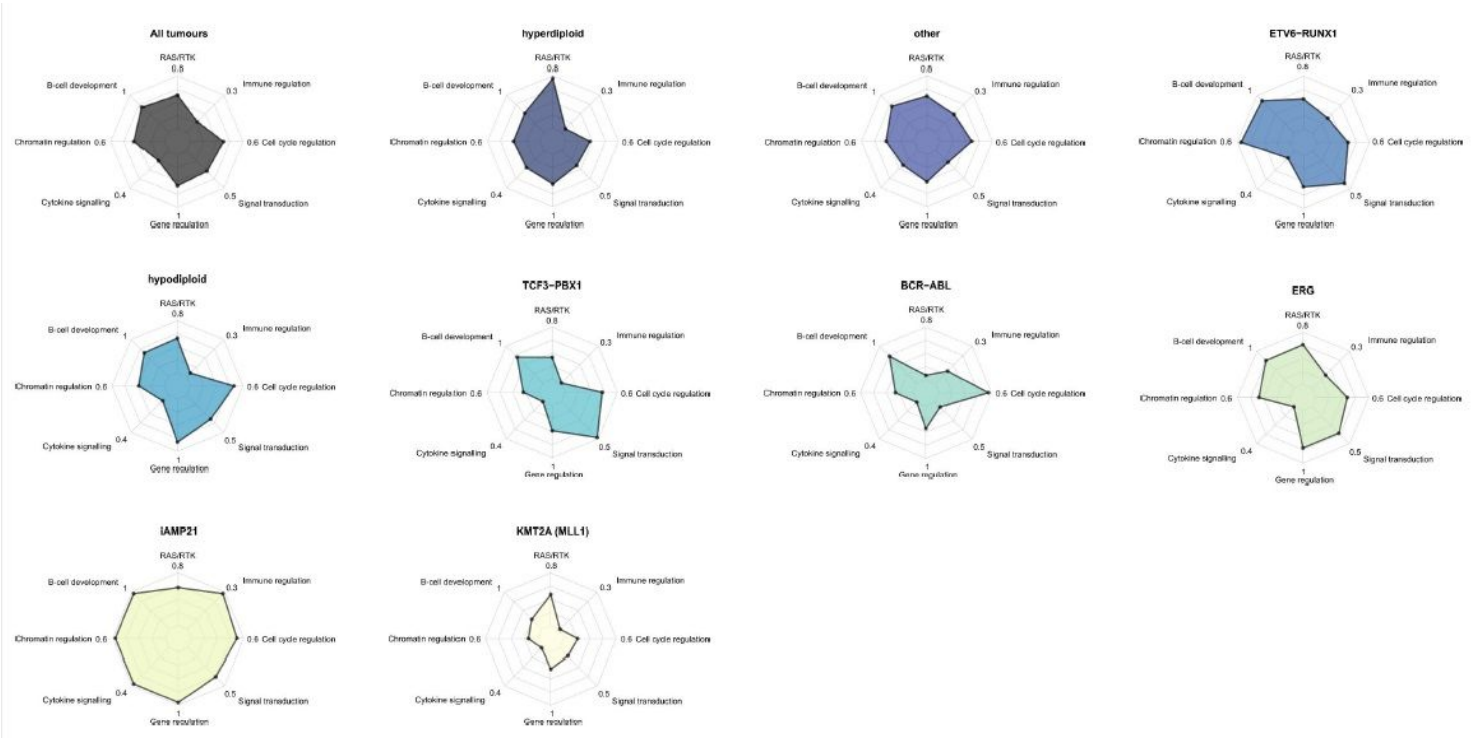

Figure 5

Pathway analysis and signature analysis. Radar plots showing the most frequently altered pathways for each subtype. Driver genes grouped according to biological pathway. Somatic alterations for a selected ALL driver genes was compiled from SNV, indel, CNV, SV, RNAseq, and LOH data (CNVs include only focal events). Subtype defining events are excluded (e.g, disruption of ETV6 or RUNX1 in ETV6-RUNX1 positive tumours). The proportion of tumours with an alteration in any gene assigned to that pathway is plotted on the radial axis. Each axis is scaled separately. Gene – pathway assignments: RAS/RTK; NRAS, KRAS, PTPN11, FLT3, NF1, ABL1. B-cell development; PAX5, IKZF1, ETV6, ZEB2, RUNX1, TCF3, RAG1, RAG2, EBF1. Chromatin regulation; SETD2, HDAC7, NSD2, CTCF, KMT2A, STAG2, histone gene cluster 1. Cytokine signalling; JAK2, IL7R, CRLF2. Gene regulation; CREBBP, MLLT1, MLLT3, AFF1, BTG1, ERG, TCF4, NCOA6. Signal transduction; TBL1XR1, TBL1X, PBX1, PAG1. Cell cycle regulation; CDKN2A, CDKN2B, RB1. Immune regulation; BTLA, HLA-DRB5.

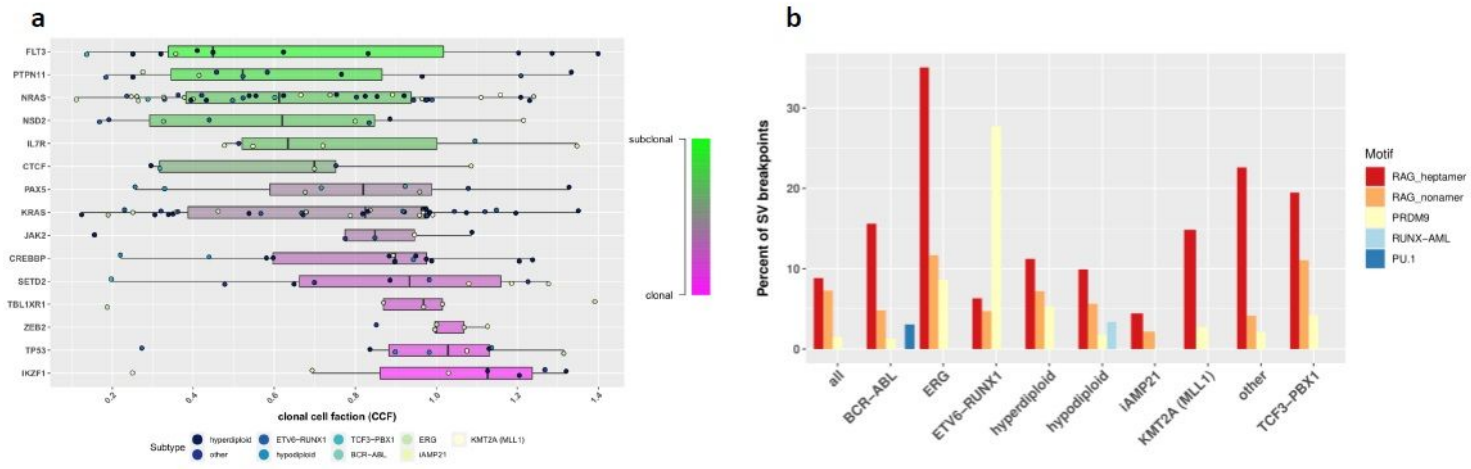

Figure 6

Driver clonality and SV breakpoint enrichment. (a) Driver gene mutation (SNV/indels) clonality. Box and whiskers plot showing the proportion clonal mutations for ALL driver genes. Each circle represents a mutation, coloured according to disease subtype, for tumours with multiple mutations in the same gene the variant with the highest clonal cell fraction is retained. (b) Structural variant motif enrichment. Bar chart showing motif enrichment at SV breakpoints. Two 100bp of sequences flaking each breakpoint of an SV were extracted and analysed using HOMER. Y-axis; percent of extracted sequences containing motifs.

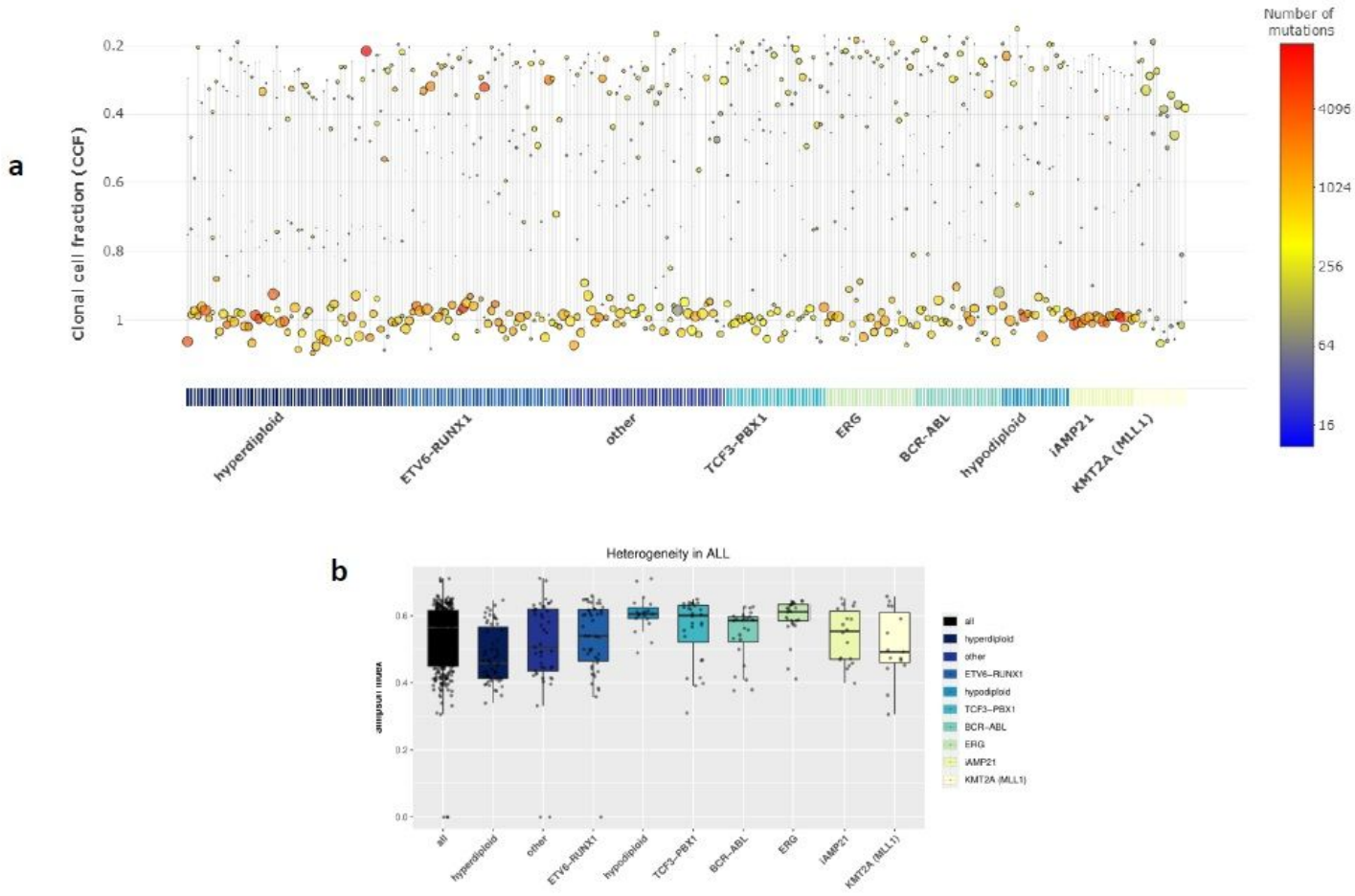

Figure 7

Clonal architecture and evolution. Variant cancer cell fractions (CCF) were calculated and variants clustered into clonal and subclonal populations. (a) Distribution of clones. Horizontal lines represent single tumours, circles represent clones; the size and colour of circles corresponding the proportion and number of variants assigned to each clone. Y-axis; clonal frequency (proportion of cell cells with a variant(s)). (b) Heterogeneity between subtypes. Box and whiskers plot of Simpson index (higher values indicative of increased heterogeneity). Each dot corresponds to a tumour.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.pdf
- SupplementaryTables.pdf