

Unique mutation and gene expression in High Hyperdiploid Acute Lymphoblastic Leukaemia with whole chromosome uniparental isodisomy

Mayur Parihar

Tata Translational Cancer Research Center/Tata Medical Center <https://orcid.org/0000-0002-8059-0914>

Debdutta Ganguli

Tata Translational Cancer Research Centre

Anindyajit Banerjee

Tata Translational Cancer Research Centre

Rubina Islam

Tata Translational Cancer Research Centre

Sangramjit Basu

Tata Translational Cancer Research Centre

Chumki Bhattacharya

Tata Translational Cancer Research Centre

Piyali Sarkar

Tata Translational Cancer Research Centre

Shivani Bhagwat

Tata Translational Cancer Research Centre

Subhoshree Ghose

Tata Translational Cancer Research Center

Debparna Saha

Tata Translational Cancer Research Centre

Uzma Zaheer

Tata Translational Cancer Research Centre

Ritam Siddhanta

Tata Translational Cancer Research Centre

Pritha Paul

Tata Translational Cancer Research Centre

Jasmeet Sidhu

Tata Translational Cancer Research Centre

Nandana Das

Tata Translational Cancer Research Centre

Manash Gogoi

Tata Translational Cancer Research Centre

Abhijit Chakraborty

La Jolla Institute for Allergy and Immunology

Binuja Varma

TCS Life science

Uma Sunderam

TCS Life science

Abhirami Ramasubramaniam

TCS Life science

Nidhan Biswas

National Institute of Biomedical Genomics <https://orcid.org/0000-0001-7960-3290>

Chitrapita Das

NATIONAL INSTITUTE OF BIOMEDICAL GENOMICS

Manish Singh

Tata Medical Center

Barun Chakraborty

Tata Medical Center

Ashish Gorantla

Tata Medical Center

Niharendu Ghara

Tata Medical Center

Deepak Mishra

Tata Medical Center

Shekhar Krishnan

Tata Medical Center/ Tata Translational Cancer Research Center

Rajgopal Srinivasan

Tata Consultancy Services (India)

Vaskar Saha (✉ v.saha@manchester.ac.uk)

Division of Cancer, Faculty of Biology, Medicine and Health Sciences, University of Manchester, Tata Medical Center/Tata Translational Cancer Research Centre

Article**Keywords:**

Posted Date: February 16th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1349722/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Whole genome copy number alterations, targeted gene-panel and RNA sequencing were used to investigate differences in high hyperdiploid (HH) B-cell acute lymphoblastic leukemia (ALL), with and without whole chromosome uniparental isodisomy (wUPD). Patients with wUPD demonstrated a higher modal number of chromosomes with gains of chromosome 5. Mutations in genes within epigenetic pathways with upregulation of genes involved in cellular response to stress and stimuli were seen in wUPD, in contrast to mutations in *RAS/RTK* pathways and upregulation of genes in RNA Polymerase III pathway in noUPD. AP-1 transcription factors and *NUDT18*, potentially implicated in *thiopurine* drug resistance, were upregulated in wUPD. wUPD was associated with female gender, higher presenting white cell count and lower end of induction minimal residual disease levels, though survival rates were similar in the two groups. Genome-wide differences in HH ALL with and without UPD suggest plausible biological explanations for the heterogeneity in therapeutic response.

Introduction:

High hyperdiploid (HH) (51–67 chromosomes) is the most common cytogenetic subtype of childhood B-Cell precursor (BCP) acute lymphoblastic leukaemia (ALL) accounting for 30–40% of all BCP-ALLs worldwide. It is characterised by non-random gains of chromosomes, usually + X, + 4, + 6, + 10, + 14, + 17, + 18 and + 21¹. Evidence suggests that HH originates during a single mitotic event², possibly after cell fusion³. With contemporary minimal residual disease (MRD) based risk stratified therapeutic protocols, HH ALL patients have outcomes close to 90%, and classified as good risk genetics⁴. Nevertheless, HH is also the largest genetic subgroup in relapsed BCP-ALL⁵. Prognostically, outcomes in relapsed ALL are significantly correlated with the time of the relapse. Early on-therapy relapses have poor outcomes even with allogeneic stem cell transplant⁶, while late off-therapy relapses often have excellent outcomes with conventional chemotherapy⁷. About 70% of HH relapses are late while the rest are early⁵. The basis for these outcome heterogeneities remains unclear. Large structural abnormalities (seen in about 50% of HH BCP-ALL), modal numbers, sex and age do not appear to influence outcomes, but + 17 and + 18 in the absence of + 5 and + 20 are associated with lower relapse rates^{8,9}. Smaller prognostically significant copy number alterations such as those that characterise the *IKZF1*-plus group¹⁰ in other BCP-ALLs are infrequent in HH^{7,11,12}. Approximately half of HH cases have mutations in genes of the receptor tyrosine kinases (RTK)/RAS pathways or histone modifiers¹³. These are frequently subclonal and gained or lost at relapse¹⁴. Mutations/deletions of *CREBBP* appear to be enriched at relapse and the co-occurrence of mutations in *KRAS* and *CREBBP* are associated with a higher incidence of relapse¹⁵. However, these do not explain disease recurrence in the majority of HH ALL. Monosomic chromosomes duplicated at mitosis give rise to acquired whole chromosome uniparental isodisomy (wUPD). Between 25–50% of HH patients have wUPD, a phenomenon almost unique to this genetic subgroup of ALL^{11,13,16,17}. Mutations appear to be rare in genes present on chromosomes with wUPD, and seen more frequently in segmental UPD's which also occur in other ALL subtypes¹⁸. The prognostic significance of wUPD is unknown. In

this report, we investigate copy number alterations, mutations and gene expression in HH with and without wUPD to investigate possible biological differences that may lead to variations in response to treatment.

Results

Chromosomal Gains in High Hyperdiploid ALL

One hundred six patients with HH were identified in 303 BCP-ALL patients sequentially analysed between 2015–2019. 58 HH patients were analysed with whole genome SNP arrays (Supplementary Table 1 and Supplementary Fig. 1). Detection of copy neutral loss of heterozygosity (LOH) across most chromosomes with an accompanying *TP53* mutation identified a patient with masked hypodiploidy who was excluded from further analyses. Similar to previous reports, the most frequent gains of chromosomes were +X, +4, +6, +10, +14, +17, +18 and +21 with modal numbers ranging from 52–67 chromosomes (Fig. 1A) ^{1,11,19–21}. Majority of gains (83%) were trisomies, except for chromosome 21, which was tetrasomic in 48 and pentasomic in two patients. All tetrasomies were duplications of two homologues (2:2 pattern), except in four patients (3 patients with tetrasomy 21 and one with tetrasomy 10). The only chromosome loss detected was a monosomy 7 seen in a UPN7. Subclonal chromosomal gains were seen in 6 patients (Supplementary Table 2.1).

Structural changes in HH ALL

Structural abnormalities > 5MB were seen in 33 (58%) patients (Table 1). The commonest were a partial gain of 1q and isochromosome 7q, similar to previous reports ^{11,19,22} (Supplementary Table 2.2). Segmental subclonal gains and losses were observed in 8 patients. A single patient showed subclonal isochromosome of 7q (UPN 47). A chromothripsis like pattern ^{23–25} was seen on chromosome 20 in UPN 9 (Fig. 1B). 45 microdeletions (< 5MB), were seen in 22 patients (median 2; range 1–5) (Supplementary Table 2.3) of which 13 had additional structural abnormalities. Twenty-eight deletions identified in 21 patients, involved genes commonly deleted in paediatric BCP-ALLs ^{26–29} (Fig. 1C). Deletions were predominantly hemizygous except those involving *CDKN2A/2B* genes, where 6/7 patients showed homozygous deletions. In UPN 54, the homozygous deletion of *CDKN2A/2B* was in a wUPD 9 ¹⁶ (Fig. 1D). In UPN 8, an *IKZF1* deletion (exon 2–8) was detected on a single homologue of a trisomic chromosome 7. A 37.1Kb deletion in 16q22.1 was detected in UPN 16. This resulted in an in-frame fusion between exon 2 of *RIPOR1* and exon 3 of *CTCF*, verified at transcript level, resulting in a premature stop codon and presumably loss of *CTCF* function (Fig. 1E). A *RIPOR1::CTCF* fusion has been previously reported in acute myeloid leukaemia ³⁰. A similar loss of *CTCF* function as a result of a downstream deletion, resulting in a *CTCF::PARD6A* fusion has been previously described in HH ALL. ¹³ Deletions in 13q12.2, as previously reported, ^{11,31} were not observed in this cohort. An AA genotype (rs10994982) of the *ARID5B* gene reported to be associated with an increased risk of HH BCP-ALL was detected in 45 (79%) patients ^{32,33} (Supplementary Table 2.4).

Frequency of Whole Chromosome Uniparental isodisomy in HH

Whole-chromosome uniparental isodisomy (wUPD) was seen in 26 (45%) patients (median 2; range 1–6 chromosomes), with subclonal wUPD of chromosome 9 in two patients (Fig. 2A). wUPD was further confirmed by whole genome sequencing in 3 patients (Fig. 2B). No germline wUPD's were detected in remission DNA samples in 22 analysed patients, confirming the somatic origin of wUPD (Supplementary Table 2.5). UPN10 was a female patient with copy neutral LOH of four copies of chromosome X. Segmental UPD's were seen in 20 patients (1–14 segments) (Supplementary Table 2.6). Frequencies of structural abnormalities/microdeletions were comparable in patients with or without wUPD. wUPD's were most commonly seen involving chromosomes 9, 15 and 19^{11,16,17}. Compared to those without (noUPD), patients with wUPD, had higher chromosomal modal numbers (Fig. 2C) similar to previous reports¹⁶. More frequent gains of chromosomes + 5, +11, + 12, +16 and + 22 not typical of classical HH were observed in wUPD (Fig. 2D). Gains of chromosomes 17 and 18 were evenly distributed in wUPD and noUPD patients with gain of chromosome 5 more frequent in wUPD patients. Given the differences in patterns of gains between HH with wUPD and noUPD we further investigated for differences in somatic mutations and gene expression between the two groups.

Distinct Mutations are present in wUPD and noUPD

Mutation analysis using a targeted gene panel (mean coverage 1311.4x, range 124.3-5122x, Supplementary Table 3, Supplementary Fig. 2) was performed on 20 noUPD, 21 wUPD and one HH patient for whom somatic copy number abnormalities (SCNA) data was not available. Two hundred seventy-three variants were predicted to be pathogenic or probably pathogenic based on the analysis done by an in-house informatics pipeline (Supplementary Fig. 3). Frameshift insertions, deletions, substitutions and stop gain variants were considered pathogenic by default. Variants with less than 10 reads were excluded. 270 single nucleotide variants (SNVs) were inspected manually with the Integrative Genomic Viewer (IGV), and 138 variants validated by Sanger sequencing and/or IGV visualisation of RNA sequencing data. Mutations not previously described, or not validated by sequencing, or not present on RNA sequencing or uncertain on IGV inspection, were excluded. Finally, 60 pathogenic mutations in 20 genes were identified in 36 patients (average 1.66 mutations per sample) including 46 missense mutations, 4 non-frame shift insertions deletions and substitution, 3 frame shift insertions and deletions and 4 stop-gain variants. (Supplementary Fig. 4 and Supplementary Table 4)

The majority of mutations, 41/60 (68%), were identified in genes with two copies. Seventeen (28%) were detected in genes with more than two copies. Mutations were infrequent in wUPD chromosomes, with only 3 (5%) detected (Fig. 3A). Two mutations were identified in samples where SCNA information was unavailable. The majority (86%) of mutations had variant allele frequencies (VAF) < 50% with a peak between 31–40% (Fig. 3A), suggesting that mutations occurred after chromosomal gains as previously reported¹³. In UPN1 with wUPD of chromosome 1, the somatic VAF of *ARID1A* (1p36.11) was 89.4%,

suggesting that this mutation preceded the development of wUPD. Transitions (N = 31) were more frequent than transversions (N = 20) (Fig. 3B). The most frequent mutational signature was C > T transition, all involving CpG dinucleotides in both groups, suggesting common endogenous mutagenic mechanisms as previously reported¹³.

As previously described in HH ALL the most frequent mutations were in genes belonging to the *RAS*, epigenetic and chromatin modifying pathway¹³ (Fig. 3C)., Though the numbers are small mutations involved in the *RAS* pathway (*KRAS*,*NRAS*) occurred more frequently in noUPD, while genes involving epigenetic and chromatin modifying pathways(*KMT2D*,*ARID1A*) were more likely to be mutated in wUPD patients (Fig. 3D). Mutations in *KRAS* and *NRAS* genes occurred primarily in described hotspot regions. The few in non-hotspot regions (K117N, A146T, E63K) have been described previously in BCP-ALL and other cancers^{13,34-38} (Fig. 3E) The *RAS* mutations were mutually exclusive except in UPN30, where *KRAS* and *NRAS* VAFs of 15% and 16% were detected. This could possibly be due to mutations present in two different subclones. Of the 7 mutations identified in *KMT2D*, there were 6 missense mutations in the disordered protein region and one non-frame shift substitution in the PHD domain (Fig. 3F). Two of the three *KMT2D* mutations in this cohort, not previously reported were verified on RNA sequencing.

Differences in gene expression between wUPD and noUPD

RNA sequencing was performed on 14 wUPD and 14 noUPD patients. From 20,260 protein coding genes identified, the 1000 most variable genes (Mahalanobis distance method) were used for further analysis. The expression levels of diploid wUPD chromosomes lay between haploid and diploid chromosomes, suggesting that the loss of allelic variation was associated with a lower gene expression in chromosomes with wUPD³⁹ (Fig. 4A). Gene expression levels are strongly influenced by SCNAs. In HH, this is at the level of whole chromosomal gains^{13,17,40,41}. Unsupervised hierarchical clustering did not distinguish between HH with or without wUPD (Fig. 4B) possibly because of the commonality of whole chromosomal copy numbers between the two groups (Fig. 2A). Indeed, the commonality in segmental SCNAs between wUPD and noUPD provides a partial explanation for the grouping in the hierarchical cluster (Fig. 4B). A tSNE analysis showed a central cluster of wUPD along with peripheral clusters of noUPD and combinations of noUPD and wUPD patients (Fig. 5A). A similar observation was made with a similarity matrix analyses (Supplementary Fig. 5) which confirmed that SCNA were mostly matched in common clusters. Peripheral clusters had in common SCNAs and/or mutations (Supplementary Table 5). The central wUPD cluster was heterogeneous for SCNAs and mutations. This suggested the possibility of a unique set of genes expressed differentially in wUPD, accounting for the cluster. To further identify genes uniquely expressed in wUPD, a differential gene expression analyses was performed with EdgeR (FDR < 0.1)⁴² with gene enrichment analysis (TOPPR). 113 and 108 genes were significantly upregulated in wUPD and noUPD respectively (Supplementary Table 6.1). The expression of these genes were also influenced by whole chromosomal gains (Fig. 5B). Restricting the analysis to log fold change of -1 to + 1, identified 71 and 18 genes significantly upregulated in wUPD and noUPD respectively (Supplementary Table 6.2 and Fig. 5C). Pathway analysis was performed using 84/89 protein coding differentially expressed genes (DEGs).As previously reported for HH ALL, in both wUPD and noUPD, pathway analyses showed an enrichment in

transcription, translation and protein metabolism⁴³ (Supplementary Table 7). Genes involved in RNA polymerase II and III transcription were enriched in noUPD. Genes involved in cell response to stress stimuli and activation of AP-1 transcription family were among those enriched in wUPD (Fig. 5D).

In wUPD, copy number adjusted TPM values showed upregulation of the AP-1 transcription factors (*FOS*, *FOSB* and *JUN*), *NUDT4* and *NUDT18* and *RAC2* (Supplementary Fig. 6). RQ-RT-PCR confirmed the differences in expression in 10 of 15 genes analysed (Fig. 5E, *NUDT15* analysed for comparison). Similar copy number adjusted TPM values suggested upregulation of *ABL1* in noUPD and *C-MYC* in wUPD. RQ-RT PCR confirmed the expression of *C-MYC* (Fig. 5F), this was not validated for *ABL1*. Unavailability of samples with high expression of select genes may have affected the RQ-RT PCR results as well as variations in expression (Supplementary Fig. 7). *NRAS* upregulation was unrelated to the activating mutations in *NRAS* (Supplementary Fig. 7).

The downregulation of cohesion complex genes and *CTCF* resulting in weakening of topologically associated domain boundaries has been reported in HH⁴³. We identified a *RIPOR1-CTCF* fusion with presumptive loss-of-function of CTCF (UPN16). In wUPD expression of *CTCF* ($p = 0.0526$) and cohesion complex gene *SMC3* ($p = 0.0357$) were lower compared to noUPD patients (Supplementary Fig. 8).

Clinical Comparison of patients with wUPD and noUPD

The demographics of the patients analysed in this study are shown in Table 2. wUPD were more likely to be female ($p = 0.0497$) and have a higher presenting WBC ($p = 0.0004$). A high end of induction MRD ($\geq 10^{-4}$) was detected in 11/29 (38%) and 4/23 (17%) in noUPD and wUPD respectively ($p = 0.1318$). Relapse rates, event free and overall survival were comparable in the two groups (Supplementary Fig. 9). Figure 6 shows the correlation of copy number alterations and mutations with therapeutic responses. There were no specific SCNA or mutation patterns seen in relapsed patients. *CREBBP* mutation/deletions^{14,44}, and in particular those associated with *KRAS* mutations¹⁵ have been described to be enriched at relapse. Of the 8 patients with a *CREBBP* mutation/deletion in this cohort, 2 have relapsed. UPN51 with both *KRAS* (VAF 14%) and *CREBBP* (VAF 39%) mutations remains in CR1. UPN 20 had a MRD of 2.3% at the end of induction with subsequent clearance of MRD after consolidation. A combined bone marrow and CNS relapse occurred 50 months after diagnosis. In this retrospective analysis we identified a homozygous deletion of *IKZF1*, *JAK2*^{R683G} mutation (VAF 56%) and a *CRLF2::P2RY8* fusion (100% cells), which meets the definition of Ph-like ALL⁴⁵. Relapse occurred in 5/16(31%)/ and 7/41(17%) of patients with MRD $\geq 10^{-4}$ and $< 10^{-4}$ respectively. *IKZF1* deletions and *NRAS* mutations were not seen more frequently in those with a MRD $\geq 10^{-4}$ as previously reported⁷.

Discussion

To date, reports on genomic characterisation of HH ALL have come primarily from analyses of western patients. In this Indian cohort, chromosomal gains, modality, structural abnormalities, frequency of microdeletions in genes recurrently deleted in BCP-ALL (*IKZF1*, *PAX5*, *CDKN2A/2B*, *ETV6*) and mutations

were similar to those previously reported^{13,28,46,47}. Recurrent deletions at 13q12.2 that result in loss of TAD borders and overexpression of *FLT3*,^{11,31} were not observed. wUPD in ALL occurs almost exclusively in HH ALL^{16,18}. Mutations occurred primarily on disomic chromosomes with noUPD. Of the 3 mutations identified in wUPD chromosome, one is likely to have occurred prior to the development of wUPD. This suggests the development of wUPDs along with gains of chromosomes are initial events in the leukaemogenic process with mutations as secondary events. The expression of *CTCF* and *SMC3* was lower in wUPD patients. The lower expression of *CTCF* and cohesion alters TAD boundaries impacting gene expression in HH ALLs⁴³, with aberrant cohesion complexes associated with increased copy number heterogeneity at the level of whole chromosome copy numbers in HH ALL⁴⁸. Thus, cohesion dysregulation is a potential mechanism for the higher modal numbers and more frequent gains of non-classical chromosomes in wUPD HH ALL. A more detailed analysis using Assay for Transposase Accessible Chromatin (ATAC) sequencing, chromatin immunoprecipitation (ChIP) chromosome conformation capture, followed by sequencing to study the chromatin interactions and gene dosage difference between wUPD and noUPD patients is required to investigate this possibility.

In HH, trisomies of +4, +10 and +17 are associated with good^{19,49,50}, and trisomies of 5 and 20 with poor outcomes^{51,52}. Presence of +17 and +18 in the absence of +5 and +20 identifies subset of HH patients with a lower relapse rate and good overall survival rate⁹. Patients with wUPD were more likely to have +5, though none of the patients in this series had a +20. Patients with noUPD were more likely to have mutations in RTK/*RAS* pathways while genes involving epigenetic and chromatin modifying pathways were more likely to be mutated in wUPD. Along with a higher frequency of *RAS* mutations, *NRAS* expression (unrelated to mutations) was higher in noUPD. A number of noUPD patients also had upregulation of *ABL1*. *NRAS* mutations⁷ and *ABL1* expression⁵³ are associated with high MRD levels in childhood ALL. This supports the observation that patients with noUPD had higher MRD levels.

Therapy for ALL patients with high MRD levels at the end of induction is intensified, as this is associated with improved outcomes⁵⁴. The intensive phase of treatment lasts 6–9 months and this is followed by maintenance therapy for 2–3 years. The main drug used during this latter phase is 6-Mercaptopurine. Intolerance or resistance to 6-Mercaptopurine is associated with late relapses occurring shortly before or after stopping therapy, and late relapse are seen in the majority of HH ALL. The AP-1 factors (*FOS*, *JUN* and *FOSB*) and *RAC2*, upregulated in wUPD, have been linked to resistance to 6-Mercaptopurine^{55,56}. *NUDT18*, a member of Nudix hydroxylase family, was upregulated in wUPD. Germline polymorphisms in *NUDT15*, another Nudix hydroxylase, results in decreased activity of the enzyme. This results in an increase in the incorporation of thiopurine metabolites during the maintenance phase of therapy leading to excessive toxicity⁵⁷. There is considerable overlap in the substrate activities of *NUDT15* and *NUDT18* including the dephosphorylation and inactivation of thiopurine metabolites⁵⁸. Arguably somatic overexpression of either gene could promote resistance to thiopurines and associate with late relapse. While higher MRD levels are associated with higher relapse rates, so is thiopurine resistance, and overall outcomes of between those with and without wUPD were comparable. Furthermore, in wUPD, but not in

noUPD, there was also an elevated expression of *C-MYC* unrelated to copy number change of 8q24 or translocation involving that region. Isolated C-MYC expression has been previously described at the protein level in BCP-ALL, but its prognostic significance is unknown⁵⁹.

Given the favourable outcomes of HH ALL, we would not expect to see a survival differences in this small cohort. Similarly, identifying differences in gene expression between the two groups would also benefit from a larger number of patients especially as the similarity in copy number variation influences expression. Nevertheless, a number of differences of potential prognostic significance were identified in patients with wUPD and noUPD. Patients with wUPD, were more likely to be females, have a higher presenting white cell count with a better MRD response. DEG analyses showed upregulation of cell cycle/mitosis genes (*CCNB1*, *CDC20* and *PLK2*) suggesting that wUPD is associated with cell proliferation. Though our investigations suggest possible biological explanations for variations in the therapeutic response in HH ALL, it is likely the complex genomic changes in HH ALL, result in therapeutically significant genetic changes in individual patients. An example in this study is UPN20 which had deletions of *IKZF1* (e2-3), and *PAR1* with a *JAK2*^{R683G} activating point mutation. He suffered late relapse and retrospective FISH analysis showed a *CRLF2* gene rearrangement. In our clinical setting, FISH and karyotype are used for initial diagnoses and treatment stratification⁶⁰ and would not be able to identify such patients.

This study provides possible explanations, based on differences in gene expression, that merit further functional investigation. Further dissection of the response variability, prognostication and alternative therapeutic interventions requires collaborative analyses to which our data obtained from a systematically treated cohort of patients is contributory.

Methods

Patients

All patients included in the study were diagnosed as precursor BCP-ALL, based on morphology and immunophenotyping, and treated on the Indian Childhood Collaborative Leukaemia Group (ICICLE)/(INPOG-ALL-15-ICiCLE-ALL-14; CTRI/2015/12/006434)⁶¹ protocol between 2015 to 2019. Consent was obtained prior to start of treatment and for collection and banking of samples. The institutional review board of the institute approved the study.

Banking and Nucleic Acid Extraction

Mononuclear (MNC) fractions were obtained with Ficoll density gradient centrifugation from bone marrow aspirates collected at diagnosis and remission. Genomic DNA and total RNA were extracted from MNC cells using QIAamp Blood kit from Qiagen and Trizol reagent (Invitrogen), respectively. The concentration and purity of the extracted nucleic acids were assessed using a Thermo Nanodrop ND-

2000 UV-Vis Spectrophotometer (Wilmington, DE, USA) and Qubit 2.0 Fluorometer (Thermo Fisher Scientific, USA) and gel electrophoresis. RNA quality was assessed using Agilent TapeStation 4200. Samples demonstrating > 6.0 RINe values were selected for further downstream analysis.

Cytogenetics

Cytogenetic analysis of pre-treatment bone marrow or peripheral blood was performed using karyotyping and FISH analysis using standard protocols⁶². High hyperdiploidy was diagnosed integrating the results of karyotyping and FISH analysis as described before⁶⁰.

SNP arrays analysis

SNP array analysis was performed using CytoScan™ HD, (ThermoFisher, USA) platform as per manufacturer's instructions. Briefly, DNA extracted from mono nuclear fractions obtained from bone marrow aspirates collected at diagnosis was digested and ligated to NSP1 adaptors followed by amplification using Titanium Taq. The PCR products were purified, fragmented, biotin labelled and hybridised to Cytoscan™ HD GeneChip. Data was analysed using the Chromosome Analysis Suite 3.1.1 (ChAS 3.1.1, ThermoFISHER) software suite based on GRCh37/hg19 as previously described^{63,64}.

Remission DNA samples from patients with one or more whole chromosome uni parental isodisomy (wUPD) at diagnosis on the Cytoscan™ HD analysis were analysed with an Infinium 850K Beadchip array SNP. The samples were processed in accordance with the manufacturers protocols. Illumina generated .GCT files were imported into BlueFuse Multi version 4.5 (Illumina) and analyzed using the default BlueFuse algorithm.

Targeted Panel Sequencing

A targeted panel of 95 genes (Supplementary Table 3), based on reported [mutations](#) in BCP-ALL patients²⁸ was designed using the Agilent SureDesign software to identify single nucleotide variants (SNV). The probe design included 102966 amplicons covering 1.977 million bases; all coding exons, intron-exon boundaries and average 1000 bp upstream promoter sequences within selected genes. Target regions were captured using Agilent HaloPlex High-Sensitivity Target Enrichment kit (Agilent technologies, Santa Clara, USA) according to the manufacturer's instructions. Library fragments were amplified by PCR followed by quality assessment using D1000 High-Sensitivity Screen-Tape analysis. Qubit fluorometry as well as a quantitative PCR-based (qPCR) approach using KAPA library quantification kit from Roche

Diagnostics were used to quantify each of the libraries before pooling them for sequencing. An expected average fragment size of 230 to 250 bp were considered as a typical library fragment to perform paired-end (2x75 bp) sequencing using Next-Seq 550 Illumina platform (Illumina, San Diego, CA, USA).

RNA Sequencing

Whole-transcriptome sequencing was performed using the Illumina TruSeq stranded Total RNA library prep protocol. Quantification of libraries was performed using Qubit 2.0. After passing the quality control parameters the libraries were sequenced using Next-Seq 550 instrument at 2x150 bp read-length. An average of 50-55 million reads were generated for each sample.

Informatics

Fastq files were quality assessed and mapped to the human genome 19 using Burrows-Wheeler aligner BWA, MEM algorithm and converted to BAM files. Alignment was refined with Broad GATK and quality controlled for miscalls and recalibrated prior to analysing variant calls. Three variant callers (VarScan, VarDict and Mutect2) were used to identify variants. Variants identified by at least 2 of the 3 tools were selected for gene-based functional annotation using ANNOVAR.

3408 variants were obtained after excluding synonymous variants. 2189 germline variants, identified from sequencing remission DNA samples, were excluded leaving 1219 single nucleotide variants (SNVs).

The 1000 genome and the dbSNP database were used to filter out single nucleotide polymorphisms as were variants with low allele frequencies (<12.25%), leaving 564 variants. PROVEAN, SIFT, FATHMM and MutationTaster (<http://www.mutationtaster.org/>) tools were used to predict the deleterious effect of mutations. The M-CAP pathogenicity classifier was used to classify rare missense variants. 273/564 variants were predicted to be pathogenic or probably pathogenic by at least two of 5 programs. Frameshift insertions, deletions substitutions and stop gain variants were considered pathogenic by default. Variants with less than 10 reads were excluded. 270 SNVs were inspected manually with the Integrative Genomic Viewer (IGV), and 138 variants validated by Sanger sequencing and IGV visualisation of RNA sequencing data (Supplementary Fig. 10). Mutations not previously described; or not validated by sequencing; or not present on RNA sequencing; or uncertain on IGV inspection, were excluded.

Transcriptomic Analysis

RNASeq samples were first mapped to the reference human genome (Grch37 assembly) using hisat2 alignment program. Aligned bam files obtained from the program were further used to generate the hisat2 specific counts for annotated human Ensembl Grch37 genes using featureCounts program. Finally, the count files were then used to identify a common set of differentially expressed (FDR < 0.1) genes from EdgeR and DESeq2 program for downstream analysis (Supplementary Table 8). We applied

independent hypothesis weighting (IHW) method⁶⁵ during multiple testing procedure that increases power compared to the method of Benjamini and Hochberg by assigning data-driven weights to each hypothesis. The principal component analysis (PCA) of gene expression profile of the patient samples was performed with the top-1000 highly variable gene set. The PCA and subsequent cluster analysis were performed using FactoMineR R package. We further estimated the copy-number normalized expression signal of each gene after dividing the raw counts of the genes by their observed copy-number values and then multiplied by their expected diploid copy-number.

Whole Genome Analysis

Fastq reads were aligned against build 37 of the human reference genome using BWA mem (0.7.15) software. The resulting files were converted to compressed binary format (BAM), sorted by coordinate, indexed, and marked for PCR duplicate reads using the Picard toolkit (bundled with GATK v4.1.8.1). Base quality scores were recalibrated with GATK to improve call accuracy and variants were called with the Haplotype Caller algorithm from GATK.

Copy Number Analysis

Chromosome coverage was computed as the median of the coverage of each individual base in the chromosome. The coverage of a position in a chromosome was computed as the number of reads mapped to the chromosome that spanned the position. In computing the coverage of a position in the chromosome, only reads with a mapping quality ≥ 20 with base quality ≥ 10 were considered. Unlike targeted exome sequencing, whole genome sequencing yields uniform coverage. Chromosome 13 was selected as reference and the ratios of median coverages of other chromosomes were evaluated with respect to the reference chromosome (Supplementary Fig. 11) to derive copy number status of each chromosome. Chromosomes that showed gains on SNP array analysis exhibited a corresponding high median coverage ratio (> 1.2) with respect to the reference.

Analysis of copy neutral loss of heterozygosity

Segments of genome with UPD are characterized by a high percentage of homozygous calls. In order to confirm the wUPD reported on SNP array analysis, variant calls with the highest confidence were evaluated in the whole genome data. Specifically, a high confidence subset of single nucleotide variants in the samples (calls marked as "PASS", with a $GQ \geq 30$) that had also been reported in 1000 genomes (phase 3) were shortlisted for the analysis. The fraction of homozygous variants per chromosome arm was computed and compared with variant calls from five 1000 genomes (phase 3) samples of similar sex and ethnicity (population code BEB) as reference (Supplementary Fig. 12). Chromosomes reported with wUPD on SNP array analysis showed significantly higher than expected fraction of homozygous SNV calls (Fig. 2B) in whole genome sequencing. wUPDs of chromosomes 2, 9, 19 in UPN25,

chromosomes 15, 19, 20 in UPN13 and in chromosome 10 in UPN24, were confirmed by analysis of whole genome data.

Structural Abnormalities (>5MB)	Number of patients (%)
1q duplication	27 (47%)
1q triplication	2(3.5%)
Isochromosome 7	4(7%)
8p deletion	2(3.5%)
16p deletion	2(3.5%)
6q deletion	1 (2)
20p deletion	1(2%)
13q duplication	1(2%)
12p duplication	1(2%)
19q duplicaiton	1(2%)
5q duplication	1(2%)
Xq duplication	1(2%)
isochromosome Xq	1(2%)
Chromopthisis like pattern involving chromosome 20	1(2%)

Table 1. Frequency of Structural abnormalities > 5 MB identified by Cytoscan -HD SNP array analysis in 57 cases of paediatric high hyperdiploid BCP-ALL. More than one structural abnormality was present in 13 patients.

Declarations

Author Contributions

VS, MP designed and conceptualised the study. VS, SK JS, NG supervised treatment and provided patient material. MP, DKM performed the diagnostic bone marrow studies. DD, ST, RS, PP, RI, DS, UZ, CB, PS performed biobanking and nucleic acid extractions. MP, MKS, BKC, ABG performed the cytogenetic studies. MP, CB, PS performed SNP array and analysed the SNP array data. DD designed the targeted gene sequencing panel. DD, RI, BV, ST performed the targeted gene panel and RNA sequencing. DD, RI, SG, CB, DS, UZ, PS, MS, BKC, ABG performed the validation experiments. AB, MP analysed the targeted panel sequencing data. AB, AC, SB, US, AR analysed the RNA-sequencing data. NKB, CD, US, AR, RS analysed

the whole genome data. AB, SB, AC, ND, MPG, US, AR, RS, MP curated and contributed to visualisation of the data. VS, MP, AB, DD wrote the original draft with input from all the authors. VS and MP supervised the study and have full access to all data and the final responsibility for publication. All authors reviewed the manuscript draft submitted for publication.

Acknowledgement

The diagnostic and clinical team at Tata Medical Center, patients and families. The study was supported in parts by grants from DBT-Wellcome Indian (grant reference IA/M/12/1/500261) to VS, DBT (grant reference BT/PR12046/MED/12/665/2014) to MP, DST (ECR/2017/002508) to DD. The Gigax foundation.

Conflict of Interest - All authors have no conflicts of interest to disclose.

Sources of support – Financial Support from DBT-Wellcome Indian Alliance (grant reference IA/M/12/1/500261), DBT (grant reference BT/PR12046/MED/12/665/2014) to MP, DST (grant reference ECR/2017/002508)

Data Availability

The Targeted sequencing data, whole genome sequencing data and RNA sequencing data and the SNP array data have been submitted in SRA and GEO databases at NCBI, Bio-project Accession ID: **PRJNA765140**

References

1. Heerema, N.A. *et al.* Specific extra chromosomes occur in a modal number dependent pattern in pediatric acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **46**, 684–93 (2007).
2. Paulsson, K. *et al.* Evidence for a single-step mechanism in the origin of hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **44**, 113–22 (2005).
3. Haas, O.A. Somatic Sex: On the Origin of Neoplasms With Chromosome Counts in Uneven Ploidy Ranges. *Front Cell Dev Biol* **9**, 631946 (2021).
4. Moorman, A.V. *et al.* A novel integrated cytogenetic and genomic classification refines risk stratification in pediatric acute lymphoblastic leukemia. *Blood* **124**, 1434–44 (2014).
5. Irving, J.A. *et al.* Integration of genetic and clinical risk factors improves prognostication in relapsed childhood B-cell precursor acute lymphoblastic leukemia. *Blood* **128**, 911–22 (2016).
6. Eckert, C. *et al.* Risk factors and outcomes in children with high-risk B-cell precursor and T-cell relapsed acute lymphoblastic leukaemia: combined analysis of ALLR3 and ALL-REZ BFM 2002 clinical trials. *Eur J Cancer* **151**, 175–189 (2021).

7. Parker, C. *et al.* Outcomes of patients with childhood B-cell precursor acute lymphoblastic leukaemia with late bone marrow relapses: long-term follow-up of the ALLR3 open-label randomised trial. *Lancet Haematol* **6**, e204-e216 (2019).
8. Moorman, A.V. *et al.* Outcome heterogeneity in childhood high-hyperdiploid acute lymphoblastic leukemia. *Blood* **102**, 2756–62 (2003).
9. Enshaei, A., Vora, A., Harrison, C.J., Moppett, J. & Moorman, A.V. Defining low-risk high hyperdiploidy in patients with paediatric acute lymphoblastic leukaemia: a retrospective analysis of data from the UKALL97/99 and UKALL2003 clinical trials. *Lancet Haematol* **8**, e828-e839 (2021).
10. Stanulla, M. *et al.* IKZF1(plus) Defines a New Minimal Residual Disease-Dependent Very-Poor Prognostic Profile in Pediatric B-Cell Precursor Acute Lymphoblastic Leukemia. *J Clin Oncol* **36**, 1240–1249 (2018).
11. Paulsson, K. *et al.* Genetic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Proc Natl Acad Sci U S A* **107**, 21719–24 (2010).
12. Steeghs, E.M.P. *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* **9**, 4634 (2019).
13. Paulsson, K. *et al.* The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Nat Genet* **47**, 672–6 (2015).
14. Waanders, E. *et al.* Mutational landscape and patterns of clonal evolution in relapsed pediatric acute lymphoblastic leukemia. *Blood Cancer Discov* **1**, 96–111 (2020).
15. Malinowska-Ozdowy, K. *et al.* KRAS and CREBBP mutations: a relapse-linked malicious liaison in childhood high hyperdiploid acute lymphoblastic leukemia. *Leukemia* **29**, 1656–67 (2015).
16. Lundin, K.B. *et al.* Patterns and frequencies of acquired and constitutional uniparental isodisomies in pediatric and adult B-cell precursor acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **55**, 472–9 (2016).
17. Zaliouva, M. *et al.* Slower early response to treatment and distinct expression profile of childhood high hyperdiploid acute lymphoblastic leukaemia with DNA index < 1.16. *Genes Chromosomes Cancer* **55**, 727–37 (2016).
18. Lundin-Strom, K.B. *et al.* Whole-exome sequencing exploration of acquired uniparental disomies in B-cell precursor acute lymphoblastic leukemia. *Leukemia* **32**, 2058–2062 (2018).
19. Moorman, A.V. *et al.* Outcome heterogeneity in childhood high-hyperdiploid acute lymphoblastic leukemia. *Blood* **102**, 2756–62 (2003).
20. Paulsson, K. *et al.* High modal number and triple trisomies are highly correlated favorable factors in childhood B-cell precursor high hyperdiploid acute lymphoblastic leukemia treated according to the NOPHO ALL 1992/2000 protocols. *Haematologica* **98**, 1424–32 (2013).
21. Vojcek, Á. *et al.* Conserved hierarchical gain of chromosome 4 is an independent prognostic factor in high hyperdiploid pediatric acute lymphoblastic leukemia. *Leuk Res* **52**, 28–33 (2017).

22. Raimondi, S.C. *et al.* Heterogeneity of hyperdiploid (51–67) childhood acute lymphoblastic leukemia. *Leukemia* **10**, 213 – 24 (1996).
23. Cai, H. *et al.* Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens. *BMC Genomics* **15**, 82 (2014).
24. Magrangeas, F., Avet-Loiseau, H., Munshi, N.C. & Minvielle, S. Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients. *Blood* **118**, 675–8 (2011).
25. Chen, C. *et al.* Next-generation-sequencing of recurrent childhood high hyperdiploid acute lymphoblastic leukemia reveals mutations typically associated with high risk patients. *Leuk Res* **39**, 990–1001 (2015).
26. Iacobucci, I. & Mullighan, C.G. Genetic Basis of Acute Lymphoblastic Leukemia. *J Clin Oncol* **35**, 975–983 (2017).
27. Kuiper, R.P. *et al.* High-resolution genomic profiling of childhood ALL reveals novel recurrent genetic lesions affecting pathways involved in lymphocyte differentiation and cell cycle progression. *Leukemia* **21**, 1258–66 (2007).
28. Mullighan, C.G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758 – 64 (2007).
29. Paulsson, K. *et al.* Microdeletions are a general feature of adult and adolescent acute lymphoblastic leukemia: Unexpected similarities with pediatric disease. *Proceedings of the National Academy of Sciences* **105**, 6708–6713 (2008).
30. Wen, H. *et al.* New fusion transcripts identified in normal karyotype acute myeloid leukemia. *PLoS One* **7**, e51203 (2012).
31. Yang, M. *et al.* 13q12.2 deletions in acute lymphoblastic leukemia lead to upregulation of FLT3 through enhancer hijacking. *Blood* **136**, 946–956 (2020).
32. Tao, R. *et al.* Genetic polymorphisms of ARID5B rs7089424 and rs10994982 are associated with B-lineage ALL susceptibility in Chinese pediatric population. *J Chin Med Assoc* **82**, 562–567 (2019).
33. Treviño, L.R. *et al.* Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat Genet* **41**, 1001–5 (2009).
34. Paulsson, K. *et al.* Mutations of FLT3, NRAS, KRAS, and PTPN11 are frequent and possibly mutually exclusive in high hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **47**, 26–33 (2008).
35. Ali, N.M. *et al.* Comprehensive Molecular Characterization of Adamantinoma and OFD-like Adamantinoma Bone Tumors. *Am J Surg Pathol* **43**, 965–974 (2019).
36. Angeles, A.K.J., Yu, R.T.D., Cutiongco-De La Paz, E.M. & Garcia, R.L. Phenotypic characterization of the novel, non-hotspot oncogenic KRAS mutants E31D and E63K. *Oncol Lett* **18**, 420–432 (2019).
37. Muñoz-Maldonado, C., Zimmer, Y. & Medová, M. A Comparative Analysis of Individual RAS Mutations in Cancer Biology. *Front Oncol* **9**, 1088 (2019).

38. Oshima, K. *et al.* Mutational landscape, clonal evolution patterns, and role of RAS mutations in relapsed acute lymphoblastic leukemia. *Proc Natl Acad Sci U S A* **113**, 11306–11311 (2016).
39. Ryland, G.L. *et al.* Loss of heterozygosity: what is it good for? *BMC Med Genomics* **8**, 45 (2015).
40. Gruszka-Westwood, A.M. *et al.* Comparative expressed sequence hybridization studies of high-hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **41**, 191–202 (2004).
41. Strefford, J.C. *et al.* Complex genomic alterations and gene expression in acute lymphoblastic leukemia with intrachromosomal amplification of chromosome 21. *Proc Natl Acad Sci U S A* **103**, 8167–72 (2006).
42. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–40 (2010).
43. Yang, M. *et al.* Proteogenomics and Hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia. *Nat Commun* **10**, 1519 (2019).
44. Inthal, A. *et al.* CREBBP HAT domain mutations prevail in relapse cases of high hyperdiploid childhood acute lymphoblastic leukemia. *Leukemia* **26**, 1797–803 (2012).
45. Roberts, K.G. *et al.* Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. *N Engl J Med* **371**, 1005–15 (2014).
46. Paulsson, K. *et al.* Identification of cryptic aberrations and characterization of translocation breakpoints using array CGH in high hyperdiploid childhood acute lymphoblastic leukemia. *Leukemia* **20**, 2002–7 (2006).
47. Sulong, S. *et al.* A comprehensive analysis of the CDKN2A gene in childhood acute lymphoblastic leukemia reveals genomic deletion, copy number neutral loss of heterozygosity, and association with specific cytogenetic subgroups. *Blood* **113**, 100–7 (2009).
48. Moura-Castro, L.H. *et al.* Sister chromatid cohesion defects are associated with chromosomal copy number heterogeneity in high hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **60**, 410–417 (2021).
49. Schultz, K.R. *et al.* Risk- and response-based classification of childhood B-precursor acute lymphoblastic leukemia: a combined analysis of prognostic markers from the Pediatric Oncology Group (POG) and Children's Cancer Group (CCG). *Blood* **109**, 926 – 35 (2007).
50. Sutcliffe, M.J. *et al.* High concordance from independent studies by the Children's Cancer Group (CCG) and Pediatric Oncology Group (POG) associating favorable prognosis with combined trisomies 4, 10, and 17 in children with NCI Standard-Risk B-precursor Acute Lymphoblastic Leukemia: a Children's Oncology Group (COG) initiative. *Leukemia* **19**, 734 – 40 (2005).
51. Heerema, N.A. *et al.* Prognostic impact of trisomies of chromosomes 10, 17, and 5 among children with acute lymphoblastic leukemia and high hyperdiploidy (> 50 chromosomes). *J Clin Oncol* **18**, 1876–87 (2000).
52. Chilton, L. *et al.* High hyperdiploidy among adolescents and adults with acute lymphoblastic leukaemia (ALL): cytogenetic features, clinical characteristics and outcome. *Leukemia* **28**, 1511–8

- (2014).
53. Biondi, A. *et al.* Imatinib after induction for treatment of children and adolescents with Philadelphia-chromosome-positive acute lymphoblastic leukaemia (EsPhALL): a randomised, open-label, intergroup study. *Lancet Oncol* **13**, 936–45 (2012).
 54. Vora, A. *et al.* Augmented post-remission therapy for a minimal residual disease-defined high-risk subgroup of children and young people with clinical standard-risk and intermediate-risk acute lymphoblastic leukaemia (UKALL 2003): a randomised controlled trial. *Lancet Oncol* **15**, 809–18 (2014).
 55. Holland, M. *et al.* RAC2, AEP, and ICAM1 expression are associated with CNS disease in a mouse model of pre-B childhood acute lymphoblastic leukemia. *Blood* **118**, 638–49 (2011).
 56. Marinkovic, G. *et al.* Inhibition of GTPase Rac1 in endothelium by 6-mercaptopurine results in immunosuppression in nonimmune cells: new target for an old drug. *J Immunol* **192**, 4370–8 (2014).
 57. Moriyama, T. *et al.* NUDT15 polymorphisms alter thiopurine metabolism and hematopoietic toxicity. *Nat Genet* **48**, 367–73 (2016).
 58. Carreras-Puigvert, J. *et al.* A comprehensive structural, biochemical and biological profiling of the human NUDIX hydrolase family. *Nat Commun* **8**, 1541 (2017).
 59. Allen, A. *et al.* C-myc protein expression in B-cell acute lymphoblastic leukemia, prognostic significance? *Leuk Res* **38**, 1061–6 (2014).
 60. Parihar, M. *et al.* A triple-probe FISH screening strategy for risk-stratified therapy of acute lymphoblastic leukaemia in low-resource settings. *Pediatr Blood Cancer* **65**, e27366 (2018).
 61. Das, N. *et al.* Protocol for ICiCLE-ALL-14 (InPOG-ALL-15-01): a prospective, risk stratified, randomised, multicentre, open label, controlled therapeutic trial for newly diagnosed childhood acute lymphoblastic leukaemia in India. *Trials* **23**, 102 (2022).
 62. Wan, T.S. Cancer cytogenetics: methodology revisited. *Ann Lab Med* **34**, 413 – 25 (2014).
 63. Choi, S.M., Papenhausen, P., Wertheim, G. & King, R.L. Near-haploid B lymphoblastic leukemia with an apparent hyperdiploid karyotype: the critical role of SNP analysis in establishing proper diagnosis. *Journal of Hematopathology* **7**, 27–32 (2014).
 64. South, S.T. *et al.* ACMG Standards and Guidelines for constitutional cytogenomic microarray analysis, including postnatal and prenatal applications: revision 2013. *Genetics in Medicine* **15**, 901–909 (2013).
 65. Ignatiadis, N., Klaus, B., Zaugg, J.B. & Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods* **13**, 577–580 (2016).

Table 2

Table 2 is available in the supplementary files section.

Figures

Figure 1

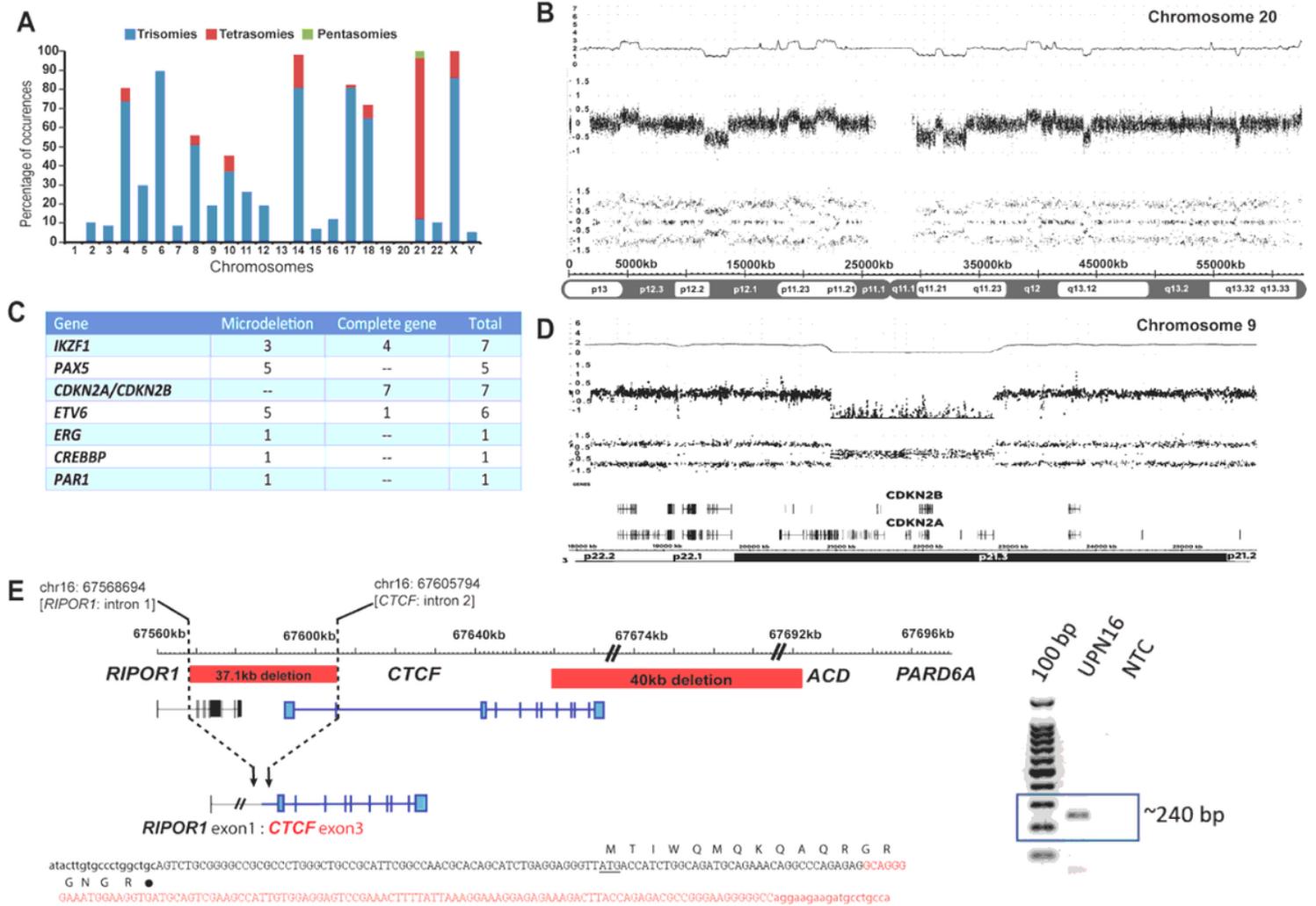


Figure 1

A. Chromosomal gains in 57 high hyperdiploid BCP ALL patients analysed by SNP array. Trisomy and tetrasomy in the sex chromosomes X and Y correspond to one or two additional copies of these chromosomes respectively. **B.** Chromophthisis like copy number pattern in chromosome 20 of UPN 9. Disomy of chromosome 20 is apparent with an average \log_2 ratio ~ 0 and A and B cell frequencies of 0, 1, and -1. Multiple segments of contiguous and concurrent gains and losses with sizes ranging from 60.707 kbps to 2059.05 kbps are visible across the short arm and long arms of the chromosome suggesting a chromophthisis like copy number change pattern. **C.** Frequently identified genetic deletions. Larger copy number abnormalities leading to whole gene deletions were seen for *IKZF1*, *CDKN2A/2B*, and *ETV6* gene. Of the 4 patients showing complete deletion of *IKZF1*, isochromosome 7q was present in 3 patients and monosomy 7 in one patient. **D.** Uniparental isodisomy of chromosome 9 in UPN 54, with a 112 kbps deletion at 9p13.2 resulting in a homozygous deletion of the *CDKN2A/2B* locus. **E.** Schematic showing a 37.1kb deletion in 16q22.1 in UPN 16. The deletion spans the intragenic regions of *RIPOR1* (intron1) and *CTCF* (intron2), resulting in juxtaposition of *RIPOR1* intron 1 to *CTCF* intron 2, with loss of the first exon and translational starting site of the *CTCF* gene (left panel). The presence of a fusion transcript was

confirmed by RT-PCR (right panel). The fusion transcript encodes for a non-functional CDS due to early termination of stop codon, resulting in loss of function of both *RIPOR1* and *CTCF* genes. The schematic also shows a deletion involving exons 6-12 of the *CTCF* gene and exons 10-12 of *ACD* resulting in a *CTCF-PARD6A* fusion reported as reported earlier (13).

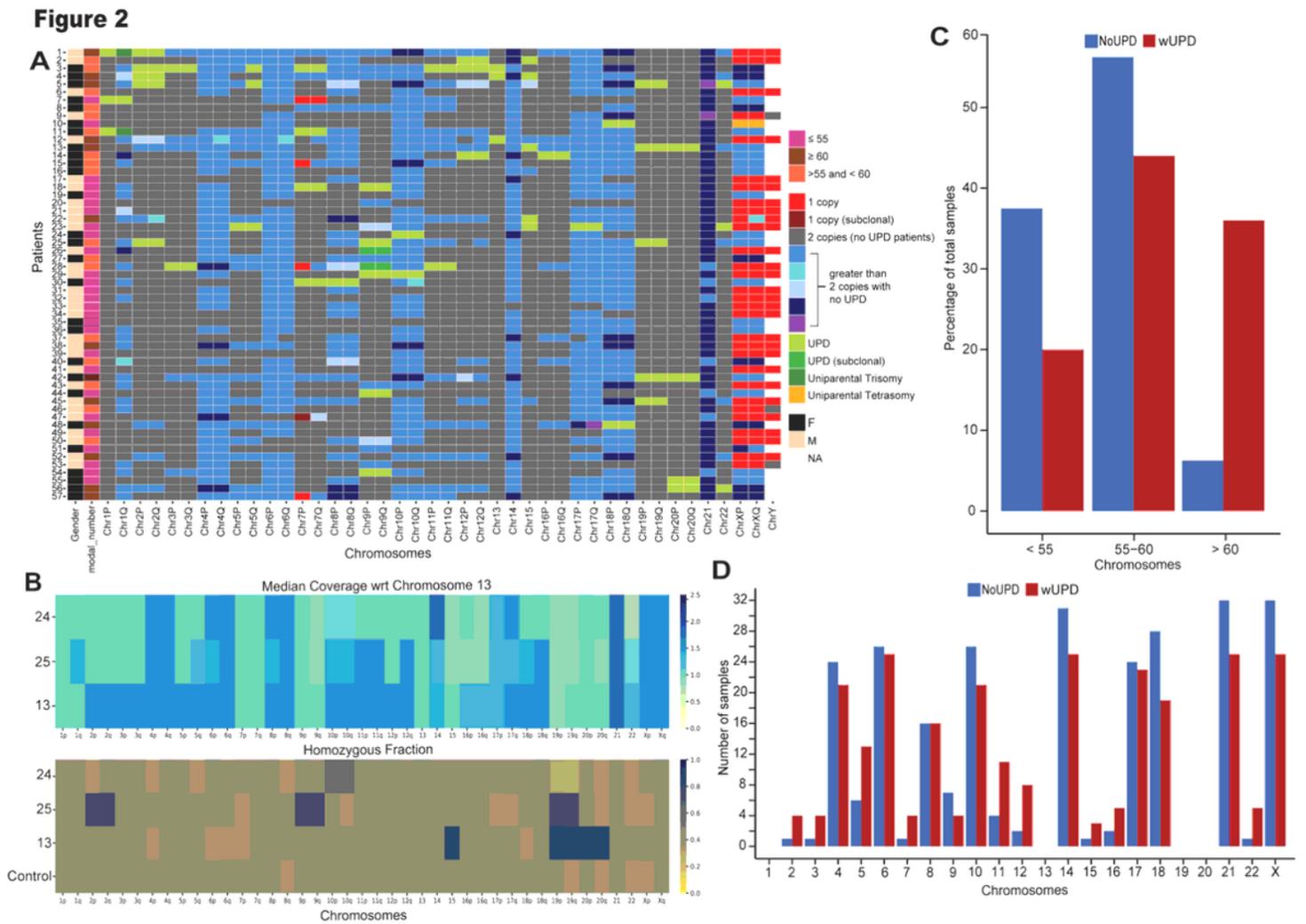


Figure 2

A. Schematic of chromosomal copy number changes and whole chromosomal uniparental isodisomy (wUPD) detected in the study cohort. Patients are in rows. Chromosomes are represented in columns, with separate adjacent columns representing p and q arms for non-acrocentric chromosomes. Copy number changes are colour coded (right panel). **B.** Whole genome analysis of three patients (UPN 13, 24 and 25). The top panel shows copy number status and the bottom panel shows heterozygosity status. Patients are in rows. Chromosomes are represented in columns, with separate adjacent columns representing p and q arms for non-acrocentric chromosomes. Top Panel: Shows copy number derived by calculating the number of reads mapped to specific chromosomes in reference to the number of reads mapped on to chromosome 13. Bottom panel: Fraction of homozygous SNV calls per chromosome arm derived by

comparing the SNVs in the samples with SNVs from 5 samples of similar sex and ethnicity procured from the 1000 genomes dataset. The copy number status and copy neutral loss of heterozygosity identified on the whole genome analysis is concurrent with SNP array analysis. **C.** Distribution of chromosomal modal numbers in patients with wUPD and noUPD. **D.** Gains of individual chromosomes in wUPD and noUPD patients. The X-axis represents the individual chromosomes and the Y-axis shows percentage of patients with gains of chromosomes.

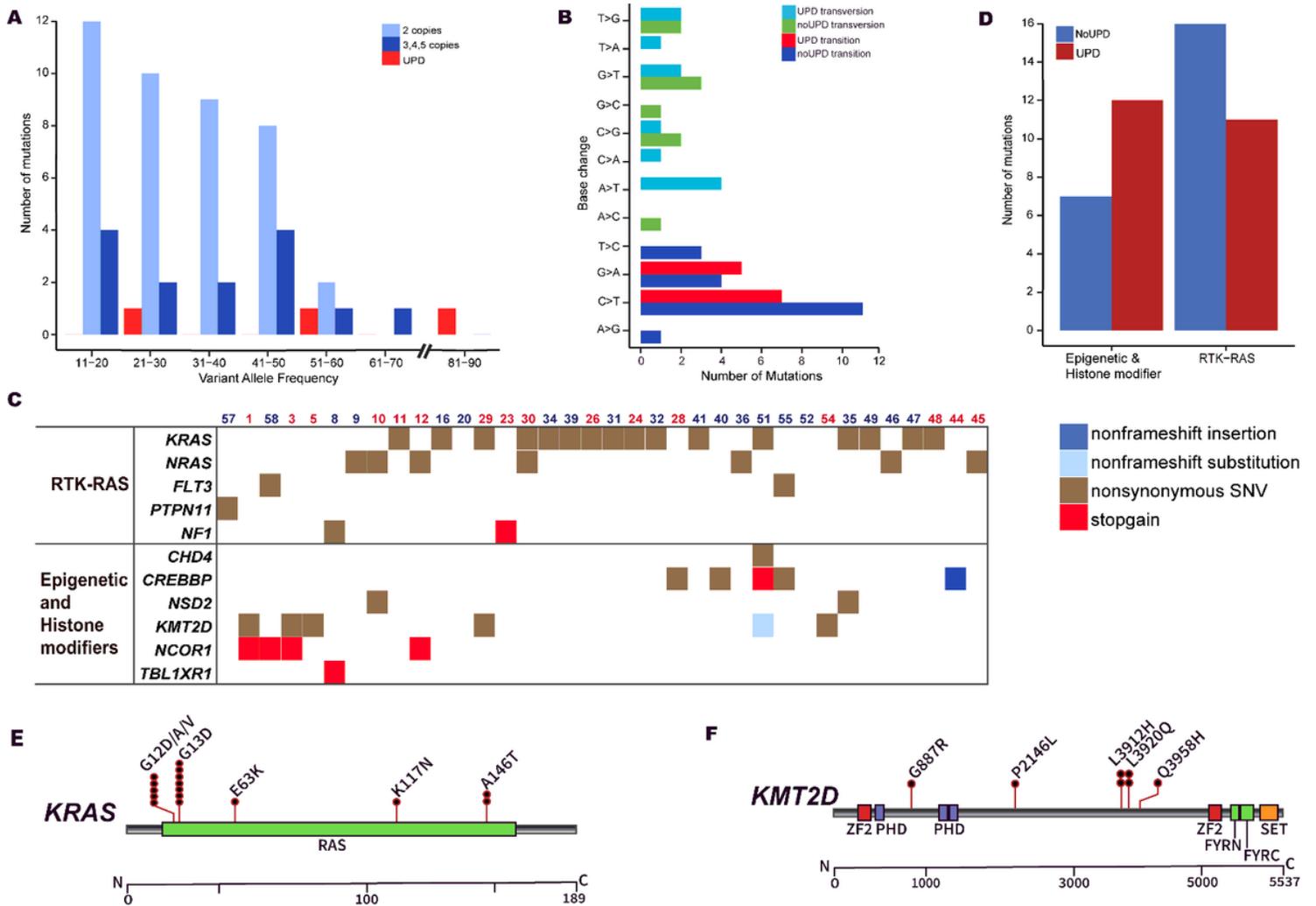
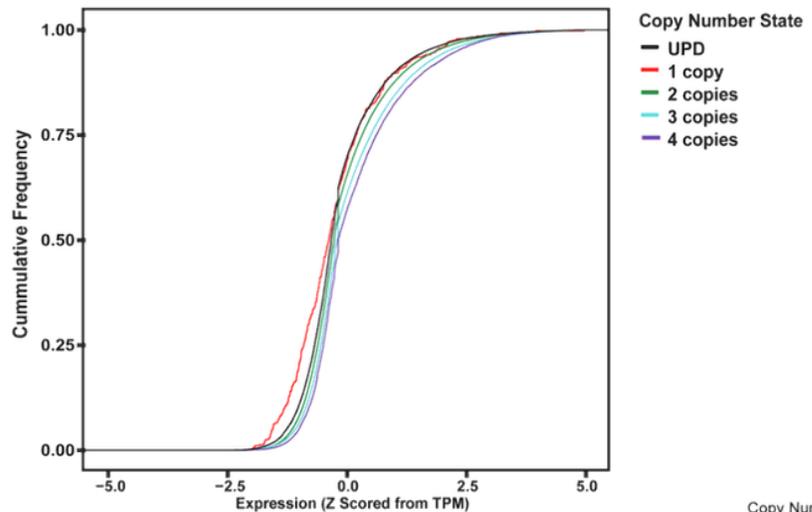


Figure 3

A Variant allele frequencies of mutations in relation to the copy number status of the chromosomes carrying the mutated gene. The variant allele frequencies are shown on X axis and the number of mutations on Y-axis. The light blue (2 copies), dark blue (> 2 copies) and red (wUPD) coloured bars represent the copy number status of the chromosome carrying the mutated gene. The majority (86%) of mutations had VAFs < 50% with a peak between 31-40%. **B** Mutational signature shows predominance of C>T transitions in a CpG context. **C** Waterfall plot showing predicted recurrent pathogenic mutations in 36 of 58 HH patients (columns). Genes involving the RAS pathway, receptor tyrosine kinase pathway, and

epigenetic and histone modifying pathways were most frequently mutated. Non-recurrent mutations have been excluded from the figure. Mutations are colour coded according to the types of the mutation. **D** Frequency of mutations in RAS-RTK pathway and epigenetic/ histone modifying pathway in wUPD and noUPD HH ALL patients. **E** Schematic diagram representing mutations in the *KRAS* gene. Types of mutations are shown along with colour coded functional domains. **F** Schematic diagram representing mutations in the *KMT2D* gene. Types of mutations are shown along with colour coded functional domains.

Figure 4
A



B

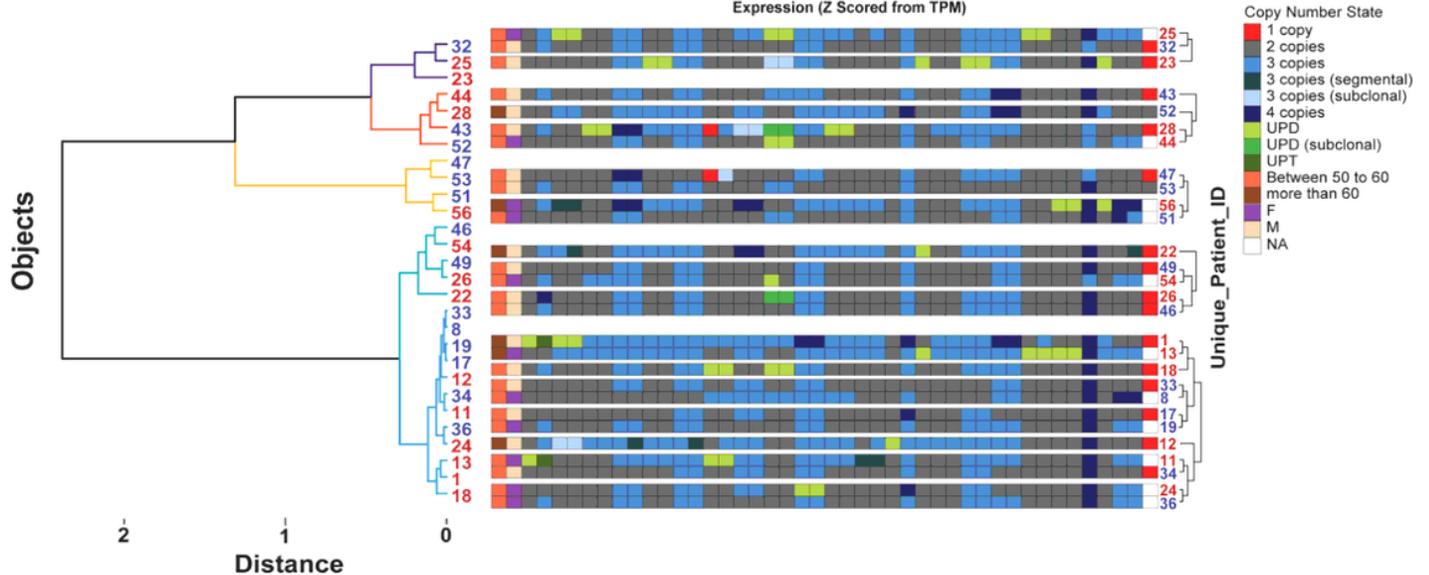


Figure 4

A Cumulative frequency distribution of gene expression values per copy number state. Gene expression values were determined from RNA sequencing data as fragments per kilobase of transcript per million reads and normalized (per gene) to mean 0 and variance 1. Copy number is correlated with gene expression. Gene expression of wUPD chromosomes is closer to single copy number. **B** Left panel: Cluster dendrogram based on unsupervised hierarchical clustering using the 1000 most variably expressed genes

derived applying the Mahalanobis distance on gene expression data obtained from RNA sequencing of 28 (14 wUPD and 14 no UPD) HH patients. The right panel shows the clustering obtained based on somatic copy number abnormalities derived from the SNP array analysis.

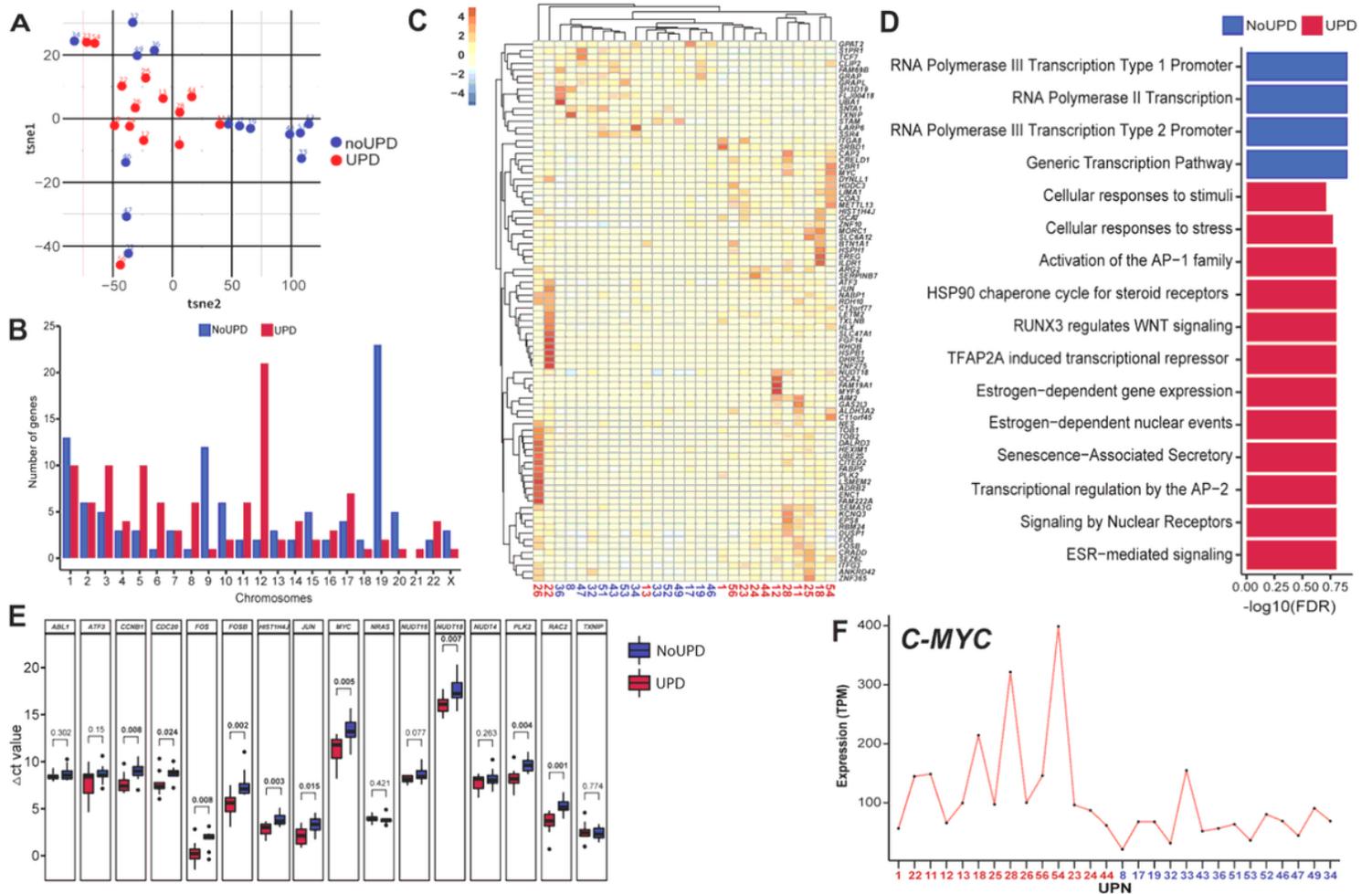


Figure 5

A tsNE plot using the 1000 most variably expressed genes. The plot shows a central cluster of wUPD patients, along with peripheral clusters of non-UPD and combinations of noUPD and wUPD. **B** Differentially expressed genes (N=221) in wUPD and noUPD mapped to the chromosomal location. Chromosomes are represented on the X-axis. The height of the bars on the Y-axis represents number of upregulated genes between wUPD and noUPD patients on each chromosome. **C** Hierarchical clustering of 28 HH ALL patients based on 89 differentially expressed genes derived using EdgeR (FDR <0.1) and gene enrichment analysis with TopPr. Heat map shows genes relatively overexpressed in wUPD HH ALL patients (in red) and genes under expressed (in blue) when compared to noUPD HH ALL patients. **D** Pathway analysis using 221 protein coding genes differentially expressed in wUPD and noUPD patients. **E** Box plot showing RQ-RT PCR validation of gene expression data in differentially expressed candidate genes between wUPD and noUPD HH ALL. Genes analysed are on the X-axis. Y-axis

shows the normalised expression (Delta Ct) with the Inter quartile range of the expression values. The center of the box plot is the median and the lower and upper hinges corresponds to the first and third quartiles respectively; the whiskers are 1.25 times of the interquartile range and data beyond this range are plotted as individual points. Non-parametric Wilcoxon rank sum test was used to calculate the p values. **F** Copy number normalised TPM values of *cMYC* gene in wUPD and noUPD patients.

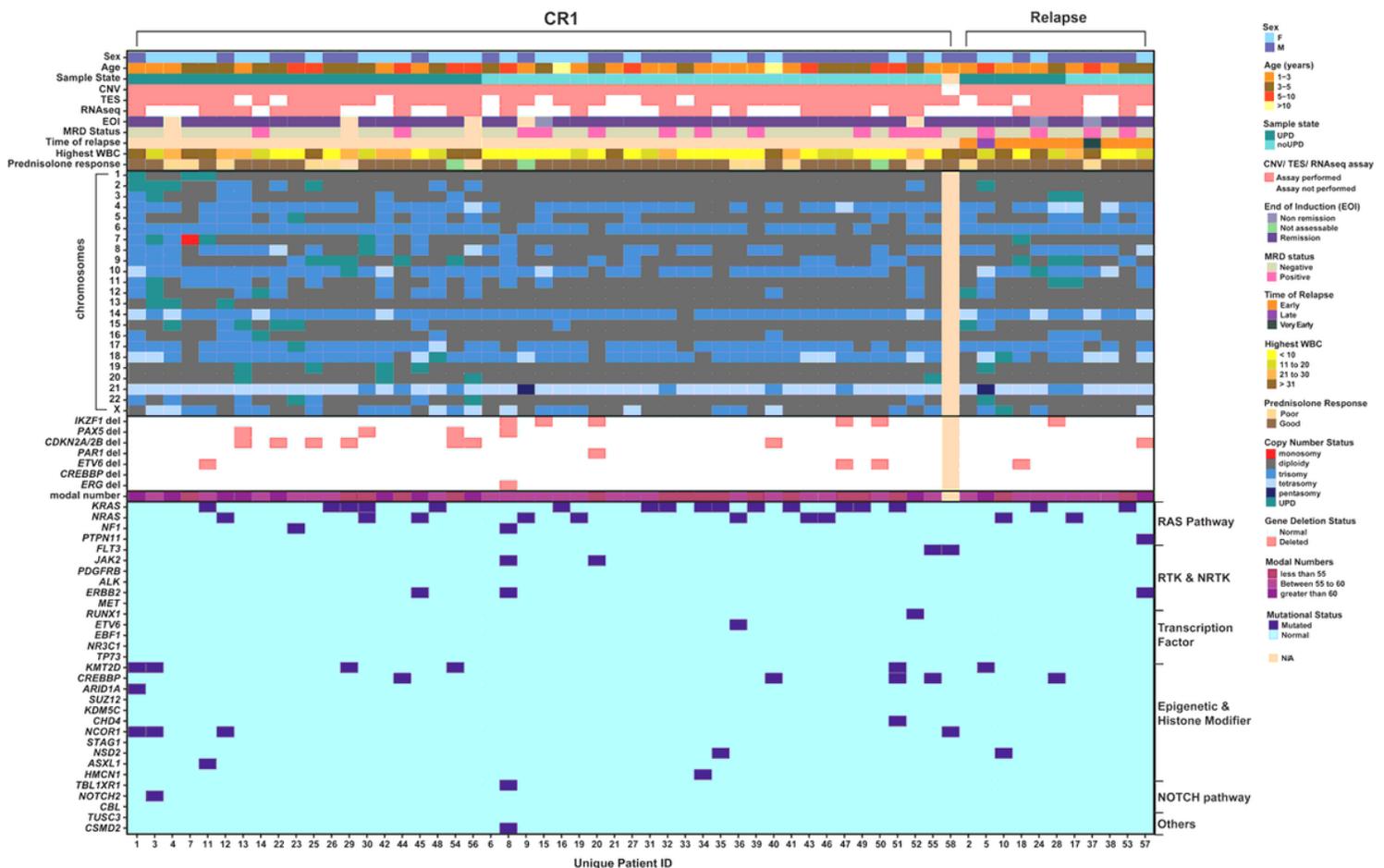


Figure 6

Summary of patient demographics, clinical data including WBC count, prednisolone response, EOI MRD, copy number status and mutations of 58 HH ALL patients.

Supplementary Files

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

- [SupplementaryTable2.xlsx](#)
- [Supplementarydata07.02.2022.pdf](#)
- [SupplementaryTable4.xlsx](#)

- [SupplementaryTable5.xlsx](#)
- [SupplementaryTable7.xlsx](#)
- [SupplementaryTable6.xlsx](#)
- [SupplementaryTable3.xlsx](#)
- [SupplementaryTable8.xlsx](#)
- [SupplementaryTable1.xlsx](#)
- [Table2.jpg](#)