nature medicine

Article

https://doi.org/10.1038/s41591-022-02113-6

# Evolutionary history of transformation from chronic lymphocytic leukemia to Richter syndrome

Received: 23 December 2021

Accepted: 28 October 2022

Published online: 9 January 2023

Check for updates

A list of authors and their affiliations appears at the end of the paper

Richter syndrome (RS) arising from chronic lymphocytic leukemia (CLL) exemplifies an aggressive malignancy that develops from an indolent neoplasm. To decipher the genetics underlying this transformation, we computationally deconvoluted admixtures of CLL and RS cells from 52 patients with RS, evaluating paired CLL–RS whole-exome sequencing data. We discovered RS-specific somatic driver mutations (including *IRF2BP2*, *SRSF1*, *B2M*, *DNMT3A* and *CCND3*), recurrent copy-number alterations beyond del(9p21)(*CDKN2A/B*), whole-genome duplication and chromothripsis, which were confirmed in 45 independent RS cases and in an external set of RS whole genomes. Through unsupervised clustering, clonally related RS was largely distinct from diffuse large B cell lymphoma. We distinguished pathways that were dysregulated in RS versus CLL, and detected clonal evolution of transformation at single-cell resolution, identifying intermediate cell states. Our study defines distinct molecular subtypes of RS and highlights cell-free DNA analysis as a potential tool for early diagnosis and monitoring.

Transformation to a high-grade malignancy accounts for therapeutic resistance and rapid disease progression across cancers[1–3]. RS, an aggressive lymphoma developing in patients with CLL, is a striking example of transformation[3]. RS is associated with median overall survival of less than 1 year, even in the modern era[3]. Despite advanced genomic characterization of CLL[4,5,6], understanding of the genetic factors driving the evolution of CLL to RS remains limited. This is partly because of the difficulties in acquiring RS tissue and paired antecedent CLL cells. These challenges have precluded comparative evolutionary analysis, and limited the ability to define the molecular events underlying transformation beyond alterations in *TP53*, *NOTCH1*, *CDKN2A/B* and *MYC*[7–10]. Although a subset of RS is believed to be clonally unrelated based on immunoglobulin heavy chain (IGHV) sequencing[3,10], a genome-wide analysis to exclude shared ancestry has not been yet performed. Finally, RS biopsies contain admixtures of RS and CLL cells, mandating development of tools for in silico deconvolution of RS and CLL genetic changes. To delineate factors contributing to high-grade transformation definitively, we analyzed exomes from matched RS and CLL DNA from 52 patients and confirmed our findings in 45 independent RS patients and 14 external RS cases[9].

## Results

### Developing an analytic framework to discover RS drivers

We assembled a discovery cohort of 53 patients with paired CLL and RS samples of predominantly diffuse large B cell histology (DLBCL), the most common form of transformation[3] (Fig. 1a and Supplementary Tables 1 and 2). Forty-five (83%) patients received previous CLL-directed therapies, with 11 (21%) having received targeted agents. Thirty-nine (72%) patients had unmutated IGHV CLL (U-CLL). Whole-exome sequencing (WES) was completed for 186 DNA samples (from 53 patients) and whole-genome sequencing (WGS) for 30 samples (11 patients; Supplementary Table 3). WES data from 42 patients originated from matched CLL–RS–germline samples (trios) and 10 from paired CLL–RS (duos). As validation, we performed WES on 45 independent RS cases, 17 of which were duos (Fig. 1b and Supplementary Table 1–3).

e-mail: gadgetz@broadinstitute.org; cwu@partners.org

https://doi.org/10.1038/s41591-022-02113-6

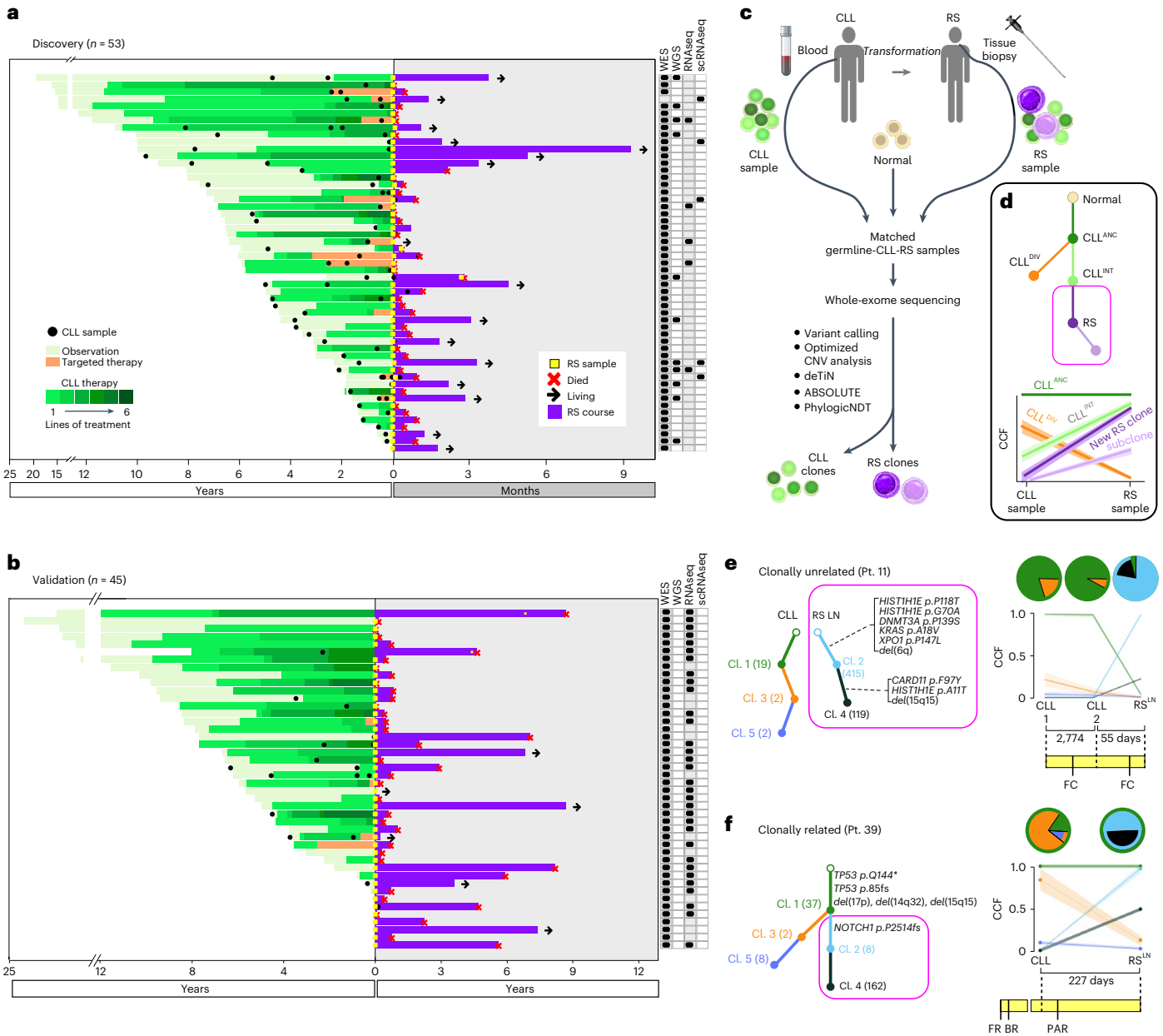



**Fig. 1 | Developing an analytic framework for detecting RS-specific clones.** **a**, Disease course of 53 RS patients from CLL diagnosis in relationship to lines of therapy and sample collection. **b**, Disease course of 44 of 45 RS validation cohort patients from CLL diagnosis in relationship to lines of therapy and sample collection (one patient with missing data). **c**, Computational schema for deciphering CLL and RS clones within RS biopsy samples. **d**, Inset shows labeled sample phylogenetic tree with associated sample cancer cell fraction (CCF) plot. Phylogenetic trees with CCF clustering, clonal abundance and associated patient disease course in representative clonally unrelated (**e**) and related (**f**) cases.

To delineate the driver events giving rise to RS, we employed established WES analysis tools and three additional steps: (1) deTiN[11] to recover somatic mutations filtered due to tumor-in-normal contamination, (2) an optimized tool to detect somatic copy-number alterations (sCNAs) and (3) PhylogicNDT[12] to establish the clonal composition per patient sample and infer the phylogenetic tree (Fig. 1c and Extended Data Fig. 1a–c). RS clones were defined as new clones arising in the RS sample, not present in the antecedent CLL sample and distinct based on somatic single nucleotide variants (sSNVs) and sCNAs. Within the CLL compartment, phylogenetic trees identified the ancestral ($CLL^{ANC}$), intermediate ($CLL^{INT}$, which expanded to give rise to RS that arose from $CLL^{ANC}$) and divergent ($CLL^{DIV}$) clones (Fig. 1d). RS was identified as related to CLL if at least one common $CLL^{ANC}$ clone was shared.

These tools were applied to infer the CLL and RS clonal structure and relatedness (Extended Data Figs. 1d, 2 and 3 and Supplementary Figs. 1 and 2). We identified instances of clonal unrelatedness to the antecedent CLL (Fig. 1e and Extended Data Fig. 2), previously classified based on IGHV sequencing in ~20% of RS[3,10]. Most RS were clonally related to the antecedent CLL ($n = 45$, 87%; Fig. 1f). Evolutionary relationships were secondarily determined by comparing the immunoglobulin gene sequence (Supplementary Table 4), largely in line with the WES-based phylogenies.

## Defining the genomic landscape of RS

To determine the pure RS genomic landscape, we identified: (1) events strictly present in RS cells through computational isolation of the RS

lineage separate from CLL$^{DIV}$ (Fig. 2a, grey outline), and (2) events newly acquired in RS clones (Fig. 2a, magenta outline). To uncover drivers of transformation, we applied MutSig2CV[13] and GISTIC2.0 (ref. 14) (Fig. 2b,c and Supplementary Table 5, Methods).

From our discovery cohort, we observed mutations in known CLL drivers (*NOTCH1*, *TP53* and *SF3B1*; Fig. 2b) and identified new candidate RS drivers (Fig. 2b,c and Extended Data Fig. 4a–x). These included mutations in *IRF2BP2* ($n = 7$), which encodes an IRF2-dependent transcriptional corepressor, which is mutated in the N1 subtype of DLBCL[15] and primary mediastinal B cell lymphoma[16] (Extended Data Fig. 4d). Inactivating mutations in the DNA methyltransferase *DNMT3A* (8%) were previously reported as a single case in RS[8]; genetically engineered mice modeling this alteration have confirmed its CLL-driving function and impact on Notch signaling[17,18] (Extended Data Fig. 4e). *B2M* loss through inactivating mutations, a mechanism of immune escape across cancers[19–22], was observed in three patients (Extended Data Fig. 4f). The detected mutations in the MYC-interacting[23] splicing factor *SRSF1* ($n = 4$) did not co-occur with mutated *SF3B1*, consistent with mutual exclusivity of splicing factor mutations across cancers[24] (Extended Data Fig. 4g). *EZH2* hotspot alterations were found in two clonally unrelated RS cases, as in DLBCL[15,19], whereas *EZH2* frameshift was seen in one related RS case (Extended Data Fig. 4h).

Strikingly, we detected numerous sCNAs (Fig. 2b–d and Extended Data Fig. 5a,b; Methods), including *del*(17p) (*TP53*, 63%) and *del*(9p21.3) (*CDKN2A/B*, 19%), with arm-level loss of 9p in five additional patients. Recurrent focal events beyond common CLL drivers included *del*(15q13.1l) (*MGA* and *B2M*, 21%), amplification (amp) of chromosome 8q24 (*MYC*, 15%), *del*(7q36) (*EZH2*, *POT1*, *KMT2C* 11.5%) and *amp*(13q31.2) (*ERCC5*, miR-17-92 12%), which have been described in high-risk CLL[25]. Changes not previously reported in CLL or RS included *amp*(9p24) (*PDL1/L2*, 8%), *del*(16q12) (11.5%), *del*(18q22) (8%) and *amp*(7q21.2) (*CDK6*, 11.5%), *del*(1p), *amp*(11q) (*POU2AF1*, *SDHD*) and *amp*(1q23). Whole-genome doubling (WGD) was noted in 15% of cases (Extended Data Figs. 2 and 3). The recurrent RS-specific gene mutations, sCNAs and WGD were confirmed in our validation cohort ($n = 45$; Fig. 2b, right bar, Supplementary Table 5 and Extended Data Fig. 4a–x) and 14 external RS genomes[9] (Extended Data Fig. 4a–x and Supplementary Table 5).

Combined analysis of our discovery and validation cohorts provided power to detect new RS drivers further, with *CCND3*, *TET2* and *BRAF* mutations and additional focal sCNAs emerging as significant (Extended Data Fig. 4v–x and Fig. 2d). Comparison of our 45 clonally related cases with prior large-scale CLL analyses[6] revealed predisposing lesions for RS, given their relative enrichment in CLL$^{ANC+INT}$, including mutated *TP53* and *NOTCH1*, *del*(17p) and *del*(14q32) but not *tri*(12), mut-*SF3B1* or *del*(11q) (all $Q < 0.05$; Fig. 2e and Supplementary Table 5). Compared to DLBCL[15,19], the driver distribution in these 45 cases was enriched for *TP53*, *del*(17p), *NOTCH1*, *del*(13q14.2), *del*(1p), *amp*(19p13.2), *SF3B1*, *EGR2* and *GNB1* (Fig. 2f, all $Q < 0.05$). Mut-*IRF2BP2*, -*MGA* and -*DNMT3A* frequency was higher in RS compared to 304 de novo DLBCLs[19].

Evaluation of the relative timing of each putative driver event in 58 related RS cases revealed *ATM* mutations, *tri*(12) mutations or *SF3B1* mutations, as already present in CLL$^{ANC}$; alterations in *TP53* (mutations and/or *del*(17p)) or *NOTCH1* and *del*(15q15.1) [*MGA*] were predominantly CLL events ($P < 0.05$; Supplementary Table 6). In contrast, *del*(9p21), *del*(9p), *del*(9q), *del*(2q37), *amp*(1q23) and *del*(6q) were most frequently observed as new RS events ($P < 0.05$; Supplementary Table 6); WGD was restricted to the RS clones (Fig. 2g, Extended Data Figs. 2 and 3 and Supplementary Figs. 1 and 2). By systematically identifying preferred genomic trajectories driving transformation, we calculated the probability for acquiring any of the RS drivers per CLL driver via network analysis (Fig. 2h and Supplementary Table 6). Significant trajectories from CLL to RS included *NOTCH1* to *del*(1p), *NOTCH1* to *del*(14q32) and *del*(14q32) to *amp*(16q23) ($P < 0.05$; $Q < 0.4$).

Overall, our findings indicate mutations of *NOTCH1*, DNA damage response and the MAPK pathway as preexisting in CLL, and alterations in epigenetics, interferon/inflammatory signaling, cell-cycle deregulation and immune evasion—whether by sSNVs or by sCNAs—as newly occurring at transformation (Fig. 3a).

## Profiling RS emerging following targeted therapies

Therapies targeting BTK, BCL2 or PI3K-delta pathways have revolutionized CLL therapy and yet have failed to prevent transformation. Since RS is a recognized mechanism of therapeutic resistance[26–28], we evaluated the 15 patients within our cohort presenting transformation to RS while receiving targeted agents. No typical resistance mutations to targeted agents were detected (*BTK*, *BCL2*), while one ibrutinib-exposed patient had *del*(8p) and one venetoclax-treated patient had *amp*(1q), both previously described sCNA drivers of resistance[29,30] (Fig. 3b)

To track disease tempo over time, we analyzed serial samples procured in the years before RS from two patients receiving targeted agents. Patient 26 illustrates the potential impact of *EZH2* inactivation to transformation while on venetoclax, since the RS specimen carried both an inactivating *EZH2* frameshift mutation and deletion of the *EZH2* locus through *del*(7q36) (Fig. 3c). Patient 3 developed nodal RS that evolved from *TP53*-mutated CLL while on ibrutinib. The RS clone emerged from an aggressive CLL subclone (clone 3) marked by focal loss of the *CDKN2A/B* locus (Fig. 3d). Newly acquired genetic changes in the RS clone included inactivating mutations in chromatin modifiers (*CHD2*, *SRSF1*), *NFKBIE* and *del*(15q15) (*MGA* loss). Transformation on targeted agents appears as an heterogenous process, distinct from acquired resistance in CLL and characterized by complex evolution marked by accumulation of multiple events.

## RS is characterized by distinct molecular subtypes

To assess the degree of similarity between RS and DLBCL, we performed unsupervised non-negative matrix factorization (NMF), clustering on our 97 RS cases along with 304 DLBCL samples[19] based on our identified RS genetic alterations together with known DLBCL drivers. Most RS (75 of 97 cases) clustered together, largely separately from DLBCL (Extended Data Fig. 5c). The DLBCL cases closest to RS comprised DLBCL C2, previously reported with biallelic *TP53* inactivation, frequent *CDKN2A/B* loss and *del*(13q14) (Rb1)[19]. Seven of eight clonally unrelated RS clustered with DLBCL (Fisher's exact test, $P = 6.75 \times 10^{-6}$), with membership across the DLBCL clusters[19], highlighting unrelated RS as a diverse entity genetically similar to de novo DLBCL.

Analysis of the combined RS validation and discovery WES data by unbiased NMF consensus clustering defined five RS molecular subtypes (Extended Data Fig. 5d). Three (RS1, RS3 and RS5) were enriched in *TP53* and/or *del*(17p) and displayed higher rates of sCNAs and genome alterations (Fig. 4a). RS1 (13.4%) was marked by WGD and fractured genomes ($P < 0.001$ Supplementary Table 7f; Methods), along with arm-level loss of 1p and 9p and *MYC* amplification. It comprised 6 of 15 M-CLL patients, highlighting WGD as a mechanism of transformation in M-CLL (Fisher's exact test, $P = 4.6 \times 10^{-3}$). RS3 (20.6%) was enriched for *del*(17p) and mutations in *TP53*, *NOTCH1* and *IRF2BP2*, and frequently contained sCNAs, including *del*(14q32.11), *del*(9q), *del*(15q15.2) (*MGA*), *amp*(16q23.2) (*IRF8*) and *del*(2q37.1). RS5 (22.7%) similarly displayed high rates of *del*(17p) and *TP53* alterations and frequent sCNAs, including *del*(16q12.1), *del*(1p35.2) and *amp*(7p), but lacked *NOTCH1* mutation. By contrast, subtypes RS2 and RS4 showed lower fraction of genome altered ($P < 0.001$ (RS2 versus 1, 3 and 5); $P < 0.005$ (RS4 versus 1, 3 and 5); Supplementary Table 7). RS2 (26.8%) was predominantly marked by *tri*(12) co-occurring with *SPEN*/*NOTCH1* and *KRAS* mutations. RS4 (16.5%) was marked by *SF3B1* and *EGR2* mutations on a background of *del*(13q).

Evaluation of differential expression between subtypes from matched transcriptomes from 36 RS cases identified distinct signatures defining RS1 ($n = 25$ genes) and RS3 ($n = 188$; Extended Data Fig. 6a and

https://doi.org/10.1038/s41591-022-02113-6

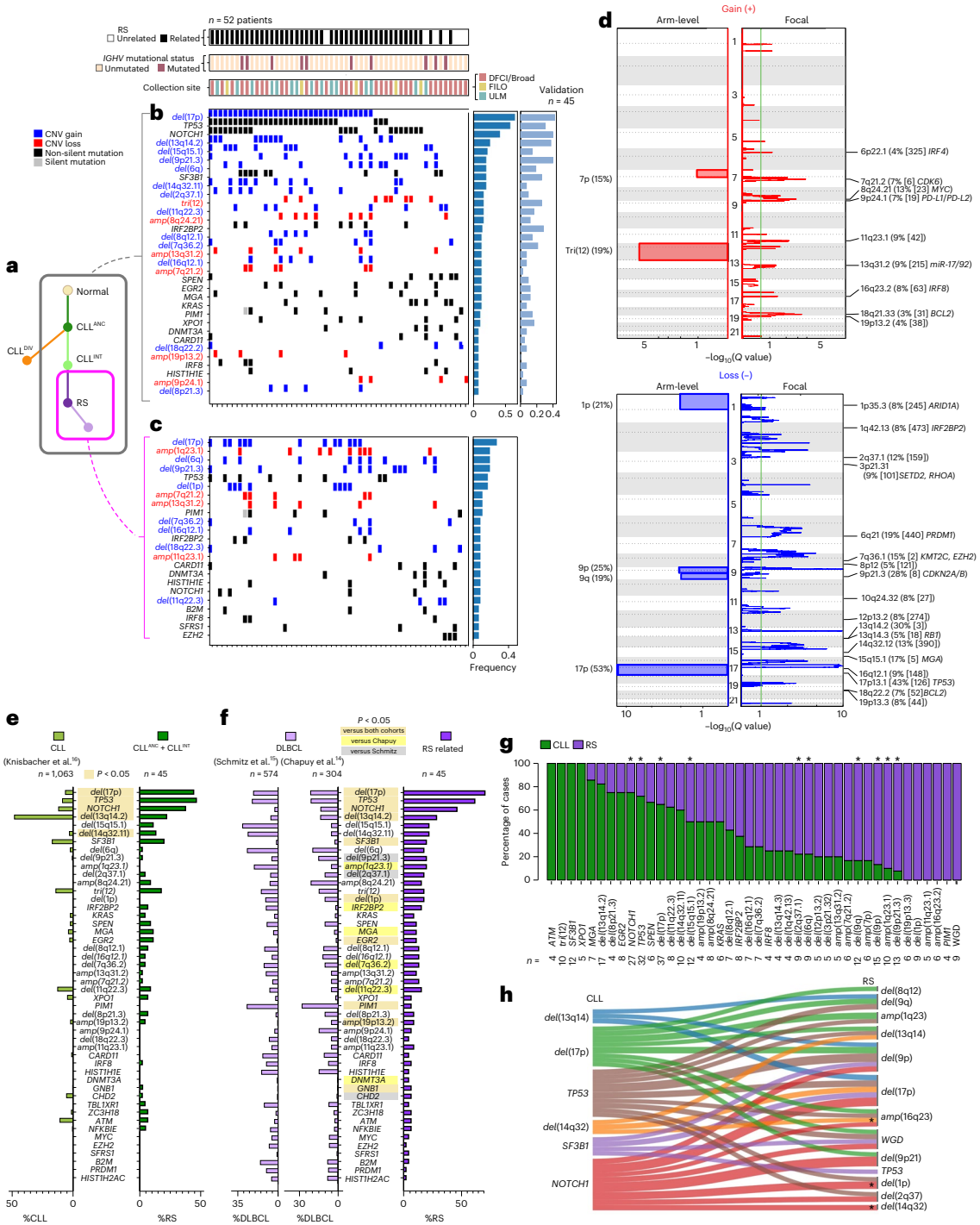



**Fig. 2 | The landscape of putative driver mutations in RS. a**, Phylogenetic tree schema demonstrating clones comprising RS history (gray box) and RS-specific clones (magenta box) (ANC, ancestor clone; INT, CLL intermediate clone; DIV, CLL divergent clone; RS, RS clone). **b**,**c**, Somatic mutation information across the putative driver genes and recurrent somatic copy-number alterations (rows) for 52 RS patients (columns) who underwent WES, ranked by frequency (right) for both RS history (**b**), alterations detected in RS cells, and RS clones (**c**), alterations acquired at transformation. Samples were annotated for sequencing site (DFCI/Broad, red; German CLL Study Group (ULM), blue; French Innovative Leukemia Organization (FILO), yellow), IGHV status (maroon, mutated; peach unmutated), and clonal relationship (black, related; white, unrelated). Light blue frequency bars adjacent to RS history represent frequency in validation cohort of each alteration ($n = 45$). **d**, GISTIC2.0 plots showing arm-level (right panel) and focal (left panel) amplifications (red, top) and deletions (blue, bottom) for RS samples in the combined discovery and validation cohorts ($n = 97$). Discovery cohort GISTIC2.0 plots are located in Extended Data Fig. 5. **e**, Frequencies of somatic alterations in CLL clones from related RS cases ($n = 45$, dark green bars) compared to CLL driver frequencies[6] using two-sided exact binomial test with Benjamini–Hochberg multiple hypothesis testing correction. **f**, RS somatic alteration frequencies (dark purple) compared to DLBCL event frequencies (light purple) from DLBCL cohorts[15,19] using two-sided exact binomial test with Benjamini–Hochberg multiple test correction. **g**, Proportion in which a recurrent driver is found as present in CLL$^{ANC+INT}$ (green) or acquired in RS (purple) across 58 related cases (only drivers affecting at least four patients are shown; Supplementary Table 6). *$P < 0.05$ (McNemar test, one-sided). **h**, Sankey plot showing trajectories from CLL driver to acquired RS driver. Only driver pairs with at least four co-occurrences across the cohort are displayed and tested for statistical significance (Supplementary Table 6). *$P < 0.05$ (Fisher's exact test, two-sided) and $Q < 0.4$.

https://doi.org/10.1038/s41591-022-02113-6

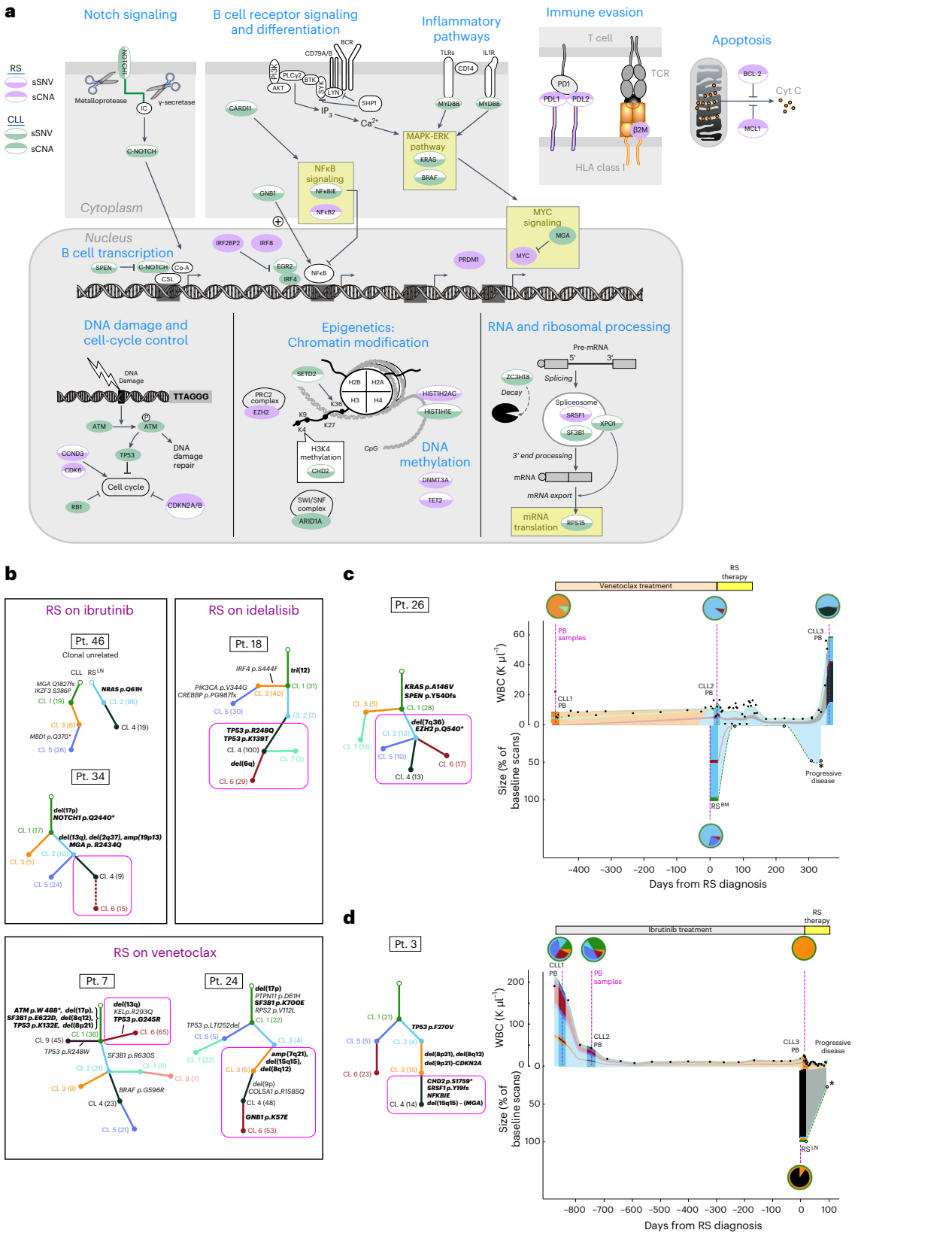



**Fig. 3 | Tracing evolution of RS on targeted agent therapy. a**, Pathways altered in CLL transformation to RS include CLL phase alterations (light green) and new drivers identified in RS (light purple). sSNV (top shading) and sCNA (bottom shading). **b**, Trees depicting clonal evolution of CLL to RS in seven select patients who developed RS on new agents. Recurrent RS drivers indicated in bold. **c**,**d**, Evolution of RS from CLL showing clonal composition and absolute tumor burden over time based on serial sampling for patient 26 (**c**) and patient 3 (**d**). Left panel: a phylogenetic tree with associated driver events (magenta square, RS clones). Right panel: relative abundance of CLL in peripheral blood by white blood cell count (1,000 cells $\mu l^{-1}$) (top) and relative abundance of RS in bottom plot (by positron emission tomography/computed tomography scan tumor metrics) with clonal evolution dynamics. Pie charts reflect composition of each sampling timepoint (pink dotted line, sampling time; top bar, treatment history; PB, peripheral blood; BM, bone marrow).

https://doi.org/10.1038/s41591-022-02113-6

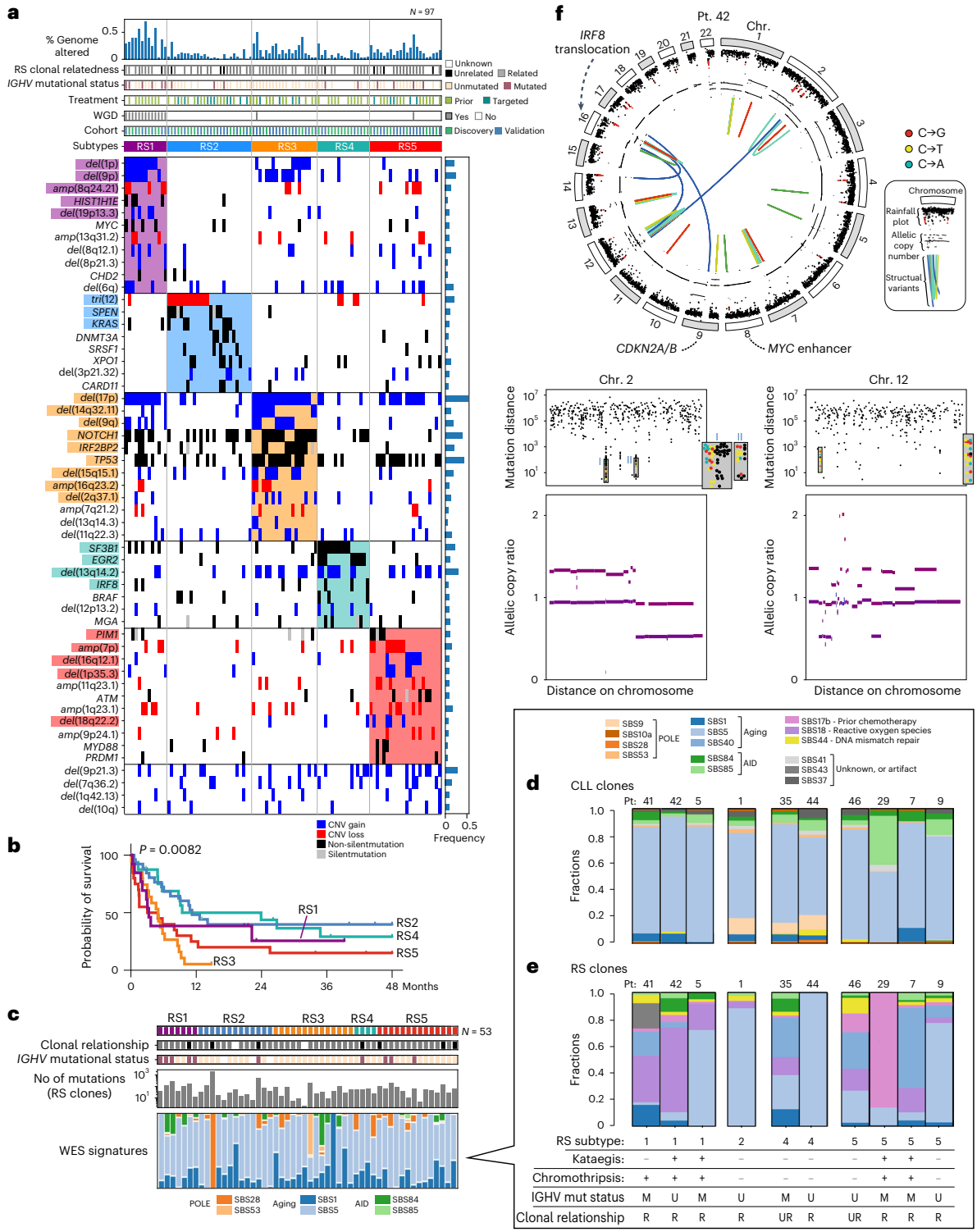



**Fig. 4 | Molecular mechanisms underlying transformation to RS. a**, Genomic classification of RS. For 97 patients (columns), five patterns of RS identified by consensus NMF clustering are depicted, with respective somatic mutations and copy-number alterations (rows). Samples are annotated for prior treatments (chemoimmunotherapy, light green; targeted agent, dark green; no prior therapy, white); IGHV status (mutated, brown; unmutated, beige; white, not determined); clonal relatedness (related, gray; unrelated, black; unknown by WES, white); and the presence of WGD (gray). Fraction genome altered per sample is shown (top). Event frequencies are indicated as blue bars on the right side for each alteration. Genes that met significance for association with a cluster by Fisher's exact test (Supplementary Table 7) are highlighted by cluster association (subtype 1, purple; subtype 2, blue; subtype 3, orange; subtype 4, green; subtype 5, red). **b**, Overall survival according to the RS genomic pattern. Kaplan–Meier curves for each subtype according to color legends. *P* value is from log-rank (Mantel Cox) testing. **c**, WES signatures for RS samples from discovery cohort ($n = 52$). **d**,**e**, WGS signatures for CLL (**d**) and RS (**e**) clones in ten evaluable patients. IGHV status (mutated, M; unmutated, UM) and clonal relationship (R, related; UR, unrelated) is indicated at bottom. **f**, Chromothripsis and kataegis in RS sample (Pt 42) with WGD. Circos plots showing structural variants (interchromosomal, blue; deletion, red; inversion, yellow; tandem duplication, green; long range, teal), allelic copy number (middle), rainfall plot with kataegis regions (red) and chromosomes (outside). Adjacent rainfall plots show kataegis regions (C to G, red; C to T, yellow; C to A, teal) with corresponding allelic copy-number ratio plot showing corresponding fragmentation.

Supplementary Table 8). RS3 displayed signatures of cell-cycle and inflammatory/interferon signaling processes in line with its enrichment for *IRF2BP2* mutations (Supplementary Table 8). Unsupervised consensus clustering identified five transcriptional clusters that associated with RS molecular subtypes (Fisher's exact test, $P = 0.038$; Extended Data Fig. 6b,c). RS2 and RS4 associated with improved overall survival (log-rank $P = 0.0082$; Fig. 4b). Clonally related cases had shorter median OS than unrelated ones (log-rank $P = 0.0094$; Extended Data Fig. 6d).

### Mutational processes underlying transformation

Evaluation of mutational profiles from the combined CLL and RS WES data revealed signatures of aging and activation-induced cytidine deaminase (AID), like previous studies[5,6,31]. We detected a dominant signature of polymerase epsilon (*POLE*) mutation in an unrelated RS case (patient 30), with deleterious *POLE* mutation and more than 2,000 sSNVs (Fig. 4c). We further analyzed WGS generated from 11 RS trios, since WGS-determined phylogenetic trees improved resolution of clones; these remained concordant with the WES phylogenies, as previously reported[32] (Extended Data Fig. 7a). CLL and RS clones of the two unrelated RS cases did not share a distant non-coding evolutionary history, definitively establishing them as unrelated malignancies (Extended Data Fig. 7b). Of 14 external WGS RS cases[9], two were clonally unrelated cases (Extended Data Fig. 7c and Supplementary Table 7). Mutational analysis of the CLL clones from 10 of 11 evaluable patients revealed signatures similar to the WES analysis (Fig. 4c). However, the RS clones revealed expanded mutational signatures, including prior chemotherapy (SBS17b), reactive oxygen species (SBS18) and defective DNA mismatch repair (SBS44, as previously reported[9]). Kataegis was recently reported in RS[9], and we indeed identified this across 4 of 11 RS genomes with clustered AID-related mutations (Fig. 4d,e and Supplementary Table 7).

From the WGS samples, we observed chromothripsis as a common defining feature of TP53-altered RS genomes (Fig. 4f and Extended Data Fig. 8a). Chromothripsis was detected in regions likely contributing to RS pathogenesis, including 7q21 (*CDK6*) (patient 41), 11q13 (*CCND1*) (patient 29) and 9p24.1 (*PDL1/L2*) (patient 41); most regions were patient specific (Extended Data Fig. 8b).

### Dynamics of transformation at single-cell resolution

We identified 292 upregulated and 111 downregulated transcripts associated with transformation from analysis of bulk RNA-seq data generated from paired high-purity RS and CLL RNA ($n = 5$; $\log_2|FC| > 1$ (where FC denotes fold change), adjusted $P < 0.05$; Fig. 5a,b and Supplementary Table 9). The larger RS cells contained more expressed transcripts and at higher abundances. Their most upregulated transcripts included regulators of mitosis, spindle assembly and cytokinesis (*AURKA*, *AURKB*, *CDK1* and *CDK2*), activation-induced cytidine deaminase (*AIDCA*) and DNA repair regulators (*BRCA1* and *XRCC2*; Fig. 5b and Supplementary Table 9). Overexpression of several of these genes has been implicated in aneuploidy in cancer[33]. By contrast, CLL showed higher relative expression of BCR signaling pathway genes.

To examine CLL transforming to RS at high resolution, we performed scRNA-seq of flow cytometry-sorted RS diagnosis biopsy specimens from five additional patients that contained clonally related RS and CLL cells within the same microenvironment (Fig. 5c and Extended Data Fig. 9a,b). Given the numerous RS-defining sCNAs in our WES data, we devised a tool, CNVSingle, to identify the expression clusters representing RS versus CLL clones based on detection of sCNA events in scRNA-seq data. CNVSingle is not dominated by reference, but rather utilizes segmentation of SNP-heterozygous sites to infer the sCNAs across a cluster of cells. This approach greatly improved the signal-to-noise ratio over other methods[34] and robustly detected tumor-specific sCNAs in malignant cells, with additional events in clusters of RS cells compared to those of CLL, and the absence of sCNAs in normal immune cells (Extended Data Fig. 9c).

Compared to CLL, the RS-identified clones across the evaluated patient samples displayed a higher number of unique molecular identifiers (UMI) per cell (that is, mean 9,000 versus 3,193 for patient 43; $P < 10^{-14}$, Wilcoxon) and genes/cell (mean 2,909 versus 1,074; $P < 10^{-14}$; Extended Data Fig. 9d). Differential expression analyses of the RS versus CLL clusters showed enrichment in pathways mapping to MYC targets, cell cycle, inflammatory response and STAT signaling pathways (Supplementary Table 9). Directional trajectories inferred using RNA velocity[35] supported a transition in cell states from CLL to RS (Extended Data Fig. 9e). The expression patterns in CLL and RS cells were sufficiently distinct that a random forest classifier could predict CLL versus RS identity of individual cells (mean $F_1 \pm \sigma = 0.92 \pm 0.01$; Methods).

Strikingly, sCNA assignments mapped to transcriptionally identified cell populations. For example, the LN cells of patient 43 (RS5 subtype) formed two groups of clusters consistent with CLL and RS (Fig. 5d, top middle). CLL cluster 2 exhibited gene expression intermediate between cluster 1 and the RS clusters, including an increase in cell-cycle genes (Fig. 5d, top right). Accordingly, the copy-number profile of cluster 1 resembled the quieter $CLL^{ANC}$ clone in WES (green), while cluster 2 showed acquisition of sCNAs of the $CLL^{INT}$ clone in WES (light blue) that subsequently gave rise to RS. RS clusters (3 and 4) displayed additional sCNAs, consistent with the chromothripsis seen in WES analysis (that is, sCNAs on chromosome 2 followed by 7, 8 and 9 with regional fragmentation; Extended Data Fig. 9f, top). Thus, intranodal cells reside on a genetic and transcriptional continuum from indolent to aggressive CLL toward RS.

Patient 10 (RS1) highlighted the rapid evolution of transformation with genomic instability in M-CLL. WES analysis established the lack of sCNAs in circulating CLL before transformation, in contrast to the abundant sCNAs and WGD in RS cells. WES of peripheral blood CLL at the time of RS diagnosis revealed new WGD but with fewer sCNAs than the LN RS WES. By flow cytometry of the LN at RS diagnosis, both CLL and RS cells were detected (Extended Data Fig. 9a, right); single-cell transcriptomes yielded three distinct populations. Cluster 1 displayed gene counts per cell consistent with CLL, whereas cluster 3 expressed much higher numbers of genes, in line with RS (Extended Data Fig. 9d); cluster 2 showed an intermediate phenotype (Fig. 5e, top). Cluster 1 demonstrated both *del*(17p) and WGD, matching the WES profile of the circulating CLL at the time of RS. Cluster 2 showed progressive genomic disorder followed by cluster 3, which highly resembled the CN profile of RS as per WES, with further sCNAs and fragmentation on chromosome 9. Therefore, in this case, *del*(17p) and WGD in an aggressive CLL clone preceded the RS transition, marked by subsequent global copy-number shifts and chromothripsis; these observations delineate the stepwise sequence of events leading to RS.

For patient 4, few RS cells were captured, but WGD and frequent sCNAs were nonetheless observed, again demonstrating the genome disorder of the RS1 subtype (Extended Data Fig. 10a). For patient 18 (RS2), clustering identified distinct early RS (clusters 3 and 4) from an RS subclone (cluster 0) containing *del*(6) that had been seen by WES (Extended Data Fig. 10b). Patient 41 (RS1) highlighted an intermediate cell state clearly residing within the coexisting forward scatter (FSC)-low CLL population (Extended Data Fig. 9b,d and Supplementary Table 8). Indeed, per CNVSingle, the intermediate state cells showed acquisition of early RS-specific events (that is, *del*(3p), *del*(4) and *del*(14q)), while expression data showed enriched cell-cycle genes (Extended Data Fig. 10c).

### Early RS clones are detectable in cell-free DNA

Given the numerous RS-associated genomic features, we assessed the feasibility of noninvasive detection of RS events through cell-free DNA (cfDNA) (Fig. 6a). We evaluated 46 plasma samples by ultralow pass WGS[36] collected from 24 patients within 3 years of RS diagnosis and through relapse (Supplementary Table 10). Samples from 17 patients were collected at the time of RS disease, including eight at

https://doi.org/10.1038/s41591-022-02113-6

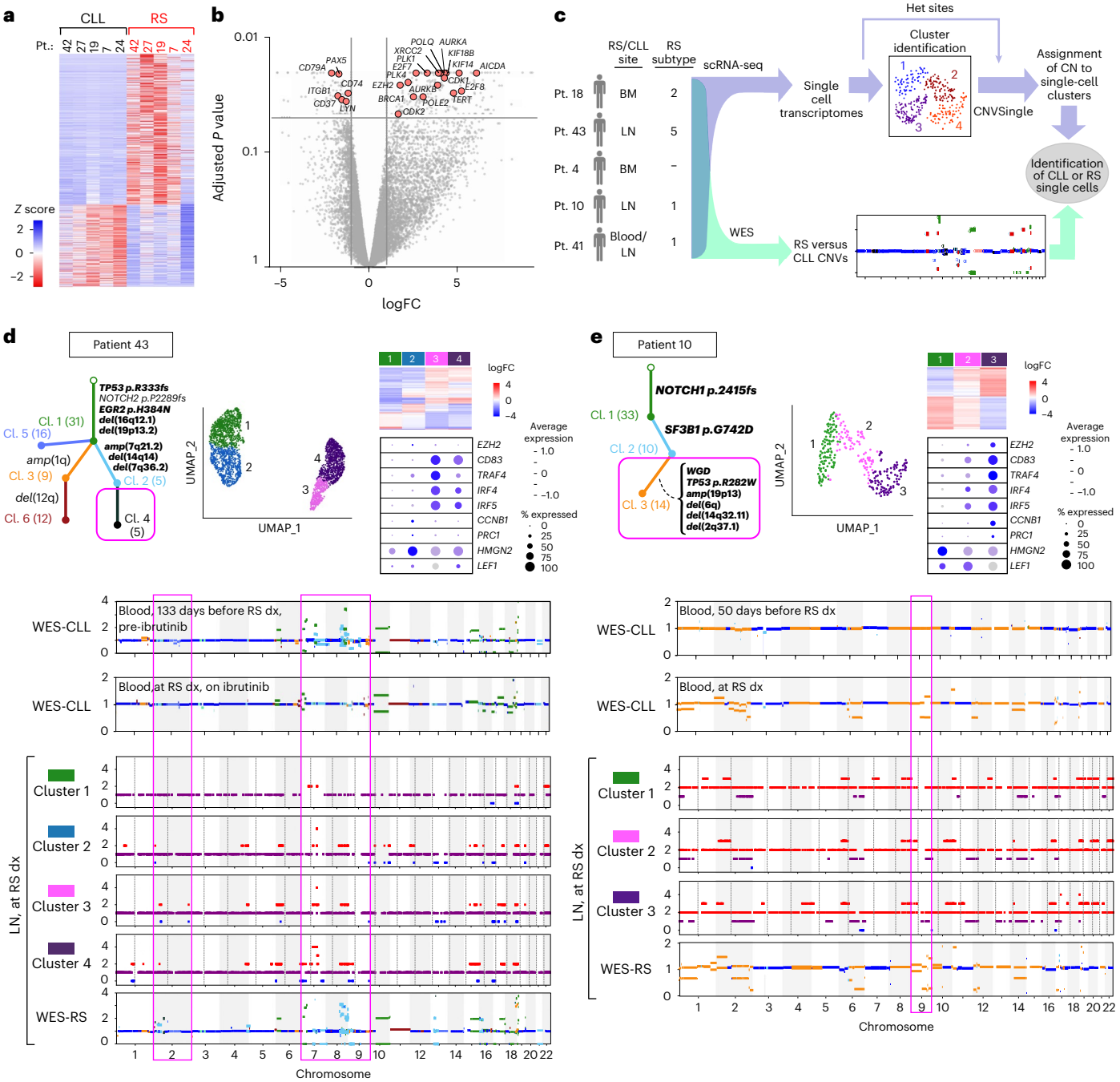



**Fig. 5 | Transformation to RS at single-cell resolution. a**, Heatmap of differentially expressed transcripts with false discovery rate (FDR) < 0.1 and absolute $\log_2$ FC > 1 in analysis between paired RS and CLL samples from patients 7, 20, 24, 27 and 42. **b**, Volcano plot of transcript expression changes in RS compared to CLL. Differentially expressed genes were assessed using limma-voom (Methods) in paired mode using sample read counts. logFC denotes $\log_2$FC, and *P* values are adjusted for multiple comparisons. Pink dots denote select relevant transcripts. **c**, Schema for assignment of copy-number changes to single cells to enable identification of CLL versus RS cells. **d**,**e**, Single-cell data show transcriptional differences between RS and CLL from patient 43 (**d**), and patient 10 (**e**), and highlight intermediate states. Phylogenetic tree showing clonal structure of RS from WES data (top left) and UMAP visualization of RS and CLL single cells (top middle). Heatmap representation of differential regulated genes between clusters (top right) and dot plot showing cluster expression of representative genes in dysregulated pathways (Supplementary Table 9) (purple shading, relative expression; dot size, percent of single-cell cells expressing transcript). Inferred allelic copy number from CNVSingle for each single-cell cluster (bottom) depicted adjacent to WES allelic copy-number plots color-coded to show copy-number events assigned to CLL and RS clones (Methods).

initial diagnosis. Ten were from the discovery cohort, and their RS characterization served as positive confirmation for detection of RS-specific alterations. Eight of these also had simultaneous (same blood draw) or contemporaneous circulating CLL cells analyzed by WES, thus offering a controlled way to evaluate the differing contributions of nodal versus circulating disease, since the cfDNA includes DNA shed from both LN and circulating CLL cells.

RS-associated genomic features were indeed detectable in plasma. WGD was observed in the cfDNA of patient 38 at the time of RS diagnosis, matching the RS WES profile, whereas circulating CLL remained

https://doi.org/10.1038/s41591-022-02113-6

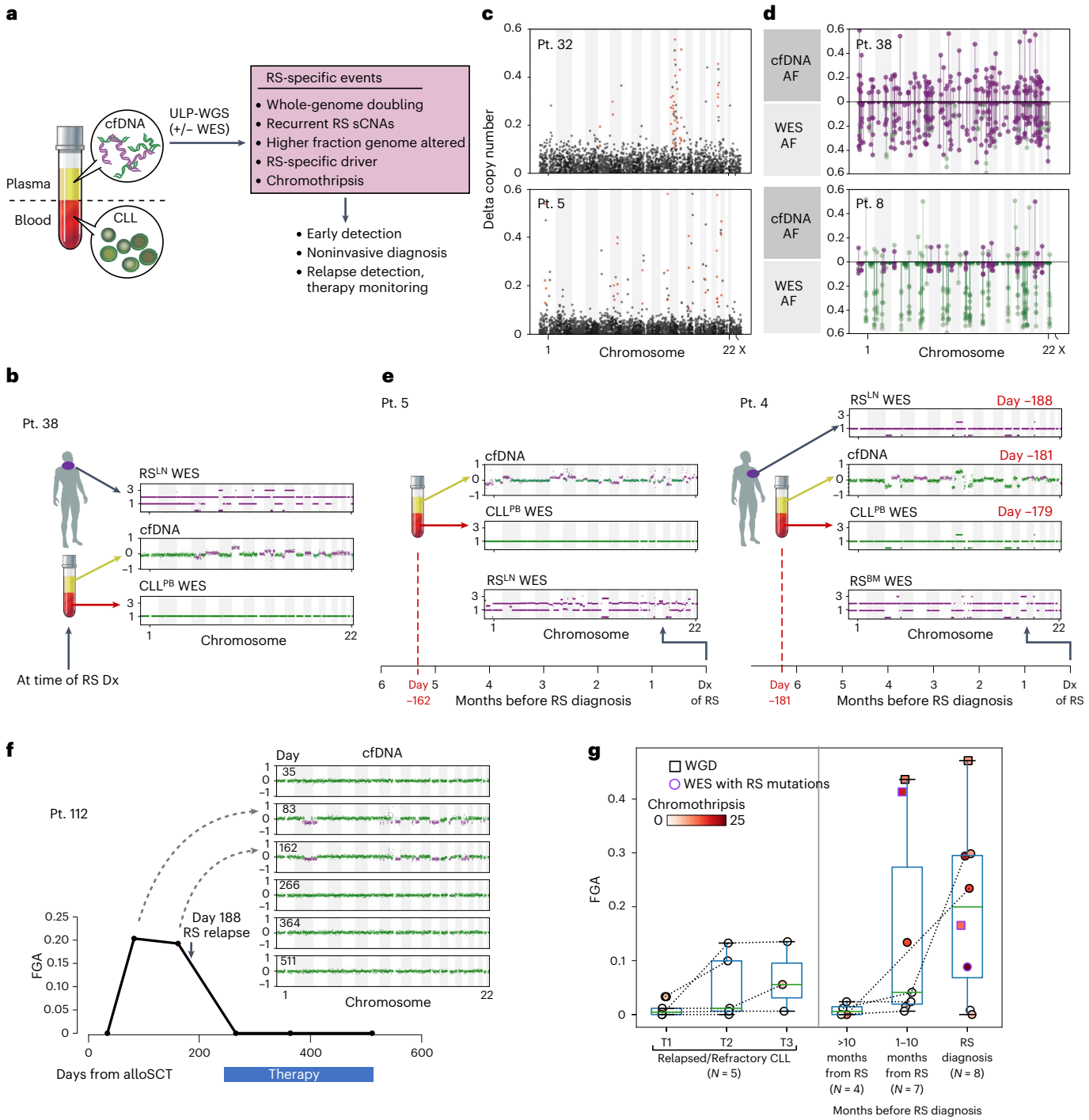



**Fig. 6 | cfDNA isolated from plasma of RS patients shows evidence of transformation. a**, Schema showing how RS-specific DNA events can be identified separately from cell-free DNA and different from circulating CLL cells. **b**, cfDNA in RS patient 38 shows WGD of clonally unrelated RS, which is not seen in circulating CLL disease at time of diagnosis. **c**, Chromothripsis is observed in cfDNA of RS patients, as demonstrated by plotting the difference between copy-number state changes across the genome (patient 32 top, patient 5 bottom). **d**, Allele frequencies for RS (purple) and CLL (green) mutations found in RS WES sample (bottom) and RS plasma sample cfDNA WES (top) for patient 38 (top panel) and patient 8 (bottom panel). **e**, Plasma from patients shows early detection of RS. Patient 5 (top) shows RS-related WGD and chromothripsis fragmentation 162 days before RS diagnosis, which is not seen in corresponding co-sampled CLL cells. Plasma from patient 20 (bottom panel) examined 181 days before RS shows RS-related WGD and sSCNVs, which are not seen in co-sampled CLL or in LN biopsy taken from prior week. **f**, sCNAs become detectable before post-transplant relapse in patient 112, as seen by plot of fraction genome altered and corresponding cfDNA samples showing emergence of new sCNVs despite continued remission of circulating and marrow CLL. **g**, Metrics of RS in cfDNA are plotted for RS samples leading up to diagnosis. *y* Axis is fragment genome altered, color scale shows presence of chromothripsis, square represents WGD sample and purple outline indicates samples for which RS mutations were detected on WES of cfDNA. CLL samples at left of figure depict 13 samples from five relapsed/refractory CLL patients. RS samples (right) show 19 samples divided by time leading up to RS in 14 RS patients. Number of samples per each category is indicated on the figure by N. Dashed lines denote serial samples from same patients. Box plots show median values as horizontal line and whiskers showing maximum and minimum values with boundaries of box showing the interquartile range.

diploid (Fig. 6b). The cfDNA of patient 44 revealed RS-associated sCNAs (*del*(9p), *amp*(13)) that were not in the CLL cells (Extended Data Fig. 10d). cfDNA analysis also highlighted RS emergence during therapy in a high-risk CLL patient with *del*(17p) (patient 99). Although the cfDNA profile at the start of CLL-directed therapy showed minimal sCNAs, the profile at the time of RS diagnosis showed abundant new sCNAs, including *amp*(8q24) (*MYC*) (Extended Data Fig. 10e). In other patients, chromothripsis was evident in plasma cfDNA (Fig. 6c and Extended Data Fig. 10f). Furthermore, for all of the four cases from our discovery cohort in which WES was additionally performed on cfDNA, RS-specific mutations were detected (Fig. 6d and Extended Data Fig. 10g).

We queried whether RS changes could be detected in cfDNA in advance of RS diagnosis. For two of seven patients whose plasma was collected 1–10 months before RS diagnosis (Supplementary Table 10), we could detect RS-associated alterations in the cfDNA, during which time they were undergoing therapies for presumed refractory CLL. In patient 5, WGD and chromothripsis (chr 6 and 16) were detected in plasma 162 days before diagnosis and were absent from CLL (Fig. 6e, left). WES of cfDNA (Extended Data Fig. 10g) further showed the presence of RS-specific mutations. In patient 20, cfDNA 181 days before RS diagnosis showed WGD and sCNAs not present in the corresponding CLL blood sample (day −179) or LN biopsy (CLL) from the prior week (Fig. 6e, right).

Finally, we probed the potential for cfDNA analysis to detect early RS relapse. We considered two patients who had achieved a state of minimal CLL involvement following allogeneic hematopoietic stem-cell transplantation (HSCT) but who subsequently relapsed with nodal RS. For patient 112, cfDNA obtained immediately following HSCT lacked evidence of RS events, but by days +83 and +162, new sCNAs, and thus increased fraction genome altered (FGA), were found, consistent with nodal disease emergence (Fig. 6f). Ultimately, biopsy-confirmed RS relapse was diagnosed on day +187. With subsequent RS response, the RS-associated cfDNA changes resolved. Patient 111 intermittently had elevated FGA in plasma following HSCT, before RS diagnosis, which resolved following RS therapy (Supplementary Table 10). Across samples, the highest levels of FGA in cfDNA were observed in RS diagnostic samples ($n = 8$), with a decreasing ratio in the preceding 1–10 months ($n = 7$) and an even lower ratio in more distant prediagnosis samples (>10 months; $n = 4$). In seven cases, FGA exceeded all values from high-risk CLL cases ($n = 14$ samples from five patients; Fig. 6g). Of the eight patients with cfDNA available at the time of biopsy-proven RS diagnosis, we confidently discerned RS-specific lesions in six (75%) using strict criteria.

## Discussion

For decades, the RS diagnosis has relied on morphologic characterization of aggressive lymphoma within the context of concurrent or known history of CLL[3]. Herein, through the implementation of advanced analytic approaches that can distinguish between the RS and CLL clones, and through integration of exome, genome and transcriptome data to the largest series of paired CLL and RS specimens to date, we have defined the distinct molecular events that precede and define the RS transition.

Of the new insights gained from this study, one was the identification of new putative driving events in RS, distinct from CLL, affecting splicing, immune evasion, epigenetics, cell-cycle regulation, interferon signaling and MYC signaling. Epigenetic remodeling has been detected in RS, impacting pathways of BCR signaling, oxidative phosphorylation, cell proliferation and MYC signaling[37]. We further identified instances of driver alterations with potential therapeutic impact, such as those affecting *CDK6* or immune checkpoints. Our study highlights major differences between RS and de novo DLBCL, despite several shared driver events. We delineated five RS subtypes and confirmed these genomic patterns associated with distinct transcriptomes and outcome.

Second, RS is marked by numerous sCNAs and features of genomic instability (that is, chromothripsis, kataegis and WGD). Near tetraploidy has been identified as an RS risk factor[38], and our detailed genomic and single-cell analysis demonstrates how this unstable state can lead to RS evolution. These features could result from mitosis defects, as suggested by RNA expression data, and WGD may confer potential therapeutic vulnerabilities[39]. We demonstrate how such instability may be used to provide an earlier and noninvasive detection of RS in cfDNA, which should be further evaluated in clinical studies as a cost-effective approach for this difficult-to-diagnose aggressive cancer[40].

Finally, we confirmed the majority of RS is unrelated to the co-occurring CLL—a facet previously only defined based on differing IGHV clonotypes[3,10] and ultradeep IGHV sequencing[41]. Unrelated RS has been previously associated with improved clinical outcomes, which suggests distinct disease biology[3,10]. We now demonstrate that by exome-level or genome-level analysis, clonally unrelated RS is a de novo DLBCL, occurring as an independent lymphoma, lacking any shared distant genetic history with the coexisting CLL. These cases tended to lack *TP53* and *NOTCH1* alterations, were enriched in M-CLL and clustered with de novo DLBCL separately from clonally related RS. These molecular insights may help identify RS patients with a more favorable prognosis.

Altogether, our comprehensive evolutionary tracing enables a molecular definition of transformation that can guide the identification, diagnosis and prognosis of RS. Our advanced molecular framework can serve as a model for studying transformed cancers.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-022-02113-6.

## References

1. Offin, M. et al. Concurrent RB1 and TP53 alterations define a subset of EGFR-mutant lung cancers at risk for histologic transformation and inferior clinical outcomes. *J. Thorac. Oncol.* **14**, 1784–1793 (2019).
2. Volta, A. D. et al. Transformation of prostate adenocarcinoma into small-cell neuroendocrine cancer under androgen deprivation therapy: much is achieved but more information is needed. *J. Clin. Oncol.* **37**, 350–351 (2019).
3. Parikh, S. A., Kay, N. E. & Shanafelt, T. D. How we treat Richter syndrome. *Blood* **123**, 1647–1657 (2014).
4. Landau, D. A. et al. Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).
5. Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
6. Knisbacher, B. A. et al. Molecular map of chronic lymphocytic leukemia and its impact on outcome. *Nat. Genet.* **54**, 1664–1674 (2022).
7. Chigrinova, E. et al. Two main genetic pathways lead to the transformation of chronic lymphocytic leukemia to Richter syndrome. *Blood* **122**, 2673–2682 (2013).
8. Fabbri, G. et al. Genetic lesions associated with chronic lymphocytic leukemia transformation to Richter syndrome. *J. Exp. Med.* **210**, 2273–2288 (2013).
9. Klintman, J. et al. Genomic and transcriptomic correlates of Richter transformation in chronic lymphocytic leukemia. *Blood* **137**, 2800–2816 (2021).
10. Rossi, D. et al. The genetics of Richter syndrome reveals disease heterogeneity and predicts survival after transformation. *Blood* **117**, 3391–3401 (2011).
11. Taylor-Weiner, A. et al. DeTiN: overcoming tumor-in-normal contamination. *Nat. Methods* **15**, 531–534 (2018).

12. Leshchiner, I. et al. Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. Preprint at *bioRxiv* 508127 (2018).
13. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
14. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
15. Schmitz, R. et al. Genetics and pathogenesis of diffuse large B-cell lymphoma. *N. Engl. J. Med.* **378**, 1396–1407 (2018).
16. Chapuy, B. et al. Genomic analyses of PMBL reveal new drivers and mechanisms of sensitivity to PD-1 blockade. *Blood* **134**, 2369–2382 (2019).
17. Biran, A. et al. Activation of Notch and Myc Signaling via B-cell-restricted depletion of Dnmt3a generates a consistent murine model of chronic lymphocytic leukemia. *Cancer Res.* **81**, 6117–6130 (2021).
18. Mahajan, V. S. et al. B1a and B2 cells are characterized by distinct CpG modification states at DNMT3A-maintained enhancers. *Nat. Commun.* **12**, 2208 (2021).
19. Chapuy, B. et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat. Med.* **24**, 679–690 (2018).
20. Challa-Malladi, M. et al. Combined genetic inactivation of beta2-Microglobulin and CD58 reveals frequent escape from immune recognition in diffuse large B cell lymphoma. *Cancer Cell* **20**, 728–740 (2011).
21. Sade-Feldman, M. et al. Resistance to checkpoint blockade therapy through inactivation of antigen presentation. *Nat. Commun.* **8**, 1136 (2017).
22. Gettinger, S. et al. Impaired HLA Class I antigen processing and presentation as a mechanism of acquired resistance to immune checkpoint inhibitors in lung cancer. *Cancer Disco.* **7**, 1420–1435 (2017).
23. Singh, K. et al. c-MYC regulates mRNA translation efficiency and start-site selection in lymphoma. *J. Exp. Med.* **216**, 1509–1524 (2019).
24. Lee, S. C. et al. Synthetic lethal and convergent biological effects of cancer-associated spliceosomal gene mutations. *Cancer Cell* **34**, 225–241 e228 (2018).
25. Edelmann, J. et al. Genomic alterations in high-risk chronic lymphocytic leukemia frequently affect cell cycle key regulators and NOTCH1-regulated transcription. *Haematologica* **105**, 1379–1390 (2020).
26. Anderson, M. A. et al. Clinicopathological features and outcomes of progression of CLL on the BCL2 inhibitor venetoclax. *Blood* **129**, 3362–3370 (2017).
27. Jain, P. et al. Long-term outcomes for patients with chronic lymphocytic leukemia who discontinue ibrutinib. *Cancer* **123**, 2268–2273 (2017).
28. Maddocks, K. J. et al. Etiology of ibrutinib therapy discontinuation and outcomes in patients with chronic lymphocytic leukemia. *JAMA Oncol.* **1**, 80–87 (2015).
29. Burger, J. A. et al. Clonal evolution in patients with chronic lymphocytic leukaemia developing resistance to BTK inhibition. *Nat. Commun.* **7**, 11589 (2016).
30. Guieze, R. et al. Mitochondrial reprogramming underlies resistance to BCL-2 inhibition in lymphoid malignancies. *Cancer Cell* **36**, 369–384 e313 (2019).
31. Kasar, S. et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).
32. Gruber, M. et al. Growth dynamics in naturally progressing chronic lymphocytic leukaemia. *Nature* **570**, 474–479 (2019).
33. Zhang, N. et al. Overexpression of Separase induces aneuploidy and mammary tumorigenesis. *Proc. Natl Acad. Sci. USA* **105**, 13033–13038 (2008).
34. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
35. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
36. Adalsteinsson, V. A. et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat. Commun.* **8**, 1324 (2017).
37. Nadeu, F. et al. Detection of early seeding of Richter transformation in chronic lymphocytic leukemia. *Nat. Med.* **28**, 1662–1671 (2022).
38. Miller, C. R. et al. Near-tetraploidy is associated with Richter transformation in chronic lymphocytic leukemia patients receiving ibrutinib. *Blood Adv.* **1**, 1584–1588 (2017).
39. Quinton, R. J. et al. Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. *Nature* **590**, 492–497 (2021).
40. Soilleux, E. J. et al. Diagnostic dilemmas of high-grade transformation (Richter's syndrome) of chronic lymphocytic leukaemia: results of the phase II National Cancer Research Institute CHOP-OR clinical trial specialist haemato-pathology central review. *Histopathology* **69**, 1066–1076 (2016).
41. Favini, C. et al. Clonally unrelated Richter syndrome are truly de novo diffuse large B-cell lymphomas with a mutational profile reminiscent of clonally related Richter syndrome. *Br. J. Haematol.* **198**, 1016–1022 (2022).

**Erin M. Parry[1,2,3,21], Ignaty Leshchiner[2,4,21], Romain Guièze[1,2,5,6,21], Connor Johnson[2], Eugen Tausch[7], Sameer A. Parikh[8], Camilla Lemvigh[1,9], Julien Broséus[10,11], Sébastien Hergalant[10], Conor Messer[2], Filippo Utro[12], Chaya Levovitz[12], Kahn Rhrissorrakrai[12], Liang Li[2], Daniel Rosebrock[2], Shanye Yin[1,3], Stephanie Deng[1], Kara Slowik[2], Raquel Jacobs[2], Teddy Huang[1,13], Shuqiang Li[1,2,13], Geoff Fell[14], Robert Redd[14], Ziao Lin[2], Binyamin A. Knisbacher[2], Dimitri Livitz[2], Christof Schneider[7], Neil Ruthen[1,13], Liudmila Elagina[2], Amaro Taylor-Weiner[2], Bria Persaud[2], Aina Martinez[2], Stacey M. Fernandes[1], Noelia Purroy[1,2,3], Annabelle J. Anandappa[1,3], Jialin Ma[2], Julian Hess[2], Laura Z. Rassenti[15], Thomas J. Kipps[15], Nitin Jain[16], William Wierda[16], Florence Cymbalista[17], Pierre Feugier[10,18], Neil E. Kay[8], Kenneth J. Livak[1,13], Brian P. Danysh[2], Chip Stewart[2], Donna Neuberg[14], Matthew S. Davids[1,3], Jennifer R. Brown[1,3], Laxmi Parida[12], Stephan Stilgenbauer[7,22], Gad Getz[2,3,19,22] ✉ & Catherine J. Wu[1,2,3,20,22] ✉**

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Harvard Medical School, Boston, MA, USA. [4]Department of Medicine, Boston University School of Medicine, Boston, MA, USA. [5]CHU de Clermont-Ferrand, Clermont-Ferrand, France. [6]Université Clermont Auvergne, EA7453 CHELTER, Clermont-Ferrand, France. [7]Division of CLL, Department of Internal Medicine III, Ulm University, Ulm, Germany. [8]Division of Hematology, Mayo Clinic, Rochester, MN, USA. [9]Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark. [10]Inserm UMRS1256 Nutrition-Génétique et Exposition aux Risques Environnementaux (N-GERE), Université de Lorraine, Nancy, France. [11]Université de Lorraine, CHRU-Nancy, service d'hématologie biologique, pôle laboratoires, Nancy, France. [12]IBM Research, Yorktown Heights, New York, NY, USA. [13]Translational Immunogenomics Lab, Dana-Farber Cancer Institute, Boston, MA, USA. [14]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. [15]Moores Cancer Center, Medicine, University of California, San Diego, La Jolla, CA, USA. [16]Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [17]Laboratoire d'hématologie, Hôpital Avicenne—AP-HP, INSERM U978- Université Sorbonne Paris Nord, Bobigny, France. [18]Université de Lorraine, CHRU Nancy, service d'hématologie clinique, Nancy, France. [19]Cancer Center and Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. [20]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [21]These authors contributed equally: Erin M. Parry, Ignaty Leshchiner, Romain Guièze. [22]These authors jointly supervised this work: Stephan Stilgenbauer, Gad Getz, Catherine J. Wu. ✉e-mail: gadgetz@broadinstitute.org; cwu@partners.org

## Methods

### Patient sample collection and processing

CLL, RS and normal germline (that is, nontumor) samples were collected from patients following written informed consent through sample collection protocols or from clinical trial NCT03619512 from the French Innovative Leukemia Organization (FILO) with the approval of the following Institutional Review Boards (IRBs): Dana-Farber Cancer Institute IRB, University of California San Diego IRB, Mayo Clinic IRB, MD Anderson Cancer Center IRB, Ethics Committee of Ulm University (Ulm, Germany) or the University Hospital of Nancy with the approval of the Comité de Protection des personnes (CPP) Ouest IV (Nantes, France). All biospecimen collection protocols were conducted in accordance with the principles of the Declaration of Helsinki and with the approval of the IRBs of the respective institutions. Patient and sample characteristics are provided (Supplementary Tables 1 and 2). Sex was self-reported.

**RS samples.** RS samples were collected from BM, LN, lymphoid tissue or peripheral blood mononuclear cells (PBMCs) and included both fresh-frozen and formalin-fixed paraffin-embedded (FFPE) samples. Freshly collected tissue samples were disaggregated by GentleMACs digestion (Miltenyi Biotec) before cryopreservation with FBS/10% DMSO and storage in liquid nitrogen or directly stored as whole tissue blocks in liquid nitrogen. Blood and BM specimens were isolated by Ficoll/Hypaque density gradient centrifugation before cryopreservation with FBS/10% DMSO and storage in liquid nitrogen. For viably frozen samples of low purity (<30% tumor), RS cells were isolated by FACS (Aria II; Becton Dickinson) based on $CD5^+$ and $CD19^+$ co-expression on cells with increased FSC (Biolegend; CD5 FITC catalog no. 364022, CD19 PE-Cy7 catalog no. 302216). For FFPE specimens, samples from each submitting center were reviewed for >50% purity before sequencing.

**CLL samples.** CLL samples were obtained from PBMCs. Samples with higher CLL purity (white blood cell count > $25 \times 10^3$ cells $\mu l^{-1}$ or absolute lymphocyte count > $20 \times 10^3$ cells $\mu l^{-1}$) were processed without CD19 selection, and PBMCs were isolated by Ficoll/Hypaque density gradient centrifugation and then cryopreserved with FBS/10% DMSO and stored in liquid nitrogen until the time of analysis. Samples with a white blood cell count < 25,000 cells $\mu l^{-1}$ or an absolute lymphocyte count < 20,000 cells $\mu l^{-1}$ underwent CD19 selection (RosetteSep Human B cell enrichment; Stem Cell Technologies) or as previously described[4] or FACS sorting to enrich for $CD5^+CD19^+$ populations.

**Germline samples.** Sources of nontumor germline DNA included saliva (Oragene Discover (ORG500 or ORG600) kit; DNA Genotek), remission BM[4] or in vitro expanded T cells. For the latter, $CD19^-CD4^+$ or $CD19^-CD3^+$ cells were collected by FACS (Aria II; Becton Dickinson; Biolegend, catalog no. 300330, catalog no. 300506, catalog no. 363006). The cells were plated and expanded in vitro in RPMI (Gibco) containing phytohemagglutinin (1.5:100), IL-7 (20 ng $ml^{-1}$), IL-2 (100 U $ml^{-1}$), 10% human serum and β-2-mercaptoethanol (1/1,000).

### Genomic DNA sequencing

**WES.** A total of 143 samples were processed and sequenced at the Broad Institute. For these fresh blood and BM samples and cryopreserved suspension cells, genomic DNA and RNA were extracted per the manufacturer's recommendations (Qiagen). DNA was quantified in triplicate using a standardized PicoGreen dsDNA Quantitation Reagent (Invitrogen) assay. The quality control identification check was performed using fingerprint genotyping of 95 common SNPs by Fluidigm Genotyping. Library construction from double-stranded DNA was performed using the KAPA Library Prep kit, with palindromic forked adapters from Integrated DNA Technologies. Libraries were pooled before hybridization. Hybridization and capture were performed using the relevant components of Illumina's Rapid Capture Enrichment Kit, with a 37-Mb target. All library construction, hybridization and capture steps were automated on the Agilent Bravo liquid handling system. After post-capture enrichment, library pools were denatured using 0.1 N NaOH on the Hamilton Starlet. Cluster amplification of DNA libraries was performed according to the manufacturer's protocol (Illumina) using HiSeq 4000 exclusion amplification chemistry and HiSeq 4000 flow cells. Flow cells were sequenced utilizing Sequencing-by-Synthesis chemistry for HiSeq 4000 flow cells. The flow cells were then analyzed using RTA v.2.7.3 or later. Each pool of whole-exome libraries was sequenced on paired 76 cycle runs with two eight-cycle index reads across the number of lanes needed to meet coverage for all libraries in the pool. Output from Illumina software was processed by the Picard data-processing pipeline to yield BAM files containing demultiplexed, aggregated aligned reads. Standard quality control metrics, including error rates, percentage-passing filter reads and total Gb produced, were used to characterize process performance before downstream analysis.

Twenty-seven samples were processed and sequenced at the University of Ulm, Germany. Exome enrichment was performed through biotinylated RNA oligomer libraries, which are part of the SureSelectXT Human All Exon V5 capture library. The preparation workflow with the SureSelectXT reagent kit included DNA Fragmentation via Covaris supersonic shearing, end repair, ligation, library hybridization, indexing, quantitative polymerase chain reaction (PCR)-based quantification and multiplexing (per protocol v.1.7). Multiple quality controls via Agilent Bioanalyzer were implemented into the process. Libraries were amplified to produce clonal clusters and sequenced using massively parallel sequencing on the Illumina HiSeq 2000 Sequencing System. Eleven samples were processed (SureSelect QXT Agilent kit) and sequenced on a HiSeq 1000 instrument at the University of Nancy, France.

A subset of our WES data had reduced coverage in the GC-rich region of *NOTCH1*. For these, targeted deep sequencing of the *NOTCH1* 3′-UTR was performed, as previously described[6].

### WGS

**Preparation of libraries for cluster amplification and sequencing (PCR-free).** Genomic DNA (350 ng in 50 µl of solution) was processed by fragmentation through acoustic shearing (Covaris focused ultrasonicator), targeting 385-bp fragments, and additional size selection was performed using a SPRI 80 cleanup. Library preparation (HyperPrep without amplification module; KAPA Biosystems, no. KK8505) was performed as above for WES. Libraries were then quantified using quantitative PCR (KAPA Biosystems) with probes specific to the ends of the adapters, normalized to 1.7 nM, and then pooled into 24 plexes.

**Preparation of libraries for cluster amplification and sequencing (PCR-plus).** An aliquot of genomic DNA (100 ng in 50 µl of solution) was used as the input into DNA fragmentation. Shearing was performed as described above in the PCR-free procedure. Library preparation was performed using a commercially available kit provided by KAPA Biosystems (KAPA HyperPrep with Library Amplification Primer Mix, product KK8504), and with palindromic forked adapters using unique eight-base index sequences embedded within the adapter (Roche). The libraries were then amplified by 10 cycles of PCR. Following sample preparation, libraries were quantified using quantitative PCR (KAPA Biosystems) with probes specific to the ends of the adapters. This assay was automated using Agilent's Bravo liquid handling platform. Based on qPCR quantification, libraries were normalized to 2.2 nM and pooled into 24 plexes.

**Cluster amplification and sequencing (NovaSeq 6000).** Sample pools were combined with NovaSeq Cluster Amp Reagents DPX1, DPX2 and DPX3 and loaded into single lanes of a NovaSeq 6000 S4 flow-cell cell using the Hamilton Starlet Liquid Handling system. Cluster amplification and sequencing occurred on NovaSeq 6000 Instruments

utilizing sequencing-by-synthesis kits to produce 151-bp paired-end reads. Output from Illumina software was processed by the Picard data-processing pipeline to yield CRAM or BAM files containing demultiplexed, aggregated aligned reads. All sample information tracking was performed by automated LIMS messaging.

**Circulating DNA sequencing.** Whole blood was collected by routine phlebotomy. Plasma was separated within 1–4 days of collection through density centrifugation and stored at −80 °C until DNA extraction (QIAsymphony DSP Circulating DNA Kit; QIAGEN), which was performed according to the manufacturer's instructions. Library preparation was performed (KAPA HyperPrep Kit with Library Amplification; KAPA Biosystems) using duplex UMI adapters (IDT), starting with 2–3 cc of plasma. Samples were normalized and pooled using equivolume pooling, with up to 95 samples per pool. Cluster amplification was performed according to the manufacturer's protocol (Illumina) using Exclusion Amplification cluster chemistry and HiSeqX flow cells. Flow cells were sequenced on v.2 Sequencing-by-Synthesis chemistry for HiSeqX flow cells. The flow cells were then analyzed using RTA v.2.7.3 or later. Each pool of ultralow pass whole-genome libraries was run on one lane using paired 151-bp runs.

## Sequence data processing and analyses

**WES/WGS alignment and quality control.** Sequencing was conducted using standard methods[6,32,42]. All DNA sequence data were processed through Broad Institute pipelines, such that data from multiple libraries and flow-cell runs were combined into a single BAM file. This file contained reads aligned to the human genome hg19 genome assembly (v.b37, using BWA-MEM (v.0.7.15-r1140)) provided by the Picard and Genome Analysis Toolkit (GATK) developed at the Broad Institute[43], a process that involves marking duplicate reads, recalibrating base qualities and realigning around indels.

**WES analysis.** Sequences were analyzed by the Broad Institute's Cancer Genome Analysis WES Characterization Pipeline, in which aligned BAM files were inputted into a standard WES somatic variant-calling pipeline[42] and included MuTect for calling somatic single nucleotide variants (sSNVs), Strelka2 (ref. 44) for calling small insertions and deletions (indels), deTiN for estimating tumor-in-normal (TiN) contamination[11], ContEst for estimating cross-patient contamination, AllelicCapSeg for calling allelic copy-number variants and ABSOLUTE for estimating tumor purity, ploidy, CCFs and absolute allelic copy number. Artifactual variants were filtered out using a token panel-of-normals (PoN) filter, a blat filter, and an oxoG filter. For tumor samples without a matching normal control, a 'no-normal' pipeline was used, as previously described[19]. Several FFPE samples exhibited lower DNA quality, resulting in noisier profiles with standard methods. For these samples, we applied an additional filtering technique of identifying the most correlated targets across a set of FFPE samples and performing tangent normalization[45] on samples that showed consistent behavior, thus excluding artifactual copy-number targets.

**WGS analysis.** WGS analysis was performed as previously described[46]. Due to the large amount of computational resources required to process cancer whole genomes efficiently, we ran these analysis pipelines on an elastic high-performance computing (HPC) cluster on Google Cloud VMs, comprising thousands of CPU cores.

For structural variation (SV) identification, our pipeline integrates evidence from three SV detection algorithms (Manta[47], SvABA[48] and dRanger[49]) to generate a list of SV events with high confidence from WGS data. Subsequently, we applied BreakPointer[50] to pinpoint the exact breakpoint at base-level resolution. Breakpoint information was aggregated per sample to identify: (1) balanced translocations, defined as those with breakpoints on reverse strands within 1 kb of each other; (2) inversions supported on both ends; and (3) complex events, based on the number of clustered events within 50 kb of each other. Breakpoints were annotated by intersection with our lists of CLL driver genes and significant sCNA regions, and with genes in the COSMIC Cancer Gene Census (v.90)[51].

**Analysis of UK WGS.** BAM files were obtained from prior analysis and realigned them to the Broad Institute's build of hg19 (known as b37: https://gatk.broadinstitute.org/hc/en-us/articles/360035890711-GRCh37-hg19-b37-humanG1Kv37-Human-Reference-Discrepancies)[46]. Out of the 17 sample trios obtained from the UK group, 14 samples completed WGS (three failed due to data quality and realignment issues). The standard pipeline as previously described[46] was applied to these FFPE samples, except for detection of sCNAs. Formalin damage results in extremely noisy read coverage profiles, confounding traditional copy-number segmentation pipelines. To mitigate this, we applied a modified sCNA calling method that relies on segmentation of allelic imbalance at germline het sites (as opposed to segmentation of total coverage) as its primary signal. Although total coverage is extremely noisy, the fraction of reads supporting alternate versus reference alleles at heterozygous sites is undistorted, allowing for clean allelic imbalance segmentation. Within each segment of allelic imbalance, we binned total coverage on a megabase scale, which is coarse enough to average over formalin-induced coverage fluctuations, which typically manifest as sharp coverage spikes at the 10–100-kb scale. SV and phylogenetic analysis were completed for 12 of 14 samples.

**Identification of regions of kataegis and chromothripsis.** In the WGS, kataegis regions were defined by genomic regions with at least six mutations within two standard deviations of the median chromosomal intermutational distance, as previously described[52]. For FFPE samples, to account for increased background sequencing artifacts, we considered only mutations with VAF > 0.15. Regions of chromothripsis were identified based on integrated evaluation of rainfall plots, allelic CN plots and SV calls.

**Determining evolutionary relationships between RS and CLL and identifying RS-specific genetic alterations.** The PhylogicNDT[12,32] suite of tools was used to generate posterior distributions on cluster positions and mutation membership to calculate the ensemble of possible trees that support the phylogenetic relationship of detected cell populations. Through applying this tool across a set of CLL and RS samples per patient, the most likely tree was identified using probabilistic modeling and thus parent–child relationships among clones. Furthermore, all mutations were assigned to clones based on the match of mutational and clone CCF distribution. The RS clone was defined as a new emerging clone first detected in the RS sample and absent in a preceding CLL sample. In rare cases without close antecedent CLL samples, RS clones were conservatively identified from distal tree branches and through integrating available information on RS purity from pathology assessment. If a shared CLL historical clone was identified between samples, the RS was determined to be clonally related. If a shared clone was not identified across samples, the RS was determined to be clonally unrelated. In WGS PhylogicNDT results only, clusters with fewer than 20 mutations were removed, along with clusters with low CCF (<15%).

**Mapping CN alterations to RS and CLL clones.** Once clonal structure was established, subclonal sCNAs were mapped to clones using PhylogicNDT CopyNumber2Tree. Posterior probability was calculated based on CN profiles and allele-fraction distributions of heterozygous SNP sites across samples to assign likelihood of each event to belong to a clone with a particular CCF. The RS and CLL specific clonal events (both sSNVs and sCNAs) were thus identified.

**Discovery of significantly mutated genes in RS and CLL clones.** MutSig2CV[13] was run to identify driver genes from the filtered WES Mutation Annotation Format (MAF) file of both the RS history and RS

clones. Divergent CLL clones were thus excluded, allowing for the identification of recurrent drivers contained within RS cells and clones. To improve power to detect known variants further, we ran MutSig2CV on a restricted set of hypotheses through utilizing list of CLL[6] and de novo DLBCL drivers[15,19]. For the validation cohort, MutSig2CV results were reported for new drivers that met significance and were present in at least one patient from the discovery cohort.

**Identification of recurrent RS focal and arm-level copy-number events.** Somatic copy-number alterations (sCNAs) were detected using the GATK4 CNV pipeline (http://github.com/gatk-workflows/gatk4-somatic-cnvs), composed of the CalculateTargetCoverage, NormalizeSomaticReadCounts and Circular Binary Segmentation (CBS) algorithms[53] for genome segmentation, with additional normalization for FFPE samples as described for WES analysis. To identify significantly amplified or deleted genomic regions in RS samples, GISTIC2.0 (ref. 14) was applied, both before and after subtracting the CLL sample segment changes, to produce a list of candidate RS sCNA driver regions. In parallel, the antecedent CLL sCNA drivers were examined through GISTIC. Significant events were reported with a $Q$ value threshold of 0.1. A force-calling process was applied to identify the presence/absence of each sCNA driver event across tumor samples (https://github.com/getzlab/GISTIC2_postprocessing). This force-calling process was then applied to all DLBCL recurrent sCNAs in RS and to identify RS recurrent sCNAs in DLBCL, both for frequency comparisons and to build a consensus matrix for clustering.

**Signature analysis.** Mutational signatures were determined using SignatureAnalyzer (https://github.com/getzlab/getzlab-SignatureAnalyzer). We furthermore compared the identified signatures with those in COSMIC (v.3.2)[51] based on cosine similarity.

**Immunogenetic analysis.** To determine the clonal relationships between CLL and RS, we inferred the DNA sequences of immunoglobulin genes from WES/WGS data, as previously described[6] (Supplementary Table 4).

### Consensus clustering of genetic alterations

**Generation of gene sample matrix.** All significantly mutated genes (MutSig2CV, $Q \leq 0.1$, and a frequency of four or more cases) and significant regions of sCNAs (GISTIC2.0, $Q \leq 0.1$, and a frequency of four or more cases) were assembled into a gene-by-sample matrix (Supplementary Table 7c). The entries in the gene-by-sample matrix represent mutations and CN events as follows: nonsynonymous mutations, 2; synonymous mutations, 1; no mutation, 0; high-grade CN gain (CN ≥ 3.4 copies), 2; low-grade CN gain (3.4 copies ≥ CN ≥ 2.1 copies), 1; CN neutral, 0; low-grade CN loss (1.1 ≤ CN ≤ 1.9 copies), 1; high-grade CN loss (CN ≤ 1.1 copies), 2; WGD, 5.

**NMF clustering.** The seven samples without genetic drivers in the gene-by-sample matrix were assigned to cluster C0. In addition, we identified marker genes differentially expressed across clusters by applying Fisher's exact test (2 × 5 table with variant present or absent as one dimension and cluster as the second dimension) and corrected the $P$ values for multiple hypothesis testing using the BH-FDR procedure (Supplementary Table 7f). Features with a $Q$ value ≤ 0.1 were selected as cluster features and visualized as a color-coded heatmap. Features were annotated with their maximally positive associated cluster, determined by computing the 2 × 2 Fisher's exact test for all five clusters (2 × 2 table with variant present or absent as one dimension and within-cluster or outside-cluster the second dimension; Supplementary Table 7f). To ensure robustness, given the sample size of 97, we performed 100 subsampling iterations by randomly removing eight patients in each iteration and calculated a sample-by-sample similarity matrix that reflects the frequency that each of two samples were clustered together in the 100 runs. Finally, we performed UPGMA hierarchical clustering using 1-similarity as a distance metric. To define the final cluster membership, we cut the resulting dendrogram based on the modal number of clusters across the 100 subsampled consensus NMF clustering runs.

**Mutual exclusivity/co-occurrence estimations.** For each gene of interest, the significance of the co-occurrence or mutual exclusivity for each pair of different events (mutations, amplification, and deletion) that affects that gene was calculated using Fisher's exact test, and then the false discovery rate was calculated using the Benjamini–Hochberg method.

**Non-negative matrix factorization consensus clustering.** To identify clusters of tumors with shared genetic features robustly, we applied a non-negative matrix consensus clustering algorithm[54] with slight modifications. Briefly, we passed the gene-by-sample matrix to the NMF consensus clustering algorithm (testing number of clusters $k = 2$–10) and skipped the matrix normalization step so that the distance is calculated directly based on the values in the gene-by-sample matrix. The consensus NMF method was run as 20 iterations of NMF, starting with different random seeds. The NMF consensus clustering algorithm provided the cluster membership of each sample, the cophenetic coefficient for $k = 2$ to $k = 10$ clusters and silhouette values for the optimal number of clusters, which was $k = 5$ (Supplementary Table 7d).

### Bulk RNA sequencing and data analyses

High-quality RNA from CLL–RS pairs was extracted, as previously described[4]. Total RNA was quantified using the Quant-iT RiboGreen RNA Assay Kit and normalized to 5 ng $\mu l^{-1}$. Following plating, 2 µl of ERCC controls (using a 1:1,000 dilution) was spiked into each sample. An aliquot of 200 ng for each sample was transferred into library preparation, which uses an automated variant of the Illumina TruSeq Stranded mRNA Sample Preparation Kit. This method preserves strand orientation of the RNA transcript. It uses oligo dT beads to select mRNA from the total RNA sample, followed by heat fragmentation and cDNA synthesis from the RNA template. The resultant 400 bp cDNA then goes through dual-indexed library preparation: 'A' base addition, adapter ligation using P7 adapters, and PCR enrichment using P5 adapters. After enrichment, the libraries were quantified using Quant-iT PicoGreen (1:200 dilution). After normalizing samples to 5 ng $\mu l^{-1}$, the set was pooled and quantified using the KAPA Library Quantification Kit for Illumina Sequencing Platforms. The entire process was in a 96-well format, and all pipetting is done by either Agilent Bravo or Hamilton Starlet. Pooled libraries were normalized to 2 nM and denatured using 0.1 N NaOH before sequencing. Flow-cell cluster amplification and sequencing were performed according to the manufacturer's protocols using either the HiSeq 2000 or HiSeq 2500 instrument. Each run generated a 101-bp paired end with an eight-base index barcode read. Data were analyzed using the Broad Picard Pipeline, which includes de-multiplexing and data aggregation.

### Bulk RNA sequencing of validation cohort

RNA was extracted with a Macherey Nagel RNA extraction kit. Total RNA-seq libraries were generated from 500 ng of total RNA using TruSeq Stranded Total RNA LT Sample Prep Kit with Ribo-Zero Gold (Illumina), according to the manufacturer's instructions. The final cDNA libraries were checked for quality and quantified using capillary electrophoresis before sequencing with HiSeq 4000 sequencing using using 1x50 bases protocol.

**RNA-seq data analyses.** RNA-seq reads were aligned to the human reference genome hg19 using STAR (v.2.4.0.1)[55]. Lowly expressed genes with CPM < 1 in all samples were filtered out. Differentially expressed (DE) genes were assessed using limma-voom[56] in paired mode using

sample read counts, with $|\log_2 FC| > 1$ and an adjusted $P$ value < 0.25 as a cutoff. To ensure robustness of the analysis, for the five pairs of RS and CLL samples analyzed, DE genes were recalculated iteratively, each time leaving out one sample pair. Genes were rank ordered by their $t$ statistic multiplied by the frequency they were found significant ($|\log_2 FC| > 1$ and an adjusted $P$ value < 0.1) in the leave-one-out analysis. This was used as input for preranked GSEA on HALLMARK pathways (1,000 permutations, weighted enrichment statistics, MsigDB v.7.4)[57].

### RNA clustering of RS samples and integration with genetic subtypes

Gene counts were preprocessed with ComBat-seq (v.3.42.0)[58] to eliminate possible batch effects, and one sample was removed as an outlier. TPMs were computed, and genes were filtered out if TPM = 0 in at least one sample, median TPM over samples ≤ 0.5 or median TPM over samples > 1,000. TPMs were then $\log_2$ transformed, and top genes by variance ($z$ score of variance > 1) were $z$-score transformed for downstream analysis. Consensus clustering[59] using the hierarchical clustering (complete linkage) with Spearman distance was used to identify the optimal number of clusters (observed as five RNA subtypes), and the resulting consensus matrix was transformed into a distance matrix for hierarchical clustering (complete linkage). The agreement between RNA subtypes and genomically identified clusters was determined by Fisher's exact test. Supervised analysis for differentially expressed genes for each genomically identified cluster was performed using limma-voom (v.3.50.3)[56] as a one-versus-other comparison. Pathway analysis of each genomically identified cluster was performed using preranked GSEA[57] with the MsigDB Hallmark (v.7.4) genesets using the LIMMA $t$ statistic to rank order genes.

### Single-cell RNA sequencing and analysis

**Sample preparation.** For suspension samples with admixture of both CLL and RS cells, cells were thawed by drop-wise addition of warmed media (RPMI 10% FCS) and stained with antibodies (Biolegend CD5 FITC catalog no. 364022, CD19 PE-Cy7 catalog no. 302216, CD3 PB catalog no. 300330 using 2–4 µl of each antibody per 100 µl test) and a viability marker (Biolegend 7-AAD catalog no. 420404 at 1:500 or Zombie Violet catalog no. 423114 at 1:1,000) before resuspension in PBS 0.04% BSA (Ultrapure NEB/Invitrogen). For patients 19 and 41, viable $CD5^+CD19^+$ cells were sorted into RS and CLL fractions by size based on the increased forward scatter (FSC) of RS cells (BD FACS Aria II). For patients 4, 10 and 43, viable cells within the lymphocyte gate were sorted for analysis.

**Sequencing.** Five to 10,000 single cells per specimen underwent transcriptome sequencing (Chromium Controller; 10X Genomics) according to the manufacturer's instructions, using either the 3′ v2 kit (patients 19 and 41) or the 5′ v2 kit with BCR and TCR sequencing (patients 4, 10 and 43). Each flow-sorted fraction was run as a separate lane on the same chip. Libraries were pooled and sequenced on HiSeqX or NovoSeqS4 (Illumina).

**Data processing of scRNA-seq libraries.** Reads were processed and aligned to the Hg19 reference genome. All data were filtered using Cell Ranger (v.2.1.1 for patient 41; v.2.0.0 for patient 19, and v.3.0.2 for patients 4, 10 and 43). Background, or ambient, RNA was removed using CellBender, with the exception of patient 41. Data from each patient were analyzed using Seurat (v.3.1.4)[60]. QC filtering was applied to remove cells with <500 UMIs, >50,000 UMIs or >10% mitochondrial reads. Potential doublets were detected using DoubletFinder (v.2.0.2)[61] using default pN and optimal pK for each sample and removed ahead of further analysis. For patient 41, cell-cycle regression was performed followed by data integration using standard methods[60].

Clustering was then performed to identify B cell clusters; these were further sub-clustered for additional analysis of malignant B cells using the presence of standard B cell markers (*CD19*, *CD20*, *IGLL5*, *CD79A* and *CD79B*) and the absence of T/NK/Myeloid markers (*CD3*, *CD4*, *CD8*, *CD56*, *CD14*, *CD16* and *CD33*). Clustree (v.0.4.2) was used to identify stable clusters beforedownstream analysis. UMI/cell and genes/cell for each cluster were calculated with Seurat, and the mean values across CLL and RS clusters were compared using a Wilcoxon test.

**Inferred copy number across single cells (CNVSingle).** We applied a new tool CNVSingle (https://github.com/broadinstitute/CNVsingle) to the above processed Seurat objects. In brief, CNVSingle utilized normalization from matched PBMC-derived B cell profiles followed by Savitzky–Golay noise reduction. These profiles alongside the per cell allele counts across common heterozygous SNP sites identified in the samples were utilized by a hidden Markov model running in allele-specific mode on subsets of cells. Thus, CNVSingle provides allele-specific copy-number profiles for all malignant cell clusters. As validation, different types of normal cells provided copy-neutral profiles. Single cell–derived allelic CN across clusters was compared to WES CN profiles and found to be highly concordant. These profiles were then used to identify clusters, as CLL, RS or transitional and cluster identities were used for subsequent differential expression testing.

**Differential expression testing.** Expression analysis was performed on CLL and RS clusters identified as those CNVSingle profiles that matched the CLL WES or RS WES samples. Clusters that showed intermediate sCNA profiles were considered potential transitional clusters. Furthermore, genes with nonzero read counts in fewer than 20 cells were removed. Gene counts for a given cluster were obtained by summing the counts across all the cells in each respective cluster. DE genes were assessed using limma-voom (v.3.50.3) in paired mode. The ranked gene list sort the $t$ statistic from the DE analysis for RS clusters in each scRNA-seq sample was submitted to preranked GSEA to analyze the HALLMARK pathways (1,000 permutations, weighted enrichment statistics, MsigDB v.7.4)[57].

**Velocity analysis.** RNA inference of directional trajectories was performed with scVelo (v.0.2.4, with fit_connected_states=False) using the dynamical model on the normalized data. Spliced and unspliced reads were computed via velocity (v.0.17.17)[35]. The result of the model was then used to estimate gene latency, which represents the cell's internal clock and is based only on its transcriptional dynamics. The root key parameter has been computed via the CellRank (v.1.5.0) library.

**Random forest analysis.** We used a random forest approach to differentiate between CLL and RS cells in the single-cell data. Data were preprocessed using the same cell/gene filtering as in the DE analysis. To reduce the impact of cell size differences between CLL and RS, we performed a $z$-score normalization per cell. We trained an RF (*n_estimators* = 1,000, sklearn v.1.0.1) on sample LNs from patients 10 and 43 and predicted on cells from patients 41 (LN or PB samples) and 18 (BM) whose cell labels were determined by FACS sorting described earlier. We ran the random forest 20 times and obtained a mean ± $\sigma$ of 0.92 ± 0.01 when looking at only the LN sample in the test set to avoid any potential microenvironment differences. When we included the PB sample in the test set, the $F_1$ only slightly decreased to 0.86 ± 0.11, while also adding in patient 18 yielded an $F_1$ of 0.66 ± 0.01. The decrease in $F_1$ score is possibly due to differences in tissues of origin and sequencing platforms. The top discriminative features are defined as genes whose gini impurity scores were at least 3$\sigma$ above the mean.

**Clinical endpoint analysis and statistical analyses.** Data analyses were carried out using GraphPad Prism v.9 and R v.4. To compare RS drivers identified with previously reported CLL ($n$ = 1,063)[6] and DLBCL ($n$ = 304)[19] and ($n$ = 574)[15] datasets, a two-sided exact binomial test was

performed with Benjamini–Hochberg multiple hypothesis testing correction. To obtain the frequency of RS events in DLBCL cohorts before this comparison, we called RS sCNAs in 304 DLBCLs[19] and the 443 primary DLBCLs[15] for which purity was >20%. Event frequencies were compared when an event was detected in both sample sets. RS and CLL drivers co-occurrences were represented by using a Sankey diagram. Significance was evaluated by calculating the probability for acquiring each of the RS drivers considering the acquisition of a given driver in CLL using Fisher's exact test. To evaluate how often a given driver initially occurs during the RS stage in the subset of related RS, we performed the McNemar test. Differences were considered significant when a $P$ value adjusted for multiplicity of testing was <0.05. Overall survival was defined as the interval between date of transformation and death or censored at last follow-up. Survival data were calculated using the method of Kaplan–Meier, and curves were compared by log-rank testing.

**cfDNA analysis.** After sequencing, plasma cfDNA samples were processed and analyzed as reported[36]. To detect RS-specific changes, we undertook the following steps. First, we analyzed delta copy-number changes between segments, assigning a positive chromothripsis score when three consecutive 1-Mb segments had CN delta ≥ 0.1, suggested locally fractured genome. Second, to assess Richter-specific aneuploidy, we evaluated the fraction of genome in non-copy-neutral state by FGA, defining a region as altered if the segment had an event as detected by iCHOR analysis and a CN change ≥ 0.1 (to filter out low confidence CN changes) and comparing to a matched CLL sample when available. Third, we assigned WGD to samples where copy-number events had allelic ratios (corrected for iCHOR estimated purities) corresponding to two levels of allele deletions (that is, 2/0, 1/1 and 2/1 copy-number states) as measured from the main balanced copy-number level (2/2). Lastly, we performed WES on cfDNA, which we then examined for RS clonal alterations detected in bulk through phylogenetic reconstruction.

### Data deposition

WES, RNA-seq, WGS and scRNA-seq data will be deposited in dbgap (accession number phs002458.v2.p1) at the time of publication.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

WES, RNA-seq, WGS, scRNA-seq and cfDNA data are available at dbgap (https://www.ncbi.nlm.nih.gov/gap/) using accession number phs002458.v2.p1 (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002458.v2.p1). RNA-seq data from validation cohort can be accessed by registering for an EGA account (https://ega-archive.org/) and contacting the Data Access Committee under study EGAS00001005495 and accession number EGAD00001007922. The following figures have associated raw data: Fig. 2b–h, Fig. 4a,f, Fig. 5a,b,d,e, Fig. 6g, Extended Data Fig. 4a–x, Extended Data Fig. 5a–c, Extended Data Fig. 6a, Extended Data Fig. 8, Extended Data Fig. 9, and Extended Data Fig. 10.

## Code availability

Code is available for CNVSingle (https://github.com/broadinstitute/CNVsingle) and PhylogicNDT CopyNumber2Tree (https://github.com/broadinstitute/PhylogicNDT).

## References


42. Parikh, A. R. et al. Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers. *Nat. Med.* **25**, 1415–1421 (2019).
43. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
44. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
45. Gao, G. F. et al. Tangent normalization for somatic copy-number inference in cancer genome analysis. *Bioinformatics* **38**, 4677–4686 (2022).
46. Consortium, I.T.P.-C.A.o.W.G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
47. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
48. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res* **28**, 581–591 (2018).
49. Bass, A. J. et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nat. Genet.* **43**, 964–968 (2011).
50. Drier, Y. et al. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res* **23**, 228–235 (2013).
51. Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
52. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
53. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
54. Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA* **101**, 4164–4169 (2004).
55. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
56. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
57. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
58. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom. Bioinform* **2**, lqaa078 (2020).
59. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52**, 91–118 (2003).
60. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 e1821 (2019).
61. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329–337 e324 (2019).


## Acknowledgements

We thank C. Hahn, E. Ten Hacken, W. Zhang, S. Gohil and L. Werner for helpful discussions. We thank C. Patterson, S. Pollock, O. Olive, C. J. Shaughnessy, F. Dao and H. Lyon for assistance in data collection and organization and S. Belkin and C. Birger for assistance in data storage. We thank T. Lehmberg, M. McDonough, C. Galler and M. Collins for assistance in sample collection and biobanking. We thank the patients, their families and the investigators of the clinical trials for providing samples and clinical data. This study was supported by NIH/NCI P01

CA206978 (to C.J.W. and G.G.) and NCI (1U10CA180861-01) (to C.J.W.). The work is partially supported by the Broad/IBM Cancer Resistance Research Project (I.L., G.G. and L.P.) and a grant from Force Hemato (R.G.). Individual support was provided by DDCF Physician-Scientist Fellowship (E.M.P.), Dana-Farber Flames FLAIR fellowship (E.M.P.), ASCO Conquer Cancer Young Investigator Award (E.M.P.), Fishman Family Fund (R.G. and C.L.), EMBO fellowship ALTF 14-2018 (B.A.K.), NCI Research Specialist Award R50CA251956 (S.L.) and NIH/NCI R21CA267527-01 (S.Y.). Additional research support was provided by NIH R01 CA213442 (J.R.B.), Melton Family Foundation (J.R.B.), NIH/NCI R01-CA236361 (T.J.K.) and the Deutsche Forschungsgemeinschaft (DFG) SFB1074 subprojects B10 (E.T.) and subprojects B1 and B2 (E.T., C.S. and S.S.)

## Author contributions

E.M.P., I.L., R.G., C.J., G.G., S.S. and C.J.W. designed and performed the experiments, analyzed the data and wrote the manuscript. E.M.P., N.P.-Z., A.J.A., T.H. and S.L. generated single-cell RNA-seq data. C. Lemvigh, E.M.P. and I.L. analyzed single-cell RNA-seq data, along with C. Levovitz, F.U. and K.R. R.G., E.T., C. Schneider, M.S.D., N.J., W.W., L.R., T.J.K, J.B., S.H., P.F., F.C., N.K., S.P., J.R.B. and S.S. provided patient samples. K.J.L. and S.L. designed targeted NGS assay for detecting the *NOTCH1* 3′-UTR mutation. N.R. performed mapping and analysis of these NGS data. D.R., F.U., C. Levovitz and S.Y. analyzed the RNA-seq data. S.D. analyzed the mutational data under the supervision of E.M.P. and C.J.W. C.M., J.M., J.H., L.L. and C. Stewart analyzed the WGS data. C.J., I.L., B.P., L.E., D.R., A.T-W., A.M., D.L., E.M.P., R.G. and C.J.W. performed sequencing data analyses, assessment of the clonal architecture and inference of phylogenies under the supervision of I.L. and G.G. E.M.P., R.G., L.R., J.B. and S.F. prepared patient samples. I.L. developed the analytic tool for determining somatic copy-number variations from FFPE samples and CNVSingle for detecting copy-number events in single-cell RNA-seq data. D.N. performed and supervised statistical analyses. R.R. performed statistical analyses. G.F. provided graphical representation of the clinical data. B.A.K. performed immunogenetic data analyses. B.P.D., K.R., C. Levovitz and L.P. helped to design and guide the research. B.P.D., R.A.J. and K.S. were involved in managing the project. I.L., C.J. and Z.L. performed cell-free DNA analyses. All authors discussed, interpreted results and approved the manuscript.

## Competing interests

I.L. serves as a consultant for PACT Pharma, Inc.; has stock in, is on the board of and serves as a consultant for ennov1 LLC.; and is on the board of and holds equity in Nord Bio, Inc. C.J.W., G.G., B.A.K. and Z.L. are inventors on a patent: 'Compositions, Panels, and Methods for Characterizing Chronic Lymphocytic Leukemia' (PCT/US21/45144). C.J.W., G.G., E.M.P., I.L. and R.G. are named as inventors on U.S. provisional patent application serial number 63/244,625, filed on 15 September 2021, and U.S. provisional patent application serial number 63/291,213, filed on 17 December 2021, both of which are entitled 'Diagnosis and Prognosis of Richter's Syndrome.' G.G. is a founder of and consultant for and holds privately held equity in Scorpion Therapeutics; receives funding support from IBM and Pharmacyclics; and is an inventor on patent applications related to MSMuTect, MSMutSig, MSIDetect, POLYSOLVER and SignatureAnalyzer-GPU. R.G. receives funding support from Abbvie, AstraZeneca, Gilead, Janssen and Roche. M.S.D. served as a consultant for Abbvie, Adaptive Biotechnologies, Ascentage Pharma, AstraZeneca, BeiGene, Bristol-Myers Squibb, Eli Lilly, Genentech/Roche, Janssen, Merck, Ono Pharmaceuticals, Pharmacyclics, Research to Practice, Takeda, TG Therapeutics, Verastem and Zentalis; receives funding support from Ascentage Pharma, AstraZeneca, Genentech/Roche, MEI Pharma, Novartis, Pharmacyclics, Surface Oncology, TG Therapeutics and Verastem; and receives funding for travel from Abbvie, BeiGene, BioAscend, Clinical Care Options, Curio Science, Imedex, ION Solutions, Janssen, MDOutlook, PeerView, PRIME Oncology, Research to Practice and TG Therapeutics. J.R.B. has served as a consultant for Abbvie, Acerta/AstraZeneca, BeiGene, Bristol-Myers Squibb/Juno/Celgene, Catapult, Eli Lilly, Genentech/Roche, Hutchmed, Janssen, MEI Pharma, MorphoSys AG, Novartis, Pfizer, Pharmacyclics and Rigel; and received research funding from Gilead, Loxo/Lilly, SecuraBio, Sun and TG Therapeutics. C.J.W. receives funding support from Pharmacyclics and holds equity in BioNTech, Inc. N.E.K. serves as an advisor for Abbvie, AstraZeneca, Beigene, Behring, Cytomx Therapy, Dava Oncology, Janssen, Juno Therapeutics, Oncotracker, Pharmacyclics and Targeted Oncology; receives funding support from Abbvie, Acerta Pharma, Bristol Meyer Squib, Celgene, Genentech, MEI Pharma, Pharmacyclics, Sunesis, TG Therapeutics and Tolero Pharmaceuticals; and participates on the Data Safety Monitoring Committee for Agios Pharm, AstraZeneca, BMS-Celgene, Cytomx Therapeutics, Dren Bio, Janssen, MorphoSys and Rigel. T.J.K. is on the advisory board and receives funding support from Abbvie and Roche and serves on the speakers' bureau for Abbvie, Janssen and Roche. E.T. serves as an advisor and is on the speakers' bureau for Abbvie, Janssen and Roche; and receives funding support from Abbvie and Roche. S.S. is on the advisory board and receives funding, travel support and speakers' fees from AbbVie, AstraZeneca, BeiGene, BMS, Celgene, Gilead, GSK, Hoffmann-La Roche, Janssen, Novartis and Sunesis. N.J. receives research funding from AbbVie, Adaptive Biotechnologies, ADC Therapeutics, Aprea Therapeutics, AstraZeneca, BMS, Cellectis, Fate Therapeutics, Genentech, Incyte, Loxo Oncology, Medisix, Mingsight, Novalgen, Pfizer, Pharmacyclics, Precision BioSciences, Servier and Takeda; and serves on the advisory board/receives honoraria from AbbVie, Adaptive Biotechnologies, ADC Therapeutics, AstraZeneca, Beigene, BMS, Cellectis, Genentech, Janssen, MEI Pharma, Pharmacyclics, Precision BioSciences, Servier and TG Therapeutics. W.G.W. reports funding from Abbvie, AstraZeneca/Acerta Pharma, Cyclacel, Genentech, Gilead Sciences, GSK/Novartis, Janssen, Juno Therapeutics, KITE Pharma, Loxo Oncology, Inc., Miragen, Oncternal Therapeutics, Inc., Pharmacyclics LLC, Sunesis and Xencor. S.A.P. has received research funding to the institution from AbbVie, Ascentage Pharma, AstraZeneca, Janssen, Merck, Pharmacyclics and TG Therapeutics for clinical studies in which S.A.P. is a principal investigator. S.A.P has received honoraria for participation in consulting activities/advisory board meetings for AbbVie, Adaptive Biotechnologies, Amgen, AstraZeneca, Genentech, GlaxoSmithKline, Merck and Pharmacyclics (no personal compensation). K.J.L. holds equity in Standard BioTools, Inc. D.N. has stock ownership in Madrigal Pharmaceuticals. C.S. serves on a speakers' bureau for AbbVie and AstraZeneca. D.L. holds stock in and consults for ennov1. N.P is currently an employee at Bristol Meyers Squibb. All remaining authors declare no competing interests.

## Additional information

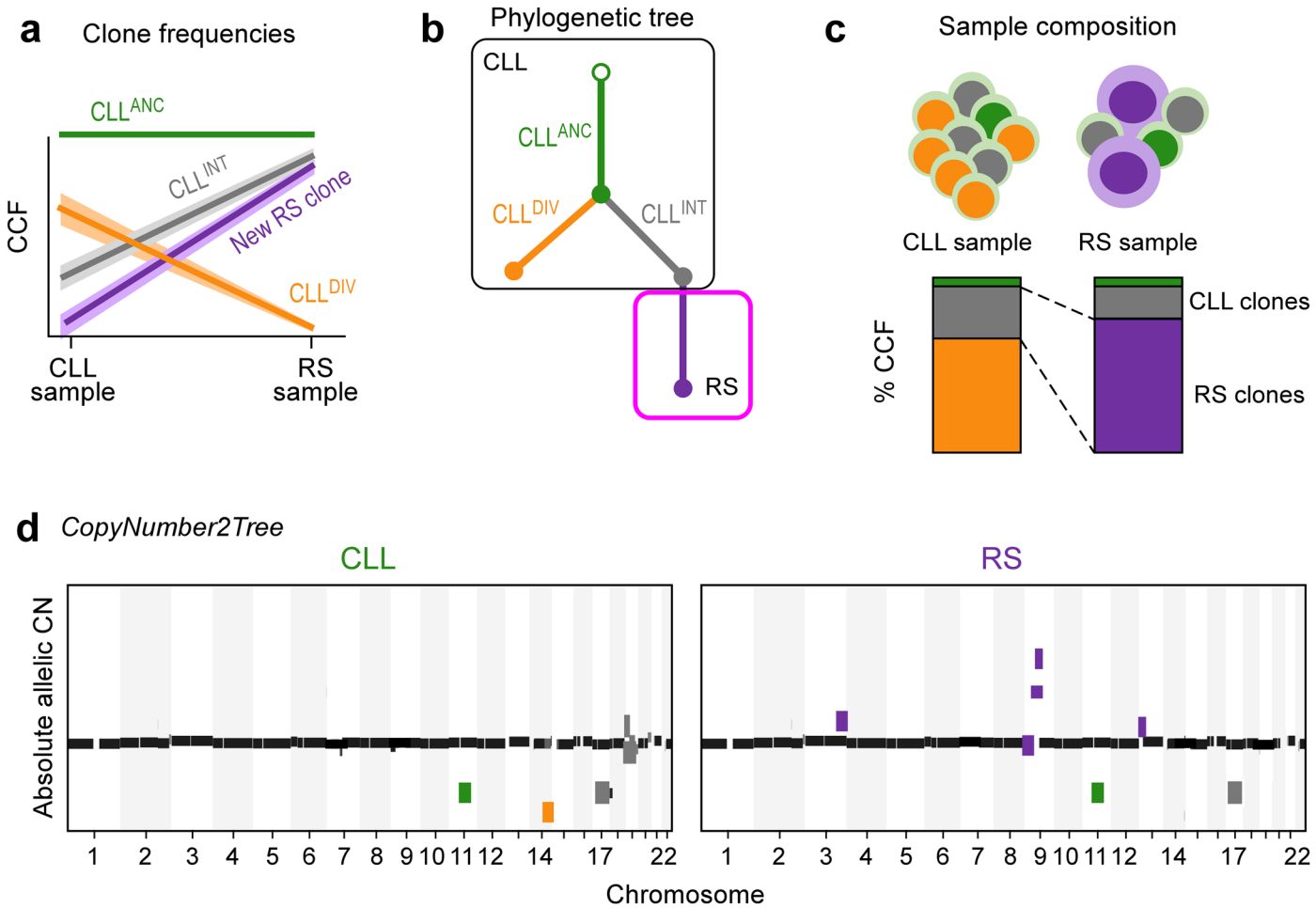
**Extended data** is available for this paper at https://doi.org/10.1038/s41591-022-02113-6.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41591-022-02113-6.

**Correspondence and requests for materials** should be addressed to Gad Getz or Catherine J. Wu.

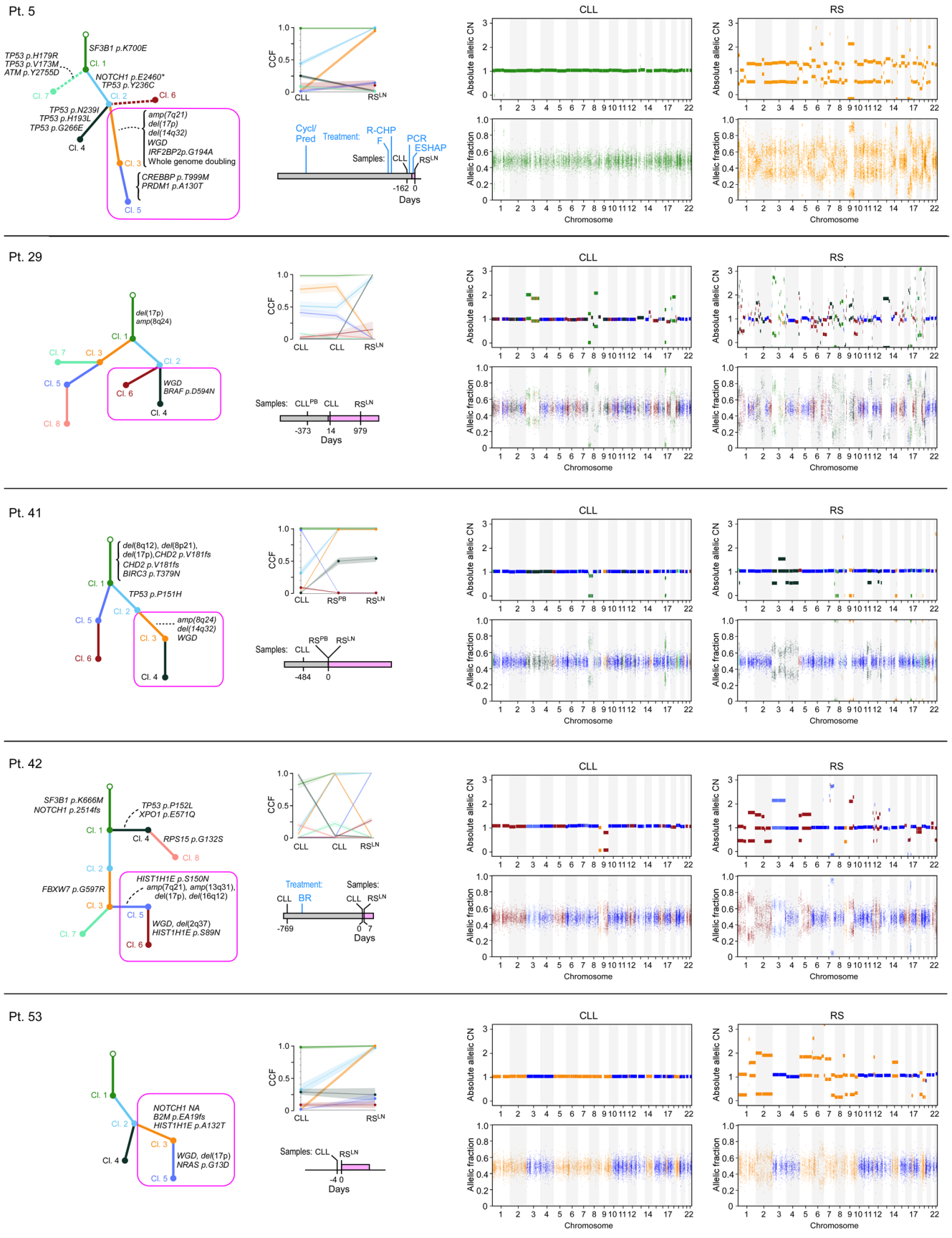
**Peer review information** *Nature Medicine* thanks Daniel Hodson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Anna Maria Ranzoni, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-022-02113-6




**Extended Data Fig. 1 | Clonal deconvolution process. a**, distinguishing RS from CLL clones after inferring subclonal composition of paired CLL and RS samples. **b**, inferring phylogenetic tree from cancer cell fraction using *PhylogicNDT*. **c**, sample composition **d**, mapping copy-number variations to clones using *CopyNumber2Tree*.

https://doi.org/10.1038/s41591-022-02113-6

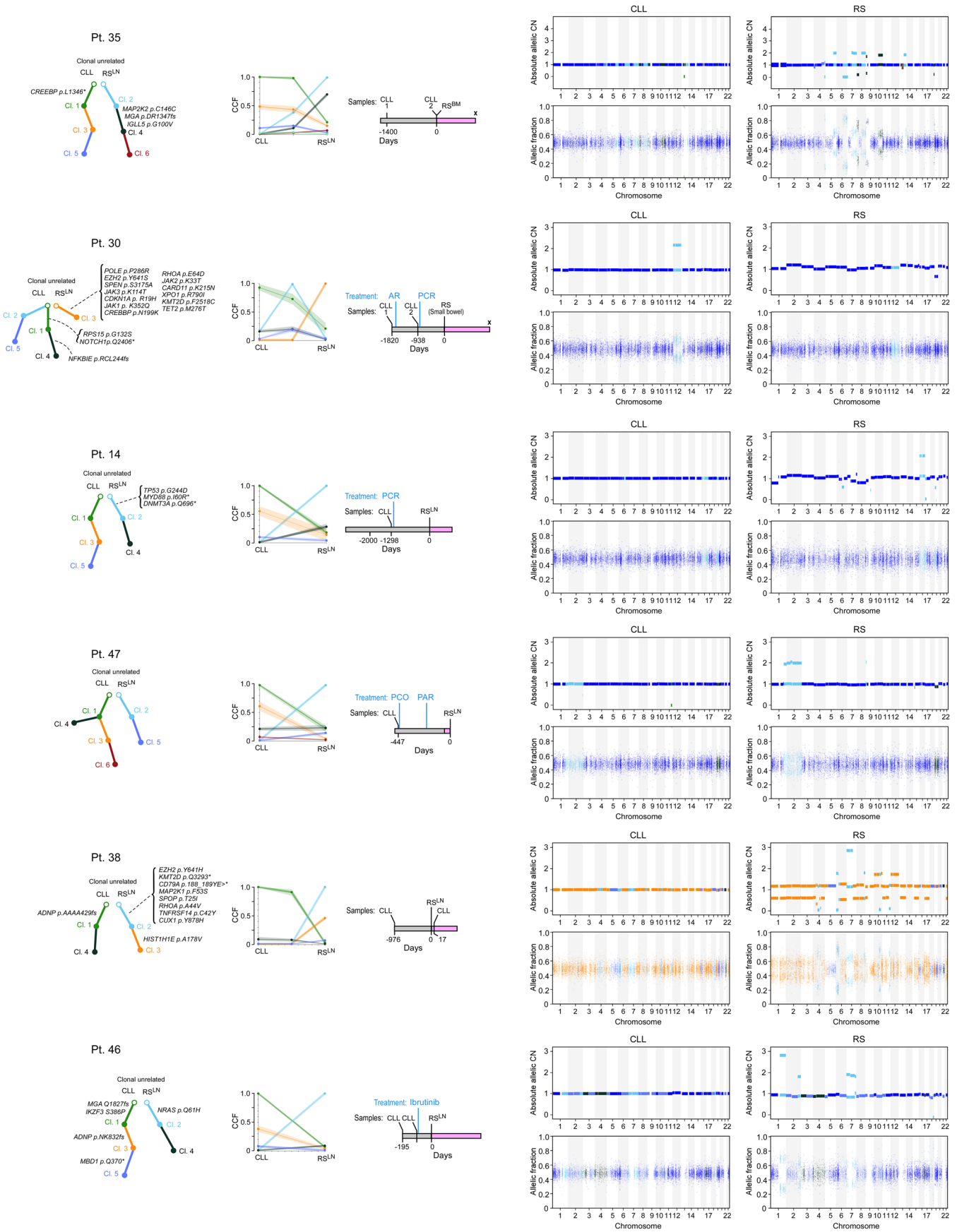



**Extended Data Fig. 2 | See next page for caption.**

https://doi.org/10.1038/s41591-022-02113-6

**Extended Data Fig. 2 | Phylogenetic reconstruction and somatic genomic alterations.** For each of the patient trios with WES data, the left panel shows the phylogenetic tree tracing the transformation history from CLL to RS. The magenta frame denotes the Richter clones. The middle top panel represents the subclonal composition inferred after clustering alterations with similar cancer cell fractions as previously reported4. The middle bottom panel indicates the timeline with RS and CLL sampling time and CLL therapeutic lines. (F, fludarabine; C, cyclophosphamide; R, rituximab; P, pentostatin; O/Ofa, ofatumumab; HDMP, high-dose methylprednisolone; A, alemtuzumab; Auto, autologous stem cell transplantation; CLB, chlorambucil; B, bendamustine; CHOP, cyclophosphamide, doxorubicin, vincristine, prednisone; ESHAP, etoposide, methylprednisolone, high-dose cytarabine, cisplatin; CHP, cyclophosphamide, doxorubicin, prednisone; Len, lenalidomide; Ob, obinutuzumab; idela; idelalisisb; D, dexamethasone; Adria, adriamycin). The right panel is composed of allelic fraction plots and allelic copy ratio plots showing clonal assignment of somatic copy-number events to CLL and RS clones. Cases with whole genome doubling in Extended Data Fig. 2 and clonal unrelated cases in Extended Data Fig. 3.

https://doi.org/10.1038/s41591-022-02113-6

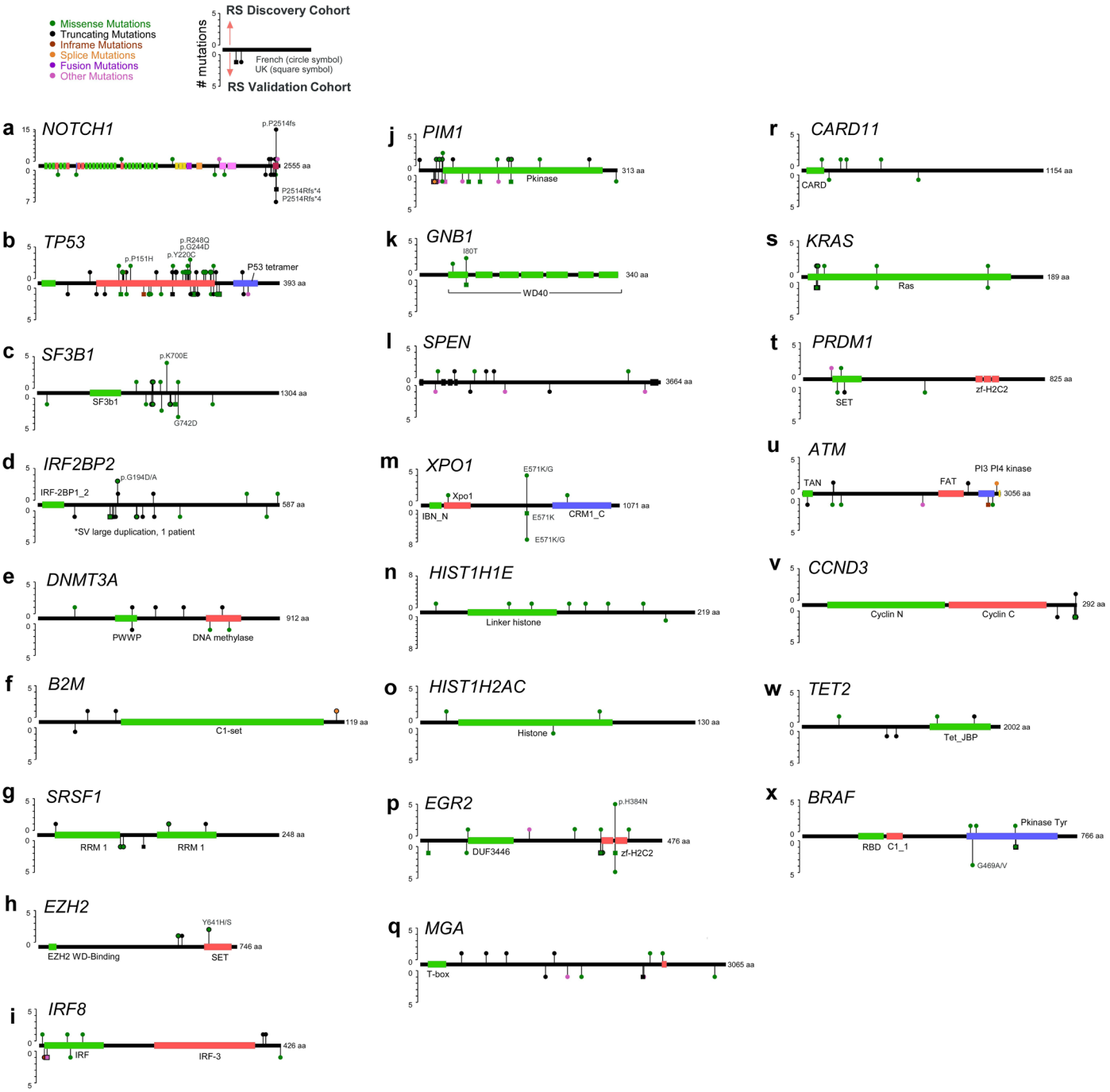



**Extended Data Fig. 3 | See next page for caption.**

https://doi.org/10.1038/s41591-022-02113-6

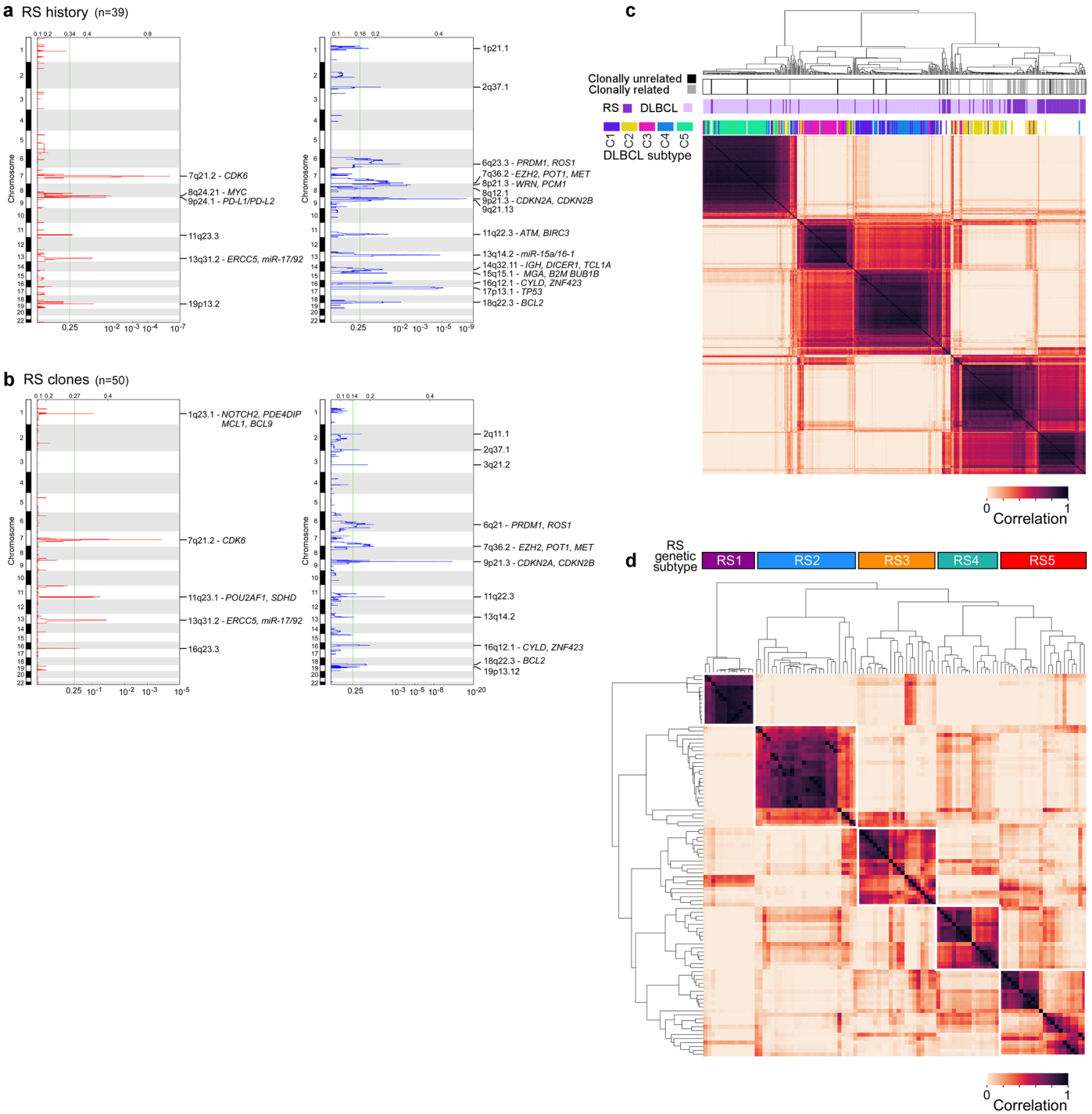
**Extended Data Fig. 3 | Phylogenetic reconstruction and somatic genomic alterations.** For each of the patient trios with WES data, the left panel shows the phylogenetic tree tracing the transformation history from CLL to RS. The magenta frame denotes the Richter clones. The middle top panel represents the subclonal composition inferred after clustering alterations with similar cancer cell fractions as previously reported4. The middle bottom panel indicates the timeline with RS and CLL sampling time and CLL therapeutic lines. (F, fludarabine; C, cyclophosphamide; R, rituximab; P, pentostatin; O/Ofa, ofatumumab; HDMP, high-dose methylprednisolone; A, alemtuzumab; Auto, autologous stem cell transplantation; CLB, chlorambucil; B, bendamustine; CHOP, cyclophosphamide, doxorubicin, vincristine, prednisone; ESHAP, etoposide, methylprednisolone, high-dose cytarabine, cisplatin; CHP, cyclophosphamide, doxorubicin, prednisone; Len, lenalidomide; Ob, obinutuzumab; idela; idelalisisb; D, dexamethasone; Adria, adriamycin). The right panel is composed of allelic fraction plots and allelic copy ratio plots showing clonal assignment of somatic copy-number events to CLL and RS clones. Cases with whole genome doubling in Extended Data Fig. 2 and clonal unrelated cases in Extended Data Fig. 3.

https://doi.org/10.1038/s41591-022-02113-6



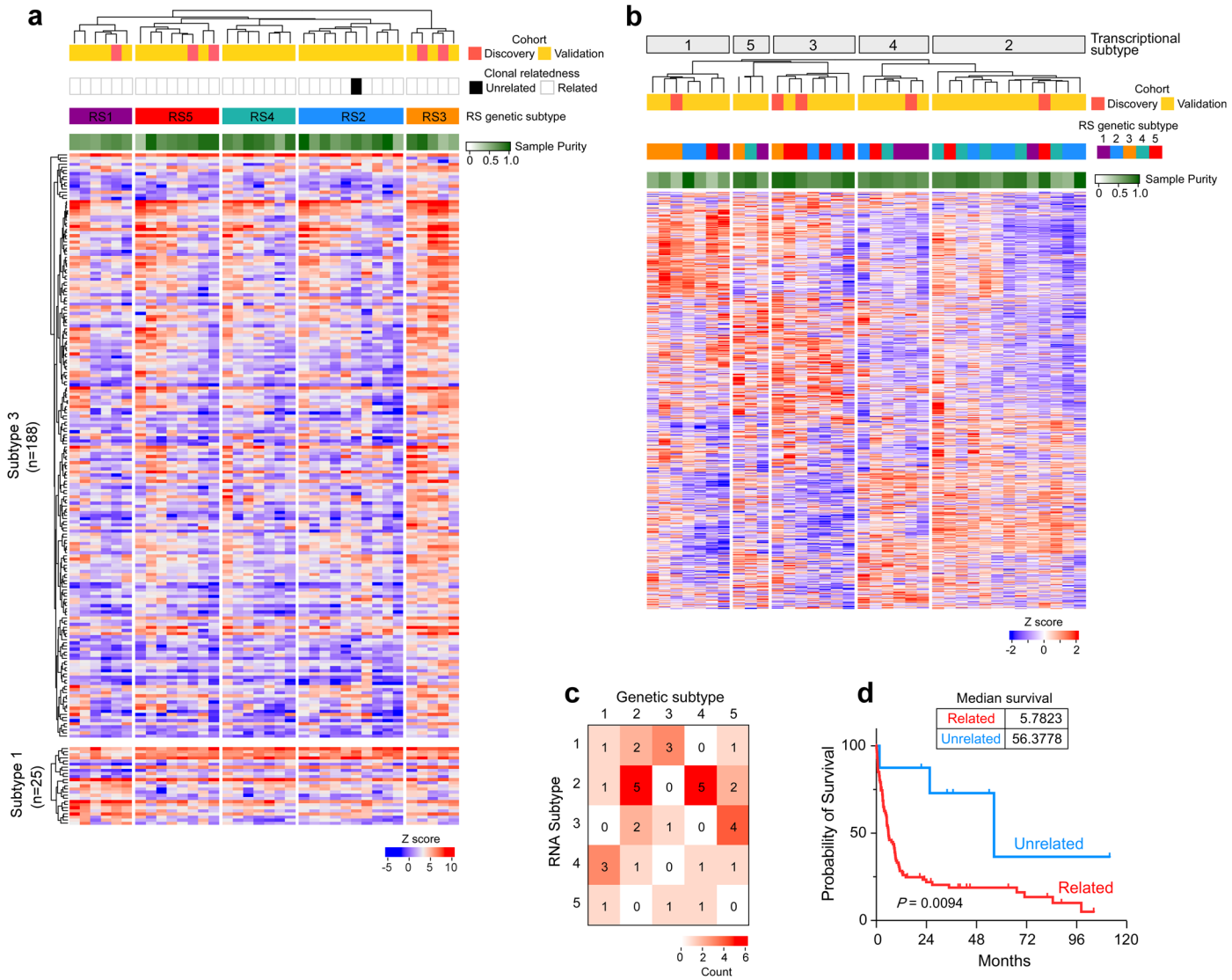

**Extended Data Fig. 4 | Putative RS driver genes. a-x**, individual protein mutation maps for selected putative Richter drivers, showing gene mutation subtype (for example, missense), position and evidence of mutational hotspots. Panels were generated by using the cBioPortal for Cancer Genomics tool.

https://doi.org/10.1038/s41591-022-02113-6



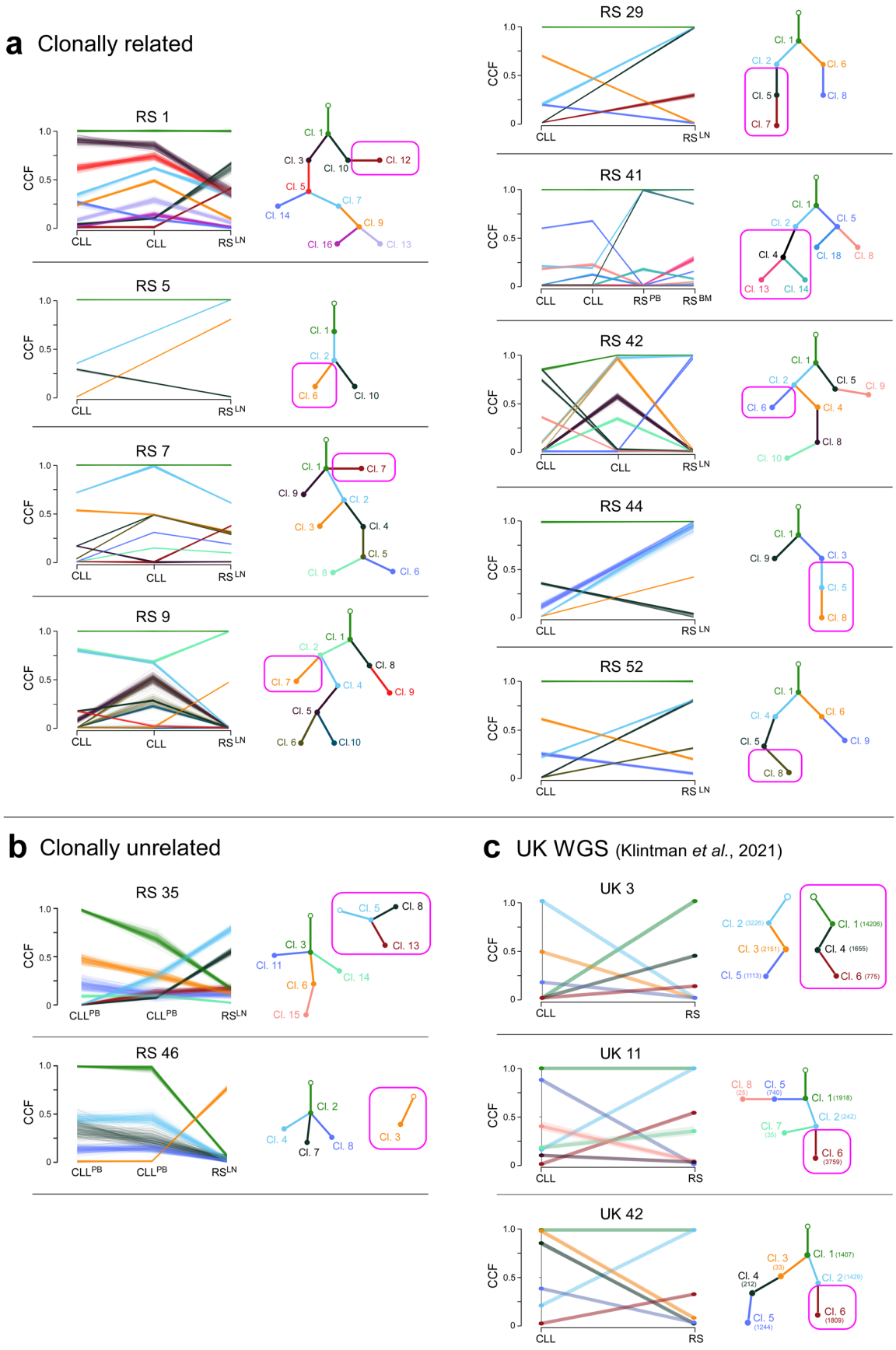

**Extended Data Fig. 5 | RS sCNAs and genomic clustering.** GISTIC2-defined recurrent copy-number gains (red, left) and losses (blue, right) are visualized for focal events for RS samples (a) and RS clones (b) (RS samples with CLL events subtracted, bottom). Chromosomes are shown on the vertical axis. Green line denotes a near significant q value of 0.25 and significant events (q<0.1) are annotated in text along with putative driver genes contained within the peak (Supplementary Table 5) **c**, NMF clustering of RS with DLBCL (304 de novo DLBCL samples[19] shows clonal related RS clusters separately from DLBCL and closes to DLBCL from C2 (ref. 19). Clonal unrelated RS clusters across DLBCL subtypes and separate from RS. Samples were annotated for clonal relationship (related RS, gray, unrelated RS, black), cohort (DLBCL, light purple; RS, dark purple) and DLBCL clusters (C1, purple; C2, yellow, C3, pink, C4, blue, C5, green)[19]. **d**, NMF clustering of RS shows 5 distinct genomic subtypes of transformation.

https://doi.org/10.1038/s41591-022-02113-6



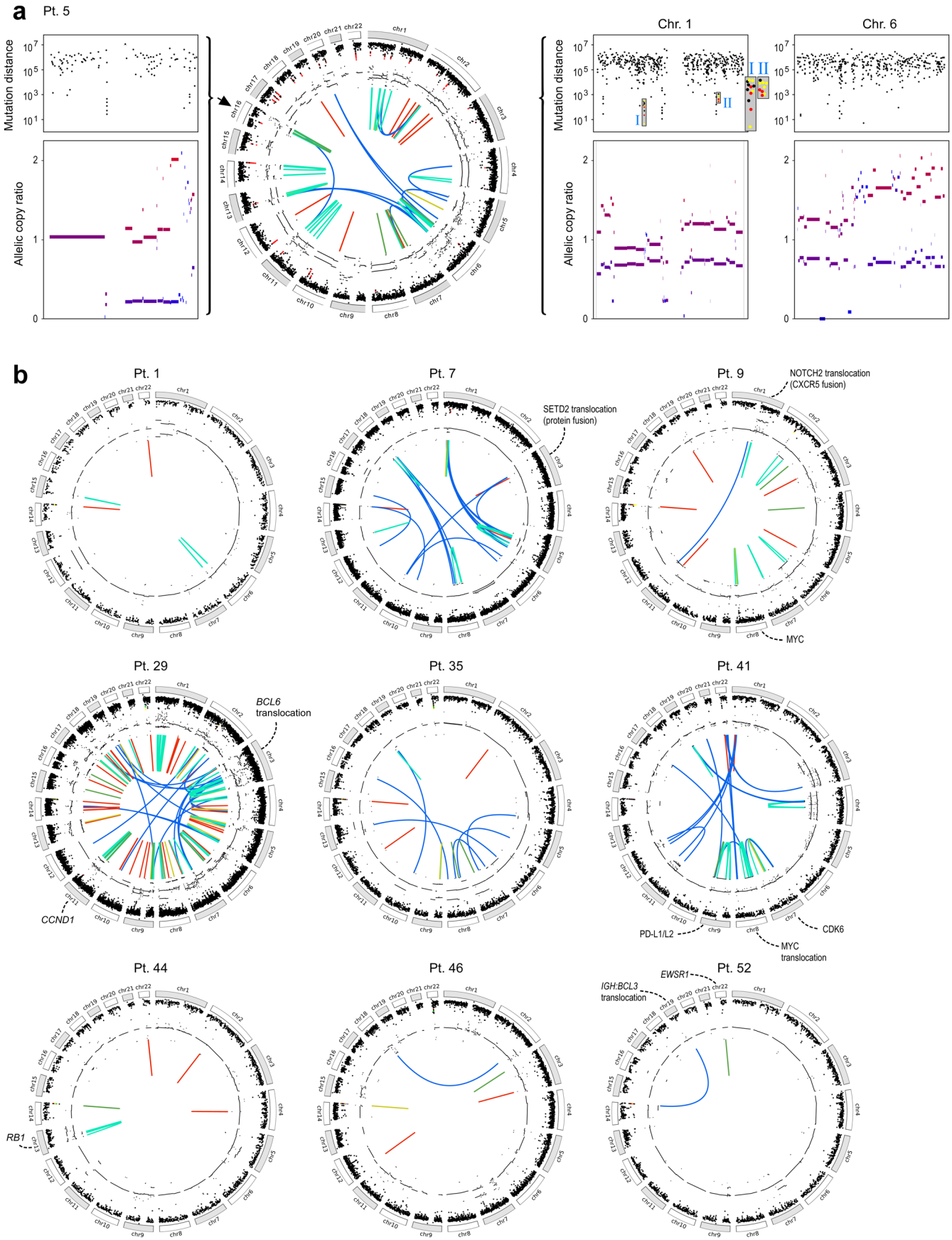

**Extended Data Fig. 6 | Transcriptome supports distinct RS molecular subtypes. a**, Supervised clustering of transcriptome data from 36 RS patients by molecular subtype highlights differentially regulated genes in subtype 1 and 3 (Supplementary Table 8). Samples are annotated for cohort (Discovery, pink; Validation, yellow), clonal relationship (unrelated, black, related, white), and sample purity by WES (green gradient). **b**, Unsupervised consensus clustering of RS transcriptome data (n=36) shows 5 clusters. (Discovery, pink; Validation, yellow), RS molecular subtype (1, purple; 2, blue; 3, orange; 4, green; and 5, pink), and sample purity by WES (green gradient). **c**, 5 × 5 table showing association between molecular subtype of RS and unsupervised transcriptome clusters (2 sided Fisher's exact test, P=0.038) **d**, Kaplan–Meier curve showing OS of clonal unrelated RS compared to clonal related RS. P value is log rank (2 sided Mantel Cox).

https://doi.org/10.1038/s41591-022-02113-6




**Extended Data Fig. 7 | Phylogenetic trees showing CLL and RS clones from WGS of paired samples. a**, Phylogenetic tree and CCF plot for 9 patients based on WGS data showing clonal related RS (magenta box). **b**, Phylogenetic tree and CCF plot for 2 patients based on WGS demonstrating clonally unrelated RS **c**, Representative phylogenetic trees and CCF plot for 3 patients from UK cohort[9] based on WGS.

https://doi.org/10.1038/s41591-022-02113-6

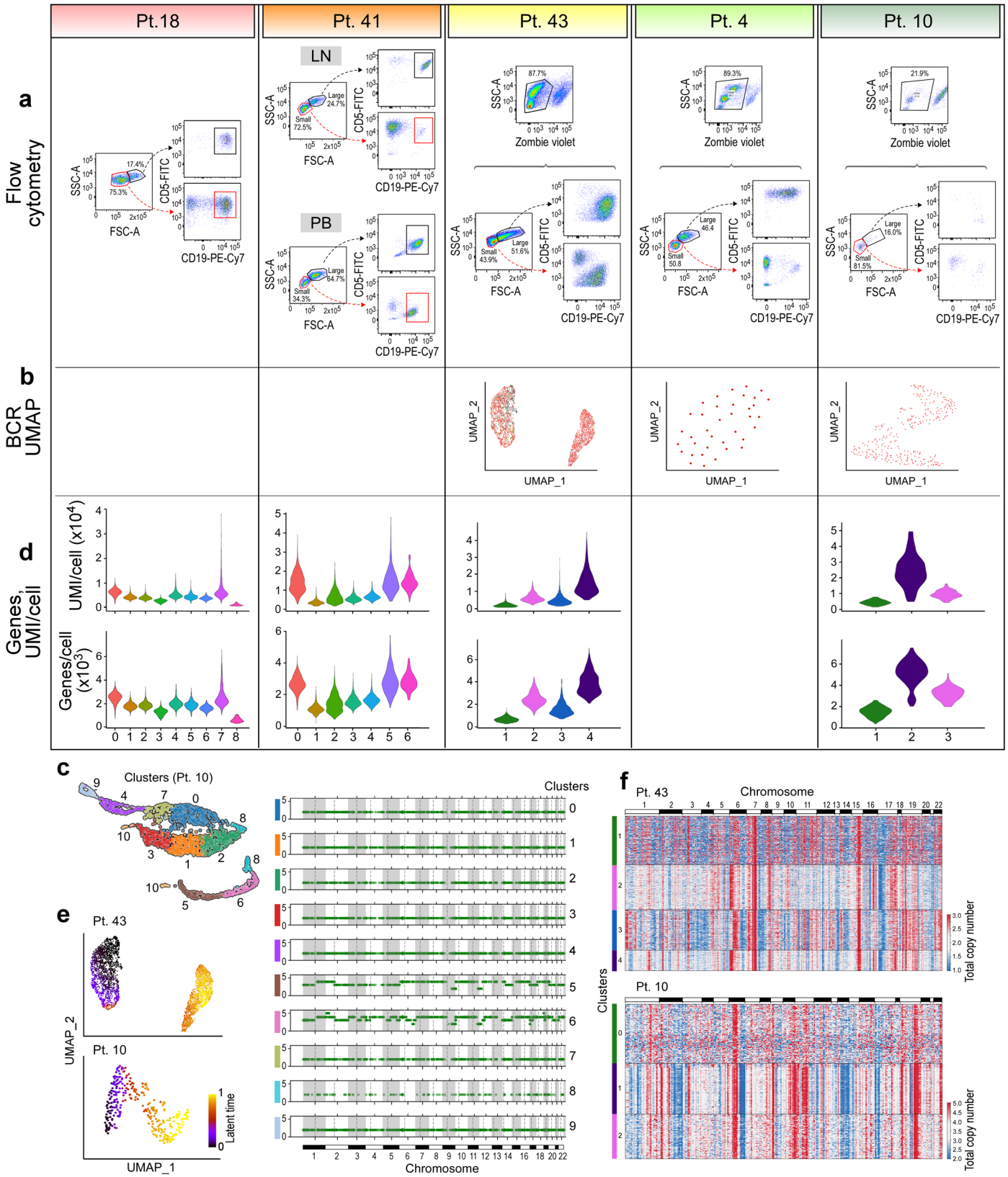



**Extended Data Fig. 8 | See next page for caption.**

https://doi.org/10.1038/s41591-022-02113-6

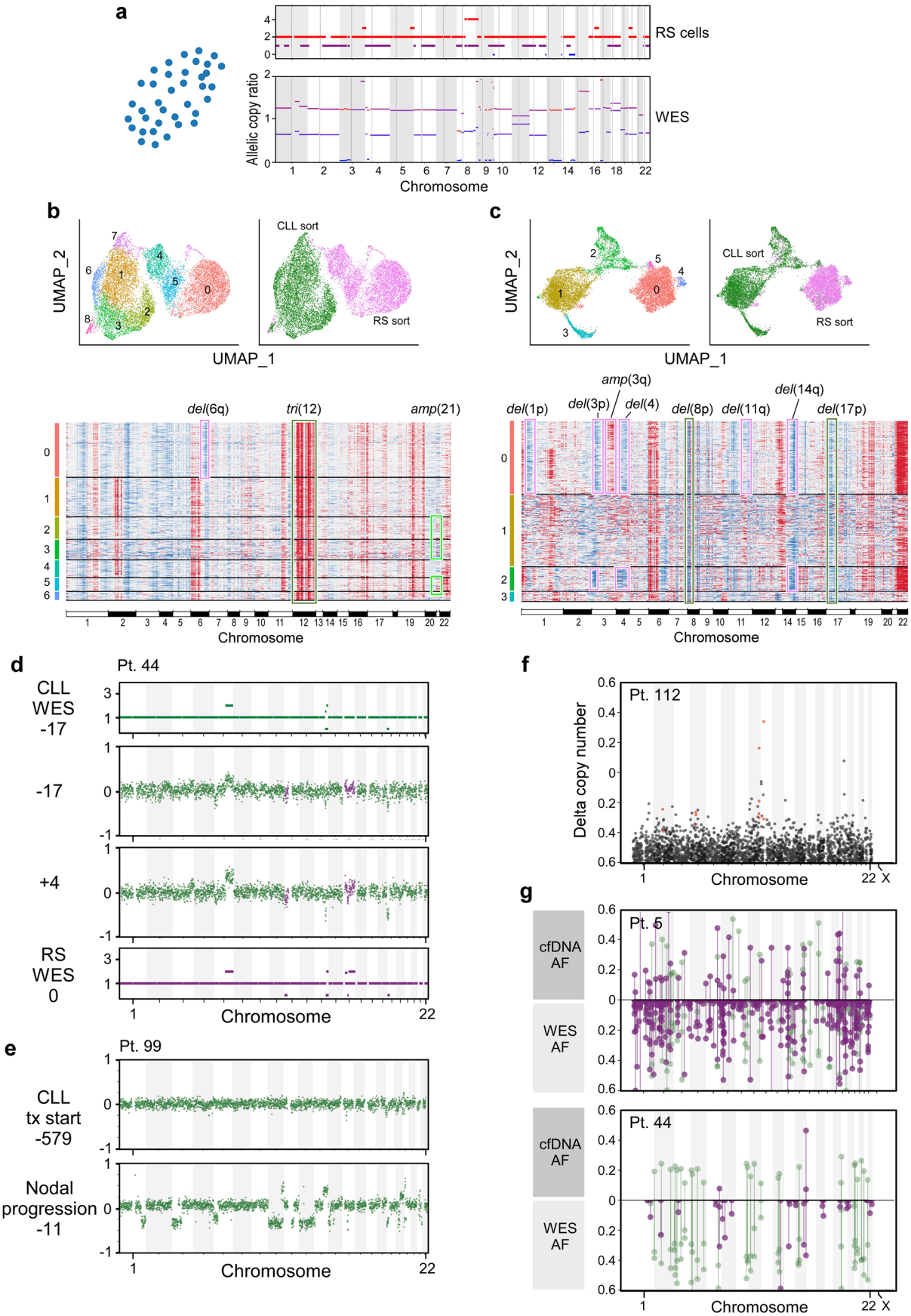
**Extended Data Fig. 8 | WGS Circos plots with or without chromothripsis.** **a**, chromothripsis and kataegis in RS sample (Pt 42) with whole genome doubling. Circos plots showing structural variants (interchromosomal, blue; deletion, red; inversion, yellow; tandem duplication, green; long range, teal), allelic copy number (middle), rainfall plot with kataegis regions (red) and chromosomes (outside). Adjacent rainfall plots show kategis regions (C to G, red; C to T, yellow; C to A, teal) with corresponding allelic copy-number fragmentation. **b**, Circos plots from RS WGS samples showing structural variants (interchromosomal, blue; deletion, red; inversion, yellow; tandem duplication, green and long range, teal), allelic copy number (middle), rainfall plot with kataegis regions (red) and chromosomes (outside). SVs impacting known genes and translocation partners are labeled (Supplementary Table 7k).

https://doi.org/10.1038/s41591-022-02113-6




**Extended Data Fig. 9 | Single cell processing and transcriptome analysis of RS samples at single cell resolution. a**, flow sorting strategy for RS single-cell samples. Flow sorting to separate RS and CLL cells by size for Patient 19 and Patient 41 (lymph node, LN; peripheral blood, PB; bone marrow, BM). Flow sorting viable cells for Pt 43, Pt 4 and Pt10. Representative flow plots below demonstrate CLL and RS cells were included in sorted population. **b**, B cell receptor (BCR) clonotypes plotted for RS and CLL clusters on UMAP visualization. **c**, Representative example from patient 10 showing CNVsingle identifies malignant B cell clusters (5 and 6) separate from immune cell clusters (0,1,2,3,4,7,9). **d**, UMI/cell and Gene/cell plots for CLL and RS single-cell clusters. RS demonstrates higher UMI/cell ($P<2.2\times10-16$ see Methods, Supplementary Table 8). **e**, RNA inference of directional trajectories is shown on UMAP visualization for Pts 43 and 10. **f**, copy-number variation heatmap inferred in each cluster from scRNA-seq data using our CNVSingle algorithm for Pts 43 and 10 (Methods).

https://doi.org/10.1038/s41591-022-02113-6




**Extended Data Fig. 10 | See next page for caption.**

https://doi.org/10.1038/s41591-022-02113-6

**Extended Data Fig. 10 | Single-cell transcriptome and copy-number analysis of RS patients. a**, UMAP visualization of single-cells from patient 4 (left) with associated allelic copy-number ratio plot inferred by CNVsingle (top right) and RS WES (bottom right). **b**, UMAP visualization of CLL and RS cells from Patient 18 (left top panel) with flow-sorting annotations (right top panel). Inferred CNAs from CNVSingle (bottom panel) are shown as heatmap with CLL (green) and RS (pink) events highlighted. **c**, UMAP visualization of CLL and RS cells from Patient 41 (left top panel) with flow-sorting annotations (right top panel). Inferred CNAs from CNVSingle (bottom panel) with CLL (green) and RS (pink) events highlighted. **d**, Plasma of patient 44 shows RS-specific sCNVs on chromosome 9 and 13 leading up to RS diagnosis, which are not reflected in circulating CLL **e**, Plasma of patient 99 at the start of CLL-directed therapy (top) and just ahead of diagnosis of RS (bottom) during CLL response. **f**, Chromothripsis in post-transplant RS plasma cfDNA at time of relapse (Pt 112). **g**, Plot showing allele frequency of RS (purple) and CLL (green) mutations in RS WES (bottom) and plasma cfDNA WES (top) for patient 5 (top) and patient 44 (bottom).

nature portfolio

Corresponding author(s): Catherine Wu

Last updated by author(s): 10/05/2022

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed |
|---|---|
| ☐ | ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ A description of all covariates tested |
| ☐ | ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☐ | ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☐ | ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | no software was used |
| Data analysis | BWA-MEM [version 0.7.15-r1140], CGA WES Characterization Pipeline v0.2 (Picard and GATK (v4.0.5.1)- including GATK CNV pipeline v4.0–CalculateTargetCoverage, NormalizeSomaticReadCounts, Circular Binary Segment algorithms, Strelka2 (v2.9.10), ContEst (GATK 3.5.0-g36282e4), MuTect (v1.1.6), DeTin (v1.8.7), AllelicCapSeg, ABSOLUTE (v1.5), IGV (v2.8.13), PhylogicNDT (v1.1), Manta (v1.6.0), SvABA (v3), dRanger, MutSig2CV (v3.1), SignatureAnalyzer (0.0.7), SvABA (v1.1), dRanger (v1.0), BreakPointer (v1.0), COSMIC Cancer Gene Census (v90), GISTIC 2.0, IgCaller v1.1, MiXCR v3.0.10, COSMIC (v3.2), STAR (v2.4.0.1), limma-voom(v3.50.3), GSEA - MSigDB v7.4, Cell Ranger Pipeline (v2.1.1, v2.0.0, v. 3.0.2), Clustree (v0.4.2), Seurat (v3.1.4), R (v 3.5.3) ComBat-seq (v3.42.0), Cellbender (v. 0.2.0), DoubletFinder (v2.0.2), GraphPad Prism Version 7, CellRank (v 1.5.0), velocity (v 0.17.17), scVelo (v 0.2.4), BD FACSDiva 8.0.1., FlowJo software (10.8.0), Copynumber2tree (https://github.com/broadinstitute/PhylogicNDT), CNVsingle (https://github.com/broadinstitute/CNVsingle) |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

WES, RNA-seq, WGS, scRNA-seq, and cfDNA data are available at dbgap (https://www.ncbi.nlm.nih.gov/gap/) using accession number phs002458.v2. RNA-seq data from validation cohort is available at EGA (https://ega-archive.org/) under study EGAS00001005495 and accession number EGAD00001007922 (http://ega-archive.org/datasets/EGAD00001007922).

## Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

| | |
|---|---|
| Reporting on sex and gender | We note that the study and analyses are conducted in accordance to the Sex and Gender Equity in Research - SAGER - guidelines. Sex was self-report and was not considered in study design. We confirm that the patients characteristics in the supplementary tables also include sex |
| Population characteristics | Data was collected on age, sex, clinical staging (Binet or Rai), IGHV mutational status, cytogenetic features, number and type of therapeutic lines, time to transformation, Richter pathology subtype, survival status at last follow-up. Statistics were performed only using overall survival data relative to genomic subtypes. |
| Recruitment | The recruitment was based on the availability of paired CLL and Richter tumor samples in the participating centers. No significant bias is expected to likely impact results. Study and analyses are conducted in accordance to the 'Sex and Gender Equity in Research - SAGER- guidelines' |
| Ethics oversight | Dana-Farber Cancer Institute (DFCI), University of Ulm in Germany, the CLL Research Consortium (including UCSD, Mayo Clinic, MD Anderson Cancer Center) and the French Innovative Leukemia Organization (FILO) group, University of Nancy approved the protocol. |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | No sample size calculation was performed. Sample sizes were based on the availability of tumor samples from participating centers |
| Data exclusions | Data failing to match the Broad Institute quality control metrics were excluded. |
| Replication | Our study was conducted on human primary tumor samples. We demonstrated excellent concordance between genomes and exomes. We also validated our initial findings from our discovery cohort in a validation cohort. |
| Randomization | This was not relevant given the exploratory nature of our work. All samples were analyzed without group allocation. |
| Blinding | Blinding is not relevant since our study does not involve any group allocation |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☐ | ☒ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☐ | ☒ Clinical data |
| ☒ | ☐ Dual use research of concern |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☐ | ☒ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

# Antibodies

Antibodies used: anti-CD19 (PE-Cy7, Biolegend cat #302216, clone HIB19), anti-CD5 (BV421, Biolegend cat #300626, clone UCHT2), anti-CD4 (FITC, Biolegend cat#300506, clone RPA-T4), anti-CD3 (Pacific Blue, Biolegend, cat#300330, clone HIT3a), anti-CD5 (FITC, Biolegend cat#364022, clone L17F12), anti-CD19 (APC, biolegend, cat #363006, clone SJ25C1), 7-AAD viability marker (Biolegend cat#420404, 1:500 dilution), Zombie Violet (Biolegend, cat#423114, 1:1000 dilution). All antibodies were used 2-4 uL per 100 uL test.

Validation: Validations performed by the manufacturer (Biolegend). Biolegend does extensive testing and quality control, as detailed on their website at biolegend.com/en-us/quality/quality-control

# Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration: NCT03619512

Study protocol: The full study protocol can be accessed by requesting to the corresponding author.

Data collection: Recruitment time period: September 6, 2017 to December 31, 2020, Observational Model: Cohort, Time perspective: Retrospective, Clinical data collection location: Central Hospital, Nancy, France

Outcomes: This was not an interventional trial.

# Flow Cytometry

## Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

## Methodology

Sample preparation:

Single-cell RNA-sequencing and analysis
Sample preparation. For suspension samples with admixture of both CLL and RS cells, cells were thawed by drop-wise addition of warmed media (RPMI 10% FCS) and stained with antibodies (Biolegend CD5 FITC cat#364022, CD19 PE-Cy7 cat#302216, CD3 PB cat#300330 using 2-4 uL of each antibody per 100 uL test ) and a viability marker (Biolegend 7-AAD cat#420404 at 1:500 or Zombie Violet cat#423114 at 1:1000) before resuspension in PBS-0.04% BSA (Ultrapure NEB/ Invitrogen). For Patients 19 and 41, viable CD5+ CD19+ cells were sorted into RS and CLL fractions by size based on the increased forward scatter (FSC) of RS cells (BD FACS Aria II). For Patients 43, 4 and 10, viable cells within the lymphocyte gate were sorted for analysis.

RS samples. RS samples were collected from BM, LN, lymphoid tissue or PBMCs and included both fresh frozen and FFPE samples. Freshly collected tissue samples were disaggregated by GentleMACs digestion (Miltenyi Biotec) before cryopreservation with FBS/10% DMSO and storage in liquid nitrogen or directly stored as whole tissue blocks in liquid nitrogen. Blood and BM specimens were isolated by Ficoll/Hypaque density gradient centrifugation prior to cryopreservation with FBS/10% DMSO and storage in liquid nitrogen. For viably frozen samples of low purity (<30% tumor), RS cells were isolated by fluorescence activated cell sorting (FACS) (Aria II instrument, Becton Dickinson) based on CD5+ and CD19+ co-expression on cells with increased forward scatter (FSC) (Biolegend, CD5-FITC cat#364022, CD19-PE-Cy7 cat#302216, 2-4 uL per 100 uL test). For FFPE specimens, samples from each submitting center were reviewed for >50% purity prior to sequencing.

| | |
|---|---|
| | Germline samples. Sources of non-tumor germline DNA included saliva (Oragene Discover [ORG500 or ORG600] kit, DNA Genotek), remission bone marrow 5 or in vitro expanded T cells. For the latter, CD19- CD4+ or CD19-CD3+ cells were collected by FACS (Aria II, BD; Biolegend, cat#300330, cat#300506, #363006). The cells were plated and expanded in vitro in RPMI (Gibco) containing phytohemagglutinin (PHA) (1.5:100), IL-7 (20 ng/mL), IL-2 (100 U/mL), 10% human serum and beta-2-mercaptoethanol (1/1000). |
| Instrument | FACS Aria II instrument (Becton Dickinson) |
| Software | BD FACSDiva software (8.0.1), FlowJo software (10.8.0) |
| Cell population abundance | Selected and representative post-sort fractions were checked by flow cytometry to assess cell purity |
| Gating strategy | FSC/SSC parameters were set to gate lymphocytes. For population including both CLL and RS cells, sorting was based on FSC-based cell size. For fluorescent assays, negative controls were used to indicate positivity/negativity boundaries. |

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.