



HAL
open science

Evolutionary history of transformation from chronic lymphocytic leukemia to Richter syndrome

Erin Parry, Ignaty Leshchiner, Romain Guièze, Connor Johnson, Eugen Tausch, Sameer Parikh, Camilla Lemvigh, Julien Broséus, Sébastien Hergalant, Conor Messer, et al.

► **To cite this version:**

Erin Parry, Ignaty Leshchiner, Romain Guièze, Connor Johnson, Eugen Tausch, et al.. Evolutionary history of transformation from chronic lymphocytic leukemia to Richter syndrome. *Nature Medicine*, 2023, 29 (1), pp.158-169. <10.1038/s41591-022-02113-6>. <hal-03950358>

HAL Id: hal-03950358

<https://hal.univ-lorraine.fr/hal-03950358v1>

Submitted on 3 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization



HHS Public Access

Author manuscript

Nat Med. Author manuscript; available in PMC 2023 July 09.

Published in final edited form as:

Nat Med. 2023 January ; 29(1): 158–169. doi:10.1038/s41591-022-02113-6.

Correspondence should be addressed to: Catherine J. Wu, MD, Department of Medical Oncology, Dana-Farber Cancer Institute, 450 Brookline Avenue, Dana Building, Room DA-520, Boston MA 02115, cwu@partners.org; Gad Getz, PhD, Broad Institute of MIT and Harvard, 415 Main St, Cambridge MA 02141, gadgetz@broadinstitute.org.

*These authors contributed equally

**These authors jointly supervised this work

AUTHOR CONTRIBUTIONS STATEMENT

E.M.P., I.L., R.G., C.J., G.G., S.S., and C.J.W. designed and performed the experiments, analyzed 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.Lem., E.M.P., and I.L. analyzed single-cell RNA seq data along with C.Lev., F.U., and K.R.

R.G., E.T., C.Sc., 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 NOTCH1 3'UTR mutation. N.R. performed mapping and analysis of this NGS data.

D.R., F.U., C.Lev., and S.Y. analyzed RNAseq data. S.D. analyzed mutational data under the supervision of E.M.P. and C.J.W.

C.M., J.M., J.H., L.L. and Chi.S. analyzed 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. provides graphical representation of clinical data.

B.A.K. performed immunogenetic data analyses.

B.P.D., K.R., C.Lev. and L.P. helped to design and guide the research.

B.P.D., R.A.J., and K.S. was involved in managing the project.

Z.L., I.L., and C.J. performed cell-free DNA analyses.

All authors discussed, interpreted results and approved the manuscript.

COMPETING INTERESTS STATEMENT

I.L. serves as a consultant for PACT Pharma Inc. and has stock, is on the board and serves as a consultant for ennov1 LLC., is on the board and holds equity in Nord Bio, Inc.; C.J.W., G.G., B.A.K., 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 September 15, 2021, and U.S. provisional patent application serial number 63/291,213, filed on December 17, 2021, both of which are entitled, "Diagnosis and Prognosis of Richter's Syndrome."; G.G. is a founder, consultant and holds privately held equity in Scorpion Therapeutics, receives funding support from: IBM and Pharmacyclics, is an inventor on patent applications related to: MSMuTect, MSMutSig, MSIDetect, POLYSOLVER, and SignatureAnalyzer-GPU; R.G. receives funding support from: Abbvie, Janssen, Gilead, AstraZeneca, and Roche; M.S.D. served as a consultant for Abbvie, Adaptive Biotechnologies, Ascentage Pharma, Astra-Zeneca, 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, Astra-Zeneca, Genentech/Roche, MEI Pharma, Novartis, Pharmacyclics, Surface Oncology, TG Therapeutics, and Verastem, 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; JRB has served as a consultant for Abbvie, Acerta/Astra-Zeneca, BeiGene, Bristol-Myers Squibb/Juno/Celgene, Catapult, Eli Lilly, Genentech/Roche, Hutchmed, Janssen, MEI Pharma, Morphosys AG, Novartis, Pfizer, Pharmacyclics, Rigil; received research funding from Gilead, Loxo/Lilly, SecuraBio, Sun, TG Therapeutics.; C.J.W. receives funding support from: Pharmacyclics; holds equity in: BioNTech, Inc; N.E.K. serves as an advisor for: Abbvie, Astra Zeneca, 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, Tolero Pharmaceuticals, participates on the DSMC (Data Safety Monitoring Committee) for: Agios Pharm, AstraZeneca, BMS-Celgene, Cytomx Therapeutics, Dren Bio, Janssen, Morpho-sys, and Rigil; T.J.K. is on the advisory board and receives funding support from: Abbvie and Roche, serves on the Speakers Bureau for Janssen, Abbvie, and Roche; E.T. serves as an advisor and on the Speakers Bureau for: Janssen, Abbvie, and Roche, receives funding support from: Abbvie and Roche; S.S. is on the advisory board and receives funding and travel support, and speaker fees from: AbbVie, AstraZeneca, BeiGene, BMS, Celgene, Gilead, GSK, Hoffmann-La Roche, Janssen, Novartis, and Sunesis; N.J. receives research funding from: Pharmacyclics, AbbVie, Genentech, AstraZeneca, BMS, Pfizer, Servier, ADC Therapeutics, Collectis, Precision BioSciences, Adaptive Biotechnologies, Incyte, Aprea Therapeutics, Fate Therapeutics, Mingsight, Takeda, Medisix, Loxo Oncology, Novalgen and serves on Advisory Board/Honoraria: Pharmacyclics, Janssen, AbbVie, Genentech, AstraZeneca, BMS, Adaptive Biotechnologies, Precision BioSciences, Servier, Beigene, Collectis, TG Therapeutics, ADC Therapeutics, MEI Pharma; W.G.W. reports funding from GSK/Novartis, Abbvie, Genentech, Pharmacyclics LLC, AstraZeneca/Acerta Pharma, Gilead Sciences, Juno Therapeutics, KITE Pharma, Sunesis, Miragen, Oncternal Therapeutics, Inc., Cyclacel, Loxo Oncology, Inc., Janssen, Xencor. S.A.P. has received research funding to the institution from Pharmacyclics, Janssen, AstraZeneca, TG Therapeutics, Merck, AbbVie, and Ascentage Pharma 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 Pharmacyclics, Merck, AstraZeneca, Genentech, GlaxoSmithKline, Adaptive Biotechnologies, Amgen, and AbbVie (no personal compensation). K.J.L. holds equity in Standard BioTools Inc. D.N. has stock ownership in Madrigal Pharmaceuticals. C.S. serves on

Evolutionary history of transformation from chronic lymphocytic leukemia to Richter syndrome

A full list of authors and affiliations appears at the end of the article.

Abstract

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*], recurrent 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 (DLBCL). 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.

Keywords

Richter syndrome; transformation; chronic lymphocytic leukemia; clonal evolution; cell free DNA; whole-exome sequencing; single cell RNA sequencing

Transformation to a high-grade malignancy accounts for therapeutic resistance and rapid disease progression across cancers^{1–3}. Richter syndrome (RS), an aggressive lymphoma developing in patients with chronic lymphocytic leukemia (CLL), is a striking example of transformation³. RS is associated with median overall survival of less than one year, even in the modern era³. Despite advanced genomic characterization of CLL^{4–6}, understanding of the genetic factors driving evolution of CLL to RS remains limited. This has been partly from 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}. While a subset of RS is believed to be clonally

speaker's bureau for Astra Zeneca and AbbVie. D.L. holds stock in and consults for ennov1. N.P is currently an employee at Bristol Meyers Squibb.

J.B. none; K.S. none; R.J. none; C.J. none; C.M. none; D.R. none; S.L. none; D.L. none; J.L. none; J.H. none; C.St. none; L.Z.R. none; C.Sc. none; S.Y. none; G.F. none; N.R. none; C.Lem. none; F.C. none; F.U. none; K.R. none; C.Lev. none; L.P. none; A.T.-W. none, T.H. none, R.R. none, L.E. none, B.P. none, J.M. none, S.F. none, A.J.A. none, S.H. none, S.D. none, L.L. none, P.F. none, and B.P.D. none.

CODE AVAILABILITY STATEMENT

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

unrelated based on 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 definitively delineate factors contributing to high-grade transformation, 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 cases⁹.

RESULTS

Developing an analytic framework to discover RS drivers

We assembled a discovery cohort of 53 patients with paired CLL and RS samples of diffuse large B-cell histology (DLBCL), the most common form of transformation³ (Fig. 1a, Supplementary Table 1–2). Forty-five (83%) patients received prior CLL-directed therapies, with 11 (21%) having received targeted agents. Thirty-nine (72%) patients had unmutated immunoglobulin heavy-chain variable region gene (*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, Supplementary Table 1–3).

To delineate the driver events giving rise to RS, we employed established WES analysis tools and 3 additional steps: (i) *deTiN*¹, to recover somatic mutations filtered due to tumor-in-normal contamination;⁴ (ii) an optimized tool to detect somatic copy number alterations (sCNAs); (iii) *PhylogicNDT*² to establish the clonal composition per patient sample and infer the phylogenetic tree (Fig. 1c; 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 Fig. 1d, 2–3; Supplementary Figure 1–2). We identified instances of clonal unrelatedness to the antecedent CLL (Fig. 1e, 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: (i) events strictly present in RS cells through computational isolation of the RS lineage separate from CLL^{DIV} (Fig. 2a-grey outline); and (ii) events newly acquired in RS clones (Fig. 2a-magenta outline). To

uncover drivers of transformation, we applied *MutSig2CV*¹³ and GISTIC2.0¹⁴ (Fig. 2b–c, Supplementary Table 5).

From our discovery cohort, we observed mutations in known CLL drivers (*NOTCH1*, *TP53*, *SF3B1*; Fig. 2b) and identified new candidate RS drivers (Fig. 2c, Extended Data Fig. 4a–x). These included mutations in *IRF2BP2* (n=7), which encodes an IRF2-dependent transcriptional corepressor, that is mutated in the N1 subtype of DLBCL¹⁵ and primary mediastinal B cell lymphoma¹⁶ (Extended Data Fig. 4d). Inactivating mutations in the DNA methyltransferase *DNMT3A* (8%) were previously reported as a single case in RS⁸; genetically engineered mice modelling 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,^{19–21} was observed in 3 patients (Extended Data Fig. 4f). The detected mutations in the MYC-interacting²³ splicing factor *SRSF1* (n=4) did not co-occur with mutated-*SF3B1*, consistent with mutual exclusivity of splicing factor mutations across cancers²⁴ (Extended Data Fig. 4g). *EZH2* hotspot alterations were found in 2 clonally unrelated RS cases, as in DLBCL^{15,19}, while *EZH2* frameshift was seen in one case (Extended Data Fig. 4h).

Strikingly, we detected numerous sCNAs (Fig. 2b–c, 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 5 additional patients. Recurrent focal events beyond common CLL drivers included *del*(15q13.11) [*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 CLL²⁵. 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 Fig. 2–3). The recurrent RS-specific gene mutations, sCNAs, and WGD were confirmed in our validation cohort (n=45) (Fig. 2b-right bar, Supplementary Table 5, Extended Data Fig. 4a–x) and 14 external RS genomes⁹ (Extended Data Fig. 4a–x; Supplementary Table 5).

Combined analysis of our discovery and validation cohorts provided power to further detect novel RS drivers, with *CCND3*, *TET2* and *BRAF* mutations and additional focal sCNAs emerging as significant (Extended Data Fig. 4v–x, Fig. 2d). Comparison of our 45 clonally related cases with prior large-scale CLL analyses⁶ 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, 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* DLBCLs¹⁹.

Evaluation of the relative timing of each putative driver event in 58 related RS cases revealed *ATM* mutations, *tri*(12) 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 Fig. 2–3, Supplementary Figure 1–2). 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, 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} (Figure 3b).

To track disease tempo over time, we analyzed serial samples procured in the years prior to RS from 2 patients receiving targeted agents. Pt 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). Pt 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 a heterogeneous 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 samples¹⁹ 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 8 clonally unrelated RS clustered with DLBCL (Fisher's exact test, $P = 6.75 \times 10^{-6}$), with membership across the DLBCL clusters¹⁹, 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 5

RS molecular subtypes (Extended Data Fig. 5d). Three (RS1, RS3, 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 vs 1, 3, 5]; $P < 0.005$ [RS4 vs 1, 3, 5]; Supplementary Table 7). RS2 (26.8%) was predominantly marked by *tr(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, 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 5 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 (Pt 30), with deleterious *POLE* mutation and >2,000 sSNVs (Figure 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 reported³² (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 cases⁹, 2 were clonally unrelated cases (Extended Data Fig. 7c, 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 reported⁹). Kataegis was recently reported in RS⁹, and we indeed identified this across 4 of 11 RS genomes with clustered AID-related mutations (Fig. 4d–e, Supplementary Table 7).

From the WGS samples, we observed chromothripsis as a common defining feature of *TP53*-altered RS genomes (Fig. 4f, Extended Data Fig. 8a). Chromothripsis was detected

in regions likely contributing to RS pathogenesis, including 7q21 (*CDK6*) (Pt 41), 11q13 (*CCND1*) (Pt 29) and 9p24.1 (*PD-L1/L2*) (Pt 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 fold change > 1 , adjusted $P < 0.05$) (Fig. 5a–b, 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*, *CDK2*), activation-induced cytidine deaminase (*AIDCA*), and DNA repair regulators (*BRCA1*, *XRCC2*) (Fig. 5b, Supplementary Table 9). Overexpression of several of these genes have been implicated in aneuploidy in cancer³³. 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 5 additional patients that contained clonally related RS and CLL cells within the same microenvironment (Fig 5c; 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 methods³⁴ 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 higher UMI/cell (*i.e.*, mean 9000 vs. 3193 for Pt 43, $P < 10^{-14}$, Wilcoxon) and genes/cell (mean 2909 vs. 1074, $P < 10^{-14}$) (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-velocity³⁵ 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 vs. 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 Pt 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, 4) displayed additional sCNAs, consistent with the chromothripsis seen in WES analysis (*i.e.*, sCNAs on chromosome 2 and 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 towards RS.

Pt 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 3 distinct populations. Cluster 1 displayed gene counts per cell consistent with CLL while 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 Pt 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 Pt 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). Pt 41 (RS1) highlighted an intermediate cell state clearly residing within the co-existing forward scatter (FSC)-low CLL population (Extended Data Fig. 9b,d; Supplementary Table 8). Indeed, per *CNVsingle*, the intermediate state cells showed acquisition of early RS-specific events (i.e. *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 non-invasive detection of RS events through cell-free DNA (cfDNA) (Fig. 6a). We evaluated 46 plasma samples by ultra-low pass (ULP)-WGS³⁶ collected from 24 patients within three years of RS diagnosis and through relapse (Supplementary Table 10). Samples from 17 patients were collected at the time of RS disease, including 8 at 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 vs 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 time of RS diagnosis, matching the RS WES profile, while circulating CLL remained 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). While the cfDNA profile at the start of CLL-directed therapy showed minimal sCNAs, that at 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; Extended Data Fig. 10f). Furthermore, for 4 of 4 cases from our discovery cohort in which WES was additionally performed on cfDNA, RS-specific mutations were detected (Fig. 6d, Extended Data Fig. 10g).

We queried whether RS changes could be detected in cfDNA in advance of RS diagnosis. For 2 of 7 patients whose plasma was collected 1–10 months prior to 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 Pt 5, WGD and chromothripsis (chr 6 and 16) were detected in plasma 162 days prior to diagnosis and were absent from CLL (Fig. 6e-left). WES of cfDNA (Extended Data Fig. 10g) further showed presence of RS-specific mutations. In Pt 20, cfDNA 181 days prior to 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 2 patients who had achieved a state of minimal CLL involvement following allogeneic hematopoietic stem cell transplantation (HSCT), but 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. Pt 111 intermittently had elevated FGA in plasma following HSCT, prior to 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 decreasing ratio in the preceding 1–10 months (n=7), and even lower ratio in more distant pre-diagnosis samples (>10 months; n=4). In 7 cases, FGA exceeded all values from high-risk CLL cases (n=14 samples from 5 patients) (Fig. 6g). Of the 8 patients with cfDNA available at the time of biopsy-proven RS diagnosis, we confidently discerned RS-specific lesions in 6 (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 CLL³. 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 novel 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 signalling³⁷. 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 5 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 (i.e. chromothripsis, kataegis, WGD). Near tetraploidy has been identified as an RS risk factor³⁸, 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 vulnerabilities³⁹. We demonstrate how such instability may be used to provide an earlier and non-invasive detection of RS in cfDNA, which should be further evaluated in clinical studies as a cost-effective approach for this difficult-to-diagnose aggressive cancer⁴⁰.

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 ultra-deep IGHV sequencing⁴¹. Unrelated RS has been previously associated with improved clinical outcomes, which suggests distinct disease biology^{3,10}. We now demonstrate that by exome- or genome-level analysis, clonal unrelated RS is a *de novo* DLBCL, occurring as an independent lymphoma, lacking any shared distant genetic history with the co-existing CLL. These cases tended to lack *TP53* and *NOTCH1* alterations, were enriched in M-CLL, and clustered with *de novo* DLBCL separately from clonal 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 identification, diagnosis and prognosis of RS. Our advanced molecular framework can serve as a model for studying transformed cancers.

ONLINE METHODS

Patient sample collection and processing

CLL, RS, and normal germline (i.e. non-tumor) 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 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 Institutional Review Boards of the respective institutions. Patient and sample characteristics are provided (Supplementary Tables 1–2). Sex was self-report.

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). For FFPE specimens, samples from each submitting center were reviewed for >50% purity prior to sequencing.

CLL samples.—CLL samples were obtained from PBMCs. Samples with higher CLL purity (WBC >25 × 10³/microliter or ALC >20 × 10³/microliter) 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 WBC <25,000/uL or ALC <20,000/uL underwent CD19 selection (RosetteSep Human B-cell enrichment, Stem Cell Technologies) or as previously described⁴ or FACS sorting to enrich for CD5+CD19+ populations.

Germline samples.—Sources of non-tumor germline DNA included saliva (Oragene Discover [ORG500 or ORG600] kit, DNA Genotek), remission bone marrow⁴ 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).

Genomic DNA sequencing

Whole-exome sequencing (WES)—A total of 143 samples were processed and sequenced at the Broad Institute (Cambridge, MA). For these fresh blood and bone marrow samples and cryopreserved suspension cells, genomic DNA and RNA was extracted per 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 (Fluidigm, San Francisco, CA). 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 prior to hybridization. Hybridization and capture were performed using the relevant components of Illumina's Rapid Capture Enrichment Kit, with a 37Mb 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 0.1N 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 flowcells. Flowcells were sequenced utilizing Sequencing-by-Synthesis chemistry for HiSeq 4000 flowcells. The flowcells 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 8 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 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, QPCR based quantification and multiplexing (per protocol version 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 HiSeq2000 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 described⁶.

Whole-genome sequencing (WGS)

Preparation of libraries for cluster amplification and sequencing (PCR-Free).

—350ng of genomic DNA in 50µL of solution was processed by fragmentation through acoustic shearing (Covaris focused ultrasonicator), targeting 385bp fragments, and additional size selection was performed using a SPRI 80 cleanup. Library preparation (Hyper Prep without amplification module, KAPA Biosystems, #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 1.7nM, and then pooled into 24-plexes.

Preparation of libraries for cluster amplification and sequencing (PCR-Plus).

—An aliquot of genomic DNA (100ng in 50µL) 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 Hyper Prep with Library Amplification Primer Mix, product KK8504), and with palindromic forked adapters using unique 8-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 2.2nM 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 flowcell cell using the Hamilton Starlet Liquid Handling system. Cluster amplification and sequencing occurred on NovaSeq 6000 Instruments

utilizing sequencing-by-synthesis kits to produce 151bp 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 flowcells. Flowcells were sequenced on v2 Sequencing-by-Synthesis chemistry for HiSeqX flowcells. The flowcells were then analyzed using RTA v.2.7.3 or later. Each pool of ultra-low pass whole genome libraries was run on one lane using paired 151bp runs.

Sequence data processing and analyses

WES/WGS alignment and quality control.—Sequencing was conducted using standard methods (Supplementary Note)^{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 (version b37, using BWA-MEM [version 0.7.15-r1140]) provided by the Picard and Genome Analysis Toolkit (GATK) developed at the Broad Institute⁴³, 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 pipeline⁴² and included *MuTect* for calling somatic single nucleotide variants (sSNVs), *Strelka2*⁴⁴ for calling small insertions and deletions (indels), *deTiN*⁴¹ for estimating tumor-in-normal (TiN) contamination, *ContEst* for estimating cross-patient contamination, *AllelicCapSeg* for calling allelic copy number variants, and *ABSOLUTE* for estimating tumor purity, ploidy, cancer cell fractions, 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 described¹⁹. 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 normalization⁴⁵ on samples that showed consistent behavior, thus excluding artifactual copy number targets.

WGS Analysis.—WGS analysis was performed as previously described⁴⁶. Due to the large amount of computational resources required to efficiently process cancer whole

genomes, 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: (i) balanced translocations, defined as those with breakpoints on reverse strands within 1 kb of each other; (ii) inversions supported on both ends; (iii) 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 (v90)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 (3 failed due to data quality and realignment issues). The standard pipeline as previously described⁴⁶ 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 kilobase scale. SV and phylogenetic analysis were completed for 12/14 samples.

Identification of regions of kataegis and chromothripsis.—In the WGS, kataegis regions were defined by genomic regions with at least 6 mutations within 2 standard deviations of the median chromosomal intermutational distance, as previously described⁵². 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*^{1,32} 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 novel emerging clone first detected in the RS sample and absent in a preceding CLL sample. In the

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 cancer cell fraction (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 allelic 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 further improve power to detect known variants, we ran *MutSig2CV* on a restricted set of hypotheses through utilizing list of CLL⁶ 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>), comprised of the CalculateTargetCoverage, NormalizeSomaticReadCounts, and Circular Binary Segmentation (*CBS*) algorithms⁵³ 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*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 (v3.2)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 described⁶. (Supplementary Table 4).

Consensus clustering of genetic alterations

Generation of gene sample matrix—All significantly mutated genes (MutSig2CV, $Q < 0.1$ and frequency ≥ 4 cases), significant regions of sCNAs (GISTIC2.0, $Q < 0.1$ and frequency ≥ 4 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: non-synonymous 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 7 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 a 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 Exact test for all 5 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 8 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, deletion) that affects that gene was calculated using Fisher's exact test, and then false discovery rate was calculated using the Benjamini-Hochberg method.

Non-negative matrix factorization consensus clustering—To robustly identify clusters of tumors with shared genetic features, we applied a non-negative matrix consensus clustering algorithm⁵⁴ with slight modifications. Briefly, we passed the gene-by-sample matrix to the NMF consensus clustering algorithm (testing number of clusters $k=2$ to 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, Supplementary Note 2).

Bulk RNA sequencing and data analyses

High-quality RNA from CLL/RS pairs was extracted, as previously described⁴. Total RNA was quantified using the Quant-iTTM RiboGreen[®] RNA Assay Kit and normalized to 5 ng/ μ L. Following plating, 2 μ L of ERCC controls (using a 1:1000 dilution) were spiked into each sample. An aliquot of 200ng for each sample was transferred into library preparation which uses an automated variant of the Illumina TruSeqTM 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 400bp 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/ μ L, 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 prior to sequencing. Flowcell 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 101bp paired-end with an eight-base index barcode read. Data was analyzed using the Broad Picard Pipeline, which includes de-multiplexing and data aggregation.

Bulk RNA-sequencing of validation cohort

RNA was extracted with Macherey Nagel RNA extraction kit (Macherey-Nagel, Düren, Germany). 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, San Diego, CA), according to manufacturer's instructions. The final cDNA libraries were checked for quality and quantified using capillary electrophoresis prior to sequencing with HiSeq 4000 sequencing using

RNA-seq data analyses.—RNA-seq reads were aligned to the human reference genome hg19 using STAR (v2.4.0.1)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 $|\log_2FC| > 1$ and adjusted p-value < 0.25 as a cutoff. To ensure robustness of the analysis, for the 5 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 ($|\log_2FC| > 1$ and adjusted $P < 0.1$) in the leave-one-out analysis. This was used as input for pre-ranked *GSEA* on HALLMARK pathways (1,000 permutations, weighted enrichment statistics, MsigDB v7.4)⁵⁷.

RNA clustering of RS samples and integration with genetic subtypes—Gene counts were pre-processed with *ComBat-seq* (v3.42.0)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 > 1000 . TPMs were then \log_2 transformed and top

genes by variance (z -score of variance > 1) were z -score transformed for downstream analysis. Consensus clustering⁵⁹ using the hierarchical clustering (complete linkage) with spearman distance was used to identify the optimal number of clusters (observed as 5 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 a Fisher's exact test. Supervised analysis for differentially expressed genes for each genomically-identified cluster was performed using *limma-voom*(v3.50.3)⁵⁶ as a one-vs-other comparison. Pathway analysis of each genomically-identified cluster was performed using Preranked-GSEA⁵⁷ with the MsigDB Hallmark (v7.4) 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 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.

Sequencing.—Five to ten thousand 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 43, 4 and 10). 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* (v2.1.1 for Patient 41; v2.0.0 for Patient 19, and v3.0.2 for Patients 4, 10 and 43). Background, or ambient, RNA was removed using *CellBender* with the exception of Patient 41. Data from each patient was analyzed using *Seurat* (v3.1.4)⁶⁰. QC filtering was applied to remove cells with fewer than 500 UMIs, $>50,000$ UMIs or more than 10% mitochondrial reads. Potential doublets were detected using *DoubletFinder* (v2.0.2)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 methods⁶⁰.

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*, *CD79B*) and the absence of T/NK/Myeloid markers (*CD3*, *CD4*, *CD8*, *CD56*, *CD14*, *CD16*, *CD33*). Clustree (v0.4.2) was used to identify stable clusters prior to downstream 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 novel 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 non-zero read counts in less 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*(v3.50.3) in paired mode. The ranked gene list sort the *t* statistic from the DE analysis for RS clusters in each scRNAseq sample was submitted to pre-ranked *GSEA* to analyze the HALLMARK pathways (1,000 permutations, weighted enrichment statistics, MsigDB v7.4)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 (RF) Analysis.—We used an RF approach to differentiate between CLL and RS cells in the single cell data. Data was 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 ($n_estimators = 1000$, *sklearn* v1.0.1) on samples LNs from Pts 10 and 43 and predicted on cells from Pts 41 (LN or peripheral blood [PB] samples) and 18 (bone marrow) whose cell labels were determined by FACS sorting described earlier. We ran the RF 20 times and obtained a mean $\pm \sigma$ 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 Pt 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 3σ above the mean.

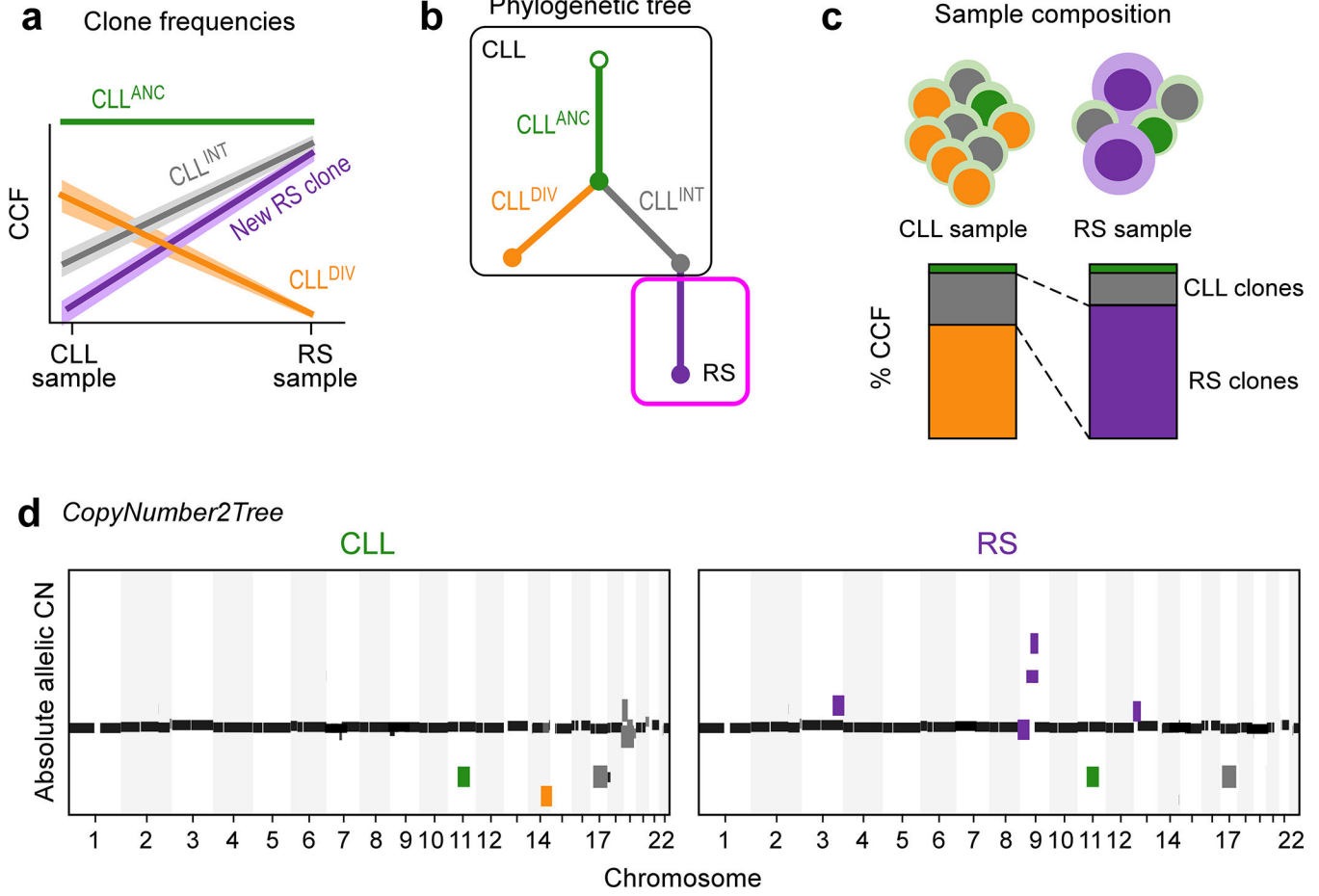
Clinical endpoint analysis and statistical analysis—Data analyses were carried out using GraphPad Prism version 9 and R software version 4. To compare RS drivers identified with previously reported CLL (n=1063)⁶ and DLBCL (n=304)¹⁹ and (n=574)¹⁵ datasets, a 2-sided exact binomial test was performed with Benjamini-Hochberg multiple hypothesis testing correction. To obtain the frequency of RS events in DLBCL cohorts prior to this comparison, we called RS sCNAs in 304 DLBCLs¹⁹ and the 443 primary DLBCLs¹⁵ 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 (OS) 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.

Cell free DNA (cfDNA) analysis—After sequencing, plasma cfDNA samples were processed and analyzed as reported³⁶. 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 3 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 fraction genome altered (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 (i.e. 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

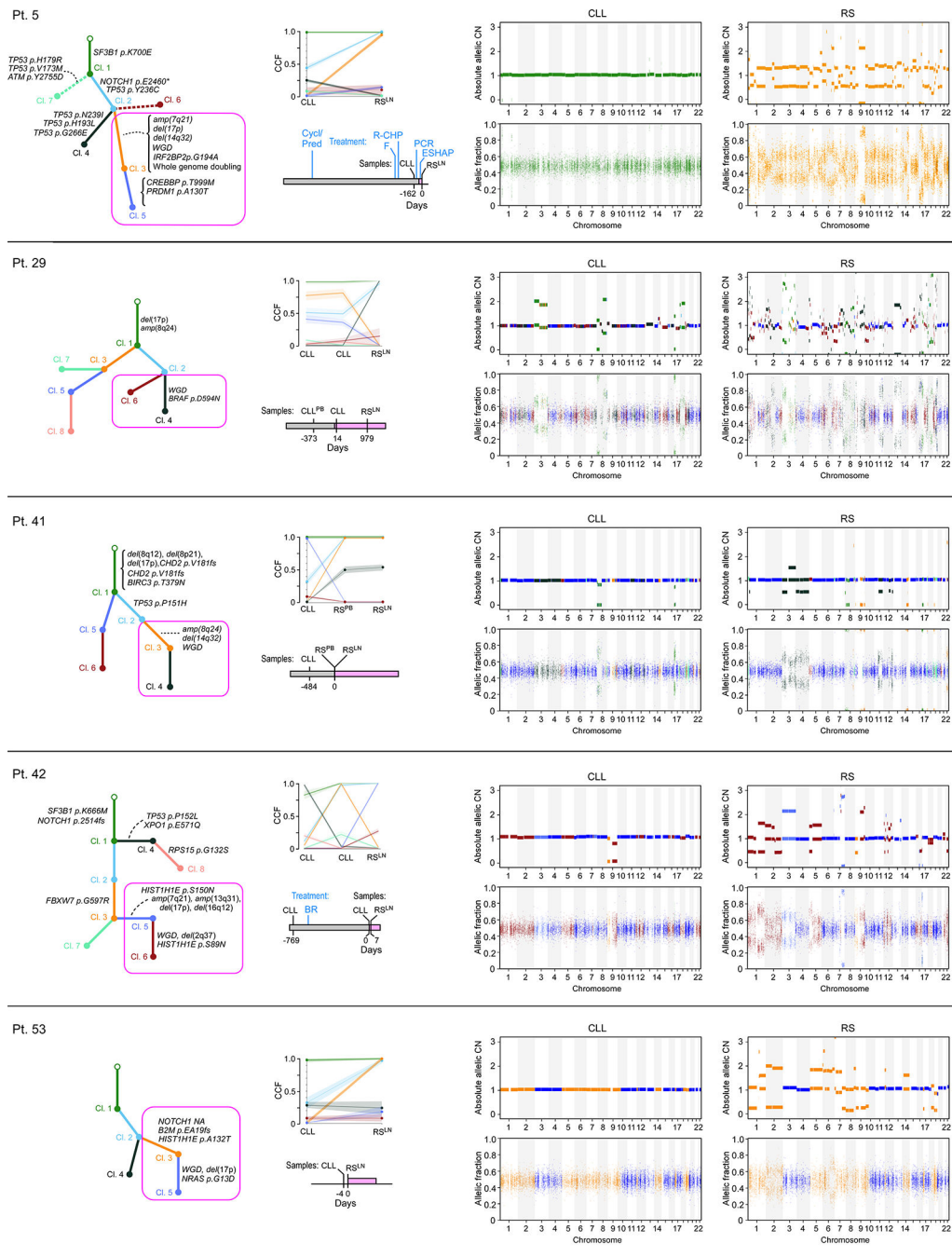
WES, RNA-seq, WGS, and scRNAseq data will be deposited in dbgap (accession number phs002458.v2.p1) at the time of publication.

Extended Data



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*.

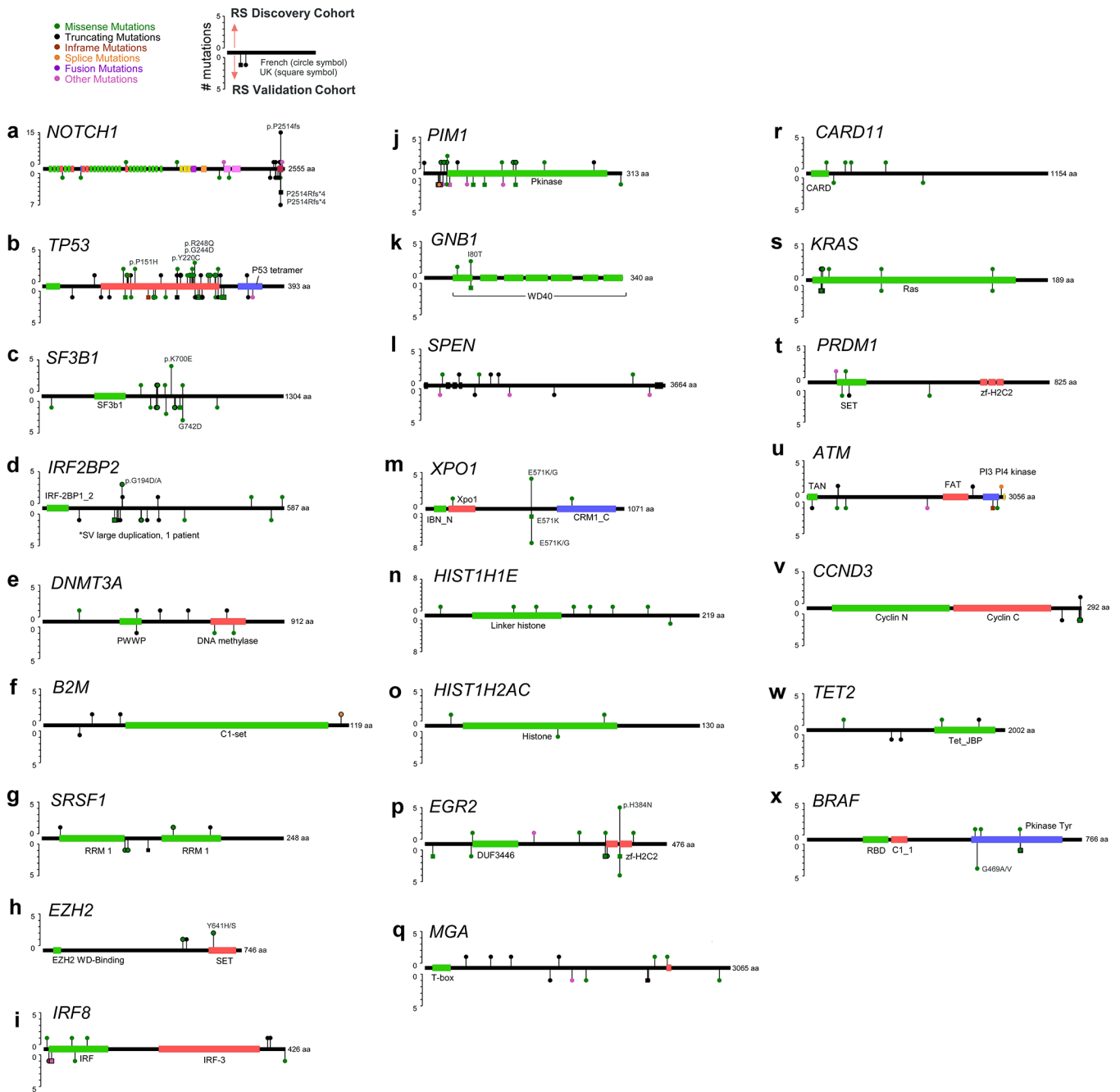


Extended Data Fig. 2. Phylogenetic reconstruction and somatic genomic alterations.

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

Extended Data Fig. 3. Phylogenetic reconstruction and somatic genomic alterations.

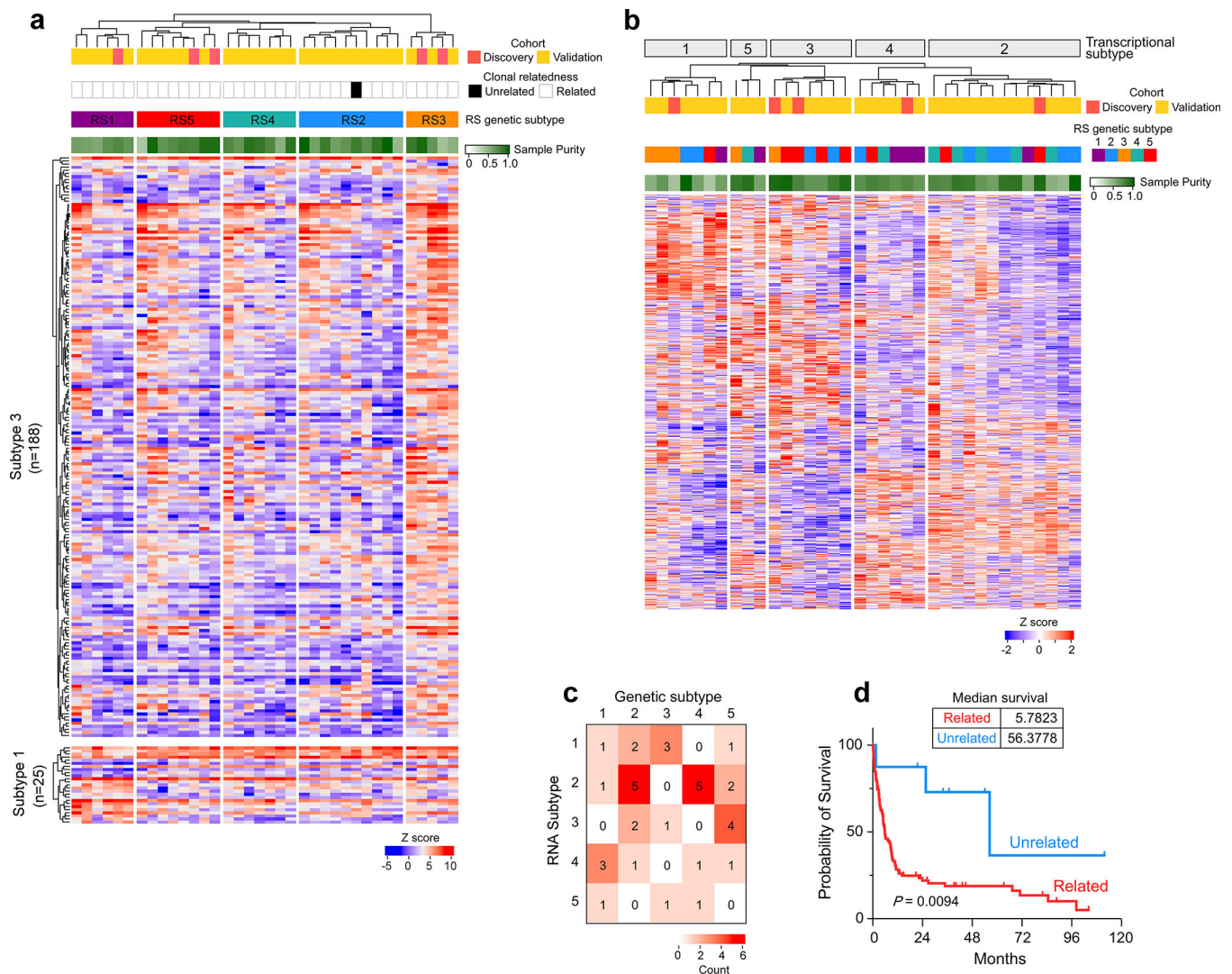
For each of the patient trios with WES data, the left panel shows the phylogenetic tree tracing the transformation history from CLL to RS. The magenta frame denotes the Richter clones. The middle top panel represents the subclonal composition inferred after clustering alterations with similar cancer 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; idelalisib; 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.



Extended Data Fig. 4. Putative RS driver genes.

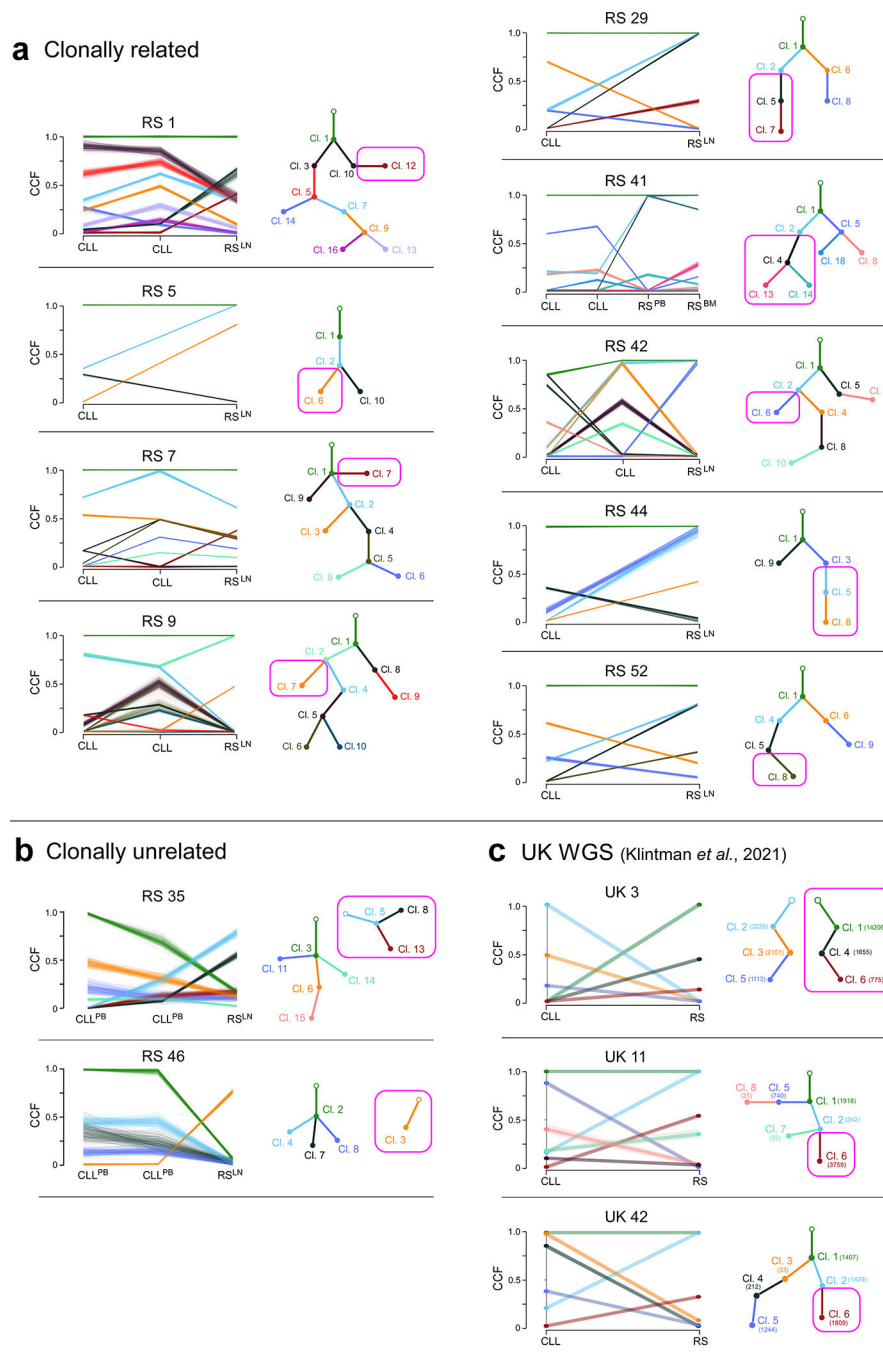
a-x, individual protein mutation maps for selected putative Richter drivers, showing gene mutation subtype (for example, missense), position and evidence of mutational hotspots. Panels were generated by using the cBioPortal for Cancer Genomics tool.

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



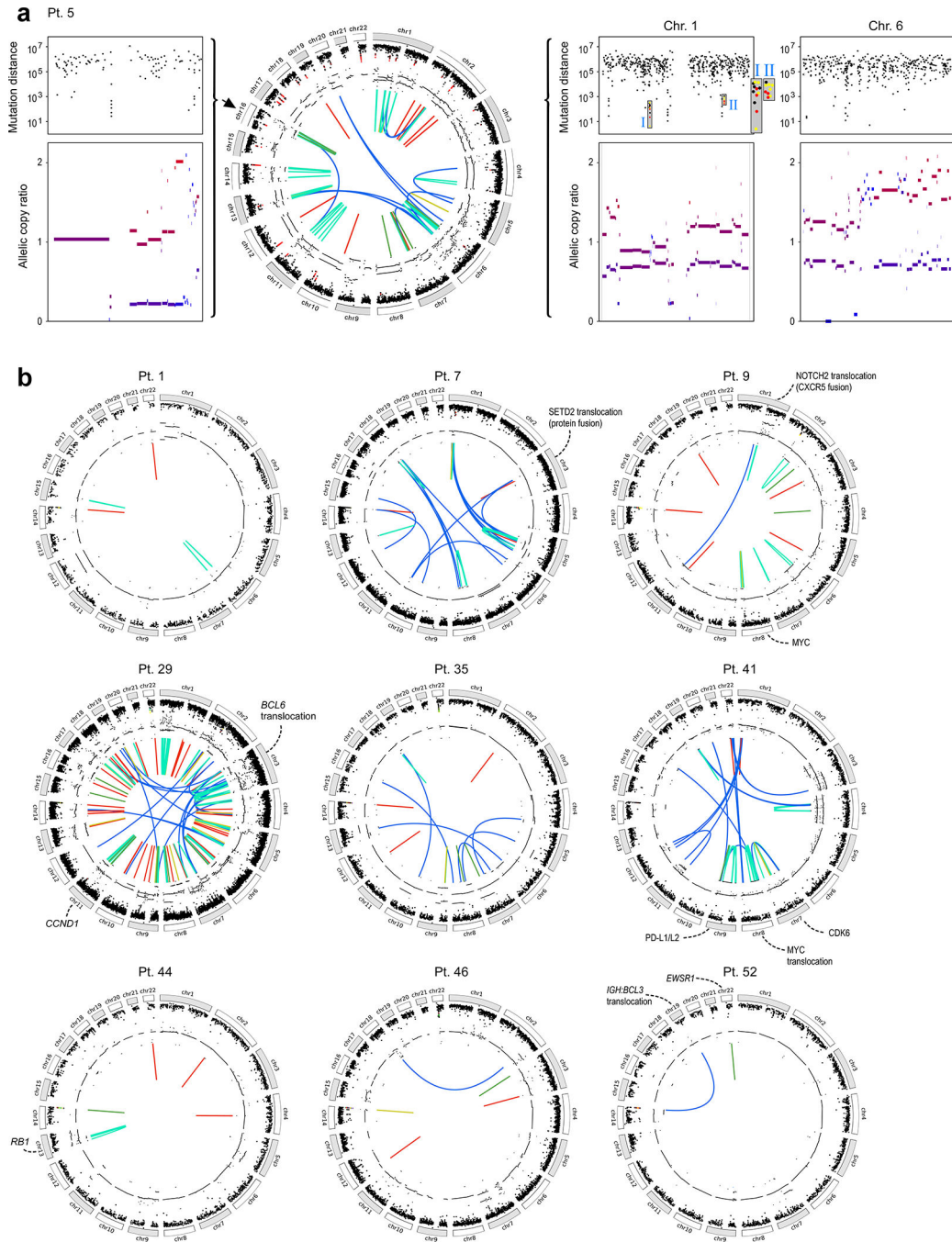
Extended Data Fig. 6. Transcriptome supports distinct RS molecular subtypes.

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).



Extended Data Fig. 7. Phylogenetic trees showing CLL and RS clones from WGS of paired samples.

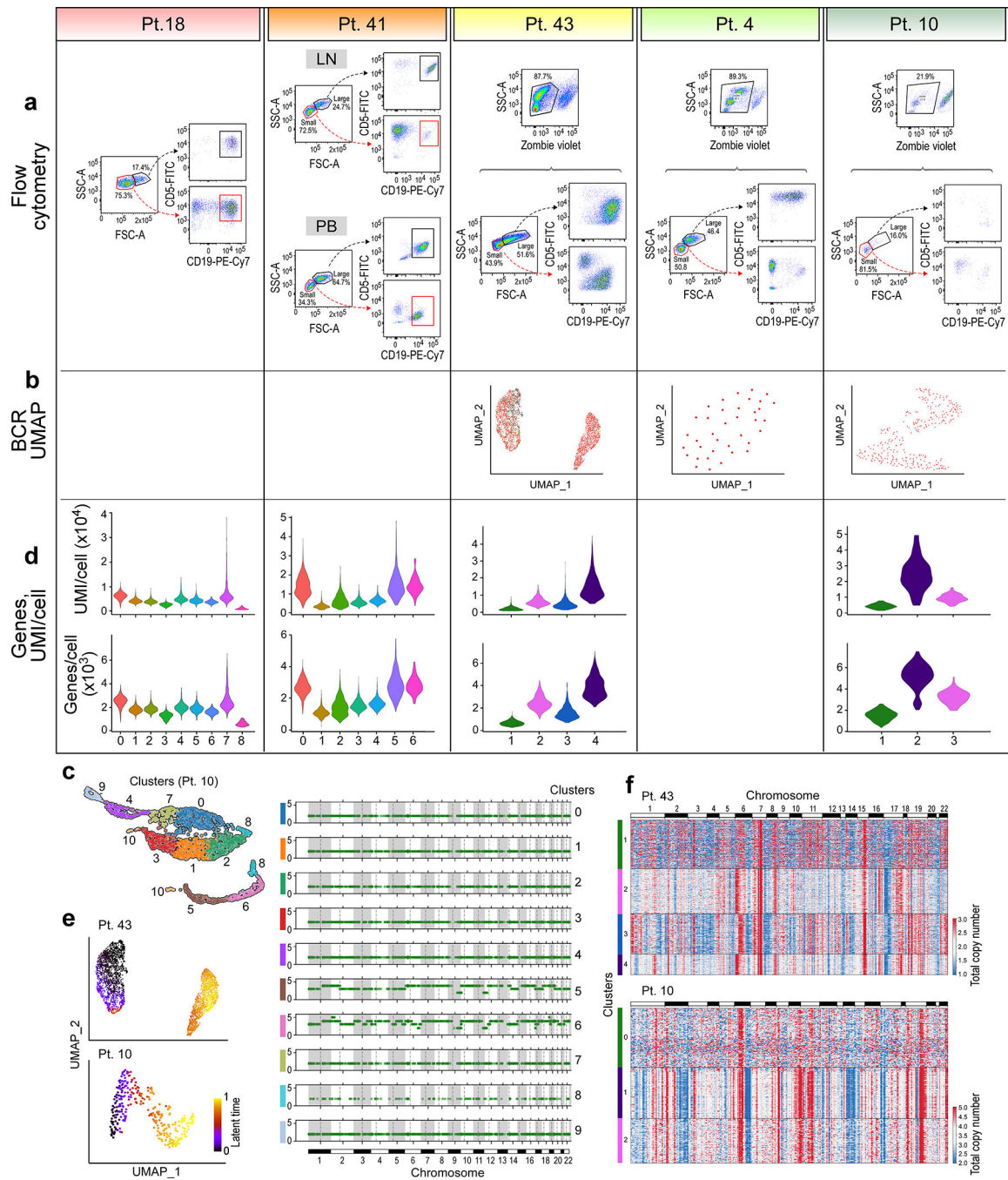
a, Phylogenetic tree and CCF plot for 9 patients based on WGS data showing clonally related RS (magenta box). **b**, Phylogenetic tree and CCF plot for 2 patients based on WGS demonstrating clonally unrelated RS **c**, Representative phylogenetic trees and CCF plot for 3 patients from UK cohort⁹ based on WGS.



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 katagis 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).



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

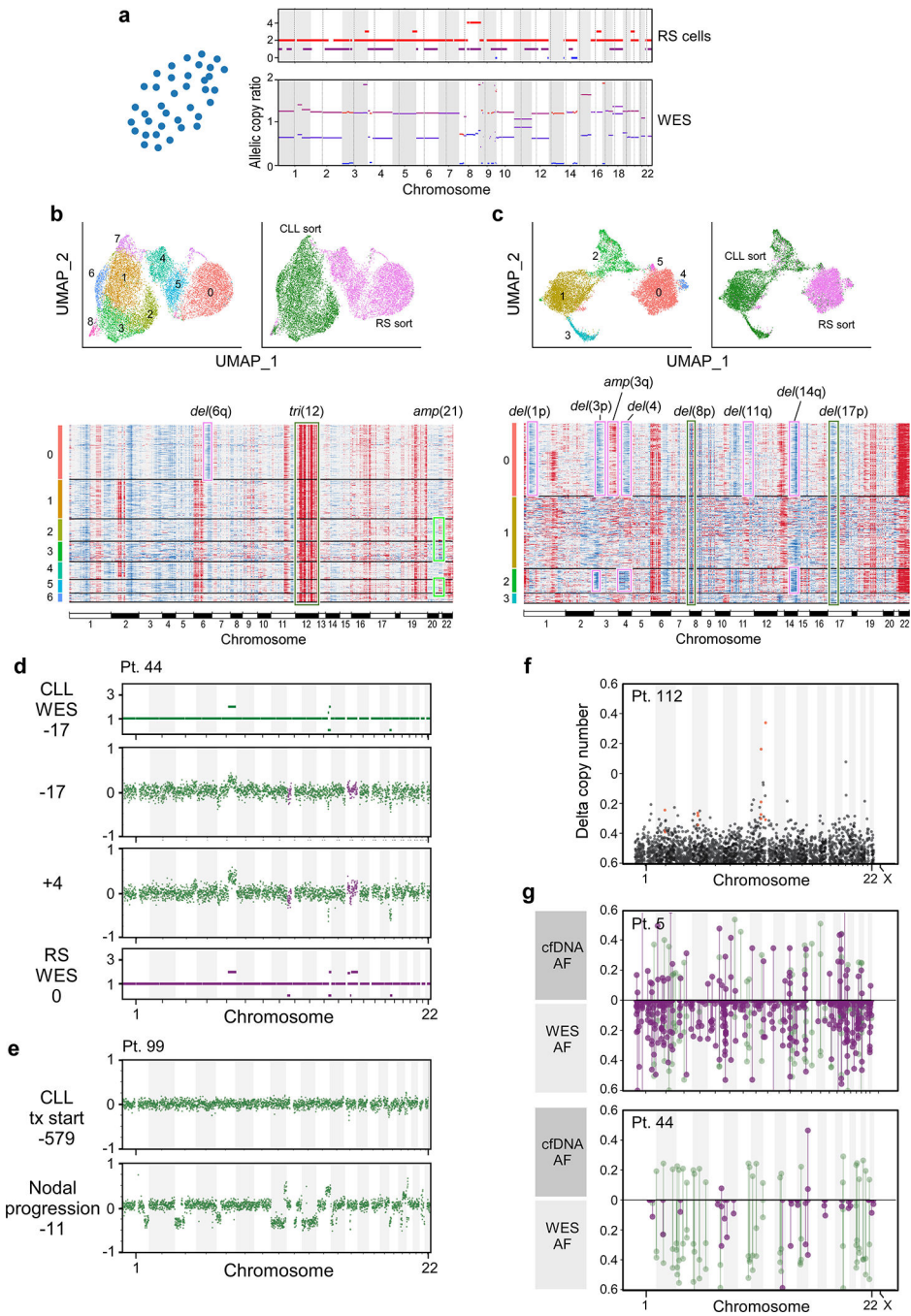
plots below demonstrate CLL and RS cells were included in sorted population. **b**, B-cell receptor (BCR) clonotypes plotted for RS and CLL clusters on UMAP visualization. **c**, Representative example from patient 10 showing CNVsingle identifies malignant B cell clusters (5 and 6) separate from immune 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 ($P < 2.2 \times 10^{-16}$ 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)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



Extended Data Fig. 10. Single-cell transcriptome and copy number analysis of RS patients. 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)

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Erin M Parry, MD PhD^{1,2,3,*}, Ignaty Leshchiner, PhD^{2,4,*}, Romain Guièze, MD PhD^{1,2,5,6,*}, Connor Johnson², Eugen Tausch, MD⁷, Sameer A. Parikh, MD⁸, Camilla Lemvigh^{1,9}, Julien Broséus, MD PhD^{10,11}, Sébastien Hergalant¹⁰, Conor Messer², Filippo Utro, PhD¹², Chaya Levovitz, MD PhD¹², Kahn Rhrissorrakrai, PhD¹², Liang Li², Daniel Rosebrock², Shanye Yin, PhD^{1,3}, Stephanie Deng¹, Kara Slowik², Raquel Jacobs², Teddy Huang^{1,13}, Shuqiang Li, PhD^{1,2,13}, Geoff Fell¹⁴, Robert Redd¹⁴, Ziao Lin², Binyamin A. Knisbacher, PhD², Dimitri Livitz², Christof Schneider⁷, Neil Ruthen^{1,13}, Liudmila Elagina², Amaro Taylor-Weiner, PhD², Bria Persaud², Aina Martinez², Stacey M Fernandes¹, Noelia Purroy, MD PhD^{1,2,3}, Annabelle J Anandappa, MD^{1,3}, Jialin Ma², Julian Hess², Laura Z Rassenti, PhD¹⁵, Thomas J Kipps, MD PhD¹⁵, Nitin Jain, MD¹⁶, William Wierda, MD PhD¹⁶, Florence Cymbalista, MD PhD¹⁷, Pierre Feugier, MD PhD^{10,18}, Neil E. Kay, MD⁸, Kenneth J. Livak, PhD^{1,13}, Brian P. Danysh, PhD², Chip Stewart, PhD², Donna Neuberg, ScD¹⁴, Matthew S. Davids, MD^{1,3}, Jennifer R. Brown, MD PhD^{1,3}, Laxmi Parida, PhD¹², Stephan Stilgenbauer, MD^{7,**}, Gad Getz, PhD^{2,3,19,**}, Catherine J. Wu, MD^{1,2,3,20,**}

Affiliations

¹Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA.

²Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.

³Harvard Medical School, Boston, MA 02215, USA.

⁴Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA.

⁵CHU de Clermont-Ferrand, F-63000, Clermont-Ferrand, France.

⁶Université Clermont Auvergne, EA7453 CHELTER, F-63000, Clermont-Ferrand, France.

⁷Division of CLL, Dept. of Internal Medicine III, Ulm University, Ulm, Germany

⁸Division of Hematology, Mayo Clinic, Rochester, MN, USA.

⁹Department of Health Technology, Technical University of Denmark, 2800 Kongens Lyngby, Denmark

¹⁰Inserm UMRS1256 Nutrition-Génétique et Exposition aux Risques Environnementaux (N-GERE), Université de Lorraine, Nancy, France.

¹¹Université de Lorraine, CHRU-Nancy, service d'hématologie biologique, pôle laboratoires, Nancy, France.

¹²IBM Research, Yorktown Heights, New York, NY, USA.

¹³Translational Immunogenomics Lab, Dana-Farber Cancer Institute, Boston, MA, USA.

¹⁴Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA.

¹⁵Moore's Cancer Center, Medicine, University of California, San Diego, La Jolla, CA, USA.

¹⁶Department of Leukemia, The University of Texas MD Anderson Cancer Center, TX, USA.

¹⁷Laboratoire d'hématologie, Hôpital Avicenne –AP-HP, INSERM U978-Université Sorbonne Paris Nord, Bobigny, France.

¹⁸Université de Lorraine, CHRU Nancy, service d'hématologie clinique, Nancy, France

¹⁹Cancer Center and Department of Pathology, Massachusetts General Hospital, Boston, MA 02114, USA.

²⁰Department of Medicine, Brigham and Women's Hospital, Boston, MA 02215, USA.

ACKNOWLEDGEMENTS

We thank Cynthia Hahn, Elisa Ten Hacken, Wandu Zhang, Satyen Gohil, and Lilian Werner for helpful discussions. We thank Candace Patterson, Sam Pollock, Oriol Olive, Conner J. Shaughnessy and Hayley Lyon for assistance in data collection and organization and Saveliy Belkin and Chet Birger for assistance in data storage. We thank Tim Lehmborg, Mikaela McDonough, Christina Galler and Mary 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 (L.L., G.G., 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), NIH/NCI K08 CA270085 (E.M.P), Fishman Family Fund (R.G., C.L.), EMBO fellowship ALTF 14-2018 (B.A.K), NCI Research Specialist Award R50CA251956 (S.L.), NIH/NCI R21CA267527-01 (S.Y.). Additional research support was provided by NIH R01 CA 213442 (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 St. St.)

DATA AVAILABILITY STATEMENT

WES, RNA-seq, WGS, scRNAseq and cfDNA data are available at dbgap (<https://www.ncbi.nlm.nih.gov/gap/>) using accession number 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 (<https://ega-archive.org/datasets/EGAD00001007922>).

The following figures have associated raw data : Fig. 2 b–h, Fig 4 a,f, Fig. 5 a,b,d,e, Fig 6h, Extended Data 4 a–x, Extended Data 5 a–c, Extended Data Fig. 6a, Extended Data Fig. 8, Extended Data Fig. 9, Extended Data Fig. 10.

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). [PubMed: 31228622]
2. Volta AD, 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). [PubMed: 30557524]
3. Parikh SA, Kay NE & Shanafelt TD How we treat Richter syndrome. *Blood* 123, 1647–1657 (2014). [PubMed: 24421328]
4. Landau DA, et al. Mutations driving CLL and their evolution in progression and relapse. *Nature* 526, 525–530 (2015). [PubMed: 26466571]
5. Puente XS, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* 526, 519–524 (2015). [PubMed: 26200345]
6. Knisbacher BA, et al. Molecular map of chronic lymphocytic leukemia and its impact on outcome. *Nat Genet* (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). [PubMed: 24004666]
8. Fabbri G, et al. Genetic lesions associated with chronic lymphocytic leukemia transformation to Richter syndrome. *J Exp Med* 210, 2273–2288 (2013). [PubMed: 24127483]
9. Klintman J, et al. Genomic and transcriptomic correlates of Richter transformation in chronic lymphocytic leukemia. *Blood* 137, 2800–2816 (2021). [PubMed: 33206936]
10. Rossi D, et al. The genetics of Richter syndrome reveals disease heterogeneity and predicts survival after transformation. *Blood* 117, 3391–3401 (2011). [PubMed: 21266718]
11. Taylor-Weiner A, et al. DeTiN: overcoming tumor-in-normal contamination. *Nat Methods* 15, 531–534 (2018). [PubMed: 29941871]
12. Leshchiner I, et al. Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. *bioRxiv*, 508127 (2018).
13. Lawrence MS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218 (2013). [PubMed: 23770567]
14. Mermel CH, 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). [PubMed: 21527027]
15. Schmitz R, et al. Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N Engl J Med* 378, 1396–1407 (2018). [PubMed: 29641966]
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). [PubMed: 31697821]
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). [PubMed: 34686499]
18. Mahajan VS, et al. B1a and B2 cells are characterized by distinct CpG modification states at DNMT3A-maintained enhancers. *Nat Commun* 12, 2208 (2021). [PubMed: 33850140]
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). [PubMed: 29713087]

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). [PubMed: 22137796]
21. Sade-Feldman M, et al. Resistance to checkpoint blockade therapy through inactivation of antigen presentation. *Nat Commun* 8, 1136 (2017). [PubMed: 29070816]
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 Discov* 7, 1420–1435 (2017). [PubMed: 29025772]
23. Singh K, et al. c-MYC regulates mRNA translation efficiency and start-site selection in lymphoma. *J Exp Med* 216, 1509–1524 (2019). [PubMed: 31142587]
24. Lee SC, et al. Synthetic Lethal and Convergent Biological Effects of Cancer-Associated Spliceosomal Gene Mutations. *Cancer Cell* 34, 225–241 e228 (2018). [PubMed: 30107174]
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). [PubMed: 31467127]
26. Anderson MA, et al. Clinicopathological features and outcomes of progression of CLL on the BCL2 inhibitor venetoclax. *Blood* 129, 3362–3370 (2017). [PubMed: 28473407]
27. Jain P, et al. Long-term outcomes for patients with chronic lymphocytic leukemia who discontinue ibrutinib. *Cancer* 123, 2268–2273 (2017). [PubMed: 28171709]
28. Maddocks KJ, et al. Etiology of Ibrutinib Therapy Discontinuation and Outcomes in Patients With Chronic Lymphocytic Leukemia. *JAMA Oncol* 1, 80–87 (2015). [PubMed: 26182309]
29. Burger JA, et al. Clonal evolution in patients with chronic lymphocytic leukaemia developing resistance to BTK inhibition. *Nat Commun* 7, 11589 (2016). [PubMed: 27199251]
30. Guieze R, et al. Mitochondrial Reprogramming Underlies Resistance to BCL-2 Inhibition in Lymphoid Malignancies. *Cancer Cell* 36, 369–384 e313 (2019). [PubMed: 31543463]
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). [PubMed: 26638776]
32. Gruber M, et al. Growth dynamics in naturally progressing chronic lymphocytic leukaemia. *Nature* 570, 474–479 (2019). [PubMed: 31142838]
33. Zhang N, et al. Overexpression of Separase induces aneuploidy and mammary tumorigenesis. *Proc Natl Acad Sci U S A* 105, 13033–13038 (2008). [PubMed: 18728194]
34. Patel AP, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, 1396–1401 (2014). [PubMed: 24925914]
35. Bergen V, Lange M, Peidli S, Wolf FA & Theis FJ Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* 38, 1408–1414 (2020). [PubMed: 32747759]
36. Adalsteinsson VA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun* 8, 1324 (2017). [PubMed: 29109393]
37. Nadeu F, et al. Detection of early seeding of Richter transformation in chronic lymphocytic leukemia. *Nat Med* 28, 1662–1671 (2022). [PubMed: 35953718]
38. Miller CR, et al. Near-tetraploidy is associated with Richter transformation in chronic lymphocytic leukemia patients receiving ibrutinib. *Blood Adv* 1, 1584–1588 (2017). [PubMed: 29296799]
39. Quinton RJ, et al. Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. *Nature* 590, 492–497 (2021). [PubMed: 33505027]
40. Soilleux EJ, 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). [PubMed: 27345622]
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* (2022).
42. Parikh AR, et al. Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers. *Nat Med* 25, 1415–1421 (2019). [PubMed: 31501609]

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). [PubMed: 20644199]
44. Kim S, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* 15, 591–594 (2018). [PubMed: 30013048]
45. Gao GF, et al. Tangent normalization for somatic copy-number inference in cancer genome analysis. *Bioinformatics* 38, 4677–4686 (2022). [PubMed: 36040167]
46. Consortium ITP-CAo.W.G. Pan-cancer analysis of whole genomes. *Nature* 578, 82–93 (2020). [PubMed: 32025007]
47. Chen X, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 32, 1220–1222 (2016). [PubMed: 26647377]
48. Wala JA, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res* 28, 581–591 (2018). [PubMed: 29535149]
49. Bass AJ, et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nat Genet* 43, 964–968 (2011). [PubMed: 21892161]
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). [PubMed: 23124520]
51. Sondka Z, et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* 18, 696–705 (2018). [PubMed: 30293088]
52. Alexandrov LB, et al. The repertoire of mutational signatures in human cancer. *Nature* 578, 94–101 (2020). [PubMed: 32025018]
53. Olshen AB, Venkatraman ES, Lucito R & Wigler M Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5, 557–572 (2004). [PubMed: 15475419]
54. Brunet JP, Tamayo P, Golub TR & Mesirov JP Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* 101, 4164–4169 (2004). [PubMed: 15016911]
55. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21 (2013). [PubMed: 23104886]
56. Law CW, Chen Y, Shi W & Smyth GK voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 15, R29 (2014). [PubMed: 24485249]
57. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102, 15545–15550 (2005). [PubMed: 16199517]
58. Zhang Y, Parmigiani G & Johnson WE ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* 2, lqaa078 (2020). [PubMed: 33015620]
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. *Machine Learning* 52, 91–118 (2003).
60. Stuart T, et al. Comprehensive Integration of Single-Cell Data. *Cell* 177, 1888–1902 e1821 (2019). [PubMed: 31178118]
61. McGinnis CS, Murrow LM & Gartner ZJ DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* 8, 329–337 e324 (2019). [PubMed: 30954475]

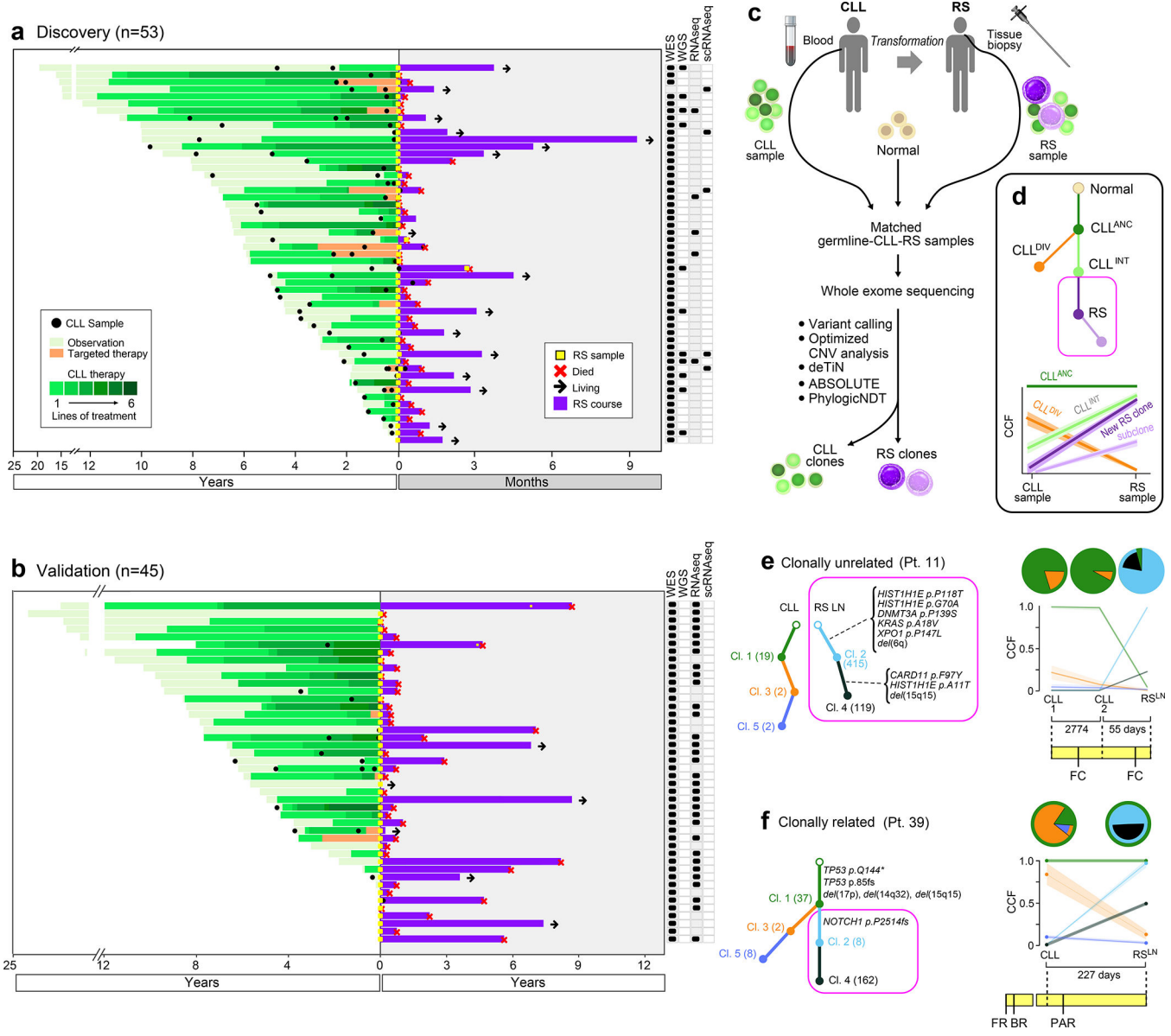


Fig. 1. Developing an analytic framework for detecting Richter Syndrome (RS)-specific clones. **a**, Disease course of 53 RS patients from CLL diagnosis in relationship to lines of therapy and sample collection. **b**, Disease course of 44 of 45 RS validation cohort patients from CLL diagnosis in relationship to lines of therapy and sample collection (1 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.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

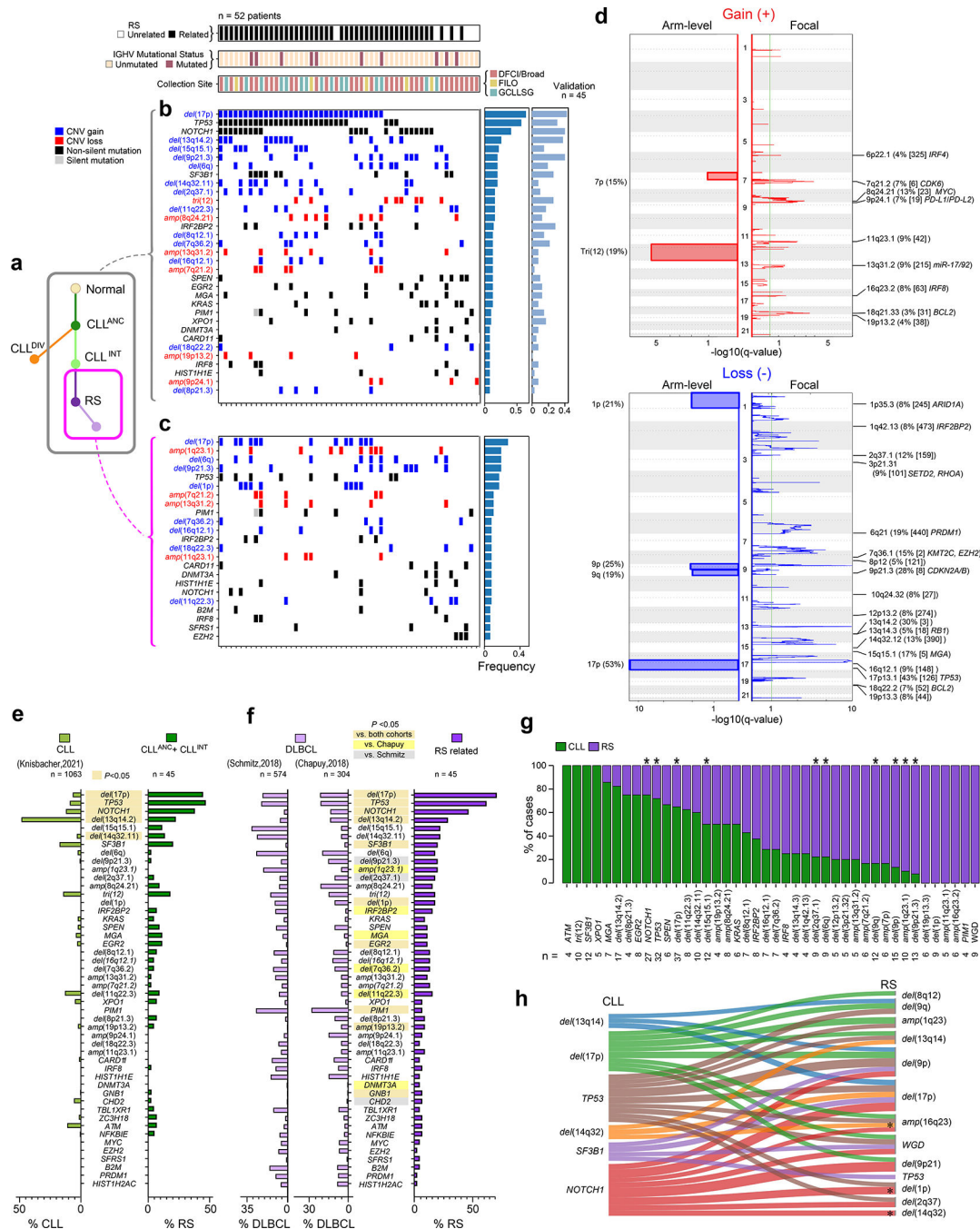


Fig. 2. The landscape of putative driver mutations in RS.

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) that underwent WES, ranked by frequency (right) for both **(b)** RS history, alterations detected in RS cells and **(c)** RS clones, alterations acquired at transformation. Samples were annotated for sequencing site (DFCI/Broad, red; German

CLL Study Group (GCLLSG), 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 Figure 5. **e**, Frequencies of somatic alterations in CLL clones from related RS cases (n=45, dark green bars) compared to CLL driver frequencies⁶ using 2 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^{15,16} using 2 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 4 patients are shown) (Supplementary Table 6). * denotes $P < 0.05$ (McNemar test, one-sided). **h**, Sankey plot showing trajectories from CLL driver to acquired RS driver. Only driver pairs with at least 4 co-occurrences across the cohort are displayed and tested for statistical significance (Supplementary Table 6). * denotes $P < 0.05$ (Fisher's exact test, two-sided) and $Q < 0.4$.

blood by white blood cell count (1000 cells/microliter) (top) and relative abundance of RS in bottom plot (by PET/CT 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)

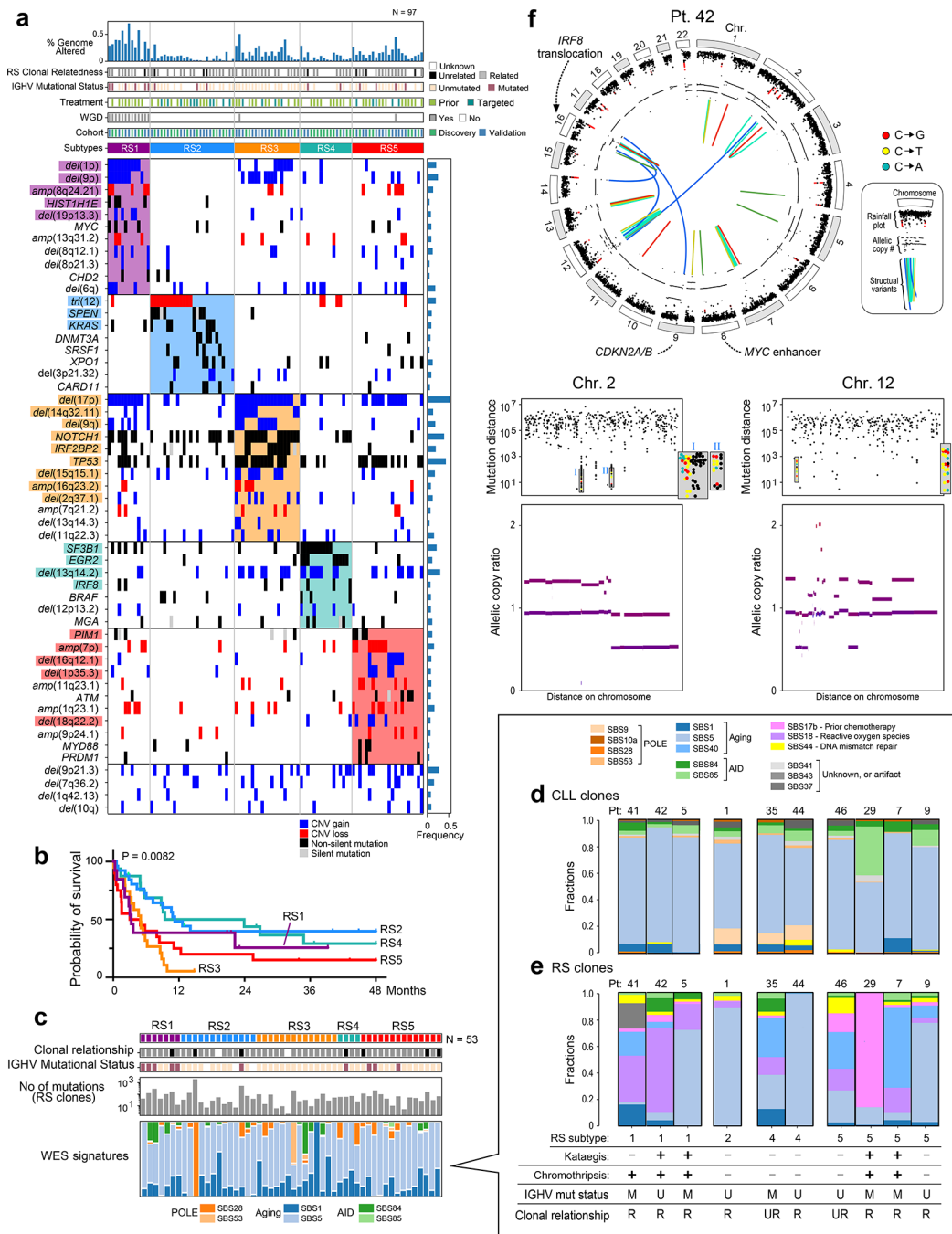


Fig. 4. Molecular mechanisms underlying transformation to RS.

a, genomic classification of RS. For 97 patients (columns), 5 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 whole genome doubling (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 (**c**) and RS (**d**) clones in 10 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 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 kataegis regions (C to G, red; C to T, yellow; C to A, teal) with corresponding allelic copy number ratio plot showing corresponding fragmentation.

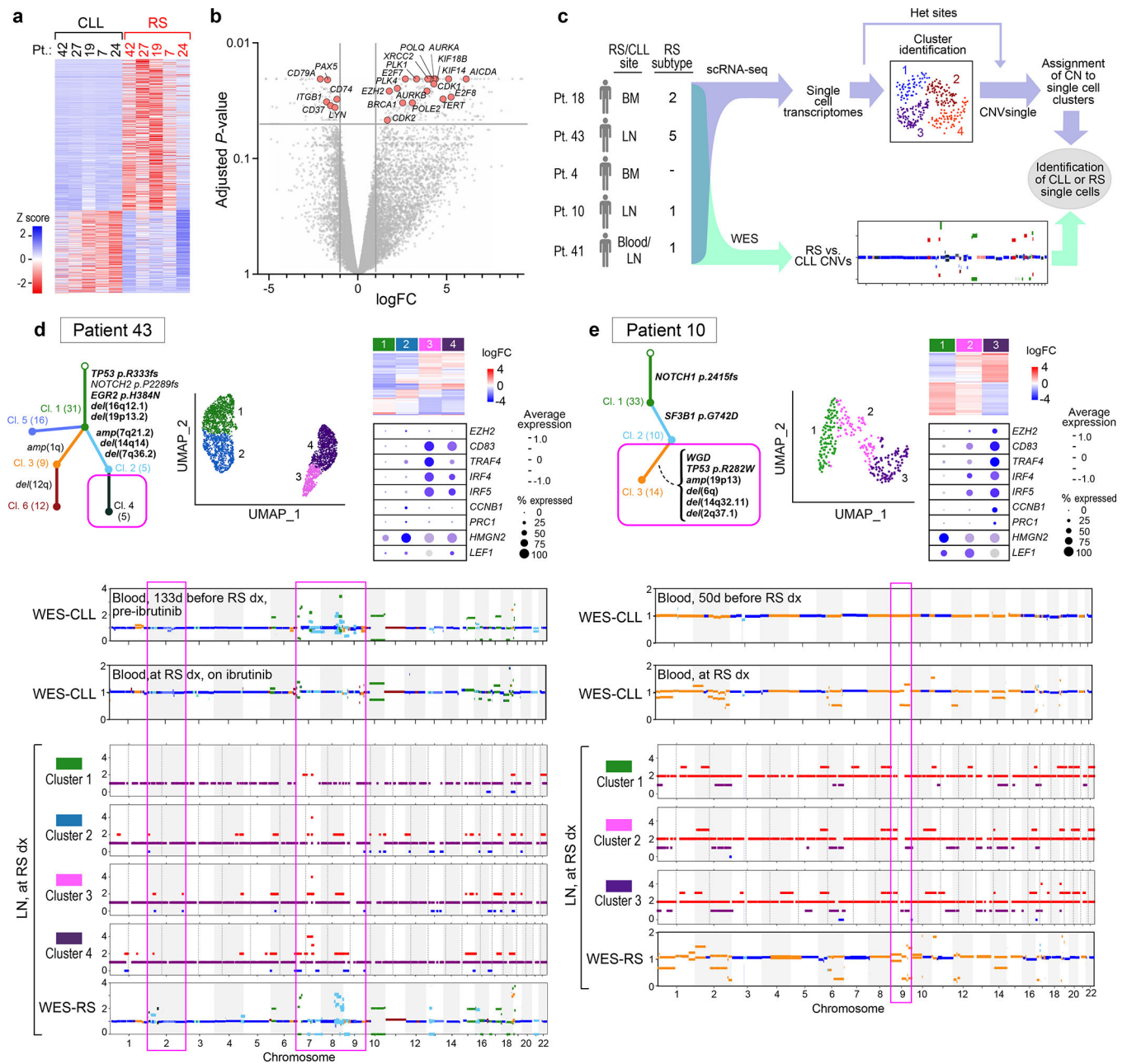


Fig. 5. Transformation to RS at single-cell resolution.

a, Heatmap of differentially expressed transcripts with FDR<0.1 and absolute log₂ fold change > 1 in analysis between paired RS and CLL samples from Pts 27, 7, 42, 20 and 24. **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 log₂FC 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 vs RS cells. **d-e**, Single-cell data shows transcriptional differences between RS and CLL from Pt 43 in **d**, and Pt 10 in **e**, and highlights 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).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

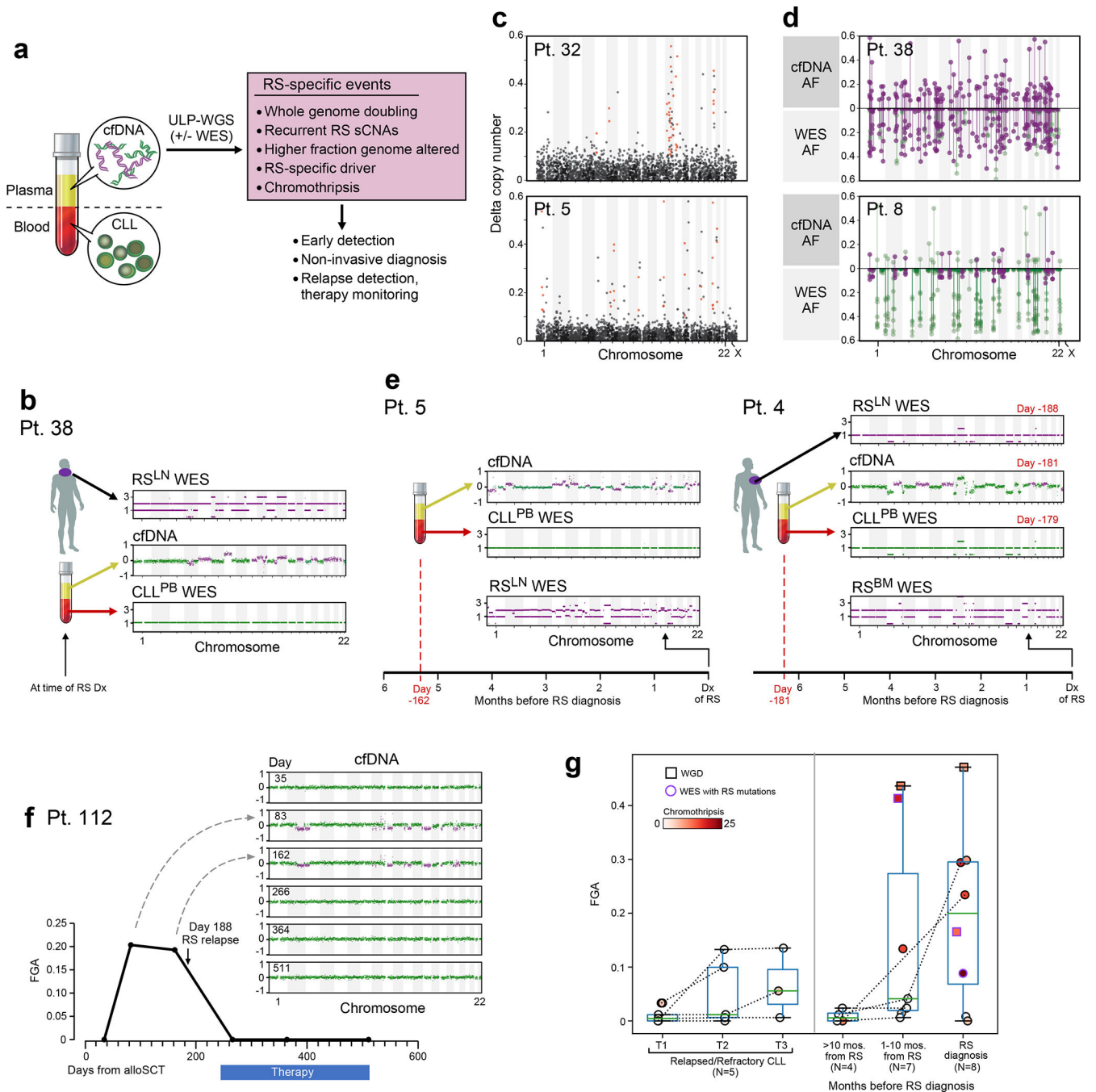


Figure 6. cfDNA isolated from plasma of RS patients shows evidence of transformation.
a, Schema showing how RS specific DNA events can be identified separately from cell-free DNA and different from circulating CLL cells. **b**, cfDNA in RS Pt 38 shows WGD of clonally unrelated RS, which is not seen in 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 (Pt 32 top, Pt 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. Pt 5 (top) shows RS-related WGD and chromothripsis fragmentation 162 days prior to RS diagnosis, which is not seen in corresponding co-sampled CLL cells. Plasma from Pt 20 (bottom panel) examined 181 days prior to RS shows RS-related WGD and sSCNVs which are not seen in co-sampled CLL or in lymph node biopsy taken from prior week. **f**, sCNAs become detectable prior to post-transplant relapse in Pt 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 fraction genome altered, color scale shows presence of chromothripsis, square represents whole genome doubled (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 5 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.