Multi-platform profiling characterizes molecular subgroups and resistance networks in Chronic Lymphocytic Leukemia

Johannes Bloehdorn[1*], Andrejs Braun[2], Amaro Taylor-Weiner[3], Billy Michael Chelliah Jebaraj[1], Sandra Robrecht[4], Julia Krzykalla[5], Heng Pan[6,7,8], Adam Giza[4], Gulnara Akylzhanova[2], Karlheinz Holzmann[9], Annika Scheffold[1], Harvey Johnston[10], Ru-Fang Yeh[11], Tetyana Klymenko[2], Eugen Tausch[1], Barbara Eichhorst[4], Lars Bullinger[12], Kirsten Fischer[4], Martin Weisser[13], Tadeusz Robak[14], Christof Schneider[1], John Gribben[2], Lekh N. Dahal[10], Mathew J. Carter[10], Olivier Elemento[6,7,8,15], Dan A. Landau[16,17], Donna S. Neuberg[18], Mark S. Cragg[10], Axel Benner[5], Michael Hallek[4], Catherine J. Wu[3,19,20,21], Hartmut Döhner[1], Stephan Stilgenbauer[1] and Daniel Mertens[1,22*]

[1]Department of Internal Medicine III, University of Ulm, Ulm, Germany
[2]Centre for Haemato-Oncology, Barts Cancer Institute, Queen Mary University of London, London, UK
[3]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA
[4]Department I for Internal Medicine and Centre for Integrated Oncology, University of Cologne, Cologne, Germany
[5]Division of Biostatistics, German Cancer Research Center, Heidelberg, Germany
[6]Caryl and Israel Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA
[7]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA
[8]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA
[9]Genomics Core Facility, Ulm University, Ulm, Germany
[10]Centre for Cancer Immunology, Cancer Sciences, Faculty of Medicine, Cancer Research UK Centre and Experimental Cancer Medicine Centre, University of Southampton, Southampton, UK
[11]Biostatistics, Genentech Inc., South San Francisco, CA, USA
[12]Medical Clinic for Hematology, Oncology and Tumor Biology, Charité University Hospital, Berlin, Germany
[13]Roche Pharma Research and Early Development, Penzberg, Germany
[14]Department of Hematology, Medical University of Lodz, Lodz, Poland
[15]Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA
[16]Cancer Genomics and Evolutionary Dynamics, Weill Cornell Medicine, New York, NY, USA
[17]New York Genome Center, New York, NY, USA
[18]Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA
[19]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA
[20]Department of Internal Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA
[21]Harvard Medical School, Boston, Massachusetts, USA
[22]German Cancer Research Center (DKFZ), Heidelberg, Germany

Running Title: Deciphering pathogenic networks in CLL

Keywords: CLL, genomic instability, EMT-like, resistance

Financial Support: This work was supported by research grants from the Else Kröner-Fresenius-Stiftung (2010_Kolleg24, 2012_A146), BMBF (PRECISE), Cancer Research UK (A24721 and A18087), Kay Kendall Leukemia Fund (KKL1101) and Blood Cancer UK (19001), European Commission / BMBF ("FIRE CLL", 01KT160), SFB 1074 (Projects B1, B2), DJCLS R 11/01.

Contact Information: Johannes Bloehdorn, Department of Internal Medicine III, Ulm University, Albert-Einstein-Allee 23, 89081 Ulm, Germany; johannes.bloehdorn@gmail.com, phone: +49 731 500 45701

Word / character counts:
Abstract: 155 words - now 150
Main text: 5082 words – now 4845
References: 73 – now 94
Figures: 6 figures, 1 consort diagram
Supplements: Figures:14; Tables:10

Abstract

Knowledge of the genomic landscape of chronic lymphocytic leukemia (CLL) grows increasingly detailed, providing challenges in contextualizing the accumulated information. To define the underlying networks, we here perform a multi-platform molecular characterization. We identify major subgroups characterized by genomic instability (GI) or activation of epithelial-mesenchymal transition (EMT)-like programs, which further subdivide into non-inflammatory and inflammatory subtypes. GI CLL exhibit disruption of genome integrity, DNA-damage response and are associated with mutagenesis mediated through activation-induced cytidine deaminase or defective mismatch repair. *TP53* wild-type and mutated/deleted cases constitute a transcriptionally uniform entity in GI CLL and show similarly poor progression-free survival at relapse. EMT-like CLL exhibit high genomic stability, reduced benefit from the addition of rituximab. EMT-like differentiation is inhibited by induction of DNA damage. This work extends the perspective on CLL biology and risk categories in *TP53* wild-type CLL. Furthermore, molecular targets identified within each subgroup provide opportunities for new treatment approaches.

Introduction

Characterization of genetic heterogeneity and its related clinical impact has provided the fundament for prognostic models in CLL and has been extended considerably in recent years.[1,2,3] However, the context in which genetic alterations arise remains to be further explored to understand disease dynamics and to refine therapeutic strategies by targeting cellular network or genetic dependencies. Alterations of the tumor suppressor genes *TP53* and *ATM* have been identified as major determinants for dysfunctional DNA-damage response (DDR), genomic instability, selection of genomically complex clones and poor response to treatment.[3,4,5,6,7,8] Treatment with genotoxic substances was found to contribute to inactivation of these tumor suppressors, acquisition of chromosomal aberrations and clonal evolution.[7,8,9,10] However, the mechanisms inducing genomic instability in cases without such lesions are incompletely characterized. Correspondingly, it remains to be proven if genomic lesions occur randomly or are specifically selected within a defined molecular or biologic framework during malignant transformation and over the natural course of CLL. Other treatment-independent factors contributing to the selection of alterations may include a heterogeneous degree of addiction to environmental stimuli or necessity to maintain the activity of certain signaling pathways.[11,12] Further, genomic lesions may evolve in a narrow spectrum depending on the related epigenetic makeup.[13,14]

In this work, we sought to delineate refined biological categories of CLL and identify cooperating pathogenic mechanisms which facilitate distinct pathways or microenvironmental interaction during disease development and evolution. We address this by performing a comprehensive characterization incorporating gene expression profiles (GEP) from two independent phase III CLL trial cohorts comprising 726 treatment naïve and relapsed patient samples, as detailed in the CONSORT diagram. Data from whole exome sequencing (WES), SNP-array analysis and protein expression is included for detailed biological characterization of the discovery cohort containing samples from untreated CLL patients enrolled on the CLL8 trial.[15] Discovered biologic subgroups are validated in the independent sample set of relapsed patients enrolled onto the REACH trial[16] and then confirmed *in vivo* using relevant genetically modified mouse models. Both the CLL8 and REACH trials were conducted as independent pivotal phase III multicenter trials to evaluate treatment with immunochemotherapy. They therefore provide an ideal basis for the correlation of biological characteristics and treatment outcome.

Results

Consensus clustering identifies distinct expression signatures associated with inflammation, genomic instability and activation of EMT-like networks

To explore tumor heterogeneity in CLL, we performed consensus clustering (CC) on CLL8 GEP data (n=337, TableS1) using 2359 variably expressed genes corresponding to a standard deviation (SD) of >0.5. First, the two initial clusters identified for k=2 and building the most distant branches of the dendrogram (protocluster C1 and C2) (Fig.S1A) were assessed for discriminatory characteristics. Using gene set enrichment analysis (GSEA), inflammatory features were identified as most prominent in segregating protocluster C1 and C2 and the subsequently derived clusters. To further decipher the underlying biology, we performed CC with an increasing number of clusters (Fig.S1A-D) and serial analytical steps by incorporating additional layers of information as schematically shown (Fig.1A, S1E). Optimal differentiation of distinct subtypes was achieved for k=6 GEP clusters (Fig.1B, S1E/F), while DNA-based class discovery approaches were insufficient to uncover similar patterns and the respective biological context (Fig.S1G). Subtypes were labeled according to the most prominent characteristics (Fig.1A) as “genomically instable, non-inflammatory” (GI) for C2 (n=133), “genomically instable with inflammatory features” ((I)GI) for C3 (n=56), “epithelial-mesenchymal-transition like, non-inflammatory” (EMT-L) for C4 (n=30) and “epithelial-mesenchymal-transition like with inflammatory features” ((I)EMT-L) for C1 (n=100). C5 cases (n=11) were labeled as “reprogrammed by early B cell factor 1” (EBF1-r), identifying tri(12) CLL as a distinct subtype with strong overexpression of *EBF1* and a transcriptional signature resembling healthy B cells. C6 (n=7) identified nuclear receptor interacting protein 1 (*NRIP1*) as specific for clusters evolving from the protocluster C1 (Fig.S2A). *NRIP1* is associated with clinical outcome in CLL and a major regulator of metabolism and coactivator of NF-kB-dependent inflammation.

Cases with *TP53* inactivation (Fig.1B, TableS2), V3-21 usage (TableS3), short telomeres (Fig.1C), high white blood cell (WBC) counts (Fig.1D, TableS2) and ZAP-70 positivity (TableS3) ($p<0.05$, Fisher`s exact test (two-sided)) were enriched in GI/(I)GI clusters. *TP53* mutated cases without concomitant del(17p) showed a near-exclusive occurrence in genomically instable cases (GI/(I)GI: n=16 (9.5%) vs. EMT-L/(I)EMT-L: n=1 (0.8%) ($p=0.002$, Pearson-chi-squared test (two-sided)). GSEA identified processes associated with genomic instability for GI/(I)GI and EMT networks for EMT-L/(I)EMT-L (Fig.2A). Furthermore, genes involved in the maintenance of genomic stability were frequently mutated (Fig.S2B/C) or overexpressed (Fig.S2D) in the GI/(I)GI cluster.

In summary, CLL can be segregated into two major biological subgroups defined by genomic instability or EMT-like networks with variable degrees of inflammatory features, further characterizing the inflammatory or non-inflammatory subtypes. GI and (I)EMT-L comprise the two largest and most distinct subtypes. Of importance, transcriptional homogeneity and consecutive co-clustering of CLL with GI/(I)GI expression signatures and cases showing *TP53* inactivation indicates that changes in genes other than *TP53* may execute similar biological effects and contribute to genomic instability.

Frequency and distribution of copy number alterations support a higher susceptibility to DNA damage in genomically instable CLL

To further investigate genomic instability, we assessed the distribution of copy number alterations (CNAs) using SNP-array. Genomic identification of significant targets in cancer (GISTIC)[17] was used to identify significantly altered regions and candidate genes representing putative targets of focal chromosomal amplification or loss (Fig.2B). Both GI and (I)GI involved frequent gains of 8q24.21 (including *MYC*) and 2p16.1 (including *XPO1, REL*). GI further showed gains of 6q22.31 and losses of 15q15.1 (including *KNSTRN*, *BUB1B*), 10q24.3 and 6q21. (I)GI showed losses of 13q14.13. (I)EMT-L had losses for 6q21 and 14q32.1 (Fig.2B). Although 11q deletions were found in all clusters at similar frequency, genes covered on 11q22.1-q22.2 (involving *YAP1* and a *MMP* cluster) were predicted by GISTIC to be specifically lost in GI (Fig.S2E-G) and showed a confirmatory underexpression (Fig.S2H). Assessing the impact of CNAs on expression, we found cluster-specific profiles of differentially expressed genes (DEGs) located in and adjacent to the minimally deleted (MDR) or minimally gained (MGR) regions, irrespective of the distribution of single CNAs (Fig.2C). This indicates that gene dosage effects for monoallelic deletions are context-dependent and modulate dominant pathogenic networks. Context-independent gene dosage effects were only observed in cases with biallelic deletion of 13q14 (Fig.2D). Significant enrichment and co-occurrence of deletions involving *RB1* (previously defined as type II deletions)[18] and losses exceeding cytoband 13q21.1, which we here define as "long distal breaks" (LDBs), were further observed in GI/(I)GI (Fig.2E). Type II deletions compared to type I deletions (not involving *RB1*) were significantly enriched in GI/(I)GI (55%) vs. (I)EMT-L/EMT-L (37.3%) (p=0.02, Fisher`s exact test (two-sided)). LDBs involving or exceeding the majority of cytoband 13q21.1 (distal of 54.7 mb) were significantly more frequent in GI/(I)GI (17.6%) vs. (I)EMT-L/EMT-L (7.1%) (p=0.03, Fisher`s exact test (two-sided)). LDBs and type II deletions showed a significant co-occurrence (87.5%) compared to LDBs and type I deletions (12.5%) (p<0.001, one-sample proportions test with continuity correction) (Fig.2E).

Taken together, the GI/(I)GI CLL subtype shows increased genomic complexity and selection of distinct chromosomal aberrations which may contribute to the disruption of genome integrity.

CLL cases with genomic instability show alterations in processes protecting genome integrity

Research into genomic instability in CLL has been primarily focused on functional loss of *TP53* or *ATM* and the consequent dysfunctional DDR. Conversely, here we observed enrichment of DNA repair signatures in GI/(I)GI (Fig.2A, 3A-D) and upregulation of *ATM* and *TP53* (Fig.2C), indicating increased DDR activation. Importantly, upregulation of p53 and phospho-p53 protein levels was confirmed in genomically instable cases without recurrent gene mutations or chromosomal aberrations other than del(13q) (Fig.3E, S3M), confirming a continuous activation independent of such lesions.

We identified numerous interdependent alterations increasing genomic instability in GI/(I)GI (summarized in Fig.3F). Critical pathogenic events involved telomere erosion (Fig.1C) and alterations of the shelterin complex. *POT1* mutations (Fig.S2C) were associated with telomeric abnormalities, chromosome breaks or fusions,[19] and *POT1* overexpression (Fig.S2D) implicated an increased requirement for telomere protection. In addition, *BUB1B and KNSTRN,* involved in correct chromosome segregation,[20,21] were identified by GISTIC as putative gene targets for del(15q15.1) (Fig.2B). We further found alterations of genes involved in regulating cell cycle checkpoints (Fig.2D/E, S3A/D/M) and numerous alterations impairing p53 and apoptosis (Fig.2B, 3H, 4D, S3A/C/M). Enrichment of DEGs for specific chromosomal regions supported complex alterations of tumor suppressors (Fig.S3E/F/G). Further, imbalanced MYC networks (Fig.3G) involved increased activation (Fig.2A/B, 3H, S3B/D/M) or loss of repressors (Fig.1B, 2B, S2C/D) of MYC family members. In line with an increased genomic instability, we observed a higher frequency and complexity of chromosomal aberrations in GI compared to (I)EMT-L cases after genotoxic stress (p=0.03, Mann-Whitney (two-sided)) (Fig.3I).

Signatures of mutational processes highlight the pathogenic role of DNA repair deficiency and activation-induced cytidine deaminase in genomically instable CLL

To better characterize pathogenic processes in CLL subtypes, we analyzed CLL8 cases with existing WES data (n=171, CD19 sorted) for mutational processes in cancer,[22] referenced in the COSMIC database. Signature projections revealed strong activation of signature 6 (defective mismatch repair (MMR), microsatellite unstable tumors), along with other mutational processes, including signature 3 (defective double-strand break (DSB)-repair) and 15 (defective MMR, observed in stomach and lung cancer), in GI (Fig.3J). Activation of signature 2, attributed to the activity of APOBEC family members, was increased in (I)GI compared to GI (p=0.01, Mann-Whitney (two-sided)) (Fig.3J, S3H), irrespective of *APOBEC* expression (Fig.S3L). This suggests that pathogenic deamination processes inducing DNA lesions are heterogeneous in genomically instable CLL. Signature 9, attributed to the activity of activation-induced cytidine deaminase (AID) (encoded by *AICDA*) during somatic hypermutation (SHM), was specific for IGHV mutated cases (p=2.5e-15, Mann-Whitney (two-sided)) (Fig.3K, S3I) and higher in GI compared to (I)EMT-L (p=0.03, Mann-Whitney (two-sided)) (Fig.S3J), while distribution of IGHV mutated cases was balanced across subtypes (GI: 43%, (I)EMT-L: 37%, (I)GI: 36%). Amplifications of *MYC* (8q24.21) were most frequently observed in GI/(I)GI cases, in line with the role of AID and possibly other APOBEC family members in mediating amplifications and translocations of this region. Moreover, IGHV mutated cases showed significantly higher activation of signatures 15 (p=0.01), 3 and 20 (p<0.005) (Mann-Whitney (two-sided)), indicating defective DNA repair (Fig.3K). Based on this observation, we considered if the error rate through defective DNA repair in GI might be exacerbated in situations of increased AID activity (and the respective AID-induced mismatches or DSB in non-Ig loci). We observed a trend for higher activation of DNA damage associated signatures 3 (defective DSB-repair) and 15 (defective MMR) and significantly higher activation of signature 6 (defective MMR) (p=0.02, Mann-Whitney (two-sided)) in IGHV mutated GI cases compared to IGHV mutated (I)EMT-L cases, while activations in IGHV unmutated cases of either subtype were low (Fig.3K, S3K). Genomic alterations implicated in genomic instability had a significantly lower incidence in IGHV mutated (I)EMT-L cases compared to IGHV mutated GI/(I)GI cases (Fig.3L), which supports a selective vulnerability in the context of AID/APOBEC activation and insufficient MMR. Conversely, when considering IGHV unmutated cases, which show a high frequency of alterations in genes like *TP53*, *ATM* or *POT1* (Fig.3L), mutational signatures associated with AID and MMR deficiency were generally low (Fig.3K). We subsequently assessed the impact of AID-mediated induction of genomic instability *in vitro* which validated such deleterious effects on the genome (Fig.S4A-I).

Together, these data show that distinct subgroups of CLL exhibit an increased accumulation of genomic lesions in association with mutational processes indicating deficient DNA repair and increased AID activity. Our observations indicate that the identified mutational processes represent independent pathogenic mechanism in GI/(I)GI subtypes.

EMT-like differentiation evolves in conjunction with inflammation and genomic stability

As described above, we observed enrichment of gene sets characteristic for EMT in cases which showed higher genomic stability (Fig.2A). During EMT, cells lose adhesion and gain invasive properties to exit from the surrounding tissue, as found for metastasis. Alterations in the EMT-like subgroup reflected central hallmarks of

EMT such as extracellular matrix remodeling (Fig.S5A) and increased cell motility (Fig.S5B).

Transcriptional signatures indicating immune signaling and TNFα-mediated inflammation were specifically upregulated in cases with EMT-like networks (Fig.2A, 4A, S5C) and correlated with upregulation of *NRIP1,* a coactivator of NF-kB-mediated inflammation (Fig.S2A). Besides inflammation, which serves as a strong EMT inducer, we were able to validate other EMT-inducing alterations like increased TGF-β signaling (Fig.2A) and *HIF1α* upregulation (Fig.S5E) for CLL cases with EMT-like networks.

Overexpression of EMT-associated transcription factors (EMT-TFs) (e.g. *ZEB1*, *SNAI1, TWIST1*) (Fig.4B) and receptor tyrosine kinases (Fig.S5D) were further confirmative for induction of EMT-like cellular programs. Notably, EMT-TFs showed a similar expression in different compartments like lymph nodes, bone marrow and peripheral blood (Fig.S5C). We also found overexpression of *EZH2*[23] and *SETD7*[24]*,* which may enhance NF-kB- (Fig.4A) and NOTCH- (Fig.4C) mediated EMT, along with other lysine methyltransferases (Fig.4D). Inflammatory features were associated with lower peripheral WBC counts (Fig.1D) supporting distinct migratory properties and environmental interaction. Notably, GEP of the CD19 negative cellular fraction (comprising the non-malignant blood component, i.e. monocytes, T and NK cells) showed unique signatures which reliably discriminated between patients belonging to the inflammatory/EMT-L or non-inflammatory/GI subtype (Fig.4E/F). These findings therefore offer evidence for a subtype-specific, CLL-mediated impact on non-malignant immune cells and altered environmental interaction.

We further identified several genes in recurrently deleted regions, including *YAP1* and a *MMP* cluster on 11q22.1-q22.2 (Fig.S2E-H) or genes residing in LDB-regions on 13q, like protocadherins (Fig.2E), which are closely linked with the EMT process.

Since the respective alterations were primarily observed in genomically instable CLL, the integrity of these regions seems indispensable for the differentiation towards EMT-like networks. In conclusion, CLL with EMT-like changes constitutes a distinct biologic subgroup with differentiated impact on the environment.

EMT-like differentiation can be induced in lymphoma and shows reciprocal inhibition with genomic instability

Since CLL cases with EMT-like networks show strong transcriptional signatures indicating inflammation and immunological response, we next aimed to validate the effects of inflammation on EMT-induction *in vivo* by utilizing a syngeneic BCL_1 lymphoma transplant mouse model. The BCL_1 tumor is a syngeneic lymphoma of BALB/c origin and transplantation results in a typical B cell leukemia/lymphoma characterized by splenomegaly, peripheral blood lymphocytosis and death of all tumor-bearing mice. This model was used as it reflects major hallmarks of the (I)EMT-L subtype: 1) lymphoma cells experience strong environmental stimulus during tumor development and migration to lymphoid organs and, 2) tumor transplantation itself is associated with a heavy inflammatory response. Notably, spleen tumor samples obtained at defined time points after transplantation showed dynamic GEP changes confirming inflammation and the associated induction of EMT-like networks as indicated by upregulation of *Vim*, the EMT-TFs *Zeb1, Snai1* and downregulation of *Cdh1*, all characteristic of EMT (Fig.4G).

Regarding the rare occurrence of alterations associated with genomic instability in cases with EMT-like networks, we next hypothesized that p53 activation and an increased DDR may inhibit induction of EMT-like programs. Specifically, (I)EMT-L cases exhibited low p53 and phospho-p53 levels (Fig.3E, S3M), and related pathway activation (Fig.2A/C) alongside inverse correlation of *TP53* and *ZEB1* expression (Fig.S5F). Lymphoma cells showed *miR-200c* induction (Fig.S5G) while the respective targets, EMT-TFs *ZEB1* and *TWIST1,* decreased after ionizing radiation (Fig.S5H), in line with the p53-*miR-200c-ZEB1* mediated suppression of EMT in solid tumors.[25] Further, increased NOTCH signaling (Fig.4C) may stabilize EMT-like networks through DDR suppression (Fig.3A-E), as previously reported.[26]

We subsequently aimed to validate the inhibitory effects of genomic instability on EMT-like differentiation *in vivo*. Among the different genes found to be involved in genomically instable cases, *TCL1A* and *MYC* were amplified and/or consistently upregulated in the GI/(I)GI subgroup and the corresponding pathway has been identified as a central element for this pathogenic network (Fig.3H, S6B). Both genes represent single oncogenic drivers and their overexpression in murine models leads to aggressive lymphomas with rapid proliferation and genomic instability.

With that in mind, we used the two corresponding models: 1) Eµ-Myc, where the c-Myc oncogene is placed under the control of the immunoglobulin enhancer to induce a highly aggressive B-lymphoid malignancy and 2) Eµ-TCL1, where the TCL1 oncogene is driven by the immunoglobulin enhancer and represents a well-established model of CLL. Tumors derived from these models were specifically assessed for c-Myc- and TCL1-induced pathway alterations by proteome profiling. Confirming the observations made in the human samples, we identified proteome profiles reflecting network activation observed for genomically instable cases while processes characteristic for the EMT-like subgroup were downregulated (Fig.4H). Alterations leading to genomic instability, therefore contribute to the inhibition of EMT-like networks.

EMT-transition *in vivo* was recently found to occur through a multistep process rather than a binary switch.[27] We assessed if such plasticity was present, and if the transition to EMT-like networks could be modulated, in aggressive Eµ-TCL1 tumors, similar to observations made for the BCL_1 model. Generating strong EMT-inducing stimuli with repetitive cycles of inflammation and environmental interaction through serial transplantations (STX) (Fig.S5I), we were able to mimic EMT plasticity in Eµ-TCL1 tumor cells (Fig.S5J/K).

These findings provide evidence that induction and maintenance of EMT-like networks requires convergence of strong EMT-inducing stimuli, while alterations associated with genomic instability and increasing aggressiveness inhibit EMT-like differentiation as depicted in our model (Fig.4I).

Pathogenic networks in CLL are not epigenetically controlled through DNA methylation

Chromatin organization influences transcriptional activity during differentiated biologic processes, and altered states can initiate or maintain pathogenic conditions such as EMT and genomic instability.[28,29,30,31] We observed characteristic profiles suggesting a heterogeneous transcriptional activity in CLL subtypes (Fig.4J/K) supported through distinct patterns of epigenetic modifiers. These included *MAFG/DNMT3B*[32] (Fig.S6A/C), *TCL1A/DNMT3A*[33] (Fig.S6B/C), TGF-β signaling[34] (Fig.2A), histone deacetylases (Fig.S6D), chromodomain-helicase-DNA-binding proteins (Fig.S6E) and others (Fig.S6F). We also observed a stringent association of these genes and the cluster hierarchy, irrespective of chromosomal alterations, particularly exemplified for tri(12) cases. Tri(12) cases exhibited highly distinct transcriptional profiles forming a single cluster (EBF1-r) or leading to clustered enrichment in GI and (I)EMT-L (Fig.1B, 4L, S6G/H). *EBF1* showed the strongest overexpression in EBF1-r (Fig.4J) and in agreement with its crucial involvement in the differentiation of B cells,[35] EBF1-r cases stringently clustered with healthy donor B cells (Fig.4M).

Notwithstanding such extensive transcriptional reprogramming towards mature B cells, tri(12) cases retained distinct profiles of epigenetic modifiers reflecting the respective cluster hierarchy (Fig.S6I). To investigate if epigenetic modification through DNA methylation may contribute to pathogenic network profiles, we analyzed reduced representation bisulfite sequencing (RRBS) data for n=182 matched cases. Robust methylation differences across groups were not observed (Fig.S6J). While processes involving AID/APOBECs and BER may actively promote demethylation in genomically instable cases, we observed significantly higher expression levels of DNA-demethylases in (I)EMT-L compared to GI (Fig.S6K).

Genomic instability is associated with poor prognosis in CLL

To define the clinical impact resulting from the underlying biology of identified CLL subtypes, we assessed the clinical course in previously untreated patients of the CLL8 trial. Progression-free survival (PFS) was shortest for chemotherapy treatment in genomically instable cases but increased considerably when fludarabine and cyclophosphamide (FC) was combined with rituximab (R), with strongest overall benefit observed for (I)GI (GI: median PFS 27.8 months (FC) vs. 42.4 months (FCR), HR: 0.55 (95%CI 0.37-0.82), p=0.004; (I)GI: median PFS 22.8 months (FC) vs. 68.1 months (FCR), HR: 0.30 (95%CI 0.15-0.60), p=0.001) (Fig.5A, S7A). In comparison, PFS was considerably longer for FC treatment in (I)EMT-L/EMT-L cases, but lacked a similar increase in efficacy when rituximab was added ((I)EMT-L: median PFS 36.1 months (FC) vs. 52 months (FCR), HR: 0.86 (95%CI 0.53-1.41), p=0.56; EMT-L: median PFS 45.5 months (FC) vs. 65.2 months (FCR), HR: 0.63 (95%CI 0.25-1.56), p=0.31) (Fig.5A, S7A).

Notably, overall survival (OS) was significantly improved in (I)GI when FC was combined with rituximab (median OS not reached (FCR) vs. 56.6 months (FC), HR: 0.32 (95%CI 0.13-0.79), p=0.013) (Fig.5A, S7B).

To further elucidate the clinical impact of the underlying biology, we performed survival analyses for genetically defined categories. We first assessed the clinical outcome in subtypes with regard to the IGHV mutation status. As previously described, identified patterns for mutational signatures support a selective vulnerability in the context of AID activation and insufficient MMR in GI/(I)GI cases (Fig.3K/L). Confirmatory, IGHV mutated (I)EMT-L cases showed the longest PFS and OS rates (Fig.5B, S8). Survival differences were especially pronounced for FC treatment, showing a considerably shorter median PFS for IGHV mutated GI cases compared to IGHV mutated (I)EMT-L cases (29.1 months vs. not reached, HR: 0.29 (95%CI 0.11-0.77), p=0.013), while median PFS in IGHV mutated GI cases was similar to IGHV unmutated cases of both the GI and (I)EMT-L subtype (24.4 and 27.8 months) (Fig S8C,TableS8).

Cases with *TP53* alterations and wild-type *TP53* show a high transcriptional homogeneity in the GI/(I)GI cluster (Fig.1B), reflecting similar biology. We next confirmed that GI compared to (I)EMT-L cases also show a poorer clinical course when segregated for chromosomal aberrations or *TP53* wild-type status (Fig.S9/S10, TableS6). Characteristics, other than *TP53* defect*,* were homogenously distributed between both groups (TableS2). The prognostic impact of biological subtypes was further validated in an extended analysis for molecularly defined subcategories. While alterations of *TP53* and *ATM* were found to associate with a poor clinical course, independent of the respective biological subtype, we observed a strong impact on outcome in *TP53* and *ATM* wild-type cases and with regard to the *SF3B1* mutation status (Fig.5C/D). The addition of rituximab considerably improved outcome in GI, whereas (I)EMT-L cases in contrast consistently lacked a similar increase of efficacy (Fig.5A/C/D, Fig.S7-S12). In line with imbalanced rates of lethal sepsis (TableS9), representing the fulminant release of cytokines from the immune system, and distinct expression profiles for CD19- non-malignant immune cells (Fig.4F), differential treatment efficacy for the addition of rituximab strongly supports a heterogeneous responsiveness of the immune system in identified biological subgroups.

Validation of CLL subtypes and prognostic impact in the REACH study cohort

We next validated the major subtypes in an independent phase III trial cohort of relapsed CLL patients (REACH, n=300) and a second internal validation set of previously unexamined, U-CLL8 samples (n=89) (Fig.6A/C, S13A-C, TableS7).
REACH was analyzed complementary to CLL8 by using consensus clustering on variably expressed genes with SD >0.5 (Fig.6A), which reliably identified the same biological categories (Fig.6B/C, S13A). Subtype-specific expression patterns were also found when hierarchical clustering was applied on the internal U-CLL8 validation set (Fig.S13B). Specific analysis of expression profiles of genes associated with increased DDR, alternative mechanisms for p53 inactivation and distribution of cases with *TP53* defect further validated the CLL subtypes in the REACH cohort (Fig.S13C, TableS7).

Increased transcriptional homogeneity and co-clustering of cases classified as "GI/(I)GI" supported the selection of unifying features after treatment (Fig.6A/C, S13C-E). The prognostic impact remained identical as observed for (I)EMT-L and GI in CLL8 (Fig.S14A/B). As hypothesized from the biological context, median PFS rates in "GI" cases without *TP53* defect resembled those of cases with *TP53* defect (19.9 vs. 17.1 months), in contrast to "(I)EMT-L" cases (median PFS 38 months, p<0.0001)(Fig.6D). Median OS was shortest for cases with *TP53* defect (35 months), followed by "GI" (68 months, p<0.0001) and "(I)EMT-L" (not reached) (Fig.6E).

Notably, the GI subtype was identified in a multivariate Cox proportional hazards regression model, along with unmutated IGHV and del(17p), as an independent adverse prognostic factor associated with short PFS in relapsed CLL cases (REACH) (TableS10).

Discussion

In this study we identified genetically and clinically distinct CLL subgroups, comprising genomic instability or activation of EMT-like networks, extending the current perspective on disease pathogenesis, progression and resistance.

While we observed higher leukocyte counts for the CLL8 discovery cohort of CD19 sorted CLL cases, likely through selection of samples with abundant material for multiple analyses, patient characteristics and especially high-risk markers showed a well-balanced distribution representative of the full trial population.

As implied from the identified mutational signatures, genomic instability may be present before malignant transformation or in the early phase of the disease and facilitated by defective DNA repair mechanism and activation of AID during SHM.[36]

Furthermore, AID may be reactivated during the disease course[37,38] and add to the acquisition of genomic lesions and clonal evolution in the GI subtype.

The specific association with the GI/(I)GI subtype further highlights a role of arginine and lysine methyltransferases in genomically instable CLL previously found to promote lymphomagenesis or induce genomic instability in cancer.[39,40,41,42,43]

While we could not detect differences for promoter or gene body methylation in identified subgroups, gene-specific regulation of methylation or demethylation dynamics in distinct regions may still impact pathogenic networks. Our data further supports a broad involvement of other epigenetic modifiers on subtype-specific chromatin organization, which itself was shown to be an important determinant of genomic stability.[28,29,30]

Development of treatment-resistant CLL has been associated with the inactivation of *TP53, ATM* and correspondingly a deficient DDR.[3,4,5,6] However, here we show rather that genomically instable CLL exhibit activation but insufficient execution of DNA-repair programs, irrespective of *TP53* status. We have further classified multiple alterations contributing to genomic instability into distinct but interdependent processes. These involve disruption of telomere maintenance, DNA and chromosome integrity, altered DDR with insufficient DNA repair, MYC pathway activation, disrupted cell cycle checkpoints and chromatin organization. Continued execution of these disordered processes therefore maintains genomic instability through the ongoing accumulation of genomic lesions.

Correspondingly, we show that therapy-associated genotoxic effects can aggravate genomic instability and worsen long-term treatment outcome in such patients when receiving sequential therapies. In particular, the poor outcome in CLL with wild-type *TP53* and *ATM* highlights the importance of alternative mechanisms for the induction of genomic instability. We show that such cases benefit considerably from p53/DDR-independent treatments like rituximab, which mediates killing through non-DDR processes, dependent on Fc and FcγR engagement.[44]

Contrasting the characteristics observed for GI, we identified (I)EMT-L cases to exhibit the most distant clustering and comprise a highly differentiated biology from this subtype. While differentiation towards EMT is a well-known phenomenon in various cancers, EMT-like changes constitute a hitherto unappreciated aspect of CLL and were unexpected as metastasis and the underlying biologic processes seem redundant for leukemic dissemination. However, EMT-like features and direct involvement of EMT-TFs were previously reported for pathogenic processes in hematologic malignancies. These involve regulation of migratory properties in myeloma[45,46], proliferative capacities and response to treatment in mantle cell lymphoma[47] and ALL[48] and regulation of immune checkpoints and aggressiveness in DLBCL[49,50]. Higher methylation levels were also found for the EMT-TF *TWIST2* in CLL with mutated IGHV[51] and in comparison to healthy donor cells[52]. Furthermore, EMT-TFs fulfill crucial roles in normal hematopoiesis and B cell maturation.[53]

Since we identify interdependent changes in malignant B cells which resemble the EMT process we have called this subgroup and the corresponding specific alterations “EMT-like”. Our observations do not put lineage-specific and phenotypic changes recognized in epithelial or mesenchymal tissues during EMT into focus, but categorize cellular properties and processes (e.g. inflammatory changes, NOTCH signaling, genomic stability) which are found likewise in either tissue context.

The EMT-like subgroup reflects various characteristics of CLL with increased environmental interaction via receptor or cytokine signaling and migratory properties.[54,55,56,57] EMT-TFs in CLL, therefore, may regulate migratory capabilities to infiltrate lymphatic tissues or other pro-survival niches, similar to other cancers.[27,58,59] Multiple studies have shown the involvement of cytokines from the microenvironment to regulate tumor inflammation and induction of EMT. Conversely, EMT-TFs themselves can induce inflammation in cancer cells and shape the microenvironment accordingly.[60] EMT-like transcriptional programs in CLL, therefore, may be activated through inflammatory, HIF1α, NOTCH1 and other signaling cascades representing a convergence from multiple pathways during lymphomagenesis. Stable integration of these signals into activated EMT-like networks with a pro-survival advantage may consolidate such differentiation.

We observed a reduced benefit for rituximab in (I)EMT-L cases occurring in conjunction with a general NOTCH pathway activation but independent of NOTCH1 mutations previously associated with rituximab resistance.[3] Systemic effects on non-tumor cells occurring in (I)EMT-L CLL may further elicit functional disruption of effector cells like macrophages, which execute treatment effects of rituximab.[61] In line with our findings, the presence of EMT-like gene expression profiles has previously been linked with immunosuppression in solid tumor studies using checkpoint inhibitors.[62,63]

We provide experimental evidence from *in vivo* and *in vitro* studies that tumor aggressiveness and the extent of DDR directly regulate EMT-like networks through suppression of EMT-TFs. However, EMT-TFs can also act in the reverse direction to suppress the DDR. EMT-TFs were shown to downregulate p53 and diminish its transcriptional activity.[64,65,66] Further, expression of the *miR-200* and *miR-34* families is regulated by p53 and both target EMT-TFs,[25,67,68,69] which themselves build tight regulatory loops with these miRs.[70,71,72] EMT-TFs protect against DNA damage in solid tumors, where *ZEB1* expression is inversely correlated with the incidence of CNAs and *TP53* mutations[73], which is again mirrored by our data. Such regulatory axes may strengthen the highly diversified biologic trajectories underlying the CLL subtypes. As recently reported, EMT-transition in solid tumors occurs though intermediate hybrid states, which exhibit distinct cellular properties and show a differentiated interaction with the microenvironment.[27] We show that EMT-like differentiation in lymphoma can be modulated, but comprises a distinct subgroup largely independent of sub-compartments like lymph node, bone marrow or peripheral blood.

In conclusion, this study extends the basis for understanding CLL pathogenesis and pathway dependencies that may be targeted by novel compounds. Identified molecular targets in a defined biologic context may further advance the development of new treatment strategies. Compound combinations targeting, for example, BCL2 and PRMT5 or XPO1, together with anti-CD20 monoclonal antibodies, may specifically synergize in genomically instable cases. Future assessment of the subtype related outcome in comprehensively characterized trial cohorts testing BCL2-, BTK- and other inhibitors in development will further elucidate the therapeutic potential of such treatment combinations.

Methods

Overview for conducted analyses and respective sample size. Specimens were collected from patients registered on the CLL8[15] (first-line) and REACH[16] (relapse) trials after informed consent. All patients suffered from progressive disease with the need for treatment. Samples were CD19 sorted (CLL8 n=337; REACH, n=300) or unsorted (U-CLL8, n=89). RNA and DNA were purified and assessed for integrity with routine protocols. Samples were used on the following experimental platforms: GEP (CLL8, n=426; REACH, n=300) on Human Exon 1.0 ST arrays (Affymetrix), SNP-array analysis (CLL8, n=309 treatment naive, of which n=18 had a paired sample at relapse) on Human SNP Arrays 6.0 (Affymetrix), analysis of mutations and mutational processes (CLL8, n=171) using WES (Illumina), RRBS for methylation analysis (CLL8, n=182), western blots for validation were performed in selected cases without recurrent alterations. FISH, IGHV and *TP53* mutation analysis was performed upon trial registration. The multiparameter analysis was conducted for CD19 sorted CLL8 samples and distribution of genetic characteristics for analyzed cases was representative for the full CLL8 trial cohort (TableS1). REACH and U-CLL8 samples were used as validation cohorts. The REACH cohort was chosen since it was designed complementary to CLL8 and ideally suited to independently validate identified transcriptional networks and related resistance mechanisms. U-CLL8 samples were chosen as internal validation set to assess the impact of lower tumor homogeneity and to exclude the possibility that transcriptional changes were induced through the process of CD19 sorting. Survival data were available with a median observation time of 5.9 (CLL8) and 4.9 (REACH) years. Pathogenic networks were validated using transgenic Eµ-Myc and Eµ-TCL1 mice or BCL_1 and Eµ-TCL1 tumor transplantation models. Details on methods, demographic and clinical characteristics are also provided in the CONSORT diagram**,** sequential workflow and analysis is visualized in Fig.1A and Fig.S1E.

Patients and samples. Analysis was conducted on peripheral blood samples from previously untreated CLL patients (n=426) from the CLL8 trial, a prospective, international, multicenter, open-label, randomized phase 3 study, comparing first-line treatment with R-FC (n=408) or FC (n=409).[15] Data on genomic aberrations and mutations, such as del(17p), del(11q), tri(12), del(13q), the IGHV, *TP53*, *SF3B1* and *NOTCH1* mutational status, was included for analysis[3]. Patient characteristics of target analysis population of n=337 CD19 sorted CLL used as discovery cohort is provided in Table S1. Samples were collected at enrolment on the CLL8 trial and selected for gene expression profiling based on availability and RNA quality.

For validation of discoveries, we further used an independent set of gene expression profiles (n=300) generated from CD19 sorted CLL patient samples of the REACH study, a prospective, international, multicenter, open-label, phase 3 study, in which patients with previously treated CLL were randomized to receive R-FC (n=276) or FC alone (n=276).[16] Details on genetic characteristics and other variables for the CLL8 gene expression cohort are listed in table S1. The primary objective of the CLL8 and REACH study was to demonstrate superiority with regards to PFS for R-FC compared with FC alone. The study protocols were approved by institutional review boards at participating centers, and all patients gave written informed consent. Details on trial design and eligibility criteria and clinical outcome have been described elsewhere.[15,16] The studies were conducted in accordance with the Declaration of Helsinki. Details for each study are further provided online at the ClinicalTrials.gov (C T G) homepage. All baseline parameters including genetics, serum parameters (such as thymidine kinase, β2-microglobulin) and cell surface markers (such as ZAP-70) were performed in a centralized manner in accredited reference laboratories of the German CLL Study Group (GCLLSG) for the CLL8 trial, as outlined in the original study protocol. The central GCLLSG genetic reference testing laboratory in Ulm conducted fluorescence in situ hybridization (FISH), mutation analysis of genes recurrently mutated in CLL (such as *TP53*, *ATM*, *NOTCH1*, *SF3B1*) by targeted resequencing and IGHV mutation status, telomere length, GEP Exon- and SNP-Array hybridization and analysis.

Trial participation, genetic testing and data analysis have been conducted after informed patient consent, with the approval of the respective local ethics committees of participating centers. Data analysis related to this study was approved by the Ulm University ethics committee.

RNA isolation and quality assessment. Ficoll density gradient centrifugation for isolation of mononuclear cells was performed on all CLL samples, immunomagnetic tumor cell enrichment via CD19 (Midi MACS, Miltenyi Biotec, Bergisch Gladbach, Germany) was performed on (n=337) samples from the CLL8 trial, (n=300) from the REACH trial and (n=5) healthy donors (2 male, 3 female) (S1 sample set). Additional 89 samples were left unsorted (S2 sample set). For n=10 CLL8 cases both the CD19 positive (+) and negative (-) compartment from CLL samples with inflammatory (I) and non-inflammatory (NI) signatures were investigated. Only cases where the CD19 negative fraction had < 3.30% (median 2.5%) contamination with CD19+CD5+ cells were used. Total RNA for mRNA profiling was extracted from whole cell lysate according to the AllPrep DNA/RNA mini kit (Qiagen). Quality control was assessed using the Agilent 2100 Bioanalyzer with the RNA 6000 Nano LabChip (Agilent Technologies). The Chip was prepared according to the manufacturer's protocol and analyzed using the 2100 Expert software (version 2.6). To ensure the best accuracy and reproducibility samples with an RNA Integrity Number (RIN) less than 7.0 were excluded from further analysis.

Gene expression profiling on Human Exon 1.0 ST Arrays or using qPCR. Samples were analyzed for mRNA expression using the Affymetrix GeneChip Human Exon 1.0 ST Array (Affymetrix, Santa Clara, CA, USA). The experiment was conducted according to the manufacturer’s protocol. In brief, 250 ng RNA per sample were amplified, transcribed to cDNA, fragmented and subsequently labeled with biotin. Array hybridization was performed at 45°C for 16-18h in the Affymetrix GeneChip Hybridization Oven 640, arrays were subsequently washed in the Fluidics Station 450 and scanned on the GeneChip scanner 3000 7G. Complete microarray data sets are available at Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/; GEO accession number: GSE58211 (REACH only); GSE126595 (full clinical data set); GSE126699 (functional data).

RNA extraction and expression analysis using qPCR: Total RNA was isolated using RNeasy or DNA/RNA AllPrep mini kit (Qiagen) as per the manufacturer’s instructions. RNA concentration was estimated using NanoDrop (Thermo Scientific) and 400ng of RNA was reverse transcribed using the Reverse Transcription Kit (Promega). The cDNA was diluted 1:10 prior to addition into the qPCR reaction mix. Sybr Green Supermix (Bio-rad) was used for the qPCR analysis as per the manufacturer´s protocol. Fold differences in gene expression was analyzed using the ΔΔ Ct method by normalizing to control sets as mentioned in the figure legends. Primers and sequence is provided in table …

Expression level of *miR-200c* (Cat. No. 002300) was analysed using the TaqManTM miRNA assays from ThermoFisher Scientific, according to the manufacturer’s instruction. *U6* snRNA (Cat. No. 001973) was used as control. In brief, total RNA was isolated using the RNeasy kit from Qiagen and 10ng of the RNA samples was reverse transcribed using the TaqMan™ MicroRNA Reverse Transcription Kit (Cat. No. 4366597) according to the manufacturer’s instructions. The reverse transcribed cDNA was diluted 1:3 and analysed using the TaqMan™ Universal PCR Master Mix, no AmpErase™ UNG (Cat. No. 4364343). The qPCR reaction was performed in a total volume of 10µl on a 384 well QuantStudio 5 Real-time PCR system from ThermoFisher Scientific.

Normalization of expression data. Raw Affymetrix Human Exon 1.0 ST Array (HuEx-1_0-st-v2) data files and the data set from Herishanu et al. (GEO ID GSE21029) have been preprocessed by the robust multichip average (RMA) algorithm using the aroma.affymetrix R package[74] (version 2.12.0). Normalized data is stored with the assigned analysis ID, raw data files include info on CD19 selection and code of the CTG registry. Besides RMA normalization, background correction and quantile normalization were applied. Aroma.affymetrix was applied to generate gene expression values summarized on the exon/probe set level and on the transcript level using the ‘core’ probe set definition according to Affymetrix. ‘Core’ refers to probe sets that are supported by the most reliable evidence from RefSeq and full-length mRNA GenBank records containing complete CDS information. We further assessed and excluded presence of potential batch effects induced by external factors including time point and location of sampling, duration of storage and time point of labeling and hybridization. Quality control was further conducted with Relative Log Expression (RLE) and Normalized Unscaled Standard Errors (NUSE), where abnormalities were not observed.

Analysis of expression data. Statistical procedures were performed with the R software environment, version 3.3.3 and 3.4.1. For GEP analysis BRB-ArrayTools Version 4.2.1-4.6.1 (available at http://linus.nci.nih.gov/BRB-ArrayTools.html and www.r-project.org) was used.

Unspecific filtering based on standard deviations (SD>0.5) was used to select transcripts with largest expression variability across all arrays for the CLL8 and REACH gene expression data set, resulting in 2359 transcripts (mRNA). Agglomerative hierarchical clustering and consensus clustering[75] on mRNA was applied using average or complete linkage and Pearson correlation distance metric with 1000 iterations, respectively. Consensus clustering was used to identify the number of clusters with best clinico-biologic segregation from k=2 up to k=10 possible clusters. Decision in favor of using k=6 was based on combined information from the delta area plot, cluster stability and clinical or biologic information as previously described. Differential expression, specifically assessed for genetic variables or defined clusters was conducted using the Class Comparison Tool from BRB-ArrayTools Version 4.2.1 with univariate permutation tests for individual genes controlling the false discovery rate (FDR) using the method of Benjamini and Hochberg. Genes for which the FDR was equal or less than 0.05

were considered significant for differential expression. Visualizing selected gene sets, such as components of a defined biological process, was conducted using the Genesis platform[76] (release 1.8.0). For depicting the cluster composition of selected gene sets hierarchical clustering was used with Pearson correlation distance and complete linkage when needed.

Specific assessment of differential expression for the most common recurrent alterations including *TP53* defect (del(17p) and/or *TP53* mutation), del(11q), tri(12), normal karyotype, del(13q), IGHV status, *SF3B1* and *NOTCH1* mutations was analyzed for respective groups defined by presence of the cytogenetic alteration or mutation of interest, irrespective of the co-occurrence of other alterations. If not otherwise detailed, for group-specific assessments focusing on genetic categories like *TP53* defect, the group of interest was calculated against the reference group containing all other cases.

Identification of biologic processes showing overrepresentation in expression clusters. Gene Set Enrichment Analysis (GSEA)[77] was used to discriminate major biologic characteristics and processes in defined clusters (release v3.0). For the analysis we used hallmark gene sets compiled at the Molecular Signatures Database, Broad Institute. GSEA was applied on mRNA expression data for respective cases of every single cluster in a comparative fashion with the remaining cases of all other clusters. Overrepresented gene sets of each single cluster with a FDR q≤0.05, which was used as stringent filter criteria, were selected for further analysis. Overrepresentation of biologic processes and resulting pattern composition for all clusters was visualized in a heatmap using the normalized enrichment scores of the overrepresented gene set as a measure for the enrichment intensity. Identified biologic processes were grouped together and labeled according to biologic similarities.

SNP-array analysis. SNP-array analysis on samples from the CLL8 trial for recurrent CNAs was conducted for n=309 previously untreated samples of which n=18 cases had a paired relapse sample[78]. All samples had GEP data with the respective cluster assignment available. In brief, for SNP-Array hybridization genomic DNA was hybridized to the Genome-Wide Human SNP Array 6.0 according to the manufacturer's protocol (Affymetrix, Santa Clara, CA, USA). SNP genotype calls were generated by applying the birdseed algorithm in Genotyping Console version 4.0 (Affymetrix) using at least 50 arrays in each analysis. DNA copy number analyses were performed using reference alignment[79], dChipSNP[80] and circular binary segmentation (CBS)[81]. Segmentation was done pairwise against intra-individual reference DNA in cases having a pure CD19 negative cell-fraction. For cases lacking matched normal material the segmentation of each sample was computed against a pool of ten gender-matched reference samples. Resulting segments with a window of at least five consecutive markers and mean log2-ratios of > 0.2 and < -0.2 were visually inspected using dChipSNP to exclude inherited copy number variants and false calls due to experimental artifacts (noise or interbatch effects). Lesions occurring in a subclone with a clone size under 25 % were revised using the aroma.affymetrix software package[82] for an exact determination of segment boundaries. Size position and location of genes were identified with the UCSC Genome Browser; assembly March 2006, NCBI36/hg18 (http://www.genome.ucsc.edu/)[83]. DNACopy version 1.44.0 and dChip version 2010.01 were used. Microarray raw data contained in the analysis have been made publicly available at Gene Expression Omnibus (GEO accession number: GSE36908 (CLL8 treatment naive) and GSE83566 (relapsed)). For visualization of deletion size for del(13q) LDBs the Integrative Genomics Viewer[84] (release 2.4.16) was used. Total number of CNAs (including non-recurrent CNAs) occurred with a mean count of 2.1 (range 0–9) per patient and showed homogenous distribution (range 1.7-2.2; GI: 2.2; EMT-L: 1.9; (I)EMT-L: 2; (I)GI: 1.7).

Genomic Identification of significant targets in cancer (GISTIC). To assess the specific enrichment of genomic amplifications and deletions within clusters identified by consensus clustering of GEP, we applied GISTIC[17] (v2.0.23) to the curated SNP array dataset. GISTIC identifies significantly amplified and deleted regions across a set of samples. Each aberration is given a G-score considering its amplitude and the frequency of its occurrence across samples within a GEP cluster. Significance of each aberration is estimated by GISTIC comparing the observed G-scores with results that would be expected by chance, using a permutation test that is based on the overall pattern of aberrations seen across the genome. To account for multiple testing, FDR estimation is done providing a consecutive *q* value for each aberration. In our analysis, a q-value cut-off below 0.25 was considered to identify significant results (vertical green line in Figure 2B).

Longitudinal analysis for CNAs. Longitudinal analysis for acquisition of CNAs before and after treatment was done on SNP-array data of cases with available baseline samples at inclusion into CLL8 trial (pre-treatment) and at the time point of relapse (post-treatment). Paired samples of cases with cluster assignment as identified through consensus clustering on GEP were available for GI baseline (n=11), GI relapse (n=11), (I)EMT-like baseline (n=7), (I)EMT-like relapse (n=7). The Wicoxon signed-rank test was applied to test for differences of aberrations in the pre- and post-treatment setting. Exemplary visualization for large representative CNAs in individual conditions was performed using dChip.

Muatation analysis and signature projections for mutational processes. Data was generated on samples from the CLL8 trial cohort. Cases used for this analysis were available for n=171 matched cases with corresponding GEP and WES data.[2] Matched cases were distributed across identified clusters in representative numbers (EBF1-r n=6, GI n=68, EMT-L n=11, (I)EMT-L n=52, (I)GI n=31, NRIP1 n=3). Libraries for WES were constructed and sequenced on an Illumina HiSeq2000 or HiSeq2500 using 76 bp paired-end reads. For targeted sequencing we used Illumina Design Studio to create custom amplicons with a size of 250 bp covering all coding regions of *TP53* and *ATM*. Library preparation was performed using TruSeq Custom Amplicon Assay Kit v1.5 (Illumina, San Diego, CA,USA) including extension and ligation steps between custom probes and adding of indices. Samples were pooled and loaded on a MiSeq flowcell in 48 sample batches and sequenced with MiSeq Reagent Kit 500v2 (Illumina) for a paired end run. Median depths of WES and targeted sequencing were 96 x and 1332 x, respectively. Software packages for bioinformatic analyses including demultiplexing, alignment to hg19 reference genome, variant calling and annotation were used.[2]

To assess the distribution and respective co-occurrence for enriched pathways and driver mutations, we used an agglomerative approach to estimate the similarity of distribution patterns for relative frequencies of mutations per cluster observed in our dataset, which was done in a hierarchial fashion (average linkage, Pearson correlation). Subsequently detailed representation of the single mutations was depicted for the respective clusters.

We used non-negative matrix factorization to assess the pathogenic processes operational in identified CLL subtypes which best explain the mutation pattern observed in corresponding cases as previously described.[22] Briefly we used a fixed matrix of signatures which were hypothesized to be involved in aging, AID related DNA damage, and DNA repair deficiencies, as reported by Alexandrov et al.[22], to perform a projection of our data on to these signatures using the NMF multiplicative update as previously described[85]. After performing this projection, we used a heuristic approach to perform signature selection. We selected the smallest set of signatures which produced a large drop in the cost function with respect to the signature sets with one more member or the same number of signatures but a different composition. This heuristic aims to perform parsimonious signature selection while still accurately representing the data. The applied SignatureAnalyzer algorithm is available at the Broad Institute homepage (https://software.broadinstitute.org/cancer/cga/msp).

Protein extraction and western blotting. Cluster specific validation of protein expression levels was performed on samples from the respective cluster, not showing chromosomal aberrations other than del(13q) or gene mutations. For total protein extraction, cells were lysed in RIPA buffer (150 mM sodium chloride, 1% IGEPAL CA-630, 0.5% sodium deoxycholate, 0.1% SDS, 50mM Tris pH 8.0) supplemented with 1mM DTT, 0.5mM PMSF and phosphatase inhibitor cocktail, for 60 min at 4°C. The amount of protein in each sample was quantified using the Protein Assay (Bio-rad). Equal concentrations of proteins were analyzed on 12% polyacrylamide gels or 4-12% Nu-Page pre-cast gels and subsequently transferred onto PVDF membranes. The western blot images were acquired using the western gel documentation system. Antibodies used for western blot analysis in CLL cases not showing recurrent alterations include the following ones (catalogue numbers are provided in brackets). From Cell Signaling: anti-AKT (#9272), anti-phospho-AKT(Thr308) (#4056), anti-phospho-p53 (ser15) (#9286), anti-PRMT5 (#2252). From Abcam: anti-c-Myc (#ab32072), anti-yH2AX (#ab26350), anti-mouse Alexa Fluor 594 (#ab150116), anti-GAPDH (#ab8245). From Santa Cruz: anti-ERK1 (k-23) (#sc-94), anti-phospho-ERK(E-4) (#sc-7383), anti-LaminB (C-20) (#sc-6216), anti-RB(C-15) (#sc-50), anti-XPO1/CRM1 (H-300) (#sc-5595), anti-ß-Actin (#sc-1615). From BD Bioscience: anti-p53(CM5) (#554293). From Thermo Fisher: HRP-conjugated anti-mouse (#A16072). Anti-ERK1 (k-23) (#sc-94) and anti-ß-Actin (#sc-1615) were diluted 1:1000 in 5% BSA+TBST 0,1%, all other antibodies were diluted 1:500 in 5% BSA+TBST 0,1%, incubation was performed at 4°C over night. For image analysis of western blots the intensities of individual bands in western blots were analyzed using Fiji ImageJ densitometry software (version 1.51j). The levels of the proteins were expressed relative to the loading controls (Actin or Lamin B). Phosphorylation levels of proteins were expressed as a relative measure compared to that of the total protein and their respective loading controls (Actin or Lamin B).

Eμ-Myc/Eμ-TCL1 transgenic and BCL$_1$ / Eμ-TCL1 transplantation mouse models. Ethics oversight: All animal experiments were performed with the approval of the respective governmental authorities and local animal experimental ethics committees in each institution. The TCL1 serial transplant mouse model was performed according to protocols approved by the state government of Baden-Wuerttemberg, following the animal welfare guidelines (Registration 1124 and 1128) and were approved by the Ulm University animal experimental ethics committee. The BCL_1 syngeneic transplant model, Eµ-Myc and Eµ-TCL1 mouse model experiments were conducted under the Home Office licenses PPL30/2964 and P4D9C89EA following approval by local ethical committees, reporting to the Home Office Animal Welfare Ethical Review Board (AWERB) at the University of Southampton. Animals were maintained and bred in a pathogen-free environment (SPF IVC barrier) with a 14/10 day and night cycle, temperature at 21 °C and humidity at 55%, as well as water and food ad libitum.

BCL_1 syngeneic transplant model to validate induction of EMT-like networks in lymphoma: To validate the potential for induction of an EMT-like program and corresponding dynamics in B cell lymphoma cells, we used a syngeneic BCL_1 tumor transplant model. For this 1 x 10^5 BCL_1 tumor cells were inoculated IV into Balb/c mice and spleens harvested at defined periods of time (day 7 (n=12), 14 (n=6), 17 (n=6) and 21 (n=6) alongside naive tumor-free mice (n=4). Spleens were snap frozen and then sectioned to provide material from which to extract RNA. Total RNA was isolated, assessed for purity using NanoDrop at 260/280nm and the Bioanalyser. Samples with RIN scores >7 were taken forward and subjected to RNA sequencing (EA[2]).

Eµ-TCL1 serial transplant mouse model to assess EMT-like plasticity in lymphoma: The Eµ-TCL1 tumors prior to the start of the experiment were transferred and expanded once in syngeneic C57BL6/J mice. For serial transfers, 10 million splenic tumor cells were transplanted by intravenous injection and the animals were sacrificed when the mice appeared critically sick, a surrogate endpoint that was defined based on a scoring for disease severity including WBC count, changes in mobility, signs of suffering, as approved by the local animal experimental ethics committee. The tumor cells were purified with ficoll and only tumors with more than 90% $CD5^+CD19^+$ CLL cells were used for serial transfers and analyses. Three rounds of serial transfers were performed and the tumors were isolated from the spleen. The EMT markers Vimentin (*Vim*), Cadherin-1 (*Cdh1*) and corresponding transcription factors *Zeb1* and *Snai1* were measured using qPCR on tumors isolated from the different serial transfers.

Eµ-Myc / Eµ-TCL1 mouse model and proteome profiling to validate GI specific networks and associated inhibition of EMT-like networks: Mass spectrometry (MS) proteomics analyses of Eµ-Myc and Eµ-TCL1 tumors was performed as described[86] and submitted to GSEA. Spontaneous tumors from female Eµ-Myc [C57BL/6J-TgN(Ighmyc)22Bri/J] hemizygous and Eµ-TCL1 [C57BL/6J-TgN(IghTCL1)22Bri/J] hemizygous mice were compared with wildtype controls aged 6 weeks and 200 days, in addition to pre-terminal model controls taken at 6 weeks of age. B cells were isolated from spleens by magnetic isolation kit (Miltenyi Biotech, Bergisch Gladbach, Germany) and snap frozen. Samples were pooled with 4 tumors from each model assigned to 2 pools of 2 tumors and non-tumor pools of 6 samples, to be accommodated in a single 8-plex. Snap frozen cell pellets were lysed in 0.5 M TEAB with 0.05% SDS, with 100 µg of protein lysate per pool TCEP-reduced, MMTS-alkylated, trypsin digested and labelled with isobaric tags for relative and absolute quantitation (iTRAQ) 8-plex according to the manufactures instructions (ABSciex, Framingham, MA). Labelled peptides were combined and pre-fractionated using a 90-minute high-pH reverse-phase C8 fractionation (2-30%

organic) collecting 69 peak-dependent fractions. Each fraction was analyzed by LC-MS/MS (Dionex Ultimate 3000 and Orbitrap Elite (Thermo Scientific)) over 200 hours of MS time using top 12 data-dependent acquisition and 120,000 resolution with reporter ions captured with HCD at 35 keV at 15,000 resolution. Raw data was analyzed by Proteome Discoverer 1.4.1.14 with SequestHT 1.1.1.11 and Percolator modules searched against the mouse UniProt Swissprot and trembl databases (downloaded 01/15). The raw data and processed outputs are available at https://www.ebi.ac.uk/pride/archive/projects/PXD004608. Relative expression was assigned from iTRAQ reporter regions and were median normalized and quality-adjusted using spiquetool.com. $\log_2$ (ratios) were generated describing each tumor sample pool relative to the two WT control pools, in addition to the 6 week pre-terminal model controls relative to the 6 week WT control. A value summarizing all 4 $\log_2$ (ratios) for each tumor model was also used (mean/(standard deviation + 1)). Lists of gene names and corresponding $\log_2$ (ratios) and summary values for all 8270 proteins were analyzed by GSEA 3.0 using the GSEA-preranked approach due to a non-standard data format enriching for the MSigDB H and C2 gene sets. All default settings were used.

Radiation induced DNA damage. The human B cell lines, MEC1, MEC2, JVM2, JVM3, LCL-WEI, EHEB and Granta were purchased from the German Collection of Microorganisms and Cell culture (Deutsche Sammlung von Mikroorganismen und Zellkulturen, DSMZ) with the certificate from the vendor and were additionally authenticated through sequencing by Multiplexion GmbH. Cell lines were cultured in RPMI medium with 10% FCS and 1% L-glutamine / were maintained in IMDM medium with 10% FCS and 1% L-glutamine. All cell lines were tested for mycoplasma contamination monthly. For induction of DNA damage, the cells were irradiated with 5Gy γ-irradiation. The cells were collected 4, 8, 16, 24 and 48 hours after ionizing irradiation with 5Gy and expression changes in *miR-200c*, *TP53, TP63, ATM, ZEB1 and TWIST1* were analyzed for individual time points and in comparison to the corresponding non-irradiated sample.

Assessment on AID induced genomic instability. Cell culture: BL2 cell line was obtained from the German Collection of Microorganisms and Cell culture (ACC 625. Deutsche Sammlung von Mikroorganismen und Zellkulturen, DSMZ). BL2 *AICDA*- cells were kindly provided by Claude-Agnes Reynaud (INSERM U1151, Paris) and were described previously.[87] Human embryonic kidney HEK293T were obtained from European Collection of Authenticated Cell Cultures (ECACC) (Culture Collections, Public Health England, Salisbury, UK). Cell lines were authenticated by DSMZ (https://www.dsmz.de/collection/catalogue/human-and-animal-cell-lines/identity-control) or Public Health England (https://www.phe-culturecollections.org.uk/media/153328/ccw5704-culture-collections-quality-policy.pdf). BL2 AICDA- cells were not authenticated. Cells were cultured in RPMI 1640 (BL2) or Dulbecco's Modified Eagle Medium (HEK293T) (Sigma-Aldrich, Dorset, UK) supplemented with 10% (v/v) fetal bovine serum, 100 U/ml penicillin, 100 U/ml streptomycin at 37 °C in a humidified atmosphere containing 5% CO_2. All cell lines were tested for mycoplasma contamination monthly. Lentiviral transfection: Packaging transfection was performed in HEK293 cells using Lipofectamine® LTX Reagent (Invitrogen®, Carlsbad, California, USA) and the following vectors: the pRSV-Rev packaging, pMDLg packaging, pMD2.G envelope, and pLenti-C-mGFP vector expressing *AICDA*

(NM_020661) Human Tagged ORF Clone (RC202949L2, OriGene, Cambridge, UK). Lentivirus-containing medium was harvested 24 or 48 hours post-transfection, filtered through 0.45 μm PES filter and concentrated using Retro-X Concentrator (Clontech, Takara Bio, USA). After overnight incubation at +4℃, the virus-containing mixture was centrifuged for 45 minutes at 500g and pellet was resuspended with the medium in 1:10 of the original volume. 107 of BL2 *AICDA*- or HEK 293T cells were then exposed to a virus containing medium for 24 hours. Infection efficiency was then determined by flow cytometry as a percentage of GFP+ cells. AID-GFP cells were then FACS sorted using BD FACSAria II Cell Sorter (BD Biosciences, Wokingham, UK). Immunofluorescence: For immunofluorescence staining, cells were harvested and cytospun onto microscope slides. Slides were fixed in -20°C methanol and washed in TBS/0.05% Tween. Primary mouse anti-yH2AX (ab26350, Abcam, Cambridge, UK) antibody, diluted 1:500 in TBS/0.1% BSA, was then applied for 1 hour at room temperature. After three washing steps with TBS/0.05%Tween, slides were stained with secondary anti-mouse Alexa Fluor 594 (ab150116, Abcam, Cambridge, UK), diluted 1:500, for 1 hour at room temperature.

Following three times with TBS/0.05% Tween. Nuclei were counterstained with DAPI and slides were mounted in ProLong Gold Antifade Reagent (Invitrogen, Life Technologies, Paisley, UK). Images were taken using Nikon Ci-L upright fluorescence microscope and Nikon NIS Elements AR software (Ver4.30.01, 64bit edition). Assessment of y-H2AX was used as universal marker of DNA damage, including DNA double strand breaks.[88] Western blot: Total protein lysates were mixed with NuPAGE® LDS Sample Buffer (Thermo Fisher Scientific, Paisley, UK), and then ten µg of protein were resolved on precast 4-12% Bis-Tris Protein Gel (Thermo Fisher Scientific, Paisley UK). Proteins were transferred by wet transfer onto Immobilon-P Membrane PVDF membrane (Merck Millipore, Billerica, Massachusetts, USA), blocked with 5% non-fat skim milk in TBS and then the membrane was incubated with appropriately diluted primary antibodies for one hour at room temperature in TBS/0.1% Tween. After three washing steps, membranes were incubated with HRP-conjugated anti-mouse antibody (1:2000, A16072, Thermo Fisher, Paisley, UK). After three washing steps, membrane chemiluminescence was analyzed by Amersham ECL Prime Western Blotting Detection Reagent (GE Healthcare®, Little Chalfont, UK) and ChemiDoc imaging system (BioRad, Hemel Hempstead, UK). The following primary antibodies were used: mouse anti-y-H2AX [9F3] (1:1500, ab26350, Abcam, Cambridge, UK), mouse monoclonal [6C5] to GAPDH (1:3000, ab8245, Abcam, Cambridge, UK) .Sister Chromatid Exchange assay: We used the sister chromatid exchange (SCE) assay to assess cellular genotoxicity and ongoing mutagenesis.[89,90] Cells were cultured in DMEM containing 10 µM 5-bromo-2'-deoxyuridine BrdU (ab142567 Abcam, Cambridge, UK) for two cell divisions cycles. After 4 hours treatment with Colcemid (0.02 µg/mL, 10295892001, Sigma-Aldrich, Gillingham, UK), cells were incubated with prewarmed 75 mM KCl solution for 20 min at 37 °C. Then, cells were spun down and fixed using Carnoy's fixative (methanol: glacial acetic acid). Mitotic cells were dropped on pre-chilled microscope slides and left to dry in the dark at room temperature. Slides were immersed in acridine orange (A1301, Thermo Fisher Scientific, Paisley, UK) for 5 min, mounted in 2×SSC buffer and covered with a coverslip. Images of chromosome spreads were obtained using a Nikon Ci-L upright fluorescence microscope and Nikon NIS Elements AR software (Ver4.30.01, 64bit edition). Sister chromatid exchanges were quantified microscopically from at least five random fields containing at least 20 metaphase spreads. Alkaline Single Cell Electrophoresis (Comet) assay: Single Cell Electrophoresis was performed using CometAssay Kit (4250-050-K, Trevigen, AMS Biotechnology, Abingdon, UK) as per manufacturer's instructions. Briefly, 105 cells were harvested by centrifugation and then resuspended in 1X PBS, mixed with 0.5% low melting agarose, and plated onto comet slides pre-coated with normal melting point agarose. Subsequently, slides were immersed in comet lysis solution for one hour at 4°C. Slides were then further treated with alkaline solution (NaOH pH 13, 200mM EDTA in H2O) submerging in electrophoresis tank. Alkaline electrophoresis was performed for 30-60 min at 21V (300mA), Nuclei and "comet tails" were subsequently stained with SYBR® Gold stain (Thermo Fisher Scientific, Paisley, UK). Slides were visualized using fluorescence microscopy, and % of tail DNA on per cell basis was determined using an open source Cell Profiler (https://cellprofiler.org/) software equipped with the Comet Assay analysis module (https://cellprofiler-examples.s3.amazonaws.com/ExampleCometAssay.zip). Triplicate slides were processed per each *AICDA*-related condition with at least 70 comets analyzed per each condition.

Telomere length analysis. Telomere length measurement was carried out using a qPCR–based technique[91]. The primers used to amplify telomere and single-copy genes (SCG) were tel1b, tel2b and HBG3, HBG4, respectively.[91] The absolute telomere length was obtained by using synthetic oligonucleotide standards for telomere (84 bp) and SCG (81 bp) PCR. Briefly, a 10 fold dilution of the telomere and SCG standard was prepared and the amount of DNA molecules in each standard was calculated as described.[92] 12ng of DNA was used per reaction (total volume of 10µl) in triplicates for the telomere and SCG PCRs and amplified using Qiagen quantitect SYBR green in 384 well plates and analyzed using 7900HT fast real-time PCR system (Applied Biosystems). Six telomere length controls with known telomere length, analyzed using terminal restriction fragment length analysis (TRF) were included in every plate to detect variations. The qPCR technique was validated by terminal restriction fragment length (TRF) analysis and Southern hybridization. 6µg of non-degraded DNA was digested overnight using Hinf I and Rsa I and resolved on a 0.8% agarose gel. In gel hybridization was carried out by drying the gel and hybridizing with a telomere specific probe, end labelled with ^{32}P. The mean telomere length was analyzed from the autoradiograph. A correlation of R^2=0.8516 was obtained upon comparison of telomere length measured using qPCR and TRF, in a control sample set (n=18). The TRF value of telomere length for each sample was calculated from the linear regression of qPCR versus TRF.

Reduced representation bisulfite sequencing and methylation analysis (RRBS). Genomic DNA from n=182 matched CLL8 samples was used to produce RRBS libraries. They were generated by digesting genomic DNA with MspI to enrich for CpG-rich fragments, and then ligated to barcoded TruSeq adapters (Illumina) to allow immediate subsequent pooling. It was followed by bisulfite conversion and PCR, as previously described[93]. Libraries were sequenced and aligned to the bisulfite-converted hg19 reference genome using Bismark (RRID: SCR_005604) v0.15.0[94].

Methylation analysis: Only CpGs with 10 or more reads were included into the analysis. Promoters were defined as the regions encompassing 2 kb upstream and downstream of the transcription start site of UCSC genes. Promoters or genes with at least 5 covered CpGs were included into the analysis. Promoter or gene methylation was calculated by the average methylation levels of all the CpGs inside.

Survival analysis. CLL8 and REACH clinical trial data were analyzed on an intention-to-treat basis (eligible subjects were analyzed as randomized). Progression free survival (PFS) was defined as the time from randomization to disease progression or death, overall survival (OS) was defined as the time between randomization and death. PFS and OS were estimated by the Kaplan–Meier method, and differences between groups were assessed using two-sided non-stratified log-rank tests. Additionally, hazard ratios (HR) and 95% confidence intervals (CI) were calculated using Cox regression modeling. With regard to overall survival (OS) and progression-free survival (PFS) multivariable Cox proportional hazards regression models were used to assess the independent prognostic value of identified major subtypes (GI, (I)GI, EMT-L, (I)EMT-L) in CLL8 and REACH. Additional prognostic factors in the models were treatment, *TP53* mutation, IGHV mutation, 17p deletion, 11q deletion, trisomy 12q and 13q deletion.

Statistical software. Statistical analysis was performed with R version 3.3.3 and 3.4.1, with R package survival, version 2.41-2; SPSS version 24-26 (IBM, NYC, NY); Prism software version 6.0h (GraphPad), MATLAB 2018b and BRB-ArrayTools Version 4.2.1-4.6.1.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability. Complete data sets are available: For GEP at Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/; GEO accession number: GSE58211 (REACH only); GSE126595 (full clinical data set); GSE126699 (functional data). For SNP-Microarray raw data at Gene Expression Omnibus (GEO accession number: GSE36908 (CLL8 treatment naive) and GSE83566 (relapsed)). CLL8 WES data is deposited in dbGaP under accession code phs000922.v1.p1. CLL8 RRBS sequencing data is available from the NCBI (GEO accession number GSE143673). The proteome profiling raw data and processed outputs are available at https://www.ebi.ac.uk/pride/archive/projects/PXD004608. All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding first author upon reasonable request. A reporting summary for this article is available as a Supplementary Information file.

References

1. Edelmann, J. *et al.* High-resolution genomic profiling of chronic lymphocytic leukemia reveals new recurrent genomic alterations. *Blood* 120, (2012).
2. Landau, D. A. *et al.* Mutations driving CLL and their evolution in progression and relapse. *Nature* 526, 525–530 (2015).
3. Stilgenbauer, S. *et al.* Gene mutations and treatment outcome in chronic lymphocytic leukemia: results from the CLL8 trial. *Blood* 123, 3247–3254.
4. Skowronska, A. *et al.* Biallelic ATM inactivation significantly reduces survival in patients treated on the United Kingdom Leukemia Research Fund Chronic Lymphocytic Leukemia 4 trial. *J. Clin. Oncol.* 30, 4524–32 (2012).
5. Stankovic, T. *et al.* Ataxia telangiectasia mutated-deficient B-cell chronic lymphocytic leukemia occurs in pregerminal center cells and results in defective damage response and unrepaired chromosome damage. *Blood* (2002) doi:10.1182/blood.V99.1.300.
6. Zenz, T. *et al.* miR-34a as part of the resistance network in chronic lymphocytic leukemia. *Blood* 113, 3801–3808 (2009).
7. Ouillette, P. *et al.* Clonal evolution, genomic drivers, and effects of therapy in chronic lymphocytic leukemia. *Clin. Cancer Res.* (2013) doi:10.1158/1078-0432.CCR-13-0138.
8. Rossi, D. *et al.* Clinical impact of small TP53 mutated subclones in chronic lymphocytic leukemia. *Blood* (2014) doi:10.1182/blood-2013-11-539726.
9. Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* (2013) doi:10.1016/j.cell.2013.01.019.
10. Knight, S. J. L. *et al.* Quantification of subclonal distributions of recurrent genomic aberrations in paired pre-treatment and relapse samples from patients with b-cell chronic lymphocytic leukemia. *Leukemia* (2012) doi:10.1038/leu.2012.13.
11. Damm, F. *et al.* Acquired initiating mutations in early hematopoietic cells of CLL patients. *Cancer Discov.* (2014) doi:10.1158/2159-8290.CD-14-0104.
12. Wang, L. *et al.* Somatic mutation as a mechanism of Wnt/β-catenin pathway activation in CLL. *Blood* (2014) doi:10.1182/blood-2014-01-552067.
13. Kulis, M. *et al.* Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat. Genet.* (2015) doi:10.1038/ng.3291.
14. Oakes, C. C. *et al.* DNA methylation dynamics during B cell maturation underlie

a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet.* (2016) doi:10.1038/ng.3488.

15. Fischer, K. *et al.* Long-term remissions after FCR chemoimmunotherapy in previously untreated patients with CLL: Updated results of the CLL8 trial. in *Blood* vol. 127 208–215 (2016).

16. Robak, T. *et al.* Rituximab plus fludarabine and cyclophosphamide prolongs progression-free survival compared with fludarabine and cyclophosphamide alone in previously treated chronic lymphocytic leukemia. *J. Clin. Oncol.* 28, 1756–65 (2010).

17. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc. Natl. Acad. Sci.* 104, 20007–20012 (2007).

18. Ouillette, P. *et al.* Integrated genomic profiling of chronic lymphocytic leukemia identifies subtypes of deletion 13q14. *Cancer Res.* 68, 1012–1021 (2008).

19. Ramsay, A. J. *et al.* POT1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. *Nat. Genet.* 45, 526–530 (2013).

20. Lee, C. S. *et al.* Recurrent point mutations in the kinetochore gene KNSTRN in cutaneous squamous cell carcinoma. *Nat. Genet.* 46, 1060–1062 (2014).

21. Baker, D. J. *et al.* Increased expression of BubR1 protects against aneuploidy and cancer and extends healthy lifespan. *Nat. Cell Biol.* 15, 96–102 (2013).

22. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* 500, 415–21 (2013).

23. Lee, S. T. *et al.* Context-specific regulation of NF-κB target gene expression by EZH2 in breast cancers. *Mol. Cell* 43, 798–810 (2011).

24. Ea, C.-K. & Baltimore, D. Regulation of NF-?B activity through lysine monomethylation of p65. *Proc. Natl. Acad. Sci.* 106, 18972–18977 (2009).

25. Chang, C. J. *et al.* P53 regulates epithelial-mesenchymal transition and stem cell properties through modulating miRNAs. *Nat. Cell Biol.* 13, 317–323 (2011).

26. Vermezovic, J. *et al.* Notch is a direct negative regulator of the DNA-damage response. *Nat. Struct. Mol. Biol.* 22, 417–424 (2015).

27. Pastushenko, I. *et al.* Identification of the tumour transition states occurring during EMT. *Nature* (2018) doi:10.1038/s41586-018-0040-3.

28. Rodriguez, J. *et al.* Chromosomal Instability Correlates with Genome-wide DNA Demethylation in Human Primary Colorectal Cancers. *Cancer Res.* 66, 8462–

9468 (2006).

29. Eden, A., Gaudet, F., Waghmare, A. & Jaenisch, R. Chromosomal Instability and Tumors Promoted by DNA Hypomethylation. *Science (80-.).* 300, 455–455 (2003).

30. Fabris, S. *et al.* Biological and clinical relevance of quantitative global methylation of repetitive DNA sequences in chronic lymphocytic leukemia. *Epigenetics* 6, 188–94 (2011).

31. Espada, J. *et al.* Regulation of SNAIL1 and E-cadherin function by DNMT1 in a DNA methylation-independent context. *Nucleic Acids Res.* (2011) doi:10.1093/nar/gkr658.

32. Fang, M., Hutchinson, L., Deng, A. & Green, M. R. Common BRAF(V600E)-directed pathway mediates widespread epigenetic silencing in colorectal cancer and melanoma. *Proc. Natl. Acad. Sci. U. S. A.* 113, 1250–5 (2016).

33. Palamarchuk, A. *et al.* Tcl1 protein functions as an inhibitor of de novo DNA methylation in B-cell chronic lymphocytic leukemia (CLL). *Proc. Natl. Acad. Sci. U. S. A.* 109, 2555–60 (2012).

34. Cardenas, H. *et al.* TGF-β induces global changes in DNA methylation during the epithelial-to-mesenchymal transition in ovarian cancer cells. *Epigenetics* 9, 1461–1472 (2014).

35. Vilagos, B. *et al.* Essential role of EBF1 in the generation and function of distinct mature B cell types. *J. Exp. Med.* 209, 775–92 (2012).

36. Liu, M. *et al.* Two levels of protection for the B cell genome during somatic hypermutation. *Nature* 451, 841–845 (2008).

37. Patten, P. E. M. *et al.* IGHV-unmutated and IGHV-mutated chronic lymphocytic leukemia cells produce activation-induced deaminase protein with a full range of biologic functions. *Blood* (2012) doi:10.1182/blood-2012-08-449744.

38. Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* (2015) doi:10.1038/ncomms9866.

39. Li, Y. *et al.* PRMT5 is required for lymphomagenesis triggered by multiple oncogenic drivers. *Cancer Discov.* 5, 288–303 (2015).

40. Kantidakis, T. *et al.* Mutation of cancer driver MLL2 results in transcription stress and genome instability. *Genes Dev.* 30, 408–20 (2016).

41. Lee, J. *et al.* A tumor suppressive coactivator complex of p53 containing ASC-2

and histone H3-lysine-4 methyltransferase MLL3 or its paralogue MLL4. *Proc. Natl. Acad. Sci. U. S. A.* 106, 8513–8 (2009).

42. Chuikov, S. *et al.* Regulation of p53 activity through lysine methylation. *Nature* 432, 353–360 (2004).

43. Huang, J. *et al.* Repression of p53 activity by Smyd2-mediated methylation. *Nature* 444, 629–632 (2006).

44. Marshall, M. J. E., Stopforth, R. J. & Cragg, M. S. Therapeutic antibodies: What have we learnt from targeting CD20 and where are we going? *Frontiers in Immunology* (2017) doi:10.3389/fimmu.2017.01245.

45. Roccaro, A. M. *et al.* CXCR4 regulates extra-medullary myeloma through epithelial-mesenchymal-transition-like transcriptional activation. *Cell Rep.* (2015) doi:10.1016/j.celrep.2015.06.059.

46. Azab, A. K. *et al.* Hypoxia promotes dissemination of multiple myeloma through acquisition of epithelial to mesenchymal transition-like features. *Blood* (2012) doi:10.1182/blood-2011-09-380410.

47. Sánchez-Tilló, E. *et al.* The EMT activator ZEB1 promotes tumor growth and determines differential response to chemotherapy in mantle cell lymphoma. *Cell Death Differ.* (2014) doi:10.1038/cdd.2013.123.

48. Thathia, S. H. *et al.* Epigenetic inactivation of TWIST2 in acute lymphoblastic leukemia modulates proliferation, cell survival and chemosensitivity. *Haematologica* (2012) doi:10.3324/haematol.2011.049593.

49. Zhao, L., Liu, Y., Zhang, J., Liu, Y. & Qi, Q. LncRNA SNHG14/miR-5590-3p/ZEB1 positive feedback loop promoted diffuse large B cell lymphoma progression and immune evasion through regulating PD-1/PD-L1 checkpoint. *Cell Death Dis.* (2019) doi:10.1038/s41419-019-1886-5.

50. Huang, W. T., Kuo, S. H., Cheng, A. L. & Lin, C. W. Inhibition of ZEB1 by miR-200 characterizes Helicobacter pylori-positive gastric diffuse large B-cell lymphoma with a less aggressive behavior. *Mod. Pathol.* (2014) doi:10.1038/modpathol.2013.229.

51. Raval, A. *et al.* TWIST2 demonstrates differential methylation in immunoglobulin variable heavy chain mutated and unmutated chronic lymphocytic leukemia. *J. Clin. Oncol.* (2005) doi:10.1200/JCO.2005.02.196.

52. Gaiti, F. *et al.* Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* (2019) doi:10.1038/s41586-019-1198-z.

53. Li, J. *et al.* The EMT transcription factor Zeb2 controls adult murine hematopoietic differentiation by regulating cytokine signaling. *Blood* (2017) doi:10.1182/blood-2016-05-714659.

54. Herishanu, Y. *et al.* The lymph node microenvironment promotes B-cell receptor signaling, NF-kappaB activation, and tumor proliferation in chronic lymphocytic leukemia. *Blood* 117, 563–574.

55. Ghia, P., Granziero, L., Chilosi, M. & Caligaris-Cappio, F. Chronic B cell malignancies and bone marrow microenvironment. *Seminars in Cancer Biology* vol. 12 149–155 (2002).

56. Valsecchi, R. *et al.* HIF-1α regulates the interaction of chronic lymphocytic leukemia cells with the tumor microenvironment. *Blood* (2016) doi:10.1182/blood-2015-07-657056.

57. Arruga, F. *et al.* Functional impact of NOTCH1 mutations in chronic lymphocytic leukemia. *Leukemia* 28, 1060–1070 (2014).

58. Kalluri, R. & Weinberg, R. A. The basics of epithelial-mesenchymal transition. *Journal of Clinical Investigation* (2009) doi:10.1172/JCI39104.

59. Brabletz, T. To differentiate or not-routes towards metastasis. *Nature Reviews Cancer* (2012) doi:10.1038/nrc3265.

60. Suarez-Carmona, M., Lesage, J., Cataldo, D. & Gilles, C. EMT and inflammation: inseparable actors of cancer progression. *Molecular Oncology* (2017) doi:10.1002/1878-0261.12095.

61. Beers, S. A. *et al.* Type II (tositumomab) anti-CD20 monoclonal antibody out performs type I (rituximab-like) reagents in B-cell depletion regardless of complement activation. *Blood* (2008) doi:10.1182/blood-2008-04-149161.

62. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* (2016) doi:10.1016/j.cell.2016.02.065.

63. Wang, L. *et al.* EMT- and stroma-related gene expression and resistance to PD-1 blockade in urothelial cancer. *Nat. Commun.* (2018) doi:10.1038/s41467-018-05992-x.

64. Wu, W. S. *et al.* Slug antagonizes p53-mediated apoptosis of hematopoietic progenitors by repressing puma. *Cell* (2005) doi:10.1016/j.cell.2005.09.029.

65. Lee, S. H. *et al.* Blocking of p53-snail binding, promoted by oncogenic K-Ras, recovers p53 expression and function. *Neoplasia* (2009) doi:10.1593/neo.81006.

66. Kajita, M., McClinic, K. N. & Wade, P. A. Aberrant Expression of the Transcription Factors Snail and Slug Alters the Response to Genotoxic Stress. *Mol. Cell. Biol.* (2004) doi:10.1128/mcb.24.17.7559-7566.2004.

67. He, L. *et al.* A microRNA component of the p53 tumour suppressor network. *Nature* 447, 1130–1134 (2007).

68. Hermeking, H. The miR-34 family in cancer and apoptosis. *Cell Death and Differentiation* (2010) doi:10.1038/cdd.2009.56.

69. Kim, T. *et al.* p53 regulates epithelial-mesenchymal transition through microRNAs targeting ZEB1 and ZEB2. *J. Exp. Med.* (2011) doi:10.1084/jem.20110235.

70. Brabletz, S. & Brabletz, T. The ZEB/miR-200 feedback loop-a motor of cellular plasticity in development and cancer? *EMBO Reports* (2010) doi:10.1038/embor.2010.117.

71. Bracken, C. P. *et al.* A double-negative feedback loop between ZEB1-SIP1 and the microRNA-200 family regulates epithelial-mesenchymal transition. *Cancer Res.* (2008) doi:10.1158/0008-5472.CAN-08-1942.

72. Burk, U. *et al.* A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells. *EMBO Rep.* (2008) doi:10.1038/embor.2008.74.

73. Morel, A. P. *et al.* A stemness-related ZEB1-MSRB3 axis governs cellular pliancy and breast cancer genome stability. *Nat. Med.* (2017) doi:10.1038/nm.4323.

74. Bengtsson, H., Simpson, K., Bullard, J. & Hansen, K. aroma.affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. *Methods* Tech Repor, 1–9 (2008).

75. Monti, S. *et al.* Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Mach. Learn.* 52, 91–118 (2003).

76. Sturn, A., Quackenbush, J. & Trajanoski, Z. Genesis: cluster analysis of microarray data. *Bioinformatics* 18, 207–208 (2002).

77. 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).

78. Edelmann, J. *et al.* Frequent evolution of copy number alterations in CLL

following first-line treatment with FC(R) is enriched with TP53 alterations: results from the CLL8 trial. *Leukemia* 31, 734–738 (2017).

79. Pounds, S. *et al.* Reference alignment of SNP microarray signals for copy number analysis of tumors. *Bioinformatics* 25, 315–321 (2009).

80. Lin, M. *et al.* dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics* 20, 1233–1240 (2004).

81. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5, 557–572 (2004).

82. Bengtsson, H., Irizarry, R., Carvalho, B. & Speed, T. P. Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* 24, 759–767 (2008).

83. Kent, W. *et al.* UCSC Genome Browser. *Hum. genome Brows. UCSC. Genome Res.* (2002).

84. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* (2011) doi:10.1038/nbt.1754.

85. Seung, H. S. & Lee, D. D. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791 (1999).

86. Johnston, H. E. *et al.* Integrated Cellular and Plasma Proteomics of Contrasting B-cell Cancers Reveals Common, Unique and Systemic Signatures. *Mol. Cell. Proteomics* (2017) doi:10.1074/mcp.M116.063511.

87. Faili, A. *et al.* AID-dependent somatic hypermutation occurs as a DNA single-strand event in the BL2 cell line. *Nat. Immunol.* (2002) doi:10.1038/ni826.

88. Rogakou, E. P., Pilch, D. R., Orr, A. H., Ivanova, V. S. & Bonner, W. M. Double-stranded Brekas Induce Histone H2AX phosphorylation on Serine 139. *J. Biol. Chem.* (1998) doi:10.1074/jbc.273.10.5858.

89. Mezzanotte, R. & Nieddu, M. A historical overview of bromo-substituted DNA and sister chromatid differentiation. in *Methods in Molecular Biology* (2014). doi:10.1007/978-1-62703-706-8_8.

90. Carrano, A. V., Thompson, L. H., Lindl, P. A. & Minkler, J. L. Sister chromatid exchange as an indicator of mutagenesis. *Nature* (1978) doi:10.1038/271551a0.

91. Cawthon, R. M. Telomere measurement by quantitative PCR. *Nucleic Acids Res.* 30, e47 (2002).

92. O’Callaghan, N., Dhillon, V., Thomas, P. & Fenech, M. A quantitative real-time PCR method for absolute telomere length. *Biotechniques* 44, 807–9 (2008).

93. Boyle, P. *et al.* Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. *Genome Biol.* 13, R92 (2012).

94. Krueger, F. & Andrews, S. R. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27, 1571–1572 (2011).

Acknowledgements: The authors thank all patients and their physicians for trial participation and donation of samples; the DCLLSG; Sabrina Schrell and Christina Galler for their excellent technical assistance.

This work was supported by research grants from the Else Kröner-Fresenius-Stiftung (2010_Kolleg24, 2012_A146), BMBF (PRECISE), Cancer Research UK (A24721 and A18087), Kay Kendall Leukemia Fund (KKL1101) and Blood Cancer UK (19001), European Commission / BMBF (“FIRE CLL”, 01KT160), SFB 1074 (Projects B1, B2), DJCLS R 11/01.

Conflict of Interest: The authors declare that there are no conflicts that interfered with the experiments and presentation of data.

Additional information:

Supplementary information is available for this paper at………

Figure Legends

Figure 1: Composition and relationship of CLL subtypes in clustered data

A) Schematic representation for analysis, identification of CLL subtypes in the CLL8 and confirmation in the REACH cohort. The four largest clusters (GI, (I)GI, EMT-L, (I)EMT-L) and associations of *NRIP1* with the inflammatory or tri(12) with the EBF1-r signature were also identified in the independent validation cohort of the REACH trial. Co-clustering of GI/(I)GI and EMT-L/(I)EMT-L cases in the REACH cohort supports the selection of subgroup-specific characteristics during treatment. B) Heatmap showing the consensus clustering for k=6 used for defining CLL subtypes (n=337). Distribution of genetic characteristics is shown below the heatmap. Significant enrichment of variables in clusters is observed for del(17p) (p=0.05), *TP53* mutation (p=0.01), tri(12) (p=7e-06), del(13q) (p=0.03) and IGHV mutation status (p=0.008) (all Fisher`s exact test (two-sided)). *TP53* frameshift mutations occur exclusively in GI and splice site mutations in EBF1-r cases. Tri(12) is strongly overrepresented in EBF1-r (72.7%). C) Telomere length is significantly different across CLL subtypes (p<0.001, Kruskal-Wallis chi-squared) and shortest length is observed in GI with median 3.8 kb (p=0.003, Mann-Whitney (two-sided), for GI vs. (I)EMT-L) (n=333). D) White blood cell counts are significantly different across CLL subtypes (p<0.0001, Kruskal-Wallis chi-squared), show decreased counts in inflammatory CLL and are lowest in (I)EMT-L with median 61.1 G/L (p<0.0001, Mann-Whitney (two-sided), for GI vs. (I)EMT-L) (n=330). For Fig. A-D, data within individual figures derives from biologically independent samples. For the boxplots, centerline, box limits, and whiskers represent the median, 25th, and 75th percentiles and 1.5x interquartile range, respectively.

Figure 2: Pathway activation and genetic alterations in CLL subtypes

A) Heatmap showing overrepresented gene sets (FDR<0.05) identified through GSEA. The intensity range of normalized enrichment scores (NES) illustrates the degree of enrichment in CLL subtypes. Gene sets are grouped together according to the biological context. B) GISTIC analysis of copy number alterations (CNA). Chromosomal positions (1-22) on the y-axis (left) indicate losses (blue, upper panels) or gains (red, lower panels) for major clusters. Affected genes representing CNA targets within biological networks (such as *YAP1*) are shown for respective peaks. Most significant chromosomal peaks for major clusters are indicated on the right of each panel. GISTIC q-values at each locus are plotted from left to right on a log scale (bottom of each panel). Altered regions with FDR $q \leq 0.25$ (vertical green line) are considered significant. GISTIC G-Scores (amplitude of the aberration x frequency of its occurrence across samples) are plotted on top of the panels. C) Heatmap showing GEP of genes located within or adjacent to the MDR or MGR of recurrent aberrations (n=337). FDRs for DEGs ((I)EMT-L vs. GI) are highly significant (q<1e-07). D) Heatmap showing GEP of genes located at/adjacent to the MDR on 13q (n=335). Blue color code indicates deletions (dark blue: biallelic; light blue: monoallelic) and absence of del(13q) (cyan blue). Genes are ordered corresponding to chromosomal positions for the region between *DLEU1* and *RB1.* E) Visualization of del(13q) per case (blue horizontal lines). Y-axis: cluster color code, x-axis: representative genes and topography for the cumulative coverage of segment breaks per cluster. Vertical black dotted line indicates the *RB1* locus. Vertical red dotted lines indicate the majority of distal losses (around 50-51 mb). Losses extending to the distal end of cytoband 13q14.3 (orange dotted line) are variably distributed. LDBs involve/exceed the majority of cytoband 13q21.1 (distal of 54.7 mb). Biallelic deletions of 13q14 mostly cover a small region and rarely occur together with larger 13q deletions. For Fig. A-E, data within individual figures derives from biologically independent samples.

Figure 3: Biological processes operational in genomically instable CLL

A-D) Heatmap showing expression profiles (n=337) for genes involved in A) mismatch repair (MMR), B) base excision repair (BER), C) nucleotide excision repair (NER), D) non-homologous end joining (NHEJ). FDRs of DEGs (GI vs. (I)EMT-L) are indicated (q). Single genes may be involved in multiple processes. E) Protein expression in CLL subtypes; p53 (n=4 each), phospho-p53 (GI/(I)EMT-L: n=11 each, (I)GI/EMT-L: n=8 each) (normalized to actin). F/G) Models summarizing alterations which contribute to F) genomic instability, G) activation of MYC family members in GI/(I)GI. Source/method (e.g. mRNA/GISTIC) and significance/frequencies are shown in grey, along with estimated mode of regulation/biological effect (red: increase/activation; blue: decrease/inactivation). H) Protein expression in CLL subtypes; PRMT5 (n=4 each), XPO1/cMYC (GI/(I)EMT-L: n=11 each, (I)GI/EMT-L: n=8 each) (normalized to actin). I) Heatmap showing CNAs (paired CLL8 cases; before treatment (pre), at relapse (post)). CNAs are frequent at relapse (*TP53* wild-type: GI/(I)EMT-L; p<0.05, Wilcoxon signed-rank test (two-sided)). GI cases show more aberrations (mean) before ((I)EMT-L: 0.83; GI: 1.71) and after treatment ((I)EMT-L: 2.17; GI: 3.43), with considerable increase after treatment when *TP53* mutations are included (GI(pre): 1.91, GI(post): 4.36; p<0.01, Wilcoxon signed-rank test (two-sided)). Arrowheads highlight *TP53* inactivation (preexisting: red; acquired: blue). GI alterations often involve chromosomes other than 13, 12, 11, 17. J) Fraction of signature activations in CLL subtypes (EBF1-r n=6, GI n=68, EMT-L n=11, (I)EMT-L n=52, (I)GI n=31, NRIP1 n=3) and K) activation levels of mutational signatures (median centered) with order/color code according to clusters and IGHV status (light orange: IGHV mutated, yellow box: GI/(I)GI/(I)EMT-L). IGHV mutated cases show higher activations for signatures 9, 3, 15, 20. L) Alterations in DNA-damage response genes are more frequent in IGHV mutated GI/(I)GI cases (57%/50% ≥ one alteration) vs. IGHV mutated (I)EMT-L cases (11% ≥ one alteration) ($p<0.05$, Mann-Whitney (two-sided)). IGHV unmutated cases (either subtype) have similar frequencies (64-70% ≥ one alteration). Only cases with WES for respective genes and known *TP53* status (excluding EBF1-r/NRIP1) were used (n=156). For Fig. A-L, data within individual figures derives from biologically independent samples. For boxplots, centerline, box limits, and whiskers represent the median, 25th, and 75th percentiles and 1.5x interquartile range, respectively.

Figure 4: Biological processes operational in CLL with EMT-like networks

A-D) Heatmap showing expression profiles (n=337) for A) genes indicating activated TNFα/NF-kB signaling, B) EMT-TFs, C) NOTCH target genes (intensity range -1:1), D) histone lysine methyltransferases. FDRs of DEGs (GI vs. (I)EMT-L) are indicated on the right (q). E) GEP of the CD19 positive (+) and negative (-) compartment from CLL samples with inflammatory (I) and non-inflammatory (NI) signatures. F) GEP of 2374 variably expressed genes for the CD19 negative fraction. G) Tumor GEP indicating activated TNFα/NF-kB signaling and induction of EMT-like programs after BCL_1 tumor transplantation. Y-axis: median centered expression, x-axis: days (d) after transplantation of individual samples ($p<0.05$ shown for d7 vs. d21, Mann-Whitney (two-sided)). H) Heatmap showing gene set enrichment, characteristic for CLL with genomic instability or activation of EMT-like programs, in Eμ-Myc and Eμ-TCL1 mice. Gene sets were identified through GSEA (FDR<0.05) on proteome profiles of splenic tumor cells from leukemic or terminal Eμ-Myc and Eμ-TCL1 mice (n=4, in two pools) compared with B cells of tumor free wild-type mice (n=12, in two pools). Color coded normalized enrichment scores (NES) illustrate the degree of enrichment (positive: yellow to red; negative: blue, light to dark). Gene sets are grouped together according to the biological context. I) Model illustrating biologic characteristics and regulatory interplay of processes in identified subgroups as specified for GI and (I)EMT-L CLL in respective results sections. J) Topographical landscape of expression profiles for 2359 variably expressed genes (SD>0.5). Genes with the highest significance (q<1e-05) for the EMT-L, EBF1-r and NRIP1 cluster are indicated. Fold change (FC) is indicated for EBF1-r specific genes (EBF1-r vs. all other). K) Piecharts illustrating global gene expression with percentages indicating over- or under-expression in relation to the median expression per gene across the dataset. L) Heatmap showing genes (n=69) with strongest differential expression (q≤0.05, FC≥2) between the EBF1-r vs. all other clusters. CD19 sorted healthy donor B cells are included (orange). Arrowheads indicate cases with tri(12). M) Agglomerative hierarchical clustering (2359 genes, Pearson complete) for n=337 CLL and n=5 healthy donor B cells (orange). For Fig. A-F/H-M, data within individual figures derives from biologically independent samples.

Figure 5: CLL subtype, genetic markers and treatment outcome in CLL8

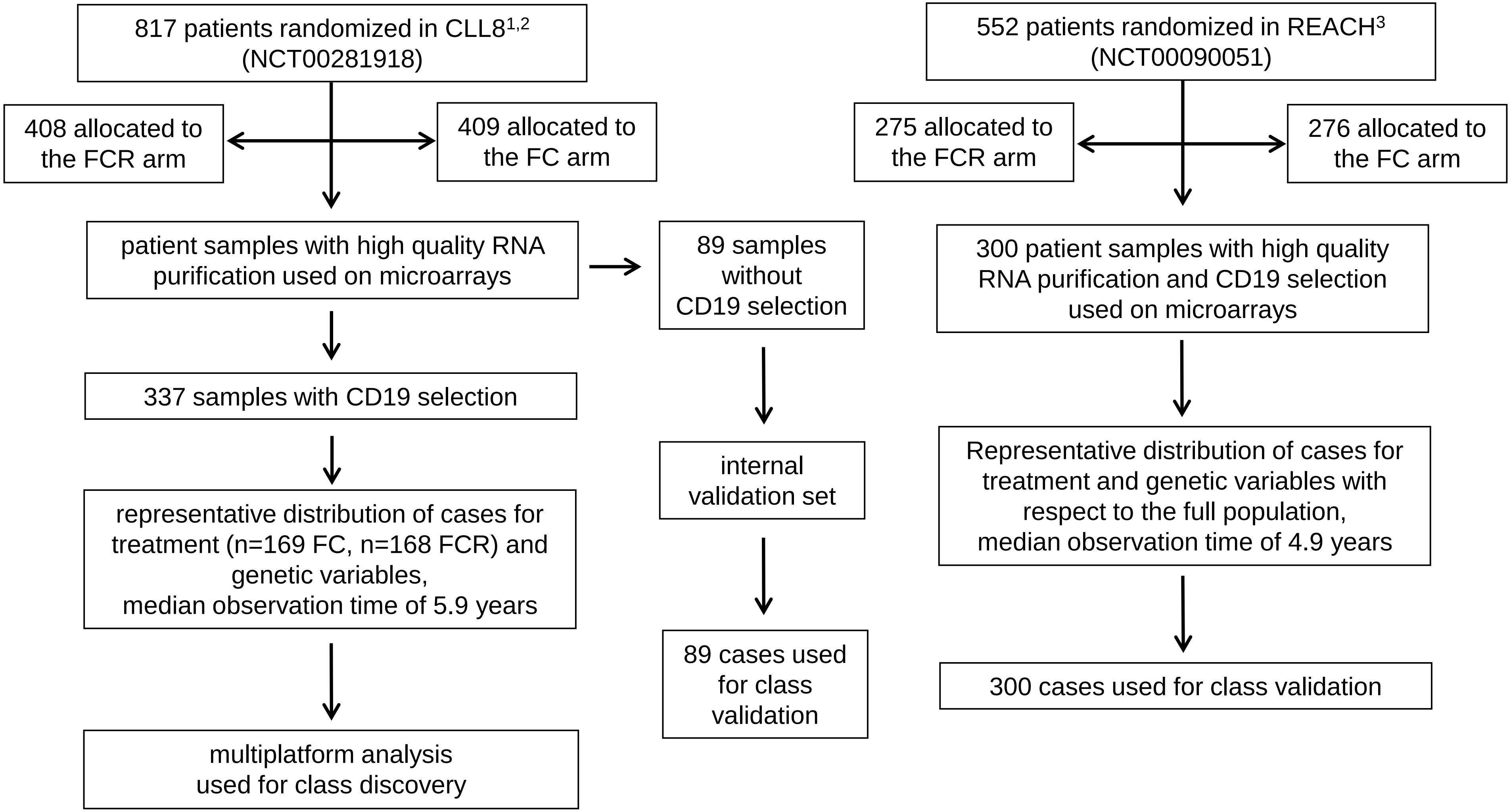
A) PFS (left) and OS (right) according to treatment arm (FC: dotted line; FCR: continuous line) and subtype (color coded) (n=319). B) PFS (left) and OS (right) according to the IGHV mutation status and subtype (color coded) (both treatment arms) (n=310). C) PFS according to subtypes; GI (dark blue) and (I)EMT-L (light blue), *TP53* and *ATM* mutation and/or deletion status (shown for all cases and individual treatment arms) (n=147). D) PFS in *TP53* wild-type cases according to *SF3B1* mutation status for GI (dark blue) or (I)EMT-L (light blue) (n=193). For Fig. A-

D, the log-rank test was used to compare the survival distributions. Data within individual figures derives from biologically independent samples.

Figure 6: Validation of CLL subtypes and prognostic impact

A) Consensus heatmap showing the 4 major CLL subtypes identified in the REACH expression dataset (n=300). Recurrent alterations are depicted below, inactivation (del/mut) of *TP53* is heterogeneously distributed (“EMT-L”: 11%, “(I)GI”: 16%, “GI”: 20%, “(I)EMT-L”: 9%). B) Heatmap depicting characteristic GEP in CLL8 (major core enrichment gene sets, 931 genes). C) Heatmap showing expression of core enrichment gene sets (as used in Figure 6B) for the REACH dataset. For better comparability, CLL8-complementary clusters (indicated by labels in quotation marks) are ordered in the same order as found by consensus clustering in CLL8. Increased biologic homogeneity is observed in “GI”/“(I)GI” cases and supported through co-clustering in Figure 6A. D) PFS in the REACH dataset for all cases with *TP53* defect (yellow), “(I)EMT-L” and “GI” cases without *TP53* defect (n=173). Significance level for GI vs. (I)EMT-L cases is calculated based on clustering from Fig.6A. E) OS in the REACH dataset for all cases with*TP53* defect (yellow), “(I)EMT-L” and “GI” cases without *TP53* defect (n=173). For Fig. D/E, the log-rank test was used to compare the survival distributions. Data within individual figures derives from biologically independent samples.

CONSORT diagram for the discovery and validation cohort



817 patients randomized in CLL8[1,2] (NCT00281918)

408 allocated to the FCR arm

409 allocated to the FC arm

patient samples with high quality RNA purification used on microarrays

337 samples with CD19 selection

representative distribution of cases for treatment (n=169 FC, n=168 FCR) and genetic variables, median observation time of 5.9 years

multiplatform analysis used for class discovery

89 samples without CD19 selection

internal validation set

89 cases used for class validation

552 patients randomized in REACH[3] (NCT00090051)

275 allocated to the FCR arm

276 allocated to the FC arm

300 patient samples with high quality RNA purification and CD19 selection used on microarrays

Representative distribution of cases for treatment and genetic variables with respect to the full population, median observation time of 4.9 years

300 cases used for class validation

1) Hallek et al., Lancet. 2010 Oct 2;376(9747):1164-74
2) Fischer et al., Blood. 2016 Jan 14;127(2):208-15.

3) Robak et al., J Clin Oncol. 2010 Apr 1;28(10):1756-65

Abbreviations: FC = chemotherapy with fludarabine and cyclophosphamide, FCR = combined chemoimmunotherapy with FC plus rituximab

Figure 1: Composition and relationship of CLL subtypes in clustered data.

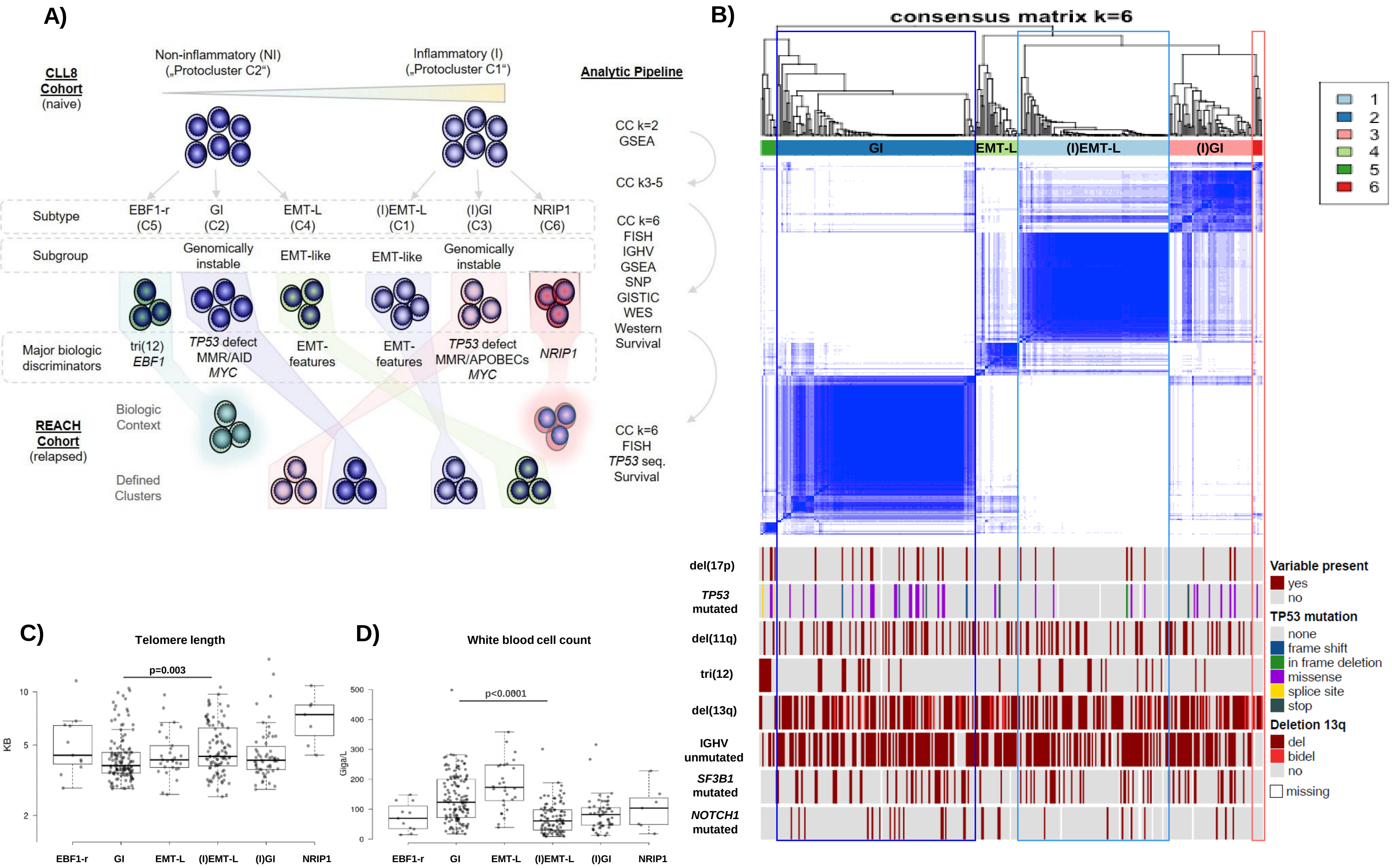


Figure 2: Pathway activation and genetic alterations in CLL subtypes.

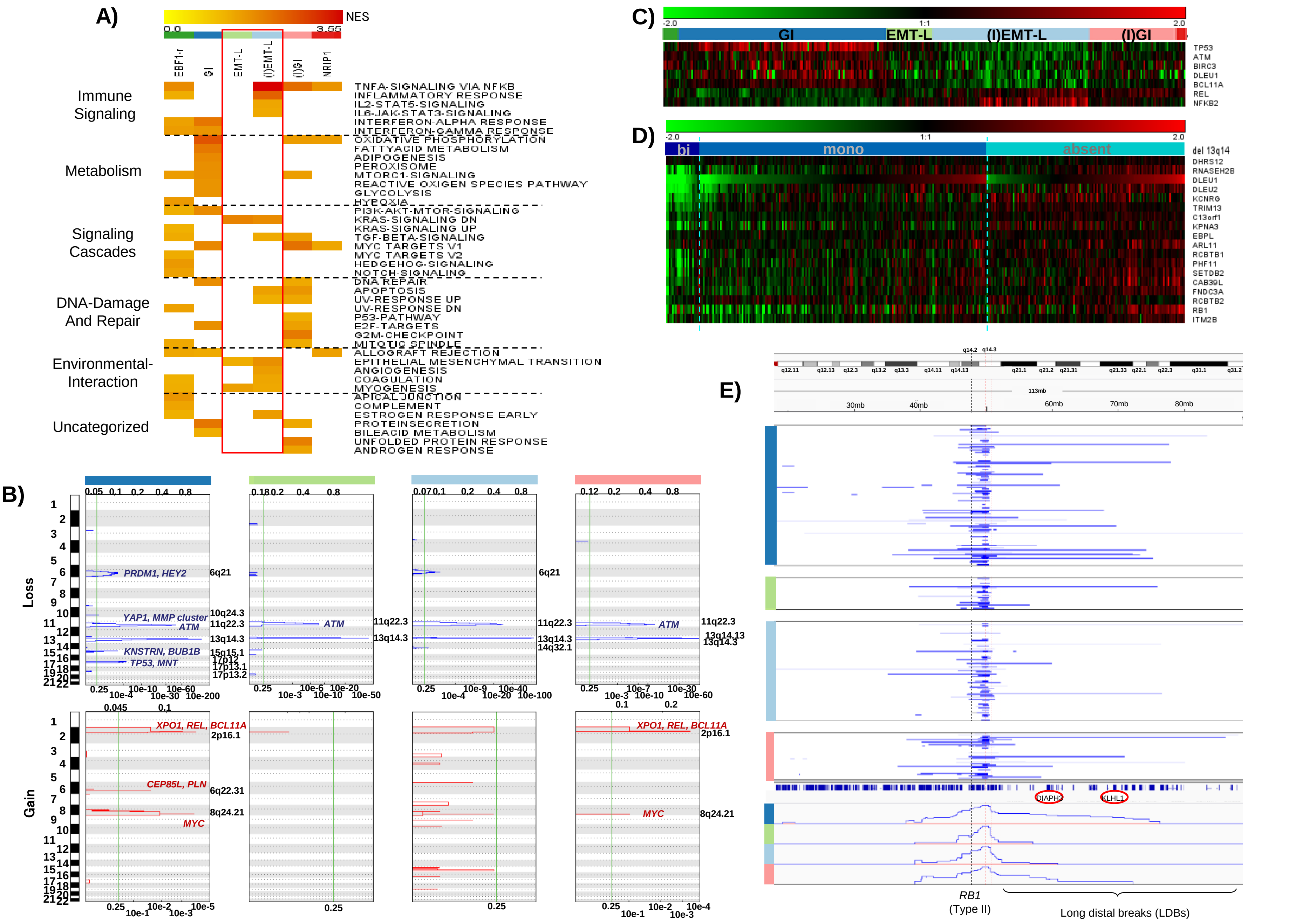


Figure 3: Biological processes operational in genomically instable CLL.

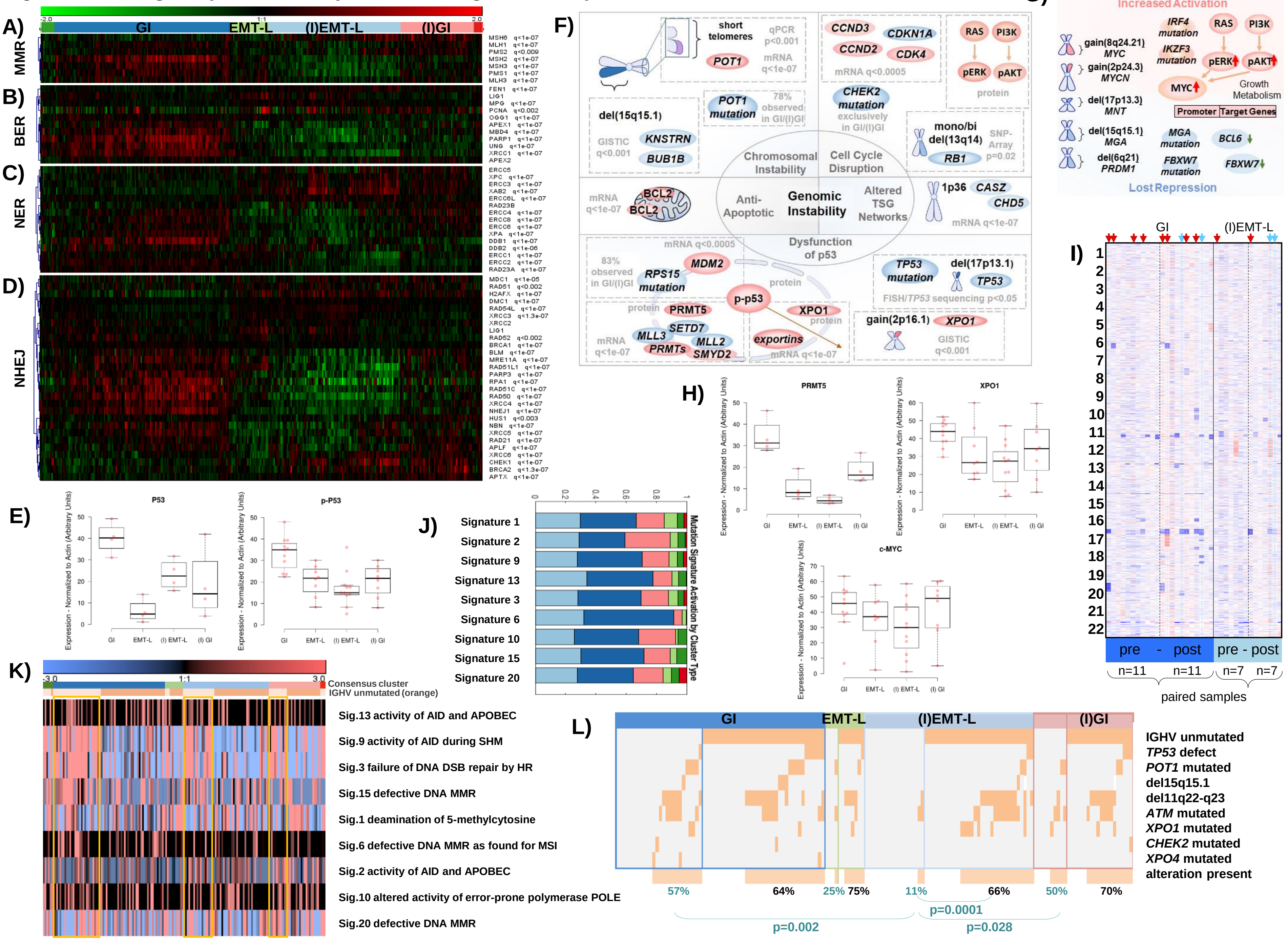


Figure 4: Biological processes operational in CLL with EMT-like networks.

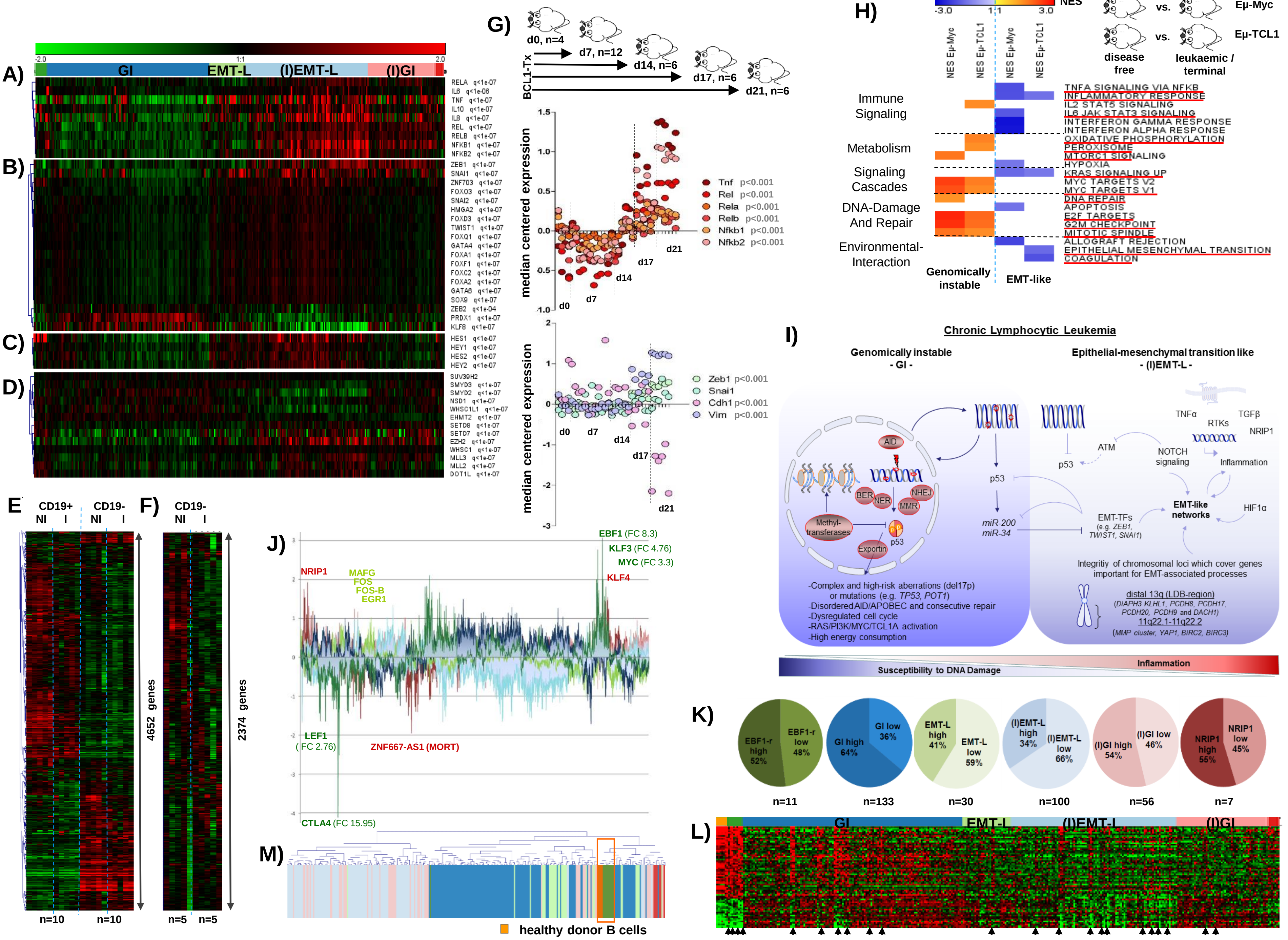


Figure 5: CLL subtype, genetic markers and treatment outcome in CLL8.

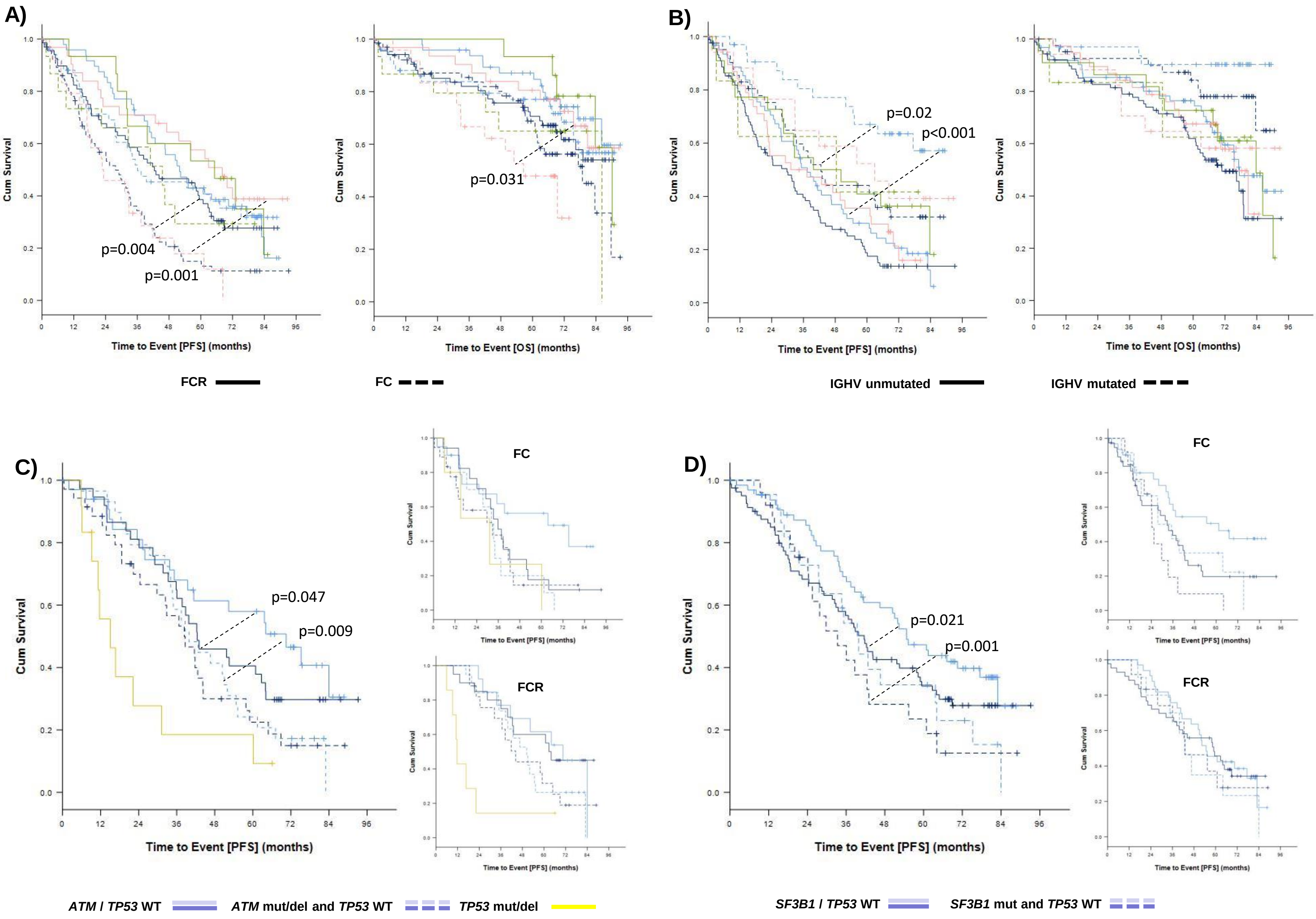
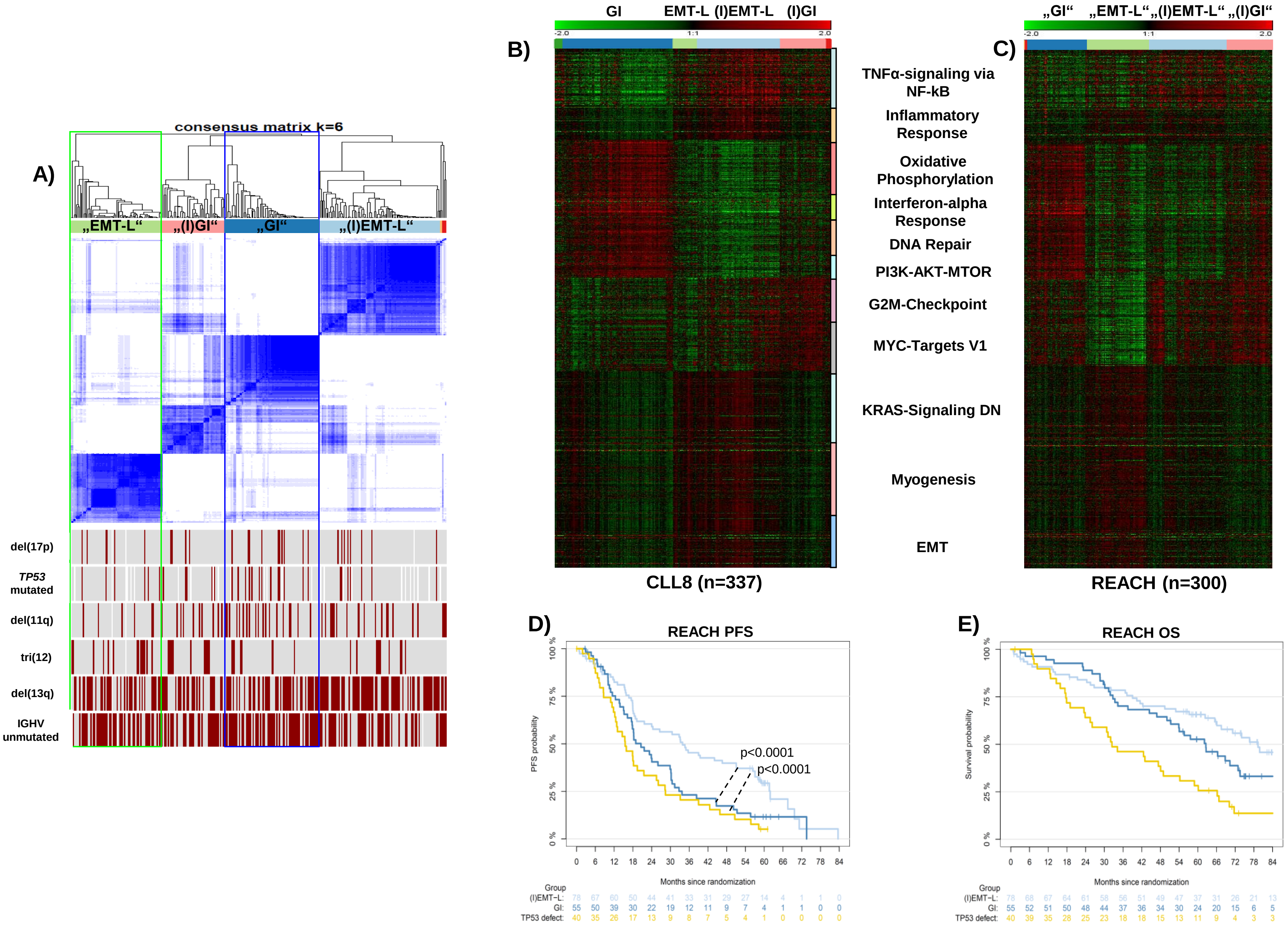


Figure 6: Validation of CLL subtypes and prognostic impact in the REACH trial cohort.


A)
consensus matrix k=6
„EMT-L“
„(I)GI“
„GI“
„(I)EMT-L“
del(17p)
TP53 mutated
del(11q)
tri(12)
del(13q)
IGHV unmutated
B)
GI
EMT-L
(I)EMT-L
(I)GI
1:1
TNFα-signaling via NF-kB
Inflammatory Response
Oxidative Phosphorylation
Interferon-alpha Response
DNA Repair
PI3K-AKT-MTOR
G2M-Checkpoint
MYC-Targets V1
KRAS-Signaling DN
Myogenesis
EMT
CLL8 (n=337)
C)
„GI“
„EMT-L“
„(I)EMT-L“
„(I)GI“
1:1
REACH (n=300)
D)
REACH PFS
PFS probability
p<0.0001
p<0.0001
Months since randomization
Group
(I)EMT−L: 78 67 60 50 44 41 33 31 29 27 14 4 1 1 0
GI: 55 50 39 30 22 19 12 11 9 7 4 1 1 0 0
TP53 defect: 40 35 26 17 13 9 8 7 5 4 1 0 0 0 0
E)
REACH OS
Survival probability
Months since randomization
Group
(I)EMT−L: 78 68 67 64 61 58 56 51 49 47 37 31 26 21 13
GI: 55 52 51 50 48 44 37 36 34 30 24 20 15 6 5
TP53 defect: 40 39 35 28 25 23 18 18 15 13 11 9 4 3 3