Recurrent mutation of the *ID3* gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing

Julia Richter[1,30], Matthias Schlesner[2,30], Steve Hoffmann[3,30], Markus Kreuz[4,30], Ellen Leich[5,30], Birgit Burkhardt[6,7,30], Maciej Rosolowski[4], Ole Ammerpohl[1], Rabea Wagener[1], Stephan H Bernhart[3], Dido Lenze[8], Monika Szczepanowski[9], Maren Paulsen[10], Simone Lipinski[10], Robert B Russell[11], Sabine Adam-Klages[12], Gordana Apic[11], Alexander Claviez[13], Dirk Hasenclever[4], Volker Hovestadt[14], Nadine Hornig[1], Jan O Korbel[15], Dieter Kube[16], David Langenberger[3], Chris Lawerenz[2], Jasmin Lisfeld[7], Katharina Meyer[17], Simone Picelli[14], Jordan Pischimarov[5], Bernhard Radlwimmer[14], Tobias Rausch[15], Marius Rohde[7], Markus Schilhabel[10], René Scholtysik[18], Rainer Spang[17], Heiko Trautmann[19], Thorsten Zenz[20–22], Arndt Borkhardt[23], Hans G Drexler[24], Peter Möller[25], Roderick A F MacLeod[24], Christiane Pott[19], Stefan Schreiber[26], Lorenz Trümper[16], Markus Loeffler[4], Peter F Stadler[27], Peter Lichter[14], Roland Eils[2,28], Ralf Küppers[18], Michael Hummel[8,31], Wolfram Klapper[9,31], Philip Rosenstiel[10,31], Andreas Rosenwald[5,31], Benedikt Brors[2,31] & Reiner Siebert[1,31], for the ICGC MMML-Seq Project[29]

Burkitt lymphoma is a mature aggressive B-cell lymphoma derived from germinal center B cells[1]. Its cytogenetic hallmark is the Burkitt translocation t(8;14)(q24;q32) and its variants, which juxtapose the *MYC* oncogene with one of the three immunoglobulin loci[2]. Consequently, MYC is deregulated, resulting in massive perturbation of gene expression[3]. Nevertheless, MYC deregulation alone seems not to be sufficient to drive Burkitt lymphomagenesis. By whole-genome, whole-exome and transcriptome sequencing of four prototypical Burkitt lymphomas with immunoglobulin gene (*IG*)-*MYC* translocation, we identified seven recurrently mutated genes. One of these genes, *ID3*, mapped to a region of focal homozygous loss in Burkitt lymphoma[4]. In an extended cohort, 36 of 53 molecularly defined Burkitt lymphomas (68%) carried potentially damaging mutations of *ID3*. These were strongly enriched at somatic hypermutation motifs. Only 6 of 47 other B-cell lymphomas with the *IG-MYC* translocation (13%) carried *ID3* mutations. These findings suggest that cooperation between *ID3* inactivation and *IG-MYC* translocation is a hallmark of Burkitt lymphomagenesis.

Burkitt lymphoma constitutes the most frequent B-cell lymphoma in children but also affects adults[1]. Endemic, sporadic and immunodeficiency-associated Burkitt lymphomas are distinguished as clinical variants[1]. Common to all of these variants is the presence of the Burkitt translocation t(8;14)(q24;q32) or its variants t(2;8) and t(8;22) in nearly all affected individuals. These chromosomal translocations juxtapose the *MYC* oncogene at 8q24 with one of the three immunoglobulin loci—*IGH* at 14q32, *IGK* at 2p12 and *IGL* at 22q11—resulting in deregulation of MYC expression[3]. *IG-MYC* translocations are not restricted to Burkitt lymphoma but also occur in other mature aggressive B-cell lymphomas, such as diffuse large B-cell lymphomas and lymphomas intermediate between Burkitt lymphoma and diffuse large B-cell lymphoma[1,4,5]. Recent gene expression profiling studies more precisely defined molecular Burkitt lymphomas (mBLs) and distinguished them from other types of mature aggressive B-cell lymphomas[5,6].

Whereas in other mature B-cell lymphomas *IG-MYC* translocations are thought to be secondary events, occurring during lymphoma progression, current models for Burkitt lymphoma assume that *IG-MYC* translocation is an early or even initiating event in this germinal center–derived B-cell lymphoma[3]. Nevertheless, deregulation of *MYC* by juxtaposition with the immunoglobulin loci seems not to be sufficient to drive lymphomagenesis. Thus, to identify genetic changes cooperating with *MYC* deregulation in Burkitt lymphomagenesis, we performed whole-genome and whole-exome sequencing in four pediatric prototypic sporadic mBLs and corresponding germline samples (**Supplementary Table 1**). Moreover, we integrated findings from transcriptome and whole-methylome bisulfite sequencing of these samples (**Supplementary Tables 2–4**).

Genomic sequencing identified the hallmark *IG-MYC* junctions and clonal *IGH* rearrangements in all lymphomas (**Supplementary Table 5**), but only few other chromosomal aberrations were found (**Supplementary Fig. 1**). Thus, in line with previous

A full list of author affiliations appears at the end of the paper.

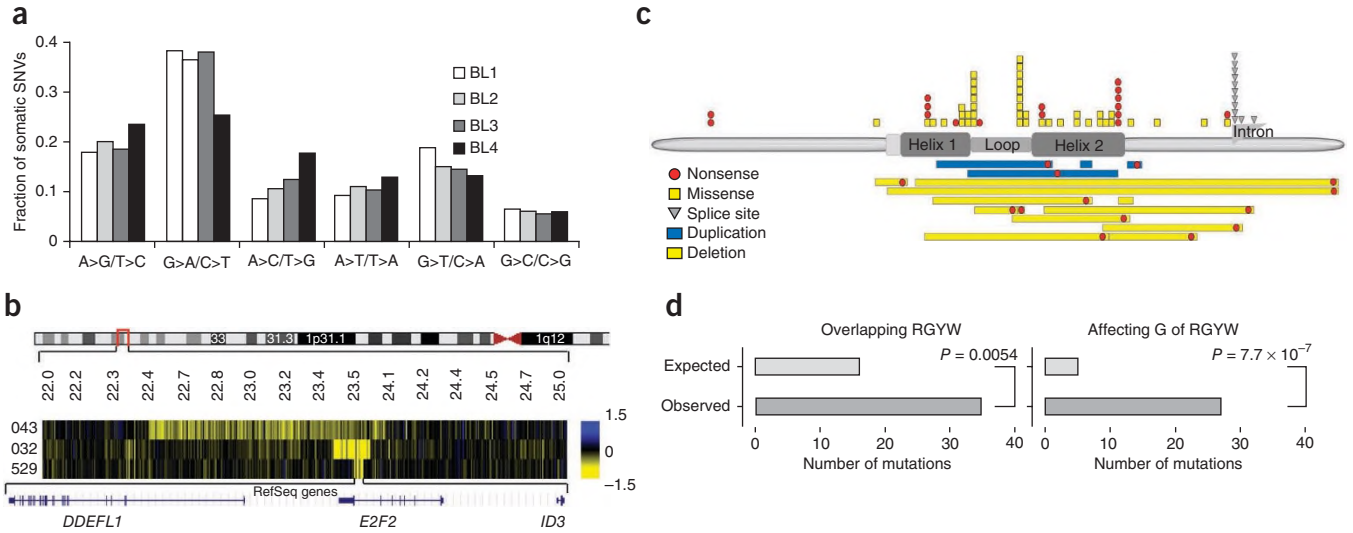

Figure 1 Spectrum of mutations in Burkitt lymphoma. (**a**) Distribution of somatic SNVs identified by integrated genomic and exomic sequencing. (**b**) Heterozygous (043) and homozygous (032, 529) deletions including the *ID3* gene were identified in Burkitt lymphoma (determined in a previous SNP array analysis)[4]. (**c**) Spectrum of *ID3* mutations in *IG-MYC* translocation–positive lymphomas and cell lines. (**d**) Mutations of *ID3* (chr. 1: 23,885,428–23,886,034) were significantly biased toward the RGYW (also known as WRCY) motif, a hotspot of the somatic hypermutation machinery[15]. *P* values were calculated using Fisher's exact test (**Supplementary Table 15**).

studies in Burkitt lymphoma[2], we show that the karyotype of Burkitt lymphoma is also quite simple at the sequence level. Gains in 1q, which are recurrently observed in mBL, might form a notable exception: two of the four Burkitt lymphoma samples had 1q gains (BL1 and BL4), which in both samples were attended by complex rearrangements of this genomic region (**Supplementary Fig. 2**).

To identify single-nucleotide variants (SNVs), we integrated data from whole-genome and whole-exome sequencing. Concordance of array- and sequence-based SNP calling exceeded 98.5% (**Supplementary Table 6**). We identified a median of 2,549.5 (range of 1,957–5,707) somatic aberrations per lymphoma (**Supplementary Tables 7** and **8**). The spectrum of somatic point mutations was conserved, with G>A/C>T transitions being the most frequent changes (**Fig. 1a** and **Supplementary Fig. 3**). Analyzing the combined data set of genome and exome data, we identified a median of 31 (range of 23–49, total of 134) potentially protein-changing somatic mutations (**Supplementary Table 9**). Of the 134 total mutations, 63 were covered by sequence reads in the transcriptome analysis (reads per kb per million mapped reads (RPKM) ≥ 1), and in 57 the mutated allele was found to be expressed (**Supplementary Table 9**). In addition to the previously characterized somatic mutations of the *MYC* and *TP53* genes, 19 out of 20 of these mutations could be confirmed by Sanger sequencing (**Supplementary Tables 10** and **11**).

The 119 genes with potentially protein-changing mutations showed heterogeneous gene expression and DNA methylation patterns (**Supplementary Fig. 4** and **Supplementary Tables 12** and **13**). Seven (*ADAMTS5*, *CDH7*, *ID3*, *NETO1*, *NHLH1*, *PHIP* and *RGL1*) were located in minimal regions of recurrent imbalance in Burkitt lymphoma recently determined by SNP array analysis[4].

We and others have previously shown that Burkitt lymphoma is seemingly a molecularly homogenous disease[5,6]. Therefore, we further focused on genes affected in at least two samples by somatic protein-changing and/or splice-site mutations (**Table 1**). Although this restriction might cause some genes frequently mutated in Burkitt lymphoma to be missed, as evidenced by *CCND3* (ref. 7 and **Supplementary Table 14**), we identified seven genes that were recurrently mutated. These included (i) genes previously known to be mutated in Burkitt lymphoma, such as *TP53* (ref. 8), *MYC*[9] and *SMARCA4* (ref. 10), (ii) genes involved in other B-cell lymphoma, such as *FBXO11* (ref. 11) and *DDX3X*[12] and (iii) *RHOA* and *ID3*. In all seven genes, the mutated allele was expressed at considerable levels (**Table 1**). Moreover, *SMARCA4* and *ID3* were among the genes with high expression in our recently established 50-gene mBL index[5].

Next, we compared the genomic location of the recurrently mutated genes with regions of imbalance in mBL identified in a recent SNP array study[4]. Of the seven recurrently mutated genes, *ID3* (gene NM_002167, transcript ENST00000374561) was the only one located in a region of minimal imbalance in mBL. More specifically, *ID3* mapped to a 138-kb minimal region of homozygous loss at 1p36.12 (**Fig. 1b** and **Table 1**). Biallelic *ID3* mutations were also observed in two mBLs studied here (BL1 and BL4), which both had a nonsense and a splice-site mutation (**Supplementary Fig. 5**). Transcriptome analysis showed that the splice-site mutation affected splicing (**Supplementary Fig. 5**, bottom). These findings suggest that recurrent biallelic inactivation of *ID3* occurs in mBL.

To test this hypothesis, we sequenced the coding region of *ID3* in additional 103 previously characterized *IG-MYC* translocation–positive mature aggressive B-cell lymphomas[5,13,14]. DNA from six samples did not amplify, and these samples were thus excluded from further analysis. *ID3* mutations were detected in 41 of 97 lymphomas (35 of 52 mBLs and 6 of 45 other aggressive B-cell lymphomas), with 19 showing 2 mutations and 5 showing even more mutations (maximum of 4), including large deletions (**Fig. 1c**, **Supplementary Figs. 6** and **7** and **Supplementary Table 14**). The high-frequency of *ID3* mutations, even in a single lymphoma, prompted us to search for potential directive mutation mechanisms. Indeed, *ID3* mutations were significantly enriched at hotspots of the somatic hypermutation machinery (RGYW motifs)[15] (**Fig. 1d**, **Supplementary Fig. 8** and **Supplementary Table 15**), which contrasts with the genome-wide distribution of observed somatic mutations that was not significantly associated with these hotspots.

With regard to protein function, the pattern of mutations in *ID3*, including 10 deletions, 5 duplications and/or insertions, 13 nonsense mutations, 37 missense mutations and 7 splice-site mutations, was skewed toward deleterious changes (**Fig. 1c**). Biallelic involvement

Table 1 Genes affected in at least two cases by high-confidence somatic protein-changing and/or splice-site mutations identified by integrated genome and exome sequence analysis in the four analyzed Burkitt lymphomas

				DNA level				RNA level			
Gene	Transcript	Chr.	Position (hg19)	Ref.	Alt.	Predicted protein change	Frequency of mutated allele	Ref.	Alt.	Ratio	Subject code
DDX3X	NM_001356	X	41205590	G	A	p.Arg475His	0.88	33	449	0.93	BL1
	NM_001193416		41203383	T	G	Splice site	0.76	100	22	0.18	BL4
FBXO11	NM_001190274	2	48036837	A	T	p.Ile783Asn	0.5	138	155	0.53	BL1
			48037502	CTA	C	Frameshift	ND	151	54	0.36	BL2
			48045966	A	G	p.Leu653Pro	0.43	112	117	0.51	BL3
ID3	NM_002167	1	23885677	G	A	p.Gln81*	0.22	2,034	991	0.33	BL1
			23885677	G	A	p.Gln81*	0.43	1,058	310	0.23	BL4
			23885616	A	G	Splice site	0.38	40	206	0.84	BL4
			23885511	C	T	Splice site	0.41	810	293	0.27	BL1
MYC	NM_002467	8	128748853	G	A	p.Arg5Gln	0.33	15	887	0.98	BL1
			128748855	G	T	p.Val6Leu	0.32	14	881	0.98	BL1
			128748858	G	A	p.Val7Met	0.31	14	864	0.98	BL1
			128750921	T	G	p.Phe153Cys	0.42	44	1,133	0.96	BL3
			128750683	C	T	p.Pro74Ser	0.39	146	330	0.69	BL4
			128748858	G	A	p.Val7Met	0.45	86	488	0.85	BL4
RHOA	NM_001664	3	49413009	C	T	p.Arg5Gln	0.53	627	799	0.56	BL2
			49412955	A	C	p.Ile23Arg	0.38	2,729	1,165	0.30	BL4
			49413009	C	T	p.Arg5Gln	0.42	2,388	878	0.27	BL4
SMARCA4	NM_003072	19	11134252	G	A	p.Arg973Gln	0.53	584	605	0.51	BL1
	NM_001128844		11105679	T	C	Splice site	0.44	3	433	0.99	BL3
TP53	NM_001126114	17	7578415	A	T	p.Val172Asp	0.5	147	177	0.55	BL2
			7578204	A	C	p.Ser215Arg	0.84	14	469	0.97	BL3

Chr., chromosome; ref., reference allele; alt., alternative allele; ND, not determined; ratio, read count alternative allele/(read count alternative allele + read count reference allele). A ratio of reference/alternative allele RNA levels of <0.5 indicates predominant expression of the wild-type allele, whereas values >0.5 indicate predominant expression of the mutated allele (assuming 100% tumor cell content).

could be formally confirmed in 16 of 27 lymphomas with at least 2 *ID3* mutations (**Supplementary Fig. 7a** and **Supplementary Table 14**) and was corroborated by deep sequencing of exon 1 of *ID3* (**Supplementary Fig. 9** and **Supplementary Table 16**). A similar pattern of mutations was detected in Burkitt lymphoma cell lines (**Supplementary Fig. 6b** and **Supplementary Tables 17** and **18**), in which we confirmed aberrant splicing (**Supplementary Fig. 7a**). Moreover, protein blot analyses showed that some of these mutations were associated with complete loss of ID3 protein expression (**Supplementary Fig. 10**).

As mentioned, *IG-MYC* translocations are not restricted to Burkitt lymphoma but also occur in other aggressive B-cell lymphomas[1,2,16]. Therefore, the 3 index samples (BL1–BL3) and 97 lymphomas from the validation cohort where gene expression data were available were analyzed together. Notably, 36 of 42 *IG-MYC* translocation–positive lymphomas with mutated *ID3* (86%) but only 19 of 58 with wild-type *ID3* (33%) showed the gene expression signature of mBL ($P < 0.001$) (**Supplementary Table 19**). In concordance, 36 of 55 mBLs but only 6 of 29 intermediate lymphomas and 0 of 16 non-mBLs carried an *ID3* mutation. Lymphomas with *ID3* mutations were strongly enriched for various epidemiological (lower age), histopathological (negative immunohistochemical staining for BCL2, positive immunohistochemical staining for BCL6 and high Ki-67 index) and genetic (absence of t(14;18) and *BCL6* breaks, low genetic complexity and low *IGHV* mutation) features of mBL and showed a significantly favorable prognosis (all with $P < 0.05$; **Supplementary Fig. 11** and **Supplementary Table 19**). Notably, the six *ID3*-mutated lymphomas with a gene expression index below the threshold defined for diagnosing mBL[5] also showed many of these features (**Supplementary Table 20**). By using a strict biological definition of mBL (mBL index > 0.95, no *IGH-BCL2* translocation and no *BCL6* breaks), we identified 53 of 100 prototypical mBLs. In examining these mBLs, we did not detect any significant histopathological or clinical differences between the 17 samples with wild-type *ID3* and the 36 with mutated *ID3* (**Supplementary Fig. 12** and **Supplementary Table 19**). Within these groups, we also did not detect differences in global gene expression patterns. Similarly, the DNA methylation pattern at the *ID3* locus,

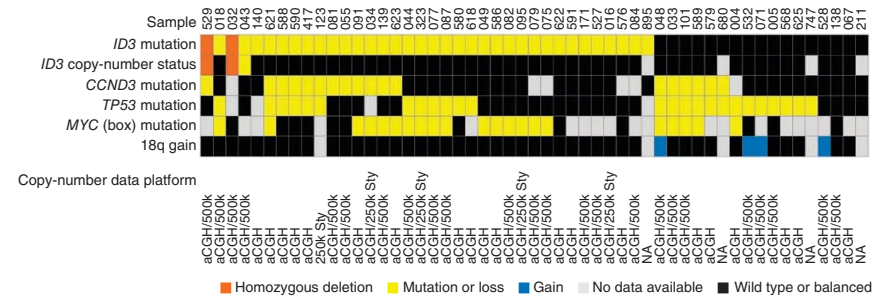

Figure 2 Mutual relationship of *ID3* mutations with *CCND3* (exon 5), *TP53* (exons 4–10) and *MYC* box mutations in mBL samples without IGH-BCL2 translocation. For ID3, only focal copy-number aberrations are shown. Because of lower resolution, sensitivity was limited for the array–comparative genomic hybridization (aCGH) platform. The mutational prevalence in mBL was as follows: *ID3* mutation, 36 of 53 (67.9%); *CCND3* (exon 5) mutation, 18 of 47 (38.3%); *TP53* mutation[27], 25 of 49 (51.0%); *MYC* box mutation, 21 of 33 (63.6%). For each affected individual, a box indicates the mutation or copy-number status for the analyzed genes and regions. Copy-number status, determined using BAC and SNP arrays, was recently published[4,5,13].

as determined by bisulfite sequencing, did not differ between samples with (BL1 and BL4) and without (BL2 and BL3) *ID3* mutations (**Supplementary Figs. 13** and **14a**) or between Burkitt lymphoma cell lines with wild-type *ID3* or monoallelic or biallelic *ID3* mutation (**Supplementary Fig. 14b**).

The lack of differences between Burkitt lymphomas with wild-type and mutated *ID3* could suggest that alternative mechanisms might substitute for *ID3* mutation. Analyzing the 53 prototypical mBLs with regard to *ID3* mutation status, we found no obvious evidence for an exclusive or concomitant occurrence of a *CCND3*, *TP53* or *MYC* mutation (**Fig. 2**). However, the number and pattern of chromosomal imbalances differed between *ID3*-mutated and wild-type mBLs (**Supplementary Fig. 12**). Most discriminating were imbalances at 18q, with a minimal region of gain at chr. 18: 51,894,728–54,354,319 (hg19) (**Supplementary Fig. 15**). In this region, only *TCF4* (alias E2-2; 7 of 8 tags) showed significantly higher gene expression in the lymphomas with wild-type *ID3* and gain of 18q compared to the *ID3*-mutated lymphomas (**Supplementary Fig. 16**).

Inhibitor of DNA binding (ID) proteins, such as ID3, can bind E proteins, such as TCF3, via the helix-loop-helix (HLH) motif common to both. Formation of such nonfunctional heterodimers prevents binding of the latter to DNA[17,18]. The results presented for Burkitt lymphoma indicate impairment of the inhibitory function of ID3 on TCF4 and probably also on the highly related TCF3 (also known as E2A), which had a somatic missense mutation (encoding a p.Arg606Gln change) affecting a conserved position next to the HLH domain in BL4. To further corroborate this hypothesis, we modeled the effect of the 42 missense mutations in *ID3* on the complex observed in cell lines and primary lymphomas (**Supplementary Fig. 17**). Indeed, 36 of 42 missense mutations affected conserved protein residues in the HLH motif (**Supplementary Fig. 18**), and several of these are predicted to affect TCF3 and/or TCF4 binding (**Supplementary Table 21**) and, thus, to impair the inhibitory effect of ID3.

ID3 is highly expressed in Burkitt lymphoma, with the notable exception of mBLs with homozygous loss of *ID3* (**Supplementary Fig. 4** and **Supplementary Tables 9** and **12**)[5]. This strong expression might be due to the fact that the *ID3* locus is a direct target of MYC[19] and due to induction of ID3 upon B-cell receptor (BCR) triggering (ref. 20 and D. Kube *et al.*, unpublished data). However, in Burkitt lymphoma with *ID3* mutations, this seems not to translate into normal ID3 protein expression or function (**Supplementary Fig. 10**). Indeed, overexpression of green fluorescent protein (GFP)-tagged wild-type ID3 in Burkitt lymphoma cell lines with mutant (BL-2) and wild-type (DG-75) *ID3* increased the number of cells in the pre-G1 fraction of the cell cycle ($P < 0.0001$) relative to cells not expressing ID3-GFP (**Fig. 3**), suggesting a selective disadvantage from high expression of wild-type ID3 in mBL.

ID3 has been implicated in a variety of functions, including regulation of cell cycle progression and B-cell differentiation. *Id3*-knockout mice show defects in humoral immunity and B-cell proliferation and develop T-cell lymphomas[17,20,21]. Our findings imply that, besides *IG-MYC* translocation, the disruption of ID3 function might be a key mechanism in the pathogenesis of Burkitt lymphoma.

This view is supported by data published during the revision of this manuscript by Schmitz and colleagues[22], who also identified *ID3* mutations in more than 50% of all Burkitt lymphomas and showed that *ID3* and/or *TCF3* mutations are present in 70% of sporadic Burkitt lymphomas. Their functional analyses suggest that *ID3*-destructive mutations and/or *TCF3*-activating mutations lead to activation of a TCF3 transcriptional program and, thereby, intensify a tonic form of BCR signaling to activate pro-survival phosphoinositide 3-kinase (PI3K) signaling[22]. Indeed, PI3K pathway activation has not only been shown to cooperate with Myc

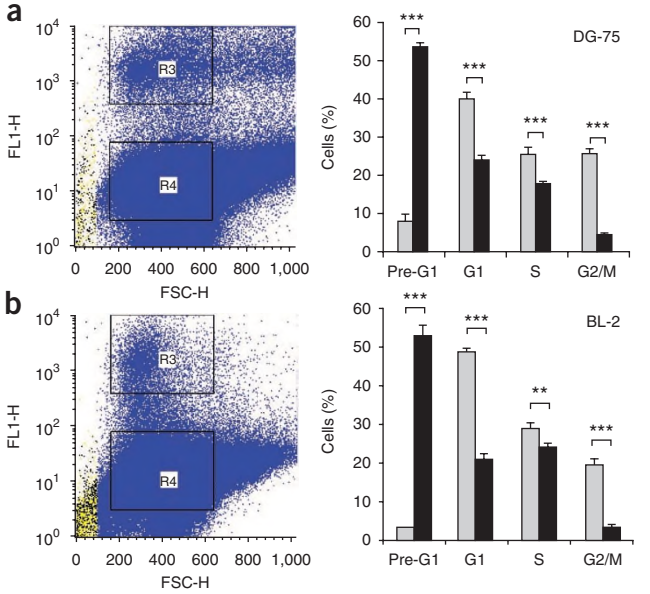

Figure 3 Cell cycle analysis of ID3-GFP–transfected cell lines by fluorescence-activated cell sorting (FACS). (**a**,**b**) The DG-75 (**a**) and BL-2 (**b**) cell lines were transfected with *ID3*-pCMV6-AC-*GFP*. Left, GFP-positive cells after 48 h incubation, the ID3-GFP–expressing (R3; 11.5–18% of DG-75 cells and 4–5% of BL-2 cells) and ID3-GFP–negative (R4) cell populations were identified by FACS analysis. Right, cell cycle analysis of ID3-GFP–positive (black) and ID3-GFP–negative (gray) cell populations was performed separately. ID3-GFP expression led to fewer cells in the G1 and G2/M phases, accompanied by significantly more cells in the pre-G1 fraction (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, *t* test). Transfections were performed in triplicate. All data are the mean ± s.d. FL1-H, fluorescence channel 1 height; FSC-H, forward scatter height.

activation in a mouse model of Burkitt lymphoma but has also been recently demonstrated in a set of Burkitt lymphomas from our cohort[5,23,24].

In conclusion, this integrative genomics study identified inactivating mutation of *ID3*, most likely frequently caused by aberrant somatic hypermutation, as a highly recurrent somatic change in *IG-MYC* translocation–positive sporadic Burkitt lymphoma, whereas these mutations are rare in other *IG-MYC* translocation–positive lymphomas. This finding indicates that the combination of *ID3* inactivation and *IG-MYC* translocation is a characteristic property of Burkitt lymphoma pathogenesis. In addition and besides the previously described changes affecting *TP53* and cell cycle function, recurrent mutation of the chromatin-remodeling complex gene *SMARCA4* (ref. 25) and the RNA-helicase encoding *DDX3X*[26] suggest the contribution of further mechanisms to Burkitt lymphoma evolution.

URLs. European Genome-phenome Archive (EGA), http://www.ebi.ac.uk/ega/; data access committee of the International Cancer Genome Consortium, http://www.icgc.org/daco; International Cancer Genome Consortium, http://www.icgc.org/; IMGT database, http://www.imgt.org/; UCSC Genome Browser, http://genome.ucsc.edu/; Picard, http://picard.sourceforge.net/; SeqPrep, https://github.com/jstjohn/SeqPrep.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Data of the validation cohort from the Molecular Mechanisms in Malignant Lymphomas (MMML) Consortium have

been deposited at the Gene Expression Omnibus (GEO) under accessions GSE4475, GSE10172 and GSE22470. All short read sequencing data have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI), under accession EGAS00001000271. Access must be awarded by the data access committee (DAC) of the International Cancer Genome Consortium (ICGC).

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This manuscript is dedicated to the memory of Karl Lennert, the founder of the Kiel classification of lymphoma, who died during the preparation of the manuscript. The authors thank O. Batic, C. Becher, C. Botz-von Drathen, A. Dietsch, J. Eils, M. Friskovec, K. Göbel, T. Grieb, S. Hengst, U. Jacobsen, T. Kaacksteen, H. Lammert, C. von der Lancken, H.-H. Müller, S. Radomski, J. Schieferstein, M. Schlapkohl, D. Schuster, S. Ölmez and L. Valles for their excellent technical support. We are grateful to G. Richter for excellent work in the coordination of the ICGC MMML-Seq Consortium and to the members of the MMML Consortium for providing extensive data of the sample analyzed within that project. We are also very grateful to all individuals who participate in this study. The project was funded by the Federal Ministry of Education and Research in Germany (BMBF) within the Program for Medical Genome Research (01KU1002A to 01KU1002J). The authors are responsible for the content of this publication. The MMML Consortium was funded by the Deutsche Krebshilfe from 2003 to 2011. Support of sequencing infrastructure by the Deutsche Forschungsgemeinschaft (DFG) and the KinderKrebsInitiative Buchholz/Holm-Seppensen is gratefully acknowledged. R.W. is the recipient of a Christoph-Schubert-Award from the KinderKrebsInitiative Buchholz/Holm-Seppensen.

AUTHOR CONTRIBUTIONS

B. Burkhardt, A.C., A.B., M. Rohde, J.L. and N.H. provided subject samples and clinical data. W.K., M.H. and P.M. coordinated and performed pathology review. D. Lenze, M. Szczepanowski, M.H. and W.K. stained and reviewed cryomaterial, prepared analytes and performed analyte quality control. D. Lenze and M.H. performed and interpreted immunoglobulin gene PCR. R.A.F.M. and H.G.D. characterized and provided cell lines. E.L., M. Schilhabel, V.H., S.P. and B.R. performed next-generation sequencing analyses, and A.R., P.R., P.L. and S.S. supervised next-generation sequencing analysis and interpreted data. C.L. coordinated transfer and data management of the sequences. M. Schlesner, B. Brors, S.H., S.H.B., P.F.S., D. Langenberger, V.H., J.O.K., T.R. and J.P. performed analysis of next-generation sequencing data, and B. Brors and R.E. conceived the statistical analysis. J.R., E.L., O.A., S.A.-K., M.P., S.L., R.B.R. and G.A. performed validation analyses. J.R. and R.W. were responsible for *ID3* mutation and expression analyses. M.K., M. Rosolowski, M.L., H.T., C.P., R. Spang, K.M., R.K., R. Scholtysik, T.Z., D.K., L.T., D.H. and R. Siebert provided and analyzed data from the MMML cohort. M.K. and M. Rosolowski performed correlative and biometric analyses of the MMML cohort. J.R., M. Schlesner, M.K., S.H., R.E. and R. Siebert interpreted data and wrote the manuscript. R.E., M.H., W.K., P.R., A.R., B. Brors, O.A. and R. Siebert designed the study and coordinated the project. All authors read and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Swerdlow, S.H. *et al. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues,* Ch. 10, 179–268 (IARC Press, Lyon, France, 2008).
2. Boerma, E.G., Siebert, R., Kluin, P.M. & Baudis, M. Translocations involving 8q24 in Burkitt lymphoma and other malignant lymphomas: a historical review of cytogenetics in the light of today's knowledge. *Leukemia* **23**, 225–234 (2009).
3. Klapproth, K. & Wirth, T. Advances in the understanding of MYC-induced lymphomagenesis. *Br. J. Haematol.* **149**, 484–497 (2010).
4. Scholtysik, R. *et al.* Detection of genomic aberrations in molecularly defined Burkitt's lymphoma by array-based, high resolution, single nucleotide polymorphism analysis. *Haematologica* **95**, 2047–2055 (2010).
5. Hummel, M. *et al.* A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *N. Engl. J. Med.* **354**, 2419–2430 (2006).
6. Dave, S.S. *et al.* Molecular diagnosis of Burkitt's lymphoma. *N. Engl. J. Med.* **354**, 2431–2442 (2006).
7. Schmitz, R. *et al.* Recurrent oncogenic mutations in *CCND3* in aggressive lymphomas. *Blood (ASH Annual Meeting Abstracts)* **118**, 435 (2011).
8. Wilda, M. *et al.* Inactivation of the *ARF–MDM-2–p53* pathway in sporadic Burkitt's lymphoma in children. *Leukemia* **18**, 584–588 (2004).
9. Johnston, J.M. & Carroll, W.L. c-*myc* hypermutation in Burkitt's lymphoma. *Leuk. Lymphoma* **8**, 431–439 (1992).
10. Love, C.L. *et al.* Whole genome and exome sequencing reveals the genetic landscape of Burkitt lymphoma. *Blood (ASH Annual Meeting Abstracts)* **118**, 433 (2011).
11. Duan, S. *et al.* FBXO11 targets BCL6 for degradation and is inactivated in diffuse large B-cell lymphomas. *Nature* **481**, 90–93 (2012).
12. Wang, L. *et al. SF3B1* and other novel cancer genes in chronic lymphocytic leukemia. *N. Engl. J. Med.* **365**, 2497–2506 (2011).
13. Klapper, W. *et al.* Molecular profiling of pediatric mature B-cell lymphoma treated in population-based prospective clinical trials. *Blood* **112**, 1374–1381 (2008).
14. Klapper, W. *et al.* Patient age at diagnosis is associated with the molecular characteristics of diffuse large B-cell lymphoma. *Blood* **119**, 1882–1887 (2012).
15. Casellas, R. *et al.* Restricting activation-induced cytidine deaminase tumorigenic activity in B lymphocytes. *Immunology* **126**, 316–328 (2009).
16. Salaverria, I. & Siebert, R. The gray zone between Burkitt's lymphoma and diffuse large B-cell lymphoma from a genetics perspective. *J. Clin. Oncol.* **29**, 1835–1843 (2011).
17. Perk, J., Iavaraone, A. & Benezra, R. Id family of helix-loop-helix proteins in cancer. *Nat. Rev. Cancer* **5**, 603–614 (2005).
18. Kee, B.L. E and ID proteins branch out. *Nat. Rev. Immunol.* **9**, 175–184 (2009).
19. Seitz, V. *et al.* Deep sequencing of MYC DNA-binding sites in Burkitt lymphoma. *PLoS ONE* **6**, e26837 (2011).
20. Pan, L. *et al.* Impaired immune responses and B-cell proliferation in mice lacking the *Id3* gene. *Mol. Cell Biol.* **19**, 5969–5980 (1999).
21. Li, J. *et al.* Mutation of inhibitory helix-loop-helix protein Id3 causes γδ T-cell lymphoma in mice. *Blood* **116**, 5615–5621 (2010).
22. Schmitz, R. *et al.* Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature* **490**, 116–120 (2012).
23. Sander, S. *et al.* Synergy between PI3K signaling and MYC in Burkitt lymphomagenesis. *Cancer Cell* **22**, 167–179 (2012).
24. Dominguez-Sola, D. & Dalla-Favera, R. Burkitt lymphoma: much more than MYC. *Cancer Cell* **22**, 141–142 (2012).
25. Medina, P.P. & Sanchez-Cespedes, M. Involvement of the chromatin-remodeling factor BRG1/SMARCA4 in human cancer. *Epigenetics* **3**, 64–68 (2008).
26. Choi, Y.J. & Lee, S.G. The DEAD-box RNA helicase DDX3 interacts with DDX5, co-localizes with it in the cytoplasm during the G2/M phase of the cycle, and affects its shuttling during mRNP export. *J. Cell Biochem.* **113**, 985–996 (2012).
27. Zenz, T. *et al.* Monoallelic *TP53* inactivation is associated with poor prognosis in chronic lymphocytic leukemia: results from a detailed genetic characterization with long-term follow-up. *Blood* **112**, 3322–3329 (2008).

[1]Institute of Human Genetics, Christian-Albrechts-University, Kiel, Germany. [2]Division of Theoretical Bioinformatics, Deutsches Krebsforschungszentrum Heidelberg (DKFZ), Heidelberg, Germany. [3]Transcriptome Bioinformatics, LIFE Research Center for Civilization Diseases, University of Leipzig, Leipzig, Germany. [4]Institute for Medical Informatics Statistics and Epidemiology, University of Leipzig, Leipzig, Germany. [5]Institute of Pathology, University of Wuerzburg, Wuerzburg, Germany. [6]Pediatric Hematology and Oncology, University Hospital Muenster, Muenster, Germany. [7]Pediatric Hematology and Oncology, University Hospital Giessen, Giessen, Germany. [8]Institute of Pathology, Charité–University Medicine Berlin, Berlin, Germany. [9]Hematopathology Section, Christian-Albrechts-University, Kiel, Germany. [10]Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany. [11]Cell Networks, Bioquant, University of Heidelberg, Heidelberg, Germany. [12]Institute of Immunology, Christian-Albrechts-University, Kiel, Germany. [13]Department of Pediatrics, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany. [14]Division of Molecular Genetics, DKFZ, Heidelberg, Germany. [15]Genome Biology Research Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [16]Department of Hematology and Oncology, Georg-Augusts-University of Göttingen, Göttingen, Germany. [17]Institute of Functional Genomics, University of Regensburg, Regensburg, Germany. [18]Institute of Cell Biology (Cancer Research), University of Duisburg-Essen, Duisburg-Essen, Medical School, Essen, Germany. [19]Department of Internal Medicine II, Hematology and Oncology, University Medical Centre, Campus Kiel, Kiel, Germany. [20]Department of Medicine V, University of Heidelberg, Heidelberg, Germany. [21]Department of Translational Oncology, National Center for Tumor Diseases (NCT) and DKFZ, Heidelberg, Germany. [22]Department of Medicine III, Ulm University, Ulm, Germany. [23]Department of Pediatric Oncology, Hematology and Clinical Immunology, Heinrich-Heine-University, Düsseldorf, Germany. [24]Department of Human and Animal Cell Cultures, German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany. [25]Institute of Pathology, Medical Faculty of the Ulm University, Ulm, Germany. [26]Department of General Internal Medicine, Christian-Albrechts-University, Kiel, Germany. [27]Bioinformatics Group, Department of Computer Science and Interdisciplinary Center of Bioinformatics, University of Leipzig, Leipzig, Germany. [28]Institute of Pharmacy and Molecular Biotechnology, Bioquant, University of Heidelberg, Heidelberg, Germany. [29]A full list of members and affiliations is provided in the **Supplementary Note**. [30]These authors contributed equally to this work. [31]These authors jointly directed this work. Correspondence should be addressed to R. Siebert (rsiebert@medgen.uni-kiel.de).

ONLINE METHODS

All working steps were performed according to the manufacturer's instructions or standard protocols unless otherwise stated. The sequences of PCR and sequencing primers are listed in **Supplementary Table 22**.

DNA and RNA extraction. DNA was extracted from tissue, blood and buffy coats using Genomic-tip 100/G (Qiagen). Extraction of total RNA, including the small RNA fraction, from frozen tissue and body fluids (blood and buffy coats) was performed using the Ambion mirVana kit and the mirVana PARIS kit (both from Life Technologies), respectively. DNA and RNA from cell lines were extracted using the Gentra Puregene Cell kit (Qiagen) and the RNeasy Mini kit (Qiagen).

Detection and sequencing of immunoglobulin gene rearrangements. Immunoglobulin heavy and light chain genes from tumor and germline samples were analyzed by multiplex PCR according to the BIOMED-2 protocol[28] with minor modifications and were separated by electrophoresis (GeneScan) on a Genetic Analyzer 3130 (Applied Biosystems). Rearrangement patterns in the tumor and corresponding germline sample were compared to exclude contamination of germline samples with tumor cells. *IGH* PCR products from tumor samples were sequenced and compared to the International Immunogenetics Information Systems (IMGT) database[29] (see URLs) to determine VH segment usage and the presence of somatic mutations.

Sequencing. *Whole-genome sequencing*. DNA libraries were prepared using the TruSeq DNA library Preparation kit Sets A and B (Illumina; estimated insert size of 343 bp). Clusters were generated with cBot, and the TruSeq PE Cluster kit v3 cBot HS (15023336_A, Illumina). Paired-end sequencing of 2 × 100 bp was performed using the TruSeq SBS kit v3-HS (200 cycles) on an Illumina HiSeq 2000 instrument.

Whole-exome sequencing. Samples were prepared with the PE DNA Sample Prep kit (Illumina) according to the manufacturer's protocol, with a few exceptions. Fragmentation was performed with the Bioruptor (Diagenode), and size selection (gel excision at ~200 bp) was carried out. Target enrichment was performed with the SeqCap EZ Human Exome Library v2.0, and clusters were generated with the cluster station (read 1) and the paired-end module (PEMIIx; read 2) using the TruSeq PE Cluster kit v5 (Illumina). We performed 76 bp paired-end sequencing on the Genome Analyzer IIx using the TruSeq SBS kit v5 (Illumina).

Transcriptome sequencing. RNA libraries were prepared using the TruSeq RNA Library Preparation kit Sets A and B (insert size of ~300 bp).

Whole-methylome bisulfite sequencing. Strand-specific MethylC-seq libraries were prepared using a previously described approach[30] with modifications. Adaptor-ligated DNA fragments with insert lengths of 200–250 bp were isolated and bisulfite converted using the EZ DNA Methylation kit (Zymo Research). PCR amplification of the fragments was performed in six parallel reactions per sample using the FastStart High Fidelity PCR kit (Roche). Library aliquots were pooled per sample and sequenced on the Illumina HiSeq 2000 platform, yielding 835 million (± 81 million (s.d.)) 101 bp paired-end reads per sample on average.

FLX amplicon sequencing of ID3. Exon 1 of *ID3* was amplified in eight primary lymphoma samples (three with two mutations, five with more mutations (maximum of four)), one cell line and three control samples using barcoded primers designed according to the 454 (Roche) Technical Bulletin (Amplicon Fusion Primer Design Guidelines for GS FLX Titanium Series Lib-A Chemistry, TCB 013-2009). PCR was performed for each sample using AccuPrime High Fidelity Tag polymerase (Invitrogen). PCR products were pooled and sequenced on the 454 platform using Roche titanium chemistry according to the manufacturer's protocol.

Bisulfite pyrosequencing of the ID3 locus. Three *ID3* regions were bisulfite pyrosequenced as previously described[31] using the Pyrosequencer ID and the DNA methylation analysis software Pyro Q-CpG 1.0.9 (Biotage).

SNP array analyses. SNP array analyses were performed using the Genome-Wide Human SNP Array 6.0 (Affymetrix).

Methylation analysis using the HumanMethylation450k BeadChip. Genomic DNA (1 µg) was bisulfite treated, applying the EZ DNA Methylation kit. DNA methylation was analyzed using the Infinium HumanMethylation450k BeadChip[32], and data were processed using GenomeStudio software (version 2011.1, Illumina), applying default settings.

Expression analyses. *ID3 RT-PCR*. RNA was reverse transcribed using the QuantiTect Reverse Transcription kit (Qiagen), and RT-PCR was performed using the primers described in **Supplementary Table 22**.

Protein blot analyses. Whole-cell lysates were prepared using RIPA buffer according to standard protocols[33]. Polyclonal antibody to ID3 (C-20, Santa Cruz Biotechnology; 1:250 dilution) in combination with a horseradish peroxidase (HRP)-conjugated secondary antibody to rabbit (1:1,000 dilution; ECL IgG HRP, linked whole antibody from donkey, GE Healthcare) was used to detect ID3. For normalization of protein loading, actin was detected using a monoclonal antibody to β-actin (1:2,500 dilution; clone AC-15, produced in mice, Sigma-Aldrich) and HRP-conjugated secondary antibody to mouse (1:1,000 dilution; ECL IgG HRP, linked whole antibody from sheep, GE Healthcare).

ID3 re-expression. Transfection of BL-2 (mutated *ID3*) and DG-75 (wild-type *ID3*) cell lines with a construct expressing ID3-GFP was performed using human ORF *ID3*-pCMV6-AC-*GFP* (OriGene) and the Cell Line Nucleofector kit V (VCA-1003, Lonza) with the Nucleofector I device. After 48 h of incubation, cells were analyzed using flow cytometry.

Cell cycle analysis. FACS-based cell cycle analyses were performed as detailed[34]. Cell cycle analysis was performed by flow cytometry using a FACSCalibur Analyzer and Cell Quest software (BD Biosciences).

Bioinformatic and statistical analysis. *DNA data processing*. Read pairs were mapped to the human reference genome (hg19, NCBI build 37.1, downloaded from the UCSC Genome Browser (see URLs)) using BWA[35] version 0.5.9-r16 (maximum insert size of 1 kb). SAMtools[36] was used to generate a coordinate-sorted BAM file, and Picard (version 1.48; see URLs) was used to merge BAM files from one sample and remove PCR duplicates.

SNV detection. For SNV calling, exome and whole-genome sequencing data were merged. Detection of SNVs was performed as described previously[37]. SNVs were functionally annotated using Annovar[38] and annotated for overlaps with SNPs (dbSNP build 135 and 1000 Genome project data) using BEDTools[39].

Indel detection. Tumor and matched control samples were analyzed with Pindel (version 0.2.4h)[40]. Events in the tumor were only considered if they were supported by at least five reads and if the number of supporting reads divided by the maximum of the read depth at the left and right breakpoint positions was larger than 0.05. The matched control sample was also analyzed by SAMtools mpileup at tumor indel positions and 10 bp up- and downstream of these sites. Variants were classified as somatic if both Pindel and SAMtools mpileup did not call a multibase variant in this region in the control sample. All somatic indel calls were manually reviewed using the Integrative Genomics Viewer[41]. Indels were annotated as described for SNVs.

Structural variant detection. Structural variants in the tumor were identified by a combination of various methods, including (i) CREST[42] using default parameters and the -g option; (ii) paired-end mapping with split-read refinement[43]; (iii) variants of >50 bp from the Pindel analysis; and (iv) read depth analysis. Structural variant candidates supported by at least two different methods were manually reviewed using the Integrative Genomics Viewer[41], and remaining calls were considered as the final structural variant set.

Detection of copy-number alterations and allelic imbalances. Allele-specific copy-number alterations were detected using an in house–developed method called allele-specific copy-number estimation from sequencing (ACE-seq; I. Bludau, B. Brors, R. Eils & M. Schlesner, unpublished data). Briefly, ACE-seq performs (i) collection of read depth from tumor and normal samples as well as allele-specific read counts in the tumor at the normal heterozygous SNP positions; (ii) genome segmentation at changes in read depth in the tumor relative to the normal sample and at changes in allelic balance using the PSCBS algorithm[44]; (iii) estimation of total copy numbers and decrease of heterozygosity; (iv) normalization according to the contamination of the tumor sample with normal cells and according to the tumor overall ploidy; and (v) calculation of allele-specific copy numbers for the genomic segments.

Furthermore, copy-number alterations and allelic imbalances were analyzed using Control-FREEC[45] and visualized using a custom R script (**Supplementary Fig. 1**). For comparison, copy-number alterations were additionally called from SNP Array 6.0 (Affymetrix) with the PennCNV[46] package.

FLX data analysis. For each sample, reads spanning exon 1 were extracted and truncated to equal length, and identical sequences were clustered. Sequence clusters were aligned pairwise against the *ID3* gene sequence using needle from the EMBOSS package[47]. To eliminate sequencing errors, all variants that had not been detected by Sanger sequencing were reverted to the reference sequence. Clusters with now identical sequences were merged, and, for the remaining clusters, the variant pattern with the number and fraction of supporting reads was reported.

ID3 mutation modeling. The structures of ID3-TCF3 and ID3-TCF4 complexes were modeled using the structures of a human ID3 dimer (Protein Data Bank (PDB) 2lfh) and a chimeric Max-TCF3 dimer (PDB 3u5v)[48]. Minor differences between the UniProt and PDB sequences were corrected using Modeller[49]. As the modeled homodimers superimposed well on each other, we constructed the heterodimer and placed it into a tetrameric and DNA-binding organization according to the structure of TCF3 (PDB 2ql2)[50] (**Supplementary Fig. 17**). We ignored the ID3 N terminus (residues 16–23), as these were disordered in the nuclear magnetic resonance (NMR) structure. We constructed a model for ID3-TCF4 in an identical fashion. These models were used to assess the impact of the mutations described in **Supplementary Table 21**.

Methylome data analysis. Sequencing reads were adaptor trimmed using SeqPrep (see URLs) and translated to a fully C-to-T converted state. Genomic alignments were performed against both *in silico* bisulfite-converted strands of the human reference genome (hg19, NCBI build 37.1) using BWA version 0.6.1-r104 (ref. 51) and the non-default parameters -q 20 -s. Previously translated bases were translated back to their original state, and reads mapping antisense to the respective reference strand were removed. Bisulfite conversion rates were estimated at >99.94% on the basis of λ phage genome spike-ins. The overall mapping rate was 88.9% on average. Single base-pair methylation ratios were determined by quantifying evidence for methylated (unconverted) and unmethylated (converted) cytosines at CpG positions. Only properly paired or singleton reads with mapping quality of ≥1 and bases with Phred-scaled quality score of ≥20 were considered.

Transcriptome data analysis. Sequence data were mapped using the segemehl algorithm[52] with default parameters and the split-read option (**Supplementary Table 4**). Small RNA sequencing data were clipped using fastx-clipper before mapping with segemehl.

Array-based gene expression analysis. For 100 lymphoma samples with known *ID3* mutational status (3 from the sequenced cases and 97 from the validation cohort), array-based gene expression data from Affymetrix GeneChip U133A were available[5]. Within mBLs[5], differentially expressed genes between samples with and without *ID3* mutation were determined using the R package limma[53]. Nominal *P* values were adjusted for multiple testing with the Benjamini-Hochberg method[54]. The smallest adjusted *P* value was equal to 0.16, and only 32 genes with a false discovery rate of <50% were found. To test whether the samples with mutated and wild-type *ID3* differed with respect to their overall gene expression pattern, three methods were used[55–57]. The *P* values, obtained using 100,000 permutations, were 0.17, 0.21 and 0.57, respectively. In the principal-component test, the first and second principal components were used to construct the test statistic. Before the analysis of differential expression, probe sets without Entrez IDs were removed from the data, and, in the case of multiple probe sets per Entrez ID, the probe set with the largest interquartile range was retained (**Supplementary Table 12**).

All gene expression (Affymetrix HG-U133A) data were jointly normalized with the VSN method[58] as described previously[4].

Copy-number analysis using aCGH and Affymetrix 250k and 500k SNP arrays. SNP arrays were processed as described[4], and aCGH data were processed as described[14]. The copy-number status of the *ID3* locus was determined using both aCGH and SNP array data.

Molecular and clinical characterization. Clinical and molecular features were compared between samples with and without *ID3* mutation. Age at diagnosis was compared using the Mann-Whitney *U* test. The gender of affected individuals, immunohistochemical staining, interphase FISH data for selected chromosomal aberrations, cell-of-origin signature (activated B cell or germinal center B cell) and molecular diagnosis were assigned as described[14] and compared using Fisher's exact test (**Supplementary Table 19**).

28. van Dongen, J.J. *et al.* Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4–CT98–3936. *Leukemia* **17**, 2257–2317 (2003).
29. Lefranc, M.P. *et al.* IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.* **27**, 209–212 (1999).
30. Lister, R. *et al.* Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**, 68–73 (2011).
31. Lamprecht, B. *et al.* Derepression of an endogenous long terminal repeat activates the *CSF1R* proto-oncogene in human lymphoma. *Nat. Med.* **16**, 571–579 (2010).
32. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295 (2011).
33. Hames, B.D. *Gel Electrophoresis of Proteins,* Vol. 3 (Oxford University Press, New York, 1998).
34. Lüschen, S. *et al.* Sensitization to death receptor cytotoxicity by inhibition of fas-associated death domain protein (FADD)/caspase signaling. Requirement of cell cycle progression. *J. Biol. Chem.* **275**, 24670–24678 (2000).
35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
36. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
37. Jones, D.T. *et al.* Dissecting the genomic complexity underlying medulloblastoma. *Nature* **488**, 100–105 (2012).
38. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
39. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
40. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
41. Robinson, J.T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
42. Wang, J. *et al.* CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* **8**, 652–654 (2011).
43. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
44. Olshen, A.B. *et al.* Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics* **27**, 2038–2046 (2011).
45. Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423–425 (2012).
46. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
47. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
48. Ahmadpour, F. *et al.* Crystal structure of the minimalist Max-E47 protein chimera. *PLoS ONE* **7**, e32136 (2012).
49. Sali, A. & Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
50. Long, A., Guanga, G.P. & Rose, R.B. Crystal structure of E47-NeuroD1/β2 bHLH domain–DNA complex: heterodimer selectivity and DNA recognition. *Biochemistry* **47**, 218–229 (2008).
51. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
52. Hoffmann, S. *et al.* Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.* **5**, e1000502 (2009).
53. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article3 (2004).
54. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. A Stat. Soc.* **57**, 289–300 (1995).
55. Goeman, J.J., van de Geer, S.A., de Kort, F. & van Houwelingen, H.C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**, 93–99 (2004).
56. Mansmann, U. & Meister, R. Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods Inf. Med.* **44**, 449–453 (2005).
57. Läuter, J., Glimm, E. & Kropf, S. New tests for data with an inherent structure. *Biometrical J.* **38**, 5–23 (1996).
58. Huber, W. *et al.* Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, S96–S104 (2002).