nature genetics

Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin

Ryan D Morin[1], Nathalie A Johnson[2], Tesa M Severson[1], Andrew J Mungall[1], Jianghong An[1], Rodrigo Goya[1], Jessica E Paul[1], Merrill Boyle[2], Bruce W Woolcock[2], Florian Kuchenbauer[2], Damian Yap[2], R Keith Humphries[2], Obi L Griffith[1], Sohrab Shah[2], Henry Zhu[3], Michelle Kimbara[3], Pavel Shashkin[3], Jean F Charlot[3], Marianna Tcherpakov[3], Richard Corbett[1], Angela Tam[1], Richard Varhol[1], Duane Smailus[1], Michelle Moksa[1], Yongjun Zhao[1], Allen Delaney[1], Hong Qian[1], Inanc Birol[1], Jacqueline Schein[1], Richard Moore[1], Robert Holt[1], Doug E Horsman[4], Joseph M Connors[2,5], Steven Jones[1], Samuel Aparicio[2], Martin Hirst[1], Randy D Gascoyne[4] & Marco A Marra[1,6]

Follicular lymphoma (FL) and the GCB subtype of diffuse large B-cell lymphoma (DLBCL) derive from germinal center B cells[1]. Targeted resequencing studies have revealed mutations in various genes encoding proteins in the NF-κB pathway[2,3] that contribute to the activated B-cell (ABC) DLBCL subtype, but thus far few GCB-specific mutations have been identified[4]. Here we report recurrent somatic mutations affecting the polycomb-group oncogene[5] *EZH2*, which encodes a histone methyltransferase responsible for trimethylating Lys27 of histone H3 (H3K27). After the recent discovery of mutations in *KDM6A* (*UTX*), which encodes the histone H3K27me3 demethylase UTX, in several cancer types[6], *EZH2* is the second histone methyltransferase gene found to be mutated in cancer. These mutations, which result in the replacement of a single tyrosine in the SET domain of the EZH2 protein (Tyr641), occur in 21.7% of GCB DLBCLs and 7.2% of FLs and are absent from ABC DLBCLs. Our data are consistent with the notion that EZH2 proteins with mutant Tyr641 have reduced enzymatic activity *in vitro*.

Advances in DNA sequencing technology have recently enabled the characterization of genomes and transcriptomes at sufficient resolution for identification of somatic point mutations[7–9]. To develop new insight into previously unidentified mutations potentially contributing to B-cell non-Hodgkin lymphomas (NHLs), we used Illumina technology to sequence genomic DNA and RNA purified from a malignant lymph node biopsy ("FL sample A") obtained from an individual with FL (Online Methods). FL Sample A was shown by immunohistochemistry to have a grade 1 FL that coexpressed CD10, BCL2 and BCL6. This sample was chosen for sequence analysis because it had an unusually simple karyotype (**Supplementary Fig. 1**), lacking the translocation t(14;18)(q32;q21) or other large-scale alterations (**Supplementary Figs. 2–5**; **Supplementary Tables 1** and **2**). We analyzed the exon sequences of this tumor for mutations in both the genome (whole-genome shotgun sequencing, WGSS) and the transcriptome (whole-transcriptome shotgun sequencing, WTSS) (**Table 1** and Online Methods). Matched constitutional DNA from the patient (FL patient A) was sequenced to reveal 'germline' sequence variants (Online Methods). We produced 25.6 aligned gigabases (Gb) from the tumor genomic library, yielding 9.47-fold redundant base coverage on average, and an additional 2.2 Gb of aligned sequence from the WTSS library, yielding 18.86-fold redundant base coverage on average within exons (**Table 1**; Online Methods). We focused our analysis on novel changes predicted to affect protein-coding sequence (Online Methods). Among these variants (**Supplementary Table 3**), we discovered a mutation affecting exon 15 of the *EZH2* gene, which encodes a portion of the EZH2 SET domain. EZH2 is the catalytic component of the PRC2 complex, which is responsible for adding methyl groups to

Table 1 Summary of exonic sequence coverage in the genome and transcriptome sequence of FL patient A

Library description	Raw reads (pairs)	Mapped reads	Total sequence coverage (bp)	Mean coverage depth of exons
FL sample A, matched germline genomic DNA	93,473,829	163,216,278	7,986,844,356	2.80
FL sample A, tumor WTSS	51,729,560	63,262,348	2,277,444,528	18.9
FL sample A, tumor genomic DNA	351,666,782	563,762,488	27,024,661,976	9.47

[1]Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada. [2]BC Cancer Agency, Vancouver, British Columbia, Canada. [3]BPS Biosciences, San Diego, California, USA. [4]Department of Pathology, [5]Division of Medical Oncology and [6]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada. Correspondence should be addressed to M.A.M. (mmarra@bcgsc.ca).

Received 10 September 2009; accepted 19 November 2009; published online 17 January 2010; doi:10.1038/ng.518

Table 2 Location and effect of all mutations in *EZH2* in FL and DLBCL determined by WTSS

Sample ID	Sample type or cell line name	Age	Sex	t(14;18)	Genomic position	Mutation[a]	Effect
HS0804	FL (sample A)	44	F	No	Chr. 7: 148139661	T→C	Tyr→His
HS0639	DLBCL	60	M	Yes	Chr. 7: 148139661	T→C	Tyr→His
HS0648	DLBCL	92	F	No	Chr. 7: 148139661	T→A	Tyr→Asn
HS0640	DLBCL	68	F	Yes	Chr. 7: 148139660	A→C	Tyr→Ser
HS0942	DLBCL	73	M	Yes	Chr. 7: 148139661	T→A	Tyr→Asn
HS0798	DB	45	M	Yes	Chr. 7: 148139661	T→A	Tyr→Asn
HS0841	KARPAS 422	73	F	Yes	Chr. 7: 148139661	T→A	Tyr→Asn
HS0900	SU-DHL-6	43	M	Yes	Chr.7: 148139661	T→A	Tyr→Asn
HS0901	WSU-DLCL2	41	M	Yes	Chr. 7: 148139660	A→T	Tyr→Phe
HS1163	OCI-LY1	NA	NA	Yes	Chr. 7: 148139661	T→A	Tyr→Asn

WTSS, whole transcriptome shotgun sequencing; FL, follicular lymphoma; DLBCL, diffuse large B-cell lymphoma; NA, not available. 'Age' and 'Sex' refer to the patient from which the sample was obtained; 'Age' is age at diagnosis.
[a]All observed mutations are heterozygous, and mutation is reported on the negative strand.

H3K27 (ref. 10), thereby repressing transcription at loci associated with histones bearing trimethylated H3K27. We established that this mutation, which is predicted to result in the replacement of Tyr641 (amino acid 641 in Q15910 and 646 in NP_004447) with a histidine, was somatic in nature by confirming its presence in tumor DNA and its absence in constitutional nontumor DNA (Online Methods). We also confirmed that the mutation was heterozygous in FL Sample AL.

Table 3 Frequency of *EZH2* Tyr641 mutations in lymphoma and benign samples

Sample type	No. samples analyzed	No. samples with *EZH2* Tyr641 mutation	Prevalence of Tyr641 mutation
FL			
Grade 1	133	10	7.5%
Grade 2	60	4	6.7%
Grade 3	28	2	7.1%
Total FL	221	16	7.2%
FL and DLBCL pairs[a]			
FL	30	2	6.7%
DLBCL	30	4	13.3%
Total FL-derived	60	6	10.0%
DLBCL			
GCB	83	18	21.7%
PMBCL	24	1	4.2%
ABC	42	0	0%
U	25	0	0%
Non-GCB	22	0	0%
Not available[b]	124	12	9.7%
Total primary DLBCL	320	31	9.7%
Other lymphoma			
MCL	25	0	0%
SLL	30	0	0%
PTCL	25	0	0%
Cell lines			
GCB cell lines[c]	7	5	71.4%
ABC cell lines[d]	2	0	0%
Benign			
Reactive lymph node	23	0	0%
Purified CD77$^+$ centroblasts[e]	8	0	0%
Total primary NHL samples	681	53	7.8%
Total cell lines	9	5	55.5%
Total benign	31	0	0%

FL, follicular lymphoma; DLBCL, diffuse large B-cell lymphoma; GCB, germinal-center B-cell subtype DLBCL defined by gene-expression profiling (GEP); PMBCL, primary mediastinal B-cell lymphoma; ABC, activated B-cell subtype of DLBCL defined by GEP; U, unclassifiable-subtype DLBCL defined by GEP; non-GCB, non–germinal center type of DLBCL defined by immunohistochemistry using the Hans criteria[23]; MCL, mantle-cell lymphoma; SLL, small lymphocytic lymphoma; PTCL, peripheral T-cell lymphoma not otherwise specified.
[a]FL and DLBCL pairs were samples derived from the same affected individual before (FL) and after transformation (DLBCL). [b]Affymetrix array analysis was not performed and hence COO information is unavailable. [c]GCB cell lines: mutated *EZH2*: DB, KARPAS 422, SU-DHL-6, WSU-DLCL2 and OCI-LY1; wild-type *EZH2*: OCI-LY7 and OCI-LY19. [d]ABC cell lines: wild-type *EZH2*: OCI-LY3 and OCI-LY10. [e]CD77$^+$ centroblasts were purified based on CD77 selection from reactive tonsils.

To determine *EZH2* mutational status in DLBCL samples, we next used WTSS to sequence the transcriptomes of 31 samples from individuals with DLBCL and 7 DLBCL cell lines. On the basis of cell-of-origin (COO) expression classification[11], the primary lymphoma samples were classified as belonging to either the ABC (n = 12), GCB (n = 15) or unclassifiable subtypes (n = 2) (Online Methods; **Supplementary Table 4**). Coverage of *EZH2* in the WTSS libraries was consistently high, ranging in these 31 samples from 5.3-fold to 187-fold redundant base-pair coverage (median 48.7-fold). Coverage of codon 641 ranged from 5-fold to 295-fold (median 46.5-fold; **Supplementary Table 4**). We identified mutations resulting in Tyr641 substitutions in 4 of these 31 samples and in 4 of the cell lines (**Table 2**). No other mutations in *EZH2* were detected, and all mutations seemed to be heterozygous. The striking recurrence of these mutations suggested that mutation of *EZH2* resulting in alteration of Tyr641 is a common feature of lymphoma. Notably, despite a median base coverage depth of 11.4-fold in the *KDM6A* (*UTX*) locus, we found no evidence for *KDM6A* mutations in these libraries.

We determined the prevalence of Tyr641-affecting mutations in both FL and DLBCL tumors by Sanger sequencing the exon containing codon 641 in 251 FL samples, of which 30 had matched DLBCL samples taken at histological transformation, and 320 primary DLBCL samples (including the original 31 samples from affected individuals) (**Supplementary Table 5**). This revealed a total of 18 FL and 35 DLBCL samples with heterozygous mutations affecting Tyr641 (**Table 3**). Of note, all such mutations detected by WTSS showed clear

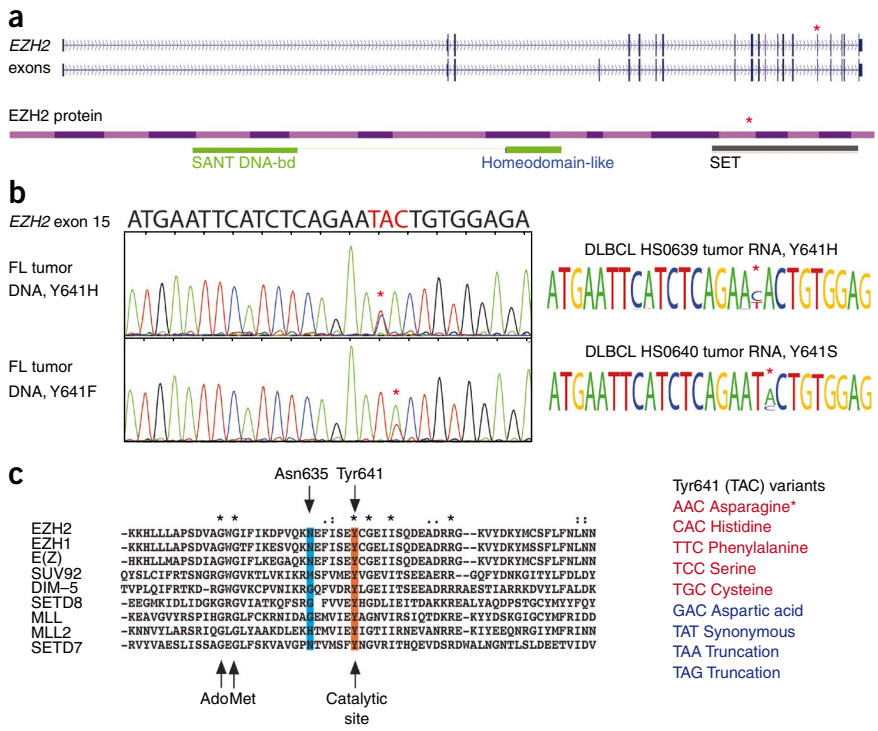

Figure 1 Recurrent mutations of Tyr641 in EZH2. (**a**) Genomic organization of the *EZH2* locus, alternative exons and protein domain structure. The location of the mutation affecting Tyr641 in exon 15 of the *EZH2* gene and protein is indicated with a red asterisk in each case. (**b**) Illustration of sequencing results. Three of the five distinct mutations and amino acid replacements in codon 641 from different lymphoma samples as detected by capillary sequencing (left) or Illumina WTSS (right). (**c**) A multiple alignment of EZH2, EZH1 (its paralog), the *Drosophila* ortholog E(Z) and six other human SET-domain proteins demonstrates the intra- and interspecies sequence conservation of SET domains. Conservation codes reported by ClustalX are shown above[24]. The predominant mutation in *EZH2* affects a key tyrosine in the catalytic site of the SET domain (orange) conserved in the *Drosophila* ortholog E(Z). With one exception, all *EZH2* mutations found in FL and DLBCL alter this amino acid. The exception was a double mutant (FL) with a second somatic mutation affecting Asn635 (blue). The mutants identified comprise five of the eight possible nonsynonymous variants of this codon (lower right, in red). Notably, the five observed amino acid changes were not found at equal frequencies. We detected a slight enrichment for Y641F (49%), with lower frequencies for Y641S (21%), Y641N (15%) and Y641H (13%), and only a single example of Y641C (2%) (**Supplementary Table 5**). Of the unobserved variants (blue), two would result in a truncated protein and the third would introduce an aspartate residue. The pattern and nature of these changes (A→G, A→T, T→G, T→A) indicated to us that these mutations are not likely to arise from activation-induced cytidine deaminase (AID)-induced somatic hypermutation at this locus[25].

evidence for expression of both alleles (**Supplementary Table 4**). To search for additional mutated sites in this gene, we also sequenced all exons of *EZH2* in tumor DNA from 24 FL samples in addition to FL sample A and found only one example of an *EZH2* mutation not affecting Tyr641 (**Fig. 1**; **Supplementary Table 5**). This mutation, affecting Asn635, was found in conjunction with a Tyr641 mutation, and we confirmed that the two mutations were in a *cis* orientation. We confirmed that these mutations were somatic in the seven individuals with FL (including 'patient A') and the two with DLBCL for whom germline DNA was available.

To exclude the possibility that such mutations can also occur in nonmalignant germinal-center B cells or in other types of lymphoma,

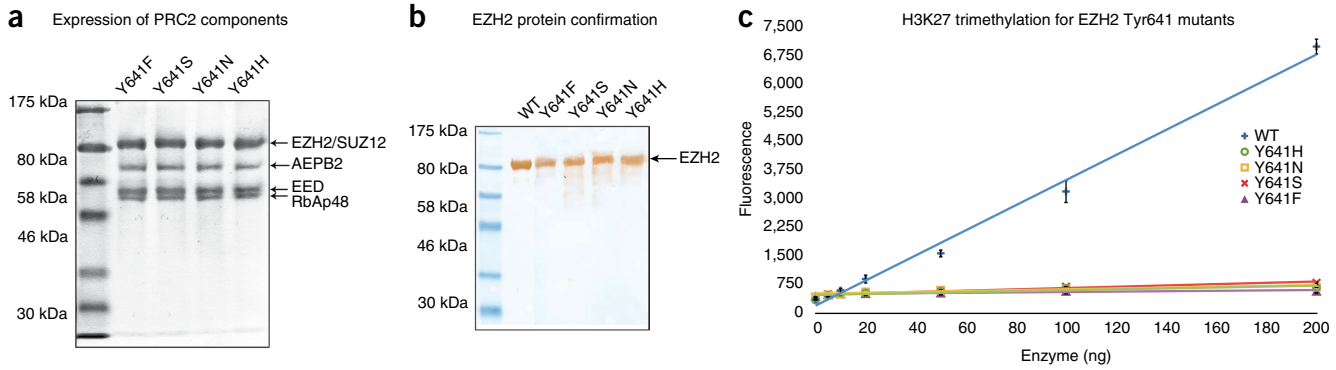

Figure 2 *In vitro* assembly and functional analysis of PRC2 with mutant and wild-type *EZH2*. (**a**) Wild-type EZH2 and each of the four Tyr641 mutants were coexpressed along with wild-type AEPB2, EED, SUZ12 and RbAp48 in Sf9 cells using a baculovirus expression system (Online Methods). Together, these five proteins associate to form an enzymatically active PRC2 complex *in vitro*. The purified complex from the Sf9 cells showed strong expression of each of these proteins and confirmed their association and assembly into PRC2. (**b**) Expression of EZH2 protein from each of the four mutant constructs was confirmed by protein blotting. (**c**) The purified complex was then assayed using biotinylated histone $H3_{21–44}$ peptide along with *S*-adenosylmethionine (in the assay buffer) to detect enzyme activity. Methylated histone H3 was measured using a highly specific antibody that recognizes only the trimethylated Lys27 residue of histone H3 (Online Methods). Europium-labeled secondary antibody was detected by time-resolved fluorescence (620 nm). PRC2 methylase activity of each mutant (and of wild-type EZH2) was tested at varying amounts of purified PRC2 (between 0 and 200 ng). The specific activities for the four mutants were calculated to be 0.001, 0.0012, 0.0011 and 0.0009 pmol/min/μg for the Y641H, Y641N, Y641S and Y641F mutants, respectively (mean = 0.00105). The wild-type enzyme (blue) showed a specific activity of 0.0071 (~6.8-fold greater). Error bars reflect the s.d. of triplicate measurements.

we sequenced this region of exon 15 in eight CD77$^+$-enriched centroblast samples from reactive tonsils and 23 reactive lymph nodes (a source of normal B cells) and 80 samples of other lymphoma types using both Sanger sequencing and targeted ultradeep Illumina re-sequencing (Online Methods; **Supplementary Fig. 6**; **Supplementary Table 6**). We also sequenced WTSS libraries generated from two additional normal centroblast samples (**Supplementary Table 4**). Consistent with our hypothesis that these mutations are unique to malignant B cells, none of these samples showed evidence for mutations affecting Tyr641 or elsewhere within the sequenced region (**Table 3**). Notably, all of the DLBCL samples for which COO was known and that were also positive for *EZH2* mutations were of the GCB subtype and not the ABC subtype. This revealed a significant enrichment of Tyr641-altering mutations among the GCB subtype of DLBCLs (**Table 3**; n = 18/83 GCB versus 0/42 ABCP = 0.00168, two-tailed Fisher's exact test).

We next assessed the effect various Tyr641 mutations would have on the structure, and potentially the function, of the EZH2 SET domain by generating a computational model (**Supplementary Fig. 7**) using the crystal structure of the highly conserved MLL1 SET domain[12] as the structural template (Online Methods). Our model indicates that Tyr641 interacts with the lysine 27 side chain of the H3 histone tail, as has been suggested in other SET domain proteins[13]. Though no *EZH2* SET-domain mutations have been reported in humans, detailed mutant phenotypes have been described in *Drosophila melanogaster*. A mutation altering the tyrosine orthologous to *EZH2* Tyr641 has been characterized in the *Drosophila* ortholog *E(z)* in an allele known as 'E(z)^1'. *Drosophila* E(z)^1 mutant protein was found to be incapable of trimethylating H3K27 *in vitro*[14].

We sought to directly determine whether EZH2 with mutant Tyr641 affects the catalytic activity of PRC2 in a cell-free methylation assay. Individual clones, each containing one of the four most frequently detected mutations (**Fig. 1**), were first expressed along with the other components of PRC2. PRC2 complexes were purified and tested *in vitro* for H3K27 methylation activity using ELISA and an antibody specific for H3K27me3 (Online Methods). The results (**Fig. 2**) indicated that, compared to wild-type EZH2, all four Tyr641 mutants consistently demonstrated a marked reduction (~7-fold) in their ability to trimethylate the H3K27 peptide. This biochemical result suggested that the four predominant Tyr641 variants observed in our sequencing study could confer reduced ability of PRC2 complexes to trimethylate H3K27 *in vivo*.

Other reports have suggested that increased abundance of *EZH2* mRNA correlates with cancer progression in tissues in which *EZH2* expression is normally low or undetectable, such as breast and prostate[6,15]. However, *EZH2* mRNA is known to be abundant in normal germinal center B cells[16], and a conditional knockout of the mouse *EZH2* ortholog indicated that the SET domain is required for early B-cell development, including rearrangement of the immunoglobulin heavy chain (*IGH*) locus[17]. Given the apparent requirement for *EZH2* in germinal center B cells, it is possible that the mechanism by which *EZH2* contributes to lymphomagenesis is distinct from the apparently straightforward increases in *EZH2* mRNA abundance observed in breast[6] and prostate[15] cancers. Expression of both *EZH2* and *BMI1* (the latter encoding the catalytic component of PRC1) has been linked to the degree of malignancy of B-cell NHLs, and perturbations in the balance of the quantities of these two proteins has been suggested as an early event in lymphomagenesis[18]. However, mutation of *EZH2* has not, to date, been implicated in B-cell malignancies or any other cancer.

Though the biological mechanism is not known, our findings suggest that Tyr641-altering mutations of *EZH2*, and possibly a reduction in H3K27 trimethylation, are involved in the pathogenesis of GCB lymphomas. The well-studied phenylalanine-tyrosine switch[19] site is known to regulate the number of methyl groups that a SET domain–containing protein can add without compromising its overall catalytic activity. Although the Tyr641 residue is distinct from the phenylalanine-tyrosine switch site, the result of our *in vitro* experiment does not rule out the possibility that these mutations may alter the product (or target) specificity of EZH2. Our finding is particularly timely in light of recent studies demonstrating enhanced DNA methylation at PRC2 targets in lymphoma as compared to normal B cells[20,21]. H3K27 trimethylation via PRC2 can be a precursor to DNA methylation and, in some cases, DNA methyltransferase may be physically coupled with PRC2 (ref. 22). Hence, Tyr641-altering mutations may contribute to the differential DNA methylation that has been observed at polycomb targets in FL[20] and DLBCL[21].

In conclusion, we have identified recurrent somatic mutations affecting a single tyrosine in the *EZH2* SET domain and have associated these with FL and DLBCL cases of only the GCB subtype. Our data indicate that mutation affecting this tyrosine is among the most frequent genetic events observed in GCB malignancies after t(14;18)(q32;q21) translocation. The altered tyrosine corresponds to a key residue in the active site of the EZH2 protein and, consistent with the results of functional studies of a comparable mutation in the *Drosophila E(z)* ortholog, we show that PRC2 complexes containing mutated EZH2 protein have reduced H3K27 trimethylation activity *in vitro*. We have shown that a wild-type copy of *EZH2* is present in all samples with Tyr641-altering mutations and have also detected expression of both alleles in the mutant samples profiled by transcriptome sequencing. This, along with the fact that all lymphomas with mutations in *EZH2* seem to have a mutation affecting Tyr641, sets *EZH2* apart from the pattern of mutational inactivation seen in the case of *KDM6A* (*UTX*), which seems to behave as a tumor suppressor gene[5]. Aside from the recurrence of inactivating mutants in UTX, EZH2 is the only protein affecting H3K27 methylation status to be identified as a target of somatic mutation in cancer, and it is the first in which recurrent mutations of the SET domain appear to be restricted to a specific lymphoma subtype.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession code. dbGaP: Sequencing reads and validated variants are available to approved investigators through NCBI dbGaP, accession number phs000235.v1.p1. These data were produced as part of the National Cancer Institute's Cancer Genome Characterization Initiative, and additional data can be obtained via the data portal at http://cgap.nci.nih.gov/cgci.html.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This study was funded in part by grants from the National Cancer Institute Office of Cancer Genomics (see below), the National Cancer Institute of Canada (NCIC) Terry Fox Foundation New Frontiers Program Project Grant (grant no. 016003/grant type 230/project title: Biology of Cancer: Follicular Lymphoma as a Model of Cancer Progression) and Genome Canada/Genome BC Grant Competition III (project title: High Resolution Analysis of Follicular Lymphoma Genomes) to J.M.C., M.A.M., R.D.G. and D.E.H. and was supported by The Terry Fox Foundation (grant no. 019001). In addition, N.A.J. is a research fellow of the Terry Fox Foundation through an award from the NCIC (019005) and the Michael Smith Foundation for Health Research (MSFHR) (ST-PDF-01793). M.A.M. is a Terry Fox Young Investigator and a Michael Smith Senior Research Scholar. A.J.M. is supported by a Fellowship Award from The Leukemia & Lymphoma

Society. R.D.M. is a Vanier Scholar (Canadian Institutes for Health Research) and is also supported by a MSFHR senior graduate fellowship. The laboratory work for this study was undertaken at the Genome Sciences Centre, British Columbia Cancer Research Centre and the Centre for Translational and Applied Genomics, a program of the Provincial Health Services Authority Laboratories. The authors thank the BC Cancer Foundation and the Lion's Club International for their support. The authors gratefully acknowledge D. Gerhard for helpful discussions. Special thanks to C. Suragh and A. Drobnies for expert project management assistance. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract no. NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

AUTHOR CONTRIBUTIONS

M.A.M., R.D.G., D.E.H. and J.M.C. conceived of the study and led the design of the experiments. R.D.M. performed the WGSS and WTSS analysis, produced **Figures 1** and **2** and, with M.A.M., wrote the manuscript. N.A.J. prepared the samples, performed sample sorting and COO analysis and contributed to the text. O.L.G. and R.D.M. analyzed gene expression data. T.M.S., A.J.M. and J.E.P. performed sequence validation experiments and visual inspection of capillary sequence data. D.S. and M.M. constructed multiplexed libraries for deep resequencing of *EZH2*. H.Z., M.K., P.S., J.F.C., D.Y. and M.T. conducted enzymatic assays. I.B. performed statistical analysis and contributed to the manuscript. J.A. and S.J. produced the model for EZH2 and contributed to the manuscript. M.B. and B.W.W. prepared samples and performed FACS. F.K. and R.K.H. validated expression findings in the RNA. A.D., H.Q., R.C. and S.S. performed copy number analysis. A.T., Y.Z., R.H., M.H. and R.M. produced the sequencing libraries and performed the sequencing. R.V. processed raw sequencing data. R.G. identified candidate mutations. J.S., M.H. and S.A. conceived of experiments and contributed to the text.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Alizadeh, A.A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
2. Compagno, M. *et al.* Mutations of multiple genes cause deregulation of NF-κB in diffuse large B-cell lymphoma. *Nature* **459**, 717–721 (2009).
3. Kato, M. *et al.* Frequent inactivation of A20 in B-cell lymphomas. *Nature* **459**, 712–716 (2009).
4. Bea, S. *et al.* Diffuse large B-cell lymphoma subgroups have distinct genetic profiles that influence tumor biology and improve gene-expression-based survival prediction. *Blood* **106**, 3183–3190 (2005).
5. Kleer, C.G. *et al.* EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc. Natl. Acad. Sci. USA* **100**, 11606–11611 (2003).
6. van Haaften, G. *et al.* Somatic mutations of the histone H3K27 demethylase gene *UTX* in human cancer. *Nat. Genet.* **41**, 521–523 (2009).
7. Morin, R. *et al.* Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45**, 81–94 (2008).
8. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
9. Ley, T.J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
10. Kirmizis, A. *et al.* Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. *Genes Dev.* **18**, 1592–1605 (2004).
11. Lenz, G. *et al.* Stromal gene signatures in large-B-cell lymphomas. *N. Engl. J. Med.* **359**, 2313–2323 (2008).
12. Southall, S.M., Wong, P.S., Odho, Z., Roe, S.M. & Wilson, J.R. Structural basis for the requirement of additional factors for MLL1 SET domain activity and recognition of epigenetic marks. *Mol. Cell* **33**, 181–191 (2009).
13. Dillon, S.C., Zhang, X., Trievel, R.C. & Cheng, X. The SET-domain protein superfamily: protein lysine methyltransferases. *Genome Biol.* **6**, 227 (2005).
14. Joshi, P. *et al.* Dominant alleles identify SET domain residues required for histone methyltransferase of Polycomb repressive complex 2. *J. Biol. Chem.* **283**, 27757–27766 (2008).
15. Varambally, S. *et al.* The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419**, 624–629 (2002).
16. Raaphorst, F.M. *et al.* Cutting edge: polycomb gene expression patterns reflect distinct B cell differentiation stages in human germinal centers. *J. Immunol.* **164**, 1–4 (2000).
17. Su, I.H. *et al.* Ezh2 controls B cell development through histone H3 methylation and Igh rearrangement. *Nat. Immunol.* **4**, 124–131 (2003).
18. van Kemenade, F.J. *et al.* Coexpression of BMI-1 and EZH2 polycomb-group proteins is associated with cycling cells and degree of malignancy in B-cell non-Hodgkin lymphoma. *Blood* **97**, 3896–3901 (2001).
19. Couture, J.F., Dirk, L.M., Brunzelle, J.S., Houtz, R.L. & Trievel, R.C. Structural origins for the product specificity of SET domain protein methyltransferases. *Proc. Natl. Acad. Sci. USA* **105**, 20659–20664 (2008).
20. O'Riain, C. *et al.* Array-based DNA methylation profiling in follicular lymphoma. *Leukemia* **23**, 1858–1866 (2009).
21. Martín-Subero, J.I. *et al.* New insights into the biology and origin of mature aggressive B-cell lymphomas by combined epigenomic, genomic, and transcriptional profiling. *Blood* **113**, 2488–2497 (2009).
22. Viré, E. *et al.* The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* **439**, 871–874 (2006).
23. Hans, C.P. *et al.* A significant diffuse component predicts for inferior survival in grade 3 follicular lymphoma, but cytologic subtypes do not predict survival. *Blood* **101**, 2363–2367 (2003).
24. Thompson, J.D., Gibson, T.J. & Higgins, D.G. Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinformatics* **00**, 2.3.1–2.3.22 (2002).
25. Pasqualucci, L. *et al.* Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature* **412**, 341–346 (2001).

ONLINE METHODS

Sample acquisition. Two samples from the affected individual initially tested (FL patient A) were used. Both had ~70% tumor content, as determined on the basis of the coexpression of CD19 and lambda by flow cytometry, and were 'fresh' frozen at source. The first was taken at the time of diagnosis and was used for WTSS and WGSS. The second was acquired at the time of progression and was flow sorted to >95% purity. It was analyzed by karyotyping and fluorescence *in situ* hybridization (FISH) for the presence of a translocation t(14;18) using the dual-color, dual-fusion probe. It was also analyzed for copy number alterations by array comparative hybridization (aCGH) and fingerprint profiling[26] and for loss of heterozygosity (LOH) by Affymetrix 500K array. All DLBCL samples profiled by WTSS were fresh-frozen biopsies having >50% tumor content as detected by flow cytometry. All other specimens used in this study were obtained at the time of diagnosis and were derived from archived fresh-frozen tissue or frozen tumor cell suspensions. Germline DNA was obtained from peripheral blood in live subjects and from CD19-negative sorted tumor cell suspensions, obtained using Miltenyi magnetic beads (Miltenyi Biotec), for deceased subjects. All lymphoma samples were diagnosed according to the World Health Organization criteria of 2008 by an expert hematopathologist (R.D.G.). Benign specimens included reactive pediatric tonsils or purified CD77-positive centroblasts sorted from reactive tonsils using Miltenyi beads (Miltenyi Biotec). The tumor specimens were collected as part of a research project approved by the University of British Columbia–British Columbia Cancer Agency Research Ethics Board (BCCA REB) and are in accordance with the Declaration of Helsinki. Informed consent was obtained from all individuals whose samples were profiled using WTSS or WGSS. Our protocols stipulate that these data will not be released into the public domain but can be made available via a tiered-access mechanism to named investigators of institutions agreeing by a materials transfer agreement to honor the same ethical and privacy principles required by the BCCA REB.

Preparation and sequencing of Illumina libraries. RNA was extracted from a total lymph node section using the AllPrep DNA/RNA Mini Kit (Qiagen) and treated with DNase I. For whole-transcriptome shotgun sequencing (WTSS/RNA-seq) analysis, we used a modified method similar to the protocol we have previously described[7]. Briefly, poly(A)$^+$ RNA was purified, using the MACS mRNA isolation kit (Miltenyi Biotec), from 5–10 μg of DNase I–treated total RNA as per the manufacturer's instructions. Double-stranded cDNA was synthesized from the purified poly(A)$^+$ RNA using the Superscript Double-Stranded cDNA Synthesis kit (Invitrogen) and random hexamer primers (Invitrogen) at a concentration of 5 μM. The cDNA was fragmented by sonication and a paired-end sequencing library prepared following the Illumina paired-end library preparation protocol (Illumina).

Genomic DNA for construction of WGSS libraries was prepared from the same biopsy material using the Qiagen AllPrep DNA/RNA Mini Kit (Qiagen). DNA quality was assessed by spectrophotometry (260 nm/280 nm and 260 nm/230 nm absorption ratios) and gel electrophoresis before library construction. DNA was sheared for 10 min using a Sonic Dismembrator 550 with a power setting of "7" in pulses of 30 s interspersed with 30 s of cooling (Cup Horn, Fisher Scientific) and then analyzed on 8% PAGE gels. The 200–300-bp DNA fraction was excised and eluted from the gel slice overnight at 4 °C in 300 μl of elution buffer (5:1 (vol/vol) LoTE buffer (3 mM Tris-HCl, pH 7.5, 0.2 mM EDTA)/7.5 M ammonium acetate) and was purified using a Spin-X Filter Tube (Fisher Scientific) and by ethanol precipitation. WGSS libraries were prepared using a modified paired-end protocol supplied by Illumina Inc. This involved DNA end-repair and formation of 3′ adenosine overhangs using the Klenow fragment of DNA polymerase I (3′–5′ exonuclease minus) and ligation to Illumina PE adaptors (with 5′ overhangs). Adaptor-ligated products were purified on QIAquick spin columns (Qiagen) and PCR-amplified using Phusion DNA polymerase (NEB) and ten cycles with the PE primer 1.0 and 2.0 (Illumina). PCR products of the desired size range were purified from adaptor ligation artifacts using 8% PAGE gels. DNA quality was assessed and quantified using an Agilent DNA 1000 series II assay (Agilent) and Nanodrop 7500 spectrophotometer (Nanodrop), and DNA was subsequently diluted to 10 nM. The final concentration was confirmed using a Quant-iT dsDNA HS assay kit and Qubit fluorometer (Invitrogen). For sequencing, clusters were generated on the Illumina cluster stations using v1 cluster reagents. Paired-end reads were generated using v3 sequencing reagents on the Illumina GA_{ii} platform following the manufacturer's instructions. Image analysis, base-calling and error calibration were performed using v1.0 of Illumina's Genome analysis pipeline. Paired-end WTSS and WGSS libraries were sequenced to 36, 50 or 76 cycles. The WGSS library comprised a mixture of 13 flow cell lanes of 36-nucleotide (nt) reads, 16 lanes of 50-nt reads and 6 lanes of 76-nt reads.

Targeted ultra-deep resequencing using read indexing. This procedure describes the individual PCR amplification of EZH2 exon 15, indexing of individual amplicons and subsequent pooling and sequencing. Individual indexes allow the deconvolution of reads deriving from individual samples in multiplexed libraries such that many samples can be concurrently sequenced in the same library. Genomic DNA from individual samples was normalized to 5 ng/μl, and 5 ng of each sample was PCR amplified using Phusion DNA polymerase (New England Biolabs) in 96-well format using gene-specific primers (Primer EZH2_015R3 and Primer EZH2_015F, **Supplementary Table 7**) to produce ~300-bp amplicons. Hot-start PCR conditions were 98 °C for 60 s, then 36 cycles of 98 °C for 10 s, 60 °C for 15 s and 72 °C for 30 s, and a final extension at 72 °C for 5 min. Amplicons were cleaned using AMPure beads (Beckman Coulter) on a Biomek F/X (Beckman Coulter) and eluted with 40-μl elution buffer EB (Qiagen). Cleaned amplicons were quality-control tested on a 1.2% SeaKem LE Agarose gel (Cambrex) using 1× TAE buffer. Bands were quantified by the QBit Fluorometer (Invitrogen) high-sensitivity assay. Approximately 500 ng of each amplicon DNA sample was then phosphorylated and end-repaired in 50-μl reactions at room temperature 20–25 °C for 30 min (5 U T4 DNA polymerase, 1 U Klenow DNA polymerase (exonuclease minus), 100 U T4 polynucleotide kinase and 0.4 mM dNTP mix (Invitrogen). End-repair reactions were cleaned using AMPure beads, and dATP was added to the 3′ ends using 5 U Klenow DNA polymerase (exonuclease minus) and 0.2 mM dATP in 1× Klenow Buffer (Invitrogen) with 30-min incubation at 37 °C in a Tetrad thermal cycler (MJ Research). DNA was again cleaned on AMPure beads using a Biomek FX. Adaptor ligation (10:1 ratio) was completed with 0.03 μM adaptor (multiplexing adaptors 1 and 2, **Supplementary Table 7**), 100 ng DNA, 5 U T4 DNA ligase, 0.2 mM ATP and 1× T4 DNA Ligase Buffer (Invitrogen) for 30 min at room temperature. Adaptor-ligated DNA was cleaned using AMPure beads on a Biomek FX. A selection of DNA samples were quantified on a QBit (Invitrogen). Phusion DNA polymerase, 15-cycle indexing enrichment PCR was performed using Primers 1.0 and 2.0 (IDT) and 96 custom indexing primers (indexes shown in **Supplementary Table 6**). The PCR program was as follows: 98 °C for 60 s, followed by 15 cycles of 98 °C for 10 s, 65 °C for 15 s and 72 °C for 30s . The PCR products were cleaned using AMPure beads and eluted in 40 μl elution buffer EB (Qiagen). Quality of product was assessed by quality-control gels with 1.75% SeaKem LE agarose in 1× TAE (0.2 μl of every amplicon) and on a Bioanalyzer-1000 (Agilent Technologies). All 96 ~400-bp amplicons from each plate (15 μl from each well) were then pooled into a separate 1.5-ml microcentrifuge tube. Hence, one tube represents a plate of 96 pooled and indexed PCR products from 96 distinct DNA templates. The 400-bp DNA size fraction was purified using 8% PAGE gels (1× TAE) and eluted from the gel slice overnight at 4 °C in 400 μl of elution buffer (5:1 (vol/vol) LoTE buffer/7.5 M ammonium acetate). Gel pieces were filtered using a Spin-X Filter Tube (Fisher Scientific). DNA was precipitated using ethanol, quantified using an Agilent DNA 1000 series II assay (Agilent Technologies) and then diluted to 10 nM. The final concentration was confirmed using a Quant-iT dsDNA HS assay kit and QBit fluorometer (Invitrogen). An individual library was constructed from each indexed sample (comprising amplicons from up to 96 distinct template DNAs). Each of these libraries was sequenced on a single flowcell lane.

SNV analysis of tumor DNA and RNA sequence. All reads were aligned to the human reference genome (hg18) or (for WTSS) to a genome file that has been augmented with a set of all exon-exon junction sequences using the MAQ aligner[27] v0.7.1. Candidate single-nucleotide variants (SNV) were identified in the aligned genomic sequence reads and the transcriptome (WTSS) reads using an approach similar to one we have previously described[7]. One key difference in our variant calling in this study was the application of a Bayesian SNV identification algorithm ('SNVmix') currently under development by

doi:10.1038/ng.518

our group (see URLs below)[28]. This approach is able to identify SNVs with a minimum coverage of two high-quality (Q20) bases. All sites assessed as being polymorphisms (SNPs) were disregarded, including variants matching a position in dbSNP or the recently sequenced personal genomes of Venter[29], Watson or the anonymous Asian[30] and Yoruban[31] individuals. Additionally, all candidate mutations also found in the genomic sequence from patient A's germline DNA were ignored. For the targeted resequencing experiment, coverage was generally greater than 1,000× read depth at codon 641. Hence, we used all unambiguously mapped reads spanning this site to determine the percentage of reads with a high-quality mismatch (Illumina base quality ≥20). These percentages are reported for each sample in the supplementary information (**Supplementary Table 6**).

Amplicon sequencing for SNV identification and Sanger sequence validation. Exon 15 of *EZH2* was PCR amplified from genomic DNA using EZH2_ASP_1 and EZH2_ASP_2. Priming sites for M13 forward -21 and M13 reverse were added to their 5′ ends to allow direct Sanger sequencing of amplicons. Unless otherwise stated, amplicons were produced from genomic DNA from both the tumor and matched germline DNA from the same patient. All capillary traces were analyzed using Mutation Surveyor, and all variants were visually inspected to confirm their presence in tumor and absence from germline traces.

Computational modeling of EZH2 wild-type and mutant SET domain. The EZH2 SET domain sequence was used to search for the structural template for homology modeling in the Protein Data Bank. The available crystal structure of the MLL1 SET domain (PDB ID 2w5z)[12] was identified as the best template (with sequence identities of 39% for the SET domain and no alignment gaps). A three-dimensional model of the EZH2 SET domain was constructed via the protein modeling server SWISS-MODEL[32]. Because MLL1 is a H3 Lys4 (H3K4) binding protein, there was some concern that the target lysine residue of EZH2 (Lys27) might not reside in the same conformation. Another concern is that the MLL1 crystal structure is in an open conformation and this conformation has reduced methyltransferase activity compared to the closed ones. The conformation change may shift the position of Tyr641. To address these concerns, we built alternative models using other structures as templates. We used the H3 Lys9 (H3K9) binding proteins EHMT1 (PDB ID 2RFI), DIM-5 (1PEG), SUV39H2 (2R3A) and G9a (2O8J), as well as the H3K36 binding protein SETD2 (3H6L). The striking overlap of the conserved tyrosine residue corresponding to Tyr641 confirms that the position of Tyr641 remains unchanged in all proteins regardless of an open or closed conformation. The cocrystallized H3 peptides in 1PEG and 2RFI helped us confirm that the conformations of Lys4 and Lys9 are quite similar in those models. Therefore, we assume that Lys27 in EZH2 will have a conformation close to that shown in the model.

***In vitro* EZH2 H3K27 trimethylation assay.** Mutant constructs were generated using site-directed mutagenesis of the RefSeq EZH2 (NM_004456) with an N-terminal histidine tag. Wild-type EZH2 and each of the four Tyr641 mutant constructs were coexpressed along with wild-type AEPB2, EED, SUZ12 and RbAp48 in *Spodoptera frugiperda* (Sf9) cells using a baculovirus expression system (pVL1392, cloned using BamHI and EcoRI). Together, these five proteins associate to form an enzymatically active PRC2 complex *in vitro*. Expression of EZH2 protein from each of the four mutant constructs was confirmed by protein blotting and detected using anti-EZH2. Assay plates are coated with biotinylated histone $H3_{21–44}$ peptide. Purified PRC2 was added to the plate along with *S*-adenosylmethionine (in the assay buffer) to detect enzyme activity. Methylated histone H3 was measured using a highly specific mouse-derived monoclonal antibody, which recognizes only the trimethylated Lys27 residue of histone H3[33] (Active Motif, catalog number 39535.). The secondary antibody, which is labeled with europium, was detected using time-resolved fluorescence (620 nm). PRC2 methylase activity of each mutant (and wild-type EZH2) was tested at varying purified PRC2 amounts (between 0 and 200 ng).

Cell lines. DB[34], KARPAS 422 (ref. 35), SU-DHL-6 (ref. 36) and WSU-DLCL2 (ref. 37) are cell lines obtained from DSMZ and all "OCI-LY"[38] lines were obtained from L. Staudt (US National Institutes of Health).

Cell-of-origin (COO) determination. Total RNA was reversed transcribed (one cycle) and hybridized to U133-2 Plus arrays according to the manufacturer's protocol (Affymetrix). CEL files were normalized using robust multichip analysis (RMA). COO was calculated using model scores for ABC and GCB derived from the 185-gene model described by Lenz *et al.*[11] and the Bayesian formula described by Wright *et al.*[39].

Copy number analysis of tumor DNA. BAC array comparative genomic hybridization was performed as previously described[40] and did not identify any significant alterations. The tumor DNA was analyzed for large copy number alterations using an Affymetrix 500K SNP array as previously described[41] and using peripheral blood as a matched normal comparator. The sequencing data were also used to directly infer the presence of large-scale deletions and amplifications. This was accomplished by probabilistic identification of deviations in the proportion of unambiguously mapped reads between the normal and tumor genomic libraries as previously described[28]. Because of the smaller number of sequence reads from the matched normal tissue, aligned reference reads were first used to define genomic bins of equal coverage containing 200 mapped reads. The sequencing depth of the normal genome provided 684,029 bins with a median size of 3,953 bp, representing 2.942 Gb of the hg18 assembly. A hidden Markov model (HMM) was used to classify and segment continuous regions into five discrete states—copy number loss (HMM 1), neutral (HMM 2), gain (HMM 3), amplification (HMM 4) and high-level amplification (HMM 5)—using methodology outlined previously[42]. All segments and their HMM states are included in **Supplementary Table 2** and listed in **Supplementary Figure 5**.

URLs. SNVmix version 0.11.7; http://compbio.bccrc.ca/?page_id=204.

26. Krzywinski, M. *et al.* A BAC clone fingerprinting approach to the detection of human genome rearrangements. *Genome Biol.* **8**, R224 (2007).
27. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
28. Shah, S.P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
29. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
30. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
31. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
32. Kopp, J. & Schwede, T. The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res.* **32**Database issue, , D230–D234 (2004).
33. Baskind, H.A. *et al.* Functional conservation of asxl2, a murine homolog for the *Drosophila* enhancer of trithorax and polycomb group gene asx. *PLoS One* **4**, e4750 (2009).
34. Beckwith, M., Longo, D.L., O'Connell, C.D., Moratz, C.M. & Urba, W.J. Phorbol ester-induced, cell-cycle-specific, growth inhibition of human B-lymphoma cell lines. *J. Natl. Cancer Inst.* **82**, 501–509 (1990).
35. Dyer, M.J., Fischer, P., Nacheva, E., Labastide, W. & Karpas, A. A new human B-cell non-Hodgkin's lymphoma cell line (Karpas 422) exhibiting both t (14;18) and t(4;11) chromosomal translocations. *Blood* **75**, 709–714 (1990).
36. Epstein, A.L. *et al.* Biology of the human malignant lymphomas. IV. Functional characterization of ten diffuse histiocytic lymphoma cell lines. *Cancer* **42**, 2379–2391 (1978).
37. Al-Katib, A.M. *et al.* Bryostatin 1 down-regulates mdr1 and potentiates vincristine cytotoxicity in diffuse large cell lymphoma xenografts. *Clin. Cancer Res.* **4**, 1305–1314 (1998).
38. Tweeddale, M.E. *et al.* The presence of clonogenic cells in high-grade malignant lymphoma: a prognostic factor. *Blood* **69**, 1307–1314 (1987).
39. Wright, G. *et al.* A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc. Natl. Acad. Sci. USA* **100**, 9991–9996 (2003).
40. Cheung, K.J. *et al.* Genome-wide profiling of follicular lymphoma by array comparative genomic hybridization reveals prognostically significant DNA copy number imbalances. *Blood* **113**, 137–148 (2009).
41. Delaney, A.D., Qian, H., Friedman, J.M. & Marra, M.A. Use of Affymetrix mapping arrays in the diagnosis of gene copy number variation. *Curr. Protoc. Hum. Genet.* **59**, 8.13.1–8.13.16 (2008).
42. Shah, S.P. *et al.* Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics* **22**, e431–e439 (2006).