Supplemental Data

Supplemental Methods

PCR amplification of *IGHV* and identification of the variable region gene segments

Genomic DNA samples were used in this study. The *IGHV* region was detected using a multiplex PCR protocol based on the European BIOMED-2 collaborative study. Multiplexed Framework 1 (FR1) primers targeting consensus sequences within each family of variable gene segments were combined with a consensus Joining (JH) primer. Leader primers were used in optimisation experiments but failed to amplify *IGHV* across all patient samples. Previous *IGHV* sequencing using Leader primers revealed a low frequency of N-gly sites in the first ~17 amino acids of the FR1 region. [3, 5] Therefore it is unlikely we have missed potential N-gly sites in this region when using the FR1 primers. Primers were purchased from Sigma and sequences can be found in the table below. Each primer had a final concentration of 10µM. The PCR mix consisted of 5µl 5x Green GoTaq Flexi buffer, 0.13µl GoTaq polymerase, 0.5µl dNTP mix (10µM), 1.5µl of magnesium chloride (25mM), 1µl FR1 primer mix (10µM), 1µl JH (10µM), 2µl of genomic DNA (50ng/µl) and 13.87µl nuclease free water to make up a final volume of 25µl. A no template control consisting of all components apart from DNA was included to exclude false positive results. PCR tubes were briefly centrifuged to collect the mix at the bottom before being placed into a thermocycler (Bio Rad Tetrad 2) using the following conditions; 95^{o}C for 2 minutes (initial denaturation) followed by an amended 38 cycles of 95^{o}C for 30 seconds, 58^{o}C for 30 seconds, 72^{o}C for 1 minute and a final extension step of 72^{o}C for 7 minutes. PCR products were separated on a 2% agarose gel (100ml of 1x TBE and 4µl of Gel Red) for 1 hour at 100 volts. A 100bp DNA ladder (exACTGene™ 100bp PCR DNA Ladder, Fisher Scientific) was used to identify bands at the 200-300bp size under UV light (Geneflash, Syngene Bio Imaging) which were then excised from the gel and purified using the Zymoclean Gel Extraction kit (Cambridge Biosciences). Purified product was eluted in 10µl of elution buffer. 5µl of eluted product (20ng/µl) with 5µl of JH primer (10µM) was sent to GATC Biotech Sanger Sequencing services in Germany for bidirectional capillary electrophoresis using the ABI PRISM 3730xl Genetic Analyser (Applied Biosystems). Sanger sequencing trace chromatograms were visualised using the BioEdit software to check for quality before sequences were aligned to all known immunoglobulin genes using the IMGT/V-QUEST web-tool (www.imgt.org/IMGT_vquest/vquest) that identifies the germline V gene segment of closest homology to the experimental sample. The PCR was then repeated as described above but with the correct V gene FR1 primer instead of the multiplex. The PCR product was then sent for sequencing with the FR1 primer to identify the specific Diversity (D) and J gene segments and therefore the entire variable region.

	Primer name	Primer sequence
Forward primers	VH1 FR1	GGCCTCAGTGAAGGTCTCCTGCAAG
	VH2 FR1	GTCTGGTCCTACGCTGGTGAAACCC
	VH3 FR1	CTGGGGGGTCCCTGAGACTCTCCTG
	VH4 FR1	CTTCGGAGACCCTGTCCCTCACCTG
	VH5 FR1	CGGGGAGTCTCTGAAGATCTCCTGT
	VH6 FR1	TCGCAGACCCTCTCACTCACCTGTG
Reverse primer	JH consensus	CTTACCTGAGGAGACGGTGACC

Biomed primer composition used to amplify the *IGHV* gene.

Selection of tumour related reads and analysis of N-glycosylation motifs

Genewiz merged high quality paired sequence reads when reads were overlapped by at least 12bp and the overlapped region was identical. A detailed description of the merging script used can be found at: https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmerge-guide/. The target sequence between highly conserved flanking regions was extracted from each merged pair and unique sequences identified following the Genewiz bioinformatics pipeline with their abundance (count) calculated.

Unique sequences of counts of 10 and above were selected from the provided excel file and converted into FASTA format. Sequences were uploaded into the IMGT/HighV-QUEST search page, the high throughput version of IMGT/V-QUEST that can analyse up to 500,000 immunoglobulin rearranged sequences per run. The IMGT reference directory aligns sequences to reference human Ig sets to obtain *IGHV* gene usage, somatic mutations and rearrangement. Following completion, the csv summary file produced by IMGT/HighV-QUEST was used to identify sequences that were similar to the major clone of the sample, previously amplified through PCR, previously identified through Sanger sequencing. This included the following criteria; In-frame rearrangement, identical V gene usage and similar/identical CDR3 regions (defined as having identical D and J gene usage and similar positioning of non-synonymous mutations in this region) and presence of CDR3 amino acid anchors; C (Cysteine at amino acid position 102) and W (Tryptophan at position 118). Sequences which did not meet this criteria were excluded from further analysis.

Selected sequences were aligned using MAFFT (multiple sequence alignment program) (Galaxy, Version 7.221.3) and cut down to cover a uniform area of the variable gene. Those with deletion or insertion of nucleotides were excluded from further analysis as we wanted to evaluate the single base substitutions of SHM only. Duplicate sequences were removed, leaving unique sequences that differ comparatively by at least 1bp. These sequences were designated subclone status.

N-glycosylation motifs are made up of an asparagine followed by any amino acid (except proline) and ending in either serine or threonine. There are 20 possible nucleotide combinations that encode motifs; aac/aat-NNN-agt/tct/tcc/tca/tcg/act/acc/aca/acg/agc. NNN represents the middle codon which can be any amino acid, except proline.

All sequences were interrogated for the above combinations in the BioEdit Sequence Alignment Editor program using the filtering option in which the 9 characters were treated as a distinct substring. Filtering was first restricted to the locations of motifs identified in the MC through Sanger sequencing, and then throughout the variable region. For the sequences that were positive for at least one of these combinations, we analysed the middle amino acid for proline codons to ensure it was a legitimate N-gly site. Any sequences with CCU, CCC, CCA, and CCG in between asparagine and serine/threonine were not regarded as N-gly motifs. To verify N-gly site findings, subclone sequences were submitted in FASTA format to the NetN-glyc 1.0 server which identifies N-gly sites. The N-gly site status of each subclone was then assigned before subclone sequences were submitted to the IgTree algorithm (described below).

All our samples were positive for N-gly sites in the heavy chain and therefore high-throughput sequencing of the light chain was omitted. Sanger sequencing of the light chain variable region revealed patients 1-3 to have no N-gly sites in the light chain of the major clone across temporal samples.

Analysis of clonal expansion using IgTree algorithm

The unique subclone sequences were saved as an NBRF/PIR file, the input file for the IgTree program. IgTree is an algorithm used to generate lineage trees from immunoglobulin variable region gene sequences. Trees are generated based on SHM profiles of the unique subclones and in relation to the root germline sequence, marked G.L. [17] The N-nucleotides identified from the V, D and J segments used were incorporated into the germline CDR3 sequence to ensure that lineage trees represented the SHM profile of the V, D and J regions.

Five folders were created in the IgTree runner folder, with the output tree file being in the .vsdot folder. The program Graphviz (gvedit.exe) was used to visualise the tree. The root of the tree is the germline sequence marked 'G.L' with experimental sequences assigned to either leaves or internal nodes of the tree. Each tree node represents a single mutation separating the sequences. Those separated by more than one mutation are indicated by the numbers on the branches. Theoretical sequences generated by the algorithm of the IgTree program are visualised as white nodes.

Supplemental Tables

Patient	Disease Event	PF Clusters	Yield (Mbases)	% >= Q30 Bases	Mean Quality Score	Estimated Coverage*
1	Diagnosis	1,319,287	662	90.75	35.65	2,198,812
	Transformation	1,365,431	685	89.14	35.26	2,275,718
2	1st relapse	1,188,214	596	90.06	35.49	1,980,357
	3rd relapse	944,471	474	90.12	35.49	1,574,118
	Transformation	1,546,380	776	89.37	35	2,577,300
3	2nd relapse	1,320,495	663	88.79	35.22	2,200,825
	3rd relapse	929,244	466	87.89	35.02	1,548,740
	Transformation	1,224,512	615	89.03	35.27	2,040,853

Supplemental Table 1. Sequencing metrics for Illumina paired-end sequencing of temporal samples derived from patients 1-3. Sequencing metrics for patient's 4-6 can be obtained from original publications. [16, 17] * Based on a 300bp amplicon size.

Patient	Disease Event	No. of aligned reads	No. of aligned reads following filtering process (total)	No of reads containing dominant VDJ rearrangement (% of total)	No of reads containing minor VDJ rearrangements (% of total)
1	Diagnosis	1,192,523	340618	337315 (99.03)	3303 (0.97)
	Transformation	1,220,139	341253	324028 (94.95)	17225 (5.05)
2	1st relapse	1,046,604	273419	259865 (95.04)	13554 (4.96)
	3rd relapse	829,744	145464	144488 (99.33)	976 (0.67)
	Transformation	1,375,958	456091	452569 (99.23)	3522 (0.77)
3	2nd relapse	1,169,453	368376	325014 (88.23)	43362 (11.77)
	3rd relapse	808,265	118836	84495 (71.10)	34341 (28.90)
	Transformation	1,085,563	256916	217889 (84.81)	39027 (15.19)

Supplemental Table 2. Read information regarding patients 1-3 generated by Illumina sequencing platform. Patients 4, 5 and 6 are excluded due to lack of raw data regarding count numbers.

Patient	Age at diagnosis	Disease Event	Major clone IgH-V3 DH JH Rearrangement identified from IMGT-HIGH/VQUEST	% Homology of major clone to IgH-V3	Biopsy Site	Histological grade	Time from diagnosis years (months)	Overall survival years (months)	Line of therapy
1	N/A	Diagnosis	V3-30, D3-16, J6	89.24	N/A	N/A	0	N/A	N/A
		Transformation	V3-30, D3-16, J6	87.89	N/A	N/A	1(6)		
2	40	1st relapse	V3-11, D3-16, J1	85.65	N/A	N/A	0(5)	4(2)	Rituximab + BEAM autograft (Jul 97)
		3rd relapse	V3-11, D3-16, J1	86.1	N/A	N/A	2(5)		Bexxar (Jul 99)
		Transformation	V3-11, D3-16, J1	85.2	N/A	N/A	4		Methotrexate (Jun 00)
3	48	2nd relapse	V3-48, D1-26, J4	79.82	Left axillary lymph node	N/A	12(3)	N/A	Expectant management (Aug 00)
		3rd relapse	V3-48, D1-26, J4	79.82	N/A	N/A	15(8)		N/A
		Transformation	V3-48, D1-26, J4	79.55	Left axillary lymph node	N/A	15(9)		R-CHOP (Mar 04)
4	43	1st relapse	V3-48, D3-10, J6	88.57	Left inguinal lymph node	Grade 2	2(8)	14(7)	CHOP, Fludarabine, Rituximab, Chlorambucil, Bortezomib
		2nd relapse	V3-48, D3-10, J6	88.57		Grade 1	4		Interferon, MethylPrednisolone
5	68	1st relapse	V3-23, D4-23, J6	89.11	Right axillary lymph node	Grade 3A	1(4)	9(2)	No treatment
		3rd relapse	V3-23, D3-3*, J5*	83.47	Right femoral lymph node	Grade 1	1(9)		
6	50	FL diagnosis	V3-23, D5-18, J6	N/A	Right cervical lymph node	Low grade	0	N/A	N/A
		tFL diagnosis	V3-23, D5-18, J6	N/A	Right cervical lymph node	N/A	4(10)		
		tFL relapse	V3-23, D5-18, J6	N/A	Retroperitoneal lymph node	N/A	6(9)		

Supplemental Table 3. Additional clinical information regarding 6 patients who underwent *IGHV* sequencing. * Although samples of patient 5 have different DJ rearrangements according to IMGT, when aligned they show highly similar CDR3 regions and share a t(14:18) breakpoint, indicating a clonal relationship.

Patient	Disease Event	IgH-V gene	% Homology of major clone to IgH-V gene	No. of N-gly sites	Region in *IGHV*	AA position of N-gly site	N-gly motif in major clone
A	Diagnosis	V4-34	90.52	1	CDR3	108	NST
	Transformation	V4-34	89.33	1	CDR3	108	NST
B	Diagnosis	V3-15	94.22	1	CDR2	56	NIT
	Transformation	V3-15	93.88	1	CDR2	56	***NKS***
C	Diagnosis	V3-23	80.69	1	CDR3	108	NFS
	Transformation	V3-23	83.33	1	CDR3	108	NFS
D	Diagnosis	V3-7	87.02	4	CDR1, CDR2, CDR2, FR3	30, 56, 62, 69	NFS, NYS, NET, NLS
	2nd relapse	V3-7	78.64	4	CDR1, CDR2, CDR2, FR3	30, 56, 62, 69	NFS, NYS, NET, NLS
	3rd relapse	V3-7	85.66	4	CDR1, CDR2, CDR2, FR3	30, 56, 62, 69	NFS, NYS, NET, NLS
	Transformation	V3-7	83.57	4	CDR1, CDR2, CDR2, FR3	30, 56, 62, 69	NFS, NYS, NET, NLS
E	1st relapse	V4-59	89.74	1	CDR2	56	NVS
	Transformation	V4-59	90.56	1	CDR2	56	NVS

Supplemental Table 4. N-gly motifs in extension cohort, patients A-E. The V4-34 gene has a germline encoded N-gly motif in the CDR2 region. However as demonstrated in Patient A, this is lost and an acquired site is gained at a different location. Italic text in red refers to differences in the location or amino acid (aa) sequence of N-gly sites, referred to as N-gly motif, within the major clone across temporal samples.

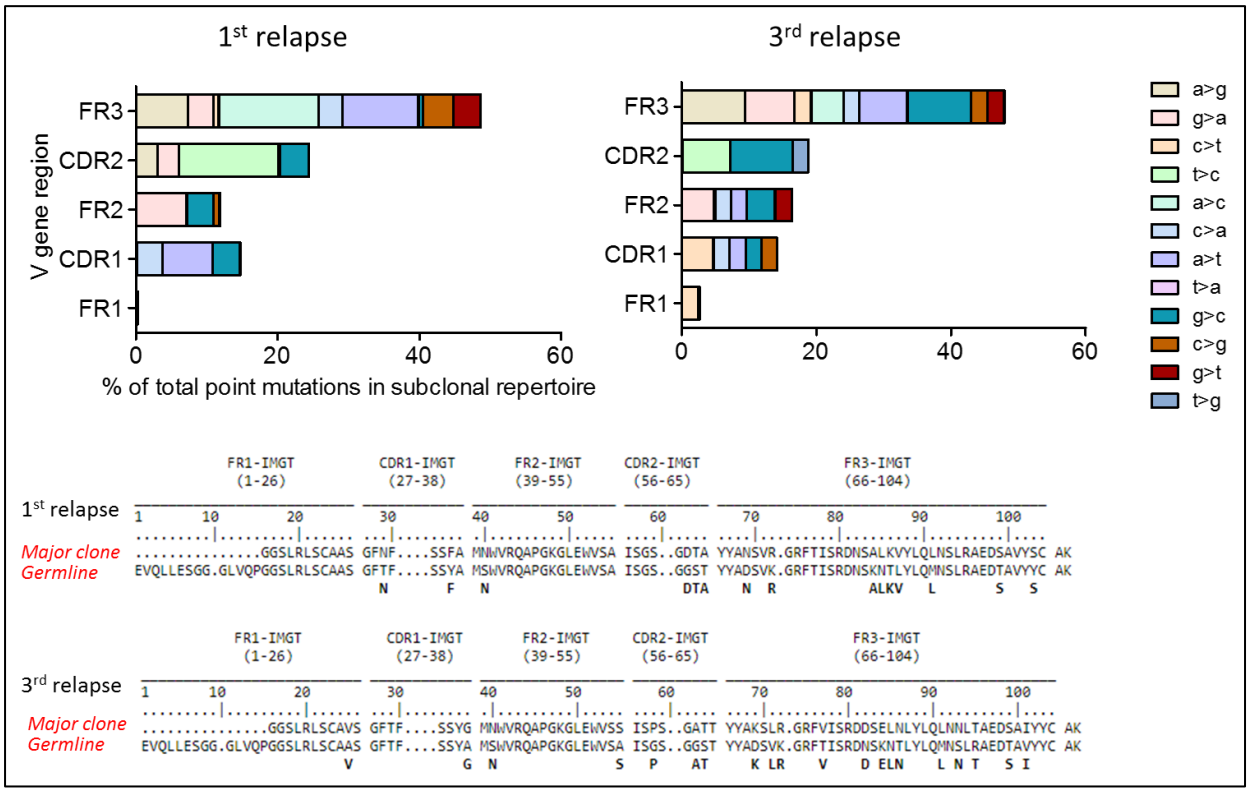
Patient	Disease Event	% of tumour related counts with all N-gly motif sites present	% of tumour related counts without all N-gly motif sites present
1	Diagnosis	99.32	0.68
	Transformation	99.42	0.58
2	1st relapse	94.7*	5.3*
	3rd relapse	94.8*	5.2*
	Transformation	86.7*	13.3*
3	2nd relapse	99.15	0.85
	3rd relapse	99.01	1
	Transformation	99.3	0.7

Supplemental Table 5. Total count numbers for disease event for patients 1-3. High % of counts containing motifs reflect how they are a significant feature of the tumour bulk. * For patient 2, the tumour related count values represent the clones in which either all four motif sites are present or clones in which ≥1 motif is absent. Patients 4, 5 and 6 are excluded due to lack of available information regarding count numbers.

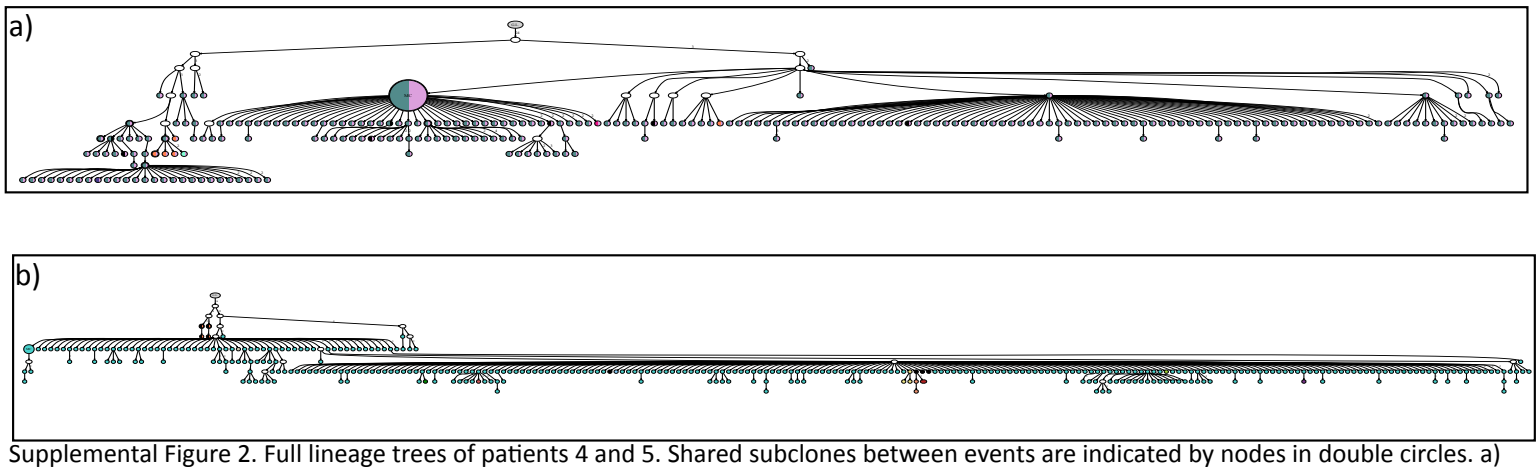
Shared negative clone	% of total tumour count in FL2	% of total tumour count in FL3
1	0.042152	0.037872
2	0.040614	0.056808
3	0.034152	0.067460
4	0.034152	0.035505
5	0.031383	0.031955
6	0.030153	0.056808
7	0.027999	0.018936
8	0.027383	0.027221
9	0.019999	0.036689
10	0.016615	0.024854
11	0.013846	0.015386
12	0.010153	0.021303
13	0.008307	0.040239
14	0.008307	0.026037
15	0.007384	0.020120
16	0.007077	0.014202
17	0.006461	0.014202
18	0.006461	0.023670
19	0.006154	0.023670
20	0.004923	0.017753

Supplemental Table 6. % of total tumour count for the negative shared subclones of patient 3.

Supplemental Figures



Supplemental Figure 1. Discordance in the SHM pattern of subclones identified in temporal samples from patient 5 across the variable region. The V gene amino acid sequence of the major clones for the serial samples is given below. Major clone and germline amino acid sequences are indicated, with amino acid changes from the germline sequence highlighted in bold letters below. The discordant amino acid changes reflect the outcome of the distinct SHM pattern in the two events. Yet a clonal relationship was previously established between the two, indicating an early divergence in evolution.[16]



Supplemental Figure 2. Full lineage trees of patients 4 and 5. Shared subclones between events are indicated by nodes in double circles. a) Lineage tree for patient 4, the 2nd relapse disease event. b) Lineage tree for patient 5, the 3rd relapse disease event.