Supplementary Methods

Clinical samples

All newly diagnosed leukemia patients at Lucile Packard Children's Hospital are asked to donate extra bone marrow from the diagnostic material for research studies. From that bank, we selected samples that were available in reverse chronological order reasoning that more recent samples would have an increased probability of having viable cells. No other criteria were used to include or exclude samples.

Flow cytometry and cell sorting

Aside from 5 samples that were only available as DNA, we performed cell sorting on all evaluable samples. BMMCs were incubated with the following antibodies from Invitrogen, BioLegend or eBiosciences for analysis by flow cytometry and cell sorting: anti-CD19-Qdot605 (clone SJ25-C1), anti-CD10-FITC (clone HI10a), anti-CD45-PE (clone HI30), anti-CD38-PE/Cy7 (clone HIT2), anti-CD81-PerCP/eFluor710 (clone 1D6), anti-CD34-Alexa647 (clone 581) and anti-CD20-APC/eFluor780 (clone 2H7). Following incubation, cells were washed and suspended in PBS/2FBS containing 4',6-diamidino-2-phenylindole (DAPI) to enable exclusion of non-viable cells. Cells were acquired and sorted using a FACSAria (BD Biosciences) instrument. From each patient sample, normal B-cells and up to two ALL leukemic populations, termed 'malignant 1' and 'malignant 2', were sorted. Normal B-cells were always defined as $CD45^{+}CD10^{-}$ $CD19^{+}CD20^{+}$. ALL leukemic cells were defined on a patient-by-patient basis at the time of sorting (see Supplemental Table 2 for marker phenotypes of each sorted population from each

patient). Sorted cells were pelleted and lysed in RLT Plus buffer for nucleic acid isolation. Analysis of flow cytometry data files was performed using FlowJo (Ashland, OR).

IgH amplification and sequencing

First stage primers were designed so as to allow for the amplification of all known alleles of the germline IgH sequences. To minimize the risk of not amplifying a specific clonotype sequence because of a somatic hypermutation at the primer hybridization site, we designed three sets of primers in the V_H segments. Therefore each V_H segment is amplified by 3 primers, decreasing the likelihood of somatic hypermutations interfering with amplification.

Primers were optimized such that each possible V_H and J_H segment was amplified at a similar rate so as to minimally skew the repertoire frequency distribution during the amplification process. Specificity of the primers was, in contrast, not optimized as the primer sequences could be mapped and removed from the eventual sequence read. Thus a given sequence may have been amplified by multiple primers. This methodology led to slightly different primer designs than have been published previously for similar IgH amplification approaches.[1] The numbers of primers and the positions of these primers are shown in Faham and Willis.[2]

At the 5' ends of V_H segment primers a universal sequence was appended. These sequences are complementary to a set of second stage PCR primers. Similarly the primers on the J_H side had a 5' tail with a universal sequence that is complementary to second stage PCR primers. The second stage PCR primers contained the sequence primer and the P5 sequence used for cluster formation in the Illumina Genome Analyzer sequencer. The primers on the V_H

side of the amplification constituted one of a set of primers, each of which had a 3' region that annealed to the overhang sequence appended in the first reaction but which further contained one of multiple 6 base pair indices that allowed for sample multiplexing on the sequencer. Each of these primers further contained a 5' tail with the P7 sequence used for cluster formation in the Illumina Genome Analyzer sequencer.

First stage PCR was carried out using a high fidelity polymerase (AccuPrime, Life Technologies) for 16 cycles. 1/100 of this amplification reaction was then used as the template for a second PCR reaction using the second stage primers including the primer containing a sample index that was unique to a particular sample. A second stage PCR was carried out for 22 cycles. Different samples were pooled to be sequenced in the same Illumina Genome Analyzer sequencing lane. The pool was then purified using the QIAquick PCR purification kit (Qiagen).

Cluster formation and sequencing was carried out per the manufacturer protocol (Illumina, Inc., La Jolla, CA). Specifically, three sequencing reactions were performed. First 115 bp were sequenced from the J_H side sufficient to sequence through the junctional sequence from J_H to V_H. At this point, the synthesized strand was denatured and washed off. A second sequencing primer was annealed that allowed the sample index to be sequenced for 6 cycles to identify the sample. At this point the reverse complement strand was generated per the Illumina protocol. A final sequencing read of 95 bp was obtained from the V_H- to-J_H direction providing ample sequence to map the V_H segment accurately.

Clonotype determination

After exclusion of low quality reads, sequence data were then analyzed to determine the clonotype sequences including mapping to germline V_H and J_H consensus sequences. First, the sample index sequences were used to identify which of the sequences originate from which of the pooled samples. Sequences whose index were not a perfect match to one of the indices used in a specific run were excluded. Next the forward read was used to map the J_H segment. Since all the sequences started from the same position of the J_H segments, all the J_H segments started at a predefined sequencing position. The first 25 bp of the J_H segments were used to map the J_H segment. Any read with more than 5 high quality mismatches to the known J_H segments was excluded from further analysis.

After J_H segment identification, V_H segments were mapped. The reverse read was used for this purpose. First, the V_H primer was mapped and excluded. Thereafter, the next 70 bases of the reverse read were mapped to the known V_H segments. Reads that did not map to J_H and V_H segments were excluded. The next step in mapping involved identifying the frame that related the forward and reverse reads and this allowed a continuous sequence from J_H to V_H to be constructed. This was done using the last 15 bases of the forward read which were reliably within the V_H segment regardless of NDN length. While these bases could be of relatively lower sequence quality as they were at the terminal end of a long read, they could be used to map within a single identified V_H segment in order to identify the position at which the two reads could be joined. Finally, the known V_H and J_H sequences to which the reads map were used to identify the point in the forward read at which the sequences at the junctions diverged from these mapped segments.

To generate a clonotype, at least two identical sequences needed to be identified. Given sequencing and PCR errors, many different but highly related clonotypes might originate from one clonotype. We therefore allowed for coalescence of highly related clonotypes. For example two sequences with one base difference but present at vastly different frequencies were consistent with sequencing or PCR error. On the other hand two sequences with two base differences and present at similar magnitudes were not likely to arise from sequencing error.

Analytical methods – Effective NDN bases

We assumed that each base of the NDN sequences was independently distributed with each base having a 25% probability of occurring. The NDN region was ordered from the end of the J_H segment toward the V_H segment, and we compared the number of consecutive matches. Six consecutive bases are required to reach significance at $p=5x10^{-4}$. The bases in the D segment come from a limited number of available alleles and as such are clearly dependent on each other. Analysis of 8235 clonotypes from one of the CLL controls using the IMGT JunctionAnalysis tool showed a 6% probability of any two clonotypes sharing a D segment. This roughly corresponded to the probability of matching 2 consecutive bases (1/16 or 0.0625), so the D segment is compressed as 2 effective NDN bases.

Supplementary Table Legends

Table S1. Patient characteristics.

Table S2. Flow cytometry summary.

Table S3. Evolution by index clonotype. Each row refers to a specific index clone. Related index clones in individual patients are noted by gray-shading. The index clone J_H and V_H segments and NDN sequence are indicated. The effective NDN (column E) is described in the supplementary methods. The frequencies (columns F-I) of the specific index clone in different cell populations are indicated. The J column indicates the number of clonotypes evolving from the index clone as well as their percentage of all clones in the sample. The frequency of these evolved clones in different cell populations are indicated in columns K-N.

Supplemental Table S4. Upstream and downstream V_H replacement in evolved clonotypes.

References

1. van Dongen JJ, Langerak AW, Bruggemann M, 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*. 2003;17(12):2257-2317.

2. M. Faham, T. D. Willis. Monitoring health and disease status using clonotype profiles. United States Patent application 2011/0207134A1, published August 25, 2011.

Supplementary Tables

Table S1. Patient characteristics.

Patient ID	Down Syndrome	Age at Diagnosis	Initial WBC	CNS	Cytogenetics	NCI Risk Group	day 29 flow MRD	Relapse (Y/N)
1		11	499	2	Normal	high-risk	<0.1%	Y
2		4	12	1	Normal	standard-risk	ND	N
3		3	3	1	Hyperdiploidy without concurrent trisomies #4, #10 and #17	standard-risk	<0.1%	N
4		4	6	2	TEL-AML	standard-risk	<0.1%	N
5		3	106	1	TEL-AML	high-risk	<0.1%	N
6		5	9	1	Hyperdiploidy without concurrent trisomies #4, #10 and #17	standard-risk	<0.1%	N
7		2	9	1	TEL-AML	standard-risk	<0.1%	N
8		5	8	1	hyperdiploid; +4, +10,+17 and TEL-AML	standard-risk	<0.1%	N
9		4	36	2	TEL-AML	standard-risk	<0.1%	N
10		8	2	2	hyperdiploid; +4, +10, +17	standard-risk	<0.1%	N
11		2	58	1	TEL-AML	high-risk	<0.1%	N
12		7	51	2	Not done	high-risk	ND	Y
13		4	4	1	TEL-AML	standard-risk	<0.1%	N
14		2	107	1	Other	high-risk	<0.1%	N
15		3	0.5	1	TEL-AML	standard-risk	<0.1%	N
16		4	43	1	hyperdiploid; +4, +10, +17	standard-risk	<0.1%	N
17		3	9	1	hyperdiploid; +4, +10, +17	standard-risk	<0.1%	N
18		4	6	1	Other	standard-risk	<0.1%	N
19		14	24	1	TEL-AML	high-risk	<0.1%	N
20		2	19	1	TEL-AML	standard-risk	<0.1%	N
21		9	9	1	Normal	standard-risk	<0.1%	N
22		17	902	3	Other	high-risk	>1%	Y
23		6	48	1	Normal	standard-risk	<0.1%	N
24		14	135	2	Ph+ (9:22)	very high-risk (pH+)	ND	N
25		12	3	1	Other	high-risk	<0.1%	N
26		4	30	1	Other	standard-risk	>0.1%, <1%	N

27		4	12	1	Hyperdiploid without concurrent trisomies #4, #10 and #17	standard-risk	<0.1%	N
28		17	4	1	Hyperdiploid clone concurrent trisomies #4, #10, #17	standard-risk	<0.1%	N
29		3	5	1	TEL-AML	standard-risk	<0.1%	N
30		8	4	1	Normal	standard-risk	>0.1%, <1%	N
31		15	7	1	Hyperdiploid without concurrent trisomies #4, #10 and #17	high-risk	<0.1%	N
32		5	17	1	hyperdiploid; +4, +10, +17	standard-risk	ND	N
33		5	19	2	Other	standard-risk	<0.1%	N
34		9	110	1	Ph+ (9:22)	very high-risk (pH+)	ND	N
35		5	75	1	TEL-AML	high-risk	<0.1%	N
36		3	3	1	Hyperdiploid without concurrent trisomies #4, #10 and #17	standard-risk	<0.1%	N
37		2	34	1	hyperdiploid; +4, +10, +17	standard-risk	<0.1%	N
38		17	56	1	Other	high-risk	<0.1%	N
39		4	2	1	Hyperdiploid without concurrent trisomies #4, #10 and #17	standard-risk	<0.1%	N
40		0.75	880	2	MLL (4:11)	MLL rearranged infant BALL	ND	N
41	Y	18	135	2	Normal (except Tri21)	high-risk	>1%	N
42		9	6	1	Hyperdiploid without concurrent trisomies #4, #10 and #17	standard-risk	<0.1%	N
43		14	844	2	Other	high-risk	ND	Y
44		3	11	1	Hyperdiploid without concurrent trisomies #4, #10 and #17	standard-risk	>1%	N
45	Y	6	5	1	Hyperdiploid without concurrent trisomies #4, #10 and #17	standard-risk	<0.1%	N
46		6	11	1	TEL-AML	standard-risk	>0.1%, <1%	N
47		14	10	1	Other	high-risk	>1%	N
48		12	75	1	Hyperdiploid without concurrent trisomies #4, #10 and #17	high-risk	<0.1%	Y
49		15	2	1	hyperdiploid; +4, +10, +17	high-risk	>0.1%, <1%	N
50		2	17	1	TEL-AML	standard-risk	<0.1%	N
51		13	2	1	Normal	high-risk	<0.1%	N

Table S2. Flow cytometry summary.

Patient ID	% viable cells (DAPI-negative; low scatter)	INPUT # of cells (Unsorted)	Total mapped sequencing reads (Unsorted)	# of normal B-cells	% normal B-cells (of viable)	% ALL (CD45low CD10+; of viable)	mal1 phenotype	# of mal1 sorted	% mal1 (of ALL population)	mal2 phenotype	# of mal2 sorted	% mal2 (of ALL population)
1	29	2,324,088	366,600	168	2.1	9.04	CD45lowCD10-	2,028,913	82.2	CD45lowCD10+	167,205	7.1
2	35.1	233,068	151,274	7,878	3.2	0.431	CD45lowCD10+CD81+	1,524	58.3	CD45lowCD10+CD81-	588	23.3
3	30.6	193,033	946,895	11,743	7.1	55.4	CD45lowCD10+CD20-	77,193	97.4	CD45lowCD10+CD20+	1,803	2.3
4	6.16	1,031	25,802	NOT sorted	8.8	4.05	NOT sorted			NOT sorted		
5	30.5	1,271,478	1,529,916	25,623	2.9	74.7	CD45lowCD10+CD34-	1,844,792	74.6	CD45lowCD10+CD34+	460,188	19.2
6	49.2	532,867	2,230,162	9,828	2.1	81.7	CD45lowCD10+	242,354	31.3	CD45-CD10+	341,639	43.2
9	43.8	1,039,865	1,572,475	21,739	1.5	77.6	CD45lowCD10+CD34-	955,901	65.1	CD45lowCD10+CD34+	490,401	34
10	45.4	711,548	1,486,112	6,802	0.7	72.5	CD45-CD10+	547,346	70.4	CD45lowCD10+	36,988	4.8
11	14.6	30,065	10,372	207	1.3	0	CD45+CD20-CD19-	10,542	78.7	CD45+CD20-CD19+	1,826	14
12	12.4	16,511	1,144,307	0	0	89.6	CD45lowCD10+CD81+	3,273	63.1	CD45lowCD10+CD81-	1,894	36.3
13	11.4	70,926	203,734	1,821	8.4	1.11	CD45lowCD10-/+	1,285	100	NOT sorted		
14	37.3	1,273,006	3,823,441	115,413	11.2	79.7	CD45lowCD10+	860,349	59.3	CD45-CD10+	236,056	16.3
15	38.1	383,631	1,224,434	8,653	1.3	77.6	CD45lowCD10+CD34+	224,129	79.9	CD45lowCD10+CD34-	53,908	19.3
16	60.1	1,285,714	1,432,888	16,913	1.7	67.6	CD45lowCD10+CD20+	646,469	53.5	CD45lowCD10+CD20-	536,608	44.7
17	3.5	101,834	1,388,158	1,895	21	23.4	CD45lowCD10+CD20+	2,210	67.1	CD45lowCD10+CD20-	1,025	31
21	25	273,139	3,700,394	5,762	3.7	76.7	CD45lowCD10+CD20+	74,066	50.4	CD45lowCD10+CD20-	65,197	44.4
22	67.3	2,204,516	3,814,667	2,036	0.1	93.7	CD45lowCD10+CD20-	2,419,690	42.7	CD45lowCD10+CD20+	1,691,163	30.4
23	27.8	287,630	2,510,472	3,630	1.7	52.7	CD45lowCD10+CD38+	74,524	50.9	CD45lowCD10+CD38-	56,453	36.8
24	41	517,362	61,514	10,604	1.8	87	CD45lowCD10+CD20-	325,583	58.8	CD45lowCD10+CD20+	233,945	37.1
25	58.3	495,173	1,039,776	1,333	0.1	54.4	CD45lowCD10+CD20-	864,473	82.4	CD45lowCD10+CD20+	169,566	17
26	57.6	766,310	1,423,941	38,693	3.4	82.5	CD45lowCD10+	681,813	46	CD45-CD10+	375,986	25.9
27	44.5	94,441	907,478	11,779	13.7	57.3	CD45lowCD10+	51,763	49.8	CD45-CD10+	11,024	10.5
28	56.3	34,071	786,060	111	0.2	94.9	CD45lowCD10+CD20-	39,586	66.1	CD45lowCD10+CD20+	19,268	32.2
29	56.1	731,887	956,595	79,723	5.5	45.5	CD45lowCD10+CD20-	313,524	65.5	CD45lowCD10+CD20+	154,037	32.3
30	62	4,109,808	1,120,684	25,451	0.9	55.5	CD45-CD10+	1,638,601	45.7	CD45lowCD10+	672,140	19.2
31	12.3	224,013	250,394	654	1.9	2.93	CD45lowCD10low	21,859	30.3	CD45-/lowCD10+	2,094	2.9

32	51.8	674,209	1,422,995	19,652	2.2	92.5	CD45lowCD10+	238,389	23.9	CD45-CD10+	672,477	66.8
33	32.3	669,093	743,752	4,881	1.3	56.6	CD45lowCD10++	201,290	36.4	CD45lowCD10+	126,203	22.4
34	51.2	1,181,758	1,30,4184	23,160	1.3	92.4	CD45lowCD10+CD20+	2,044,886	79.6	CD45lowCD10+CD20-	472,041	19.2
35	35	1,109,850	1,319,312	55,232	4.3	69.1	CD45lowCD10+CD20-	502,244	62	CD45lowCD10+CD20+	286,977	35.9
36	32	42,434	398,735	6,418	21.7	31.5	CD45lowCD10+	8,576	84.4	CD45-CD10+	2,958	8.4
37	17.7	35,305	237,205	2,872	15.3	50.1	CD45lowCD10+CD20-	4,557	53.3	CD45lowCD10+CD20+	3,832	44.7
38	47.8	1,071,429	1,433,991	5,696	0.6	91.1	CD45lowCD10+	792,248	61	CD45-CD10+	333,365	25.9
39	8.4	215,566	1,173,519	60	0.2	61.8	CD45-CD10+	18,970	52.5	CD45lowCD10+	5,922	16.5
40	42.8	4,742,090	1,344,122	47,407	3.9	0.646	CD45lowCD10-	1,119,569	48.8	CD45lowCD10+	20,060	0.7
41	51.8	3,114,843	235,941	10,748	0.3	97.8	CD45+CD10+	963,654	29.3	CD45lowCD10+	620,389	20
42	53	554,740	1,517,738	4,907	0.9	90.8	CD45-CD10+	555,229	75.2	CD45lowCD10+	71,440	9.7
43	58.9	741,701	1,283,825	6	0.002	94.9	CD45lowCD10+CD34-	1,423,570	92.6	CD45lowCD10+CD34+	44,215	3
44	52.9	2,094,325	1,751,011	34,262	2.6	82.9	CD45-CD10+	1,283,992	65.8	CD45lowCD10+	300,647	15.8
45	45.5	761,356	1,490,199	3,145	0.5	50.3	CD45lowCD10+	342809	40.5	CD45lowCD10low	25817	3
46	25.1	369,189	394,627	1,099	0.4	93.5	CD45lowCD10+CD20-	192,611	86.4	CD45lowCD10+CD20+	27,916	12.5
47	36.8	470,142	1,387,823	1,489	0.9	52.8	CD45lowCD10+CD81+	253,696	75.3	CD45lowCD10+CD81-	64,024	19.7
48	39.6	344,951	1,592,498	2,224	0.9	88.9	CD45lowCD10+CD34+	119,415	37.1	CD45lowCD10+CD34-	167,542	52.5
49	50.4	969,010	1,440,857	10,115	0.8	93.9	CD45lowCD10+	613,087	43.4	CD45-CD10+	553,427	39.2
50	45.5	835,928	919,046	56,654	4.4	66.1	CD45lowCD10+CD34-	605,581	77.8	CD45lowCD10+CD34+	144,941	18.9
51	40.4	181,202	73,049	2,484	1.2	80.2	CD45lowCD10+CD34+	116,628	80.1	CD45lowCD10+CD34-	23,418	16.2

Input # of cells (unsorted) indicates the total number of cell equivalent that were used in the IgH DNA amplification reaction.

Table S3. Evolution by index clonotype.

Patient ID	DC J-segment allele	DC V-segment allele	NDNsequence	Number of effective NDN bases	Frequency (%) of reads mapping to index clone	Frequency (%) in sort for normal B-cells	Frequency (%) in first gate for malignant cells	Frequency (%) in second gate for malignant cells	Total # (and %) non-index evolved clones	Total read frequency of evolved clones (%)	Total read frequency of evolved clones in sort for normal B-cells (%)	Total read frequency of evolved clones in first gate for malignant cells (%)	Total read frequency of evolved clones in second gate for malignant cells (%)
3	IGHJ5*02	IGHV1-2*02	ACGAGTCAACGCTCGAGTAGCAGCTACCACCACTACAACCG	19	90.3	0.5	98.3	83.3	14 (0.3%)	0.08	0	0.2	0
5	IGHJ4*02	IGHV2-26*01	AACTCCCCGAACCATAGTAGCCC	6	75.9	3	93.7	41.1	25 (1.2%)	0.01	0	0.005	0.03
5	IGHJ4*02	IGHV3-30-3*01	AAAGATCCCCAGTTAGCCATATCGGGGGGCCTTATCTCCTAGG	34	10.2	0.1	1.4	21.1	613 (30.5%)	0.7	0.05	0.09	1.8
5	IGHJ3*02	IGHV3-30-3*01	TGGGCTTCGTCCATC	7	9.6	0.1	4.1	32.7	30 (1.5%)	0.03	0	0.03	0.1
6	IGHJ5*02	IGHV1-17*02	ATATAACCAGTCAAAACGGAGAAAGGAAA	NA	33.3	0.1	2.2	35.7	3 (0.2%)	0.003	0	0.007	9.00E-04
6	IGHJ5*02	IGHV4-31*03	ATATAACCAGTCAAAATATCCCCTC	9	9.2	0.2	43.5	2.6	2 (0.1%)	0.002	0	0.006	7.00E-04
6	IGHJ5*02	IGHV1-14*01	CAGAAAAGGGAGAG	10	18.3	0.09	2	25	22 (1.5%)	0.04	0	0.003	0.05
6	IGHJ5*02	IGHV3-13*01	ACGGCGAGCTGCTATATGCCCTCG	14	10.5	0.2	50	3.1	1 (0.07%)	8.00E-04	0	0.003	0
6	IGHJ5*02	IGHV6-1*01	CCCTCCACATAAGCAGCTGGTACTACGACCCTCC	18	9.4	0.07	0.8	18	857 (59.9%)	13.6	0.08	0.7	11.1
7	IGHJ6*02	IGHV3-64*01	TAACTCCCCGAACCATTCTAATCTCTCCCCCCCGTCTCGCGGCC	24	88.7	NA	NA	NA	25 (3.4%)	0.4	NA	NA	NA
8	IGHJ4*02	IGHV2-70*01	TAACCACTACTATCATACTTTCCCTGGGG	14	78.7	NA	NA	NA	279 (57.6%)	1.4	NA	NA	NA
8	IGHJ4*02	IGHV3-30-3*01	TAACCACTACTATCATAGTATCCAC	7	17.3	NA	NA	NA	22 (4.5%)	0.1	NA	NA	NA
9	IGHJ4*02	IGHV3-11*01	TCCATCCTCCCGGCCCTCGA	11	63.2	3.6	52.1	58.1	42 (8.8%)	0.1	0	0.06	0.1
9	IGHJ6*02	IGHV3-15*01	GGGAGGGTAGTAATAACCACTACTATCATAGTAATACGTCCCTT	15	34.9	3.7	46.7	40.7	0 (0.0%)	0	0.02	0	0
10	IGHJ4*02	IGHV2-5*01	CAGTAGTCACCGTAGAGCCCAAGCCCT	16	47	0.1	40.9	33.4	2 (0.8%)	4.00E-04	0	7.00E-04	0.006
10	IGHJ5*02	IGHV3-30-3*01	TGATGGCCTGGGGCAAGCTACCACCACTACAACACGGACCCTTTCTGAACCTC	38	30.6	0.2	33.5	33.2	12 (4.6%)	0.04	0	0.06	0.07
10	IGHJ5*02	IGHV3-43*01	TGATGGCCTGGGGCAAGCACTCCC	17	21.7	0.1	25.4	32.6	15 (5.7%)	0.06	0	0.08	0.08
12	IGHJ5*02	IGHV1-3*01	TCTCGGCCACTCCAAAAATCGTAATGGAGGGGTGAAG	20	53.1	NA	50.3	51.3	8 (40.0%)	0.08	NA	0	0
12	IGHJ4*02	IGHV5-51*01	CTCCCCGAACCATAGTAACCCTACC	9	46.4	NA	49.3	48	4 (20.0%)	0.02	NA	0	0
14	IGHJ6*02	IGHV4-31*03	GGAGTCTTGGTAA	6	92.8	0.2	98.1	97.8	22 (0.3%)	1	0	0.7	0.7
15	IGHJ4*02	IGHV4-59*08	GGGGCTCCCTTGGACTCACCGTAGGAGGGG	23	22.9	0.2	18.2	10.6	3930 (83.5%)	12.8	0	12.9	14.6
15	IGHJ4*02	IGHV4-59*01	GGGGCTCCCTTGGACTCACCGTT	18	20.6	0.3	31	38	3930 (83.5%)	12.8	0	12.9	14.6
15	IGHJ4*02	IGHV3-64*01	GGGGCTCCCTTGGACTCACCGTCCTTCGG	19	16.9	13.6	19.3	27.1	2121 (45.1%)	9	0	8.5	7.8
15	IGHJ4*02	IGHV4-59*08	GGGGCTCCCTTGGACTCACCCCCTAGCCC	19	13	0	7.4	0.8	3930 (83.5%)	12.8	0	12.9	14.6
15	IGHJ4*02	IGHV3-64*01	GGGGCTCCCTTGGACTCACCGTAGGAGGGGGTCCCC	29	5.1	0	3.9	2.4	2121 (45.1%)	9	0	8.5	7.8
16	IGHJ6*02	IGHV3-33*01	TAGCAGCTGGTACTACTACAATATCACCCCCCCGGGAAACGA	19	63	4.6	67.4	66.7	2 (0.4%)	0.09	0	0.1	0.2
16	IGHJ4*02	IGHV3-23*03	ACATAGCAATCACCACGGGGGCCAATCTATAGCAGCTGGTACTACCGCTCCA	36	35.4	2.1	31.9	32.6	2 (0.4%)	0.009	0	0.005	0.004

17	IGHJ6*02	IGHV1-3*01	TCACGCCCAGTAGGAGGGTC	11	39	0	42.3	42.1	11 (2.0%)	0.01	0	0	0
17	IGHJ5*02	IGHV4-4*02	GTGGGGACC	2	35.9	0	34.7	34.8	0 (0.0%)	0	0	0	0
17	IGHJ1*01	IGHV3-13*04	CCTACTCCCAGCTGCTGCTATAAGGTGG	15	23.8	0	22.1	21.8	2 (0.4%)	4.00E-04	0	0	0
18	IGHJ5*02	IGHV7-4-1*01	GTAATAACTCCCCGAACCATAGTACCCATCTCTCAGGATG	19	58.8	NA	NA	NA	366 (42.8%)	15.2	NA	NA	NA
18	IGHJ5*02	IGHV2-5*10	GTAATAACTCCCCGAACCATAGTACCCATCTCTCAGGATGCTCGATAGGAGGCCTGT	36	7.7	NA	NA	NA	143 (16.7%)	3.1	NA	NA	NA
18	IGHJ5*02	IGHV3-11*01	GTAATAACTCCCCGAACCATAGTACCCATCTCTCAGGATAGCC	22	5.5	NA	NA	NA	335 (39.1%)	14.6	NA	NA	NA
19	IGHJ4*02	IGHV2-5*03	ATAATAACCAGTCAAAATATCGTAATACGGGCTACGGCGGGA	NA	75.2	NA	NA	NA	0 (0.0%)	0	NA	NA	NA
20	IGHJ6*03	IGHV4-30-2*01	GCCCAAGACCAGTCAAAATATCGTC	10	49.7	NA	NA	NA	89 (5.8%)	0.7	NA	NA	NA
20	IGHJ6*03	IGHV3-30-3*01	GCCCAAGACCAGTCAAAATATCGTCCTCCCCGT	18	9.2	NA	NA	NA	63 (4.1%)	0.4	NA	NA	NA
20	IGHJ4*02	IGHV3-30*18	ACTACCGGGTCCCAGGGGCCAA	12	30.2	NA	NA	NA	371 (24.1%)	1.7	NA	NA	NA
21	IGHJ4*02	IGHV5-a*01	ATGGA	5	97.3	0.2	98.1	94.4	0 (0.0%)	0	0	0	0
22	IGHJ5*02	IGHV3-21*01	GCGTTCCAGTTGTTCCCAAAC	8	96.4	41.5	94.3	99	35 (18.1%)	3.1	0.4	5.2	0.7
23	IGHJ4*02	IGHV3-13*01	CCGTCCCAGCCACTGCTATCT	6	30.3	0.03	9	39.2	2475 (32.4%)	22	0	17.7	38.8
23	IGHJ4*02	IGHV3-30*02	CCGTCCCAGCCACTGCTATCCGCTGA	11	21.3	0	35.4	2	125 (1.6%)	0.1	0	0.1	0.2
23	IGHJ4*02	IGHV3-20*01	CCGTCCCAGCCACTGCTATCTTCTTGCCCCCCTCCTCA	23	13.2	0	23.9	1.3	2457 (32.1%)	18	0	15.8	29.5
25	IGHJ5*02	IGHV3-30*18	GCCTATTTCCTGGTAGCGACACCTT	15	99.9	8.4	99.9	99.9	7 (12.7%)	0.005	0	0.002	0
26	IGHJ4*02	IGHV1-3*01	CGAGCTGCCTCGAGGGCC	12	59.1	0.8	61.7	59.5	37 (1.5%)	0.04	0	0.03	0.04
26	IGHJ4*02	IGHV3-23*01	AGGAGCAGCTGGTACTACTACAATATCCTAGGGGG	11	38.2	0.8	37.9	40	159 (6.6%)	0.1	0	0.1	0.1
27	IGHJ6*02	IGHV3-33*01	GGACTTCATAATAACCACTCCAAAAATCGTAAG	10	57.7	0	56.4	59.4	0 (0.0%)	0	0	0	0
27	IGHJ4*02	IGHV3-7*01	TCGCCAGCCACTGCTATACGCGAGCTCCCCTAGA	20	39.4	0	39.4	40.5	22 (1.3%)	0.02	0	0.007	0
28	IGHJ6*02	IGHV1-69*04	CGAATCTTTCGAACGTAGTAATAACCACTACTCCCTAAAGAGA	27	37.8	NA	36.9	37.3	8 (4.4%)	0.005	NA	0.006	0.002
28	IGHJ6*02	IGHV4-4*01	CGAATCTTTCGAACGTAGTAATAACCACTACTATCATAG	16	25.2	NA	24.6	24.2	9 (4.9%)	0.005	NA	0.006	0.003
28	IGHJ6*04	IGHV3-21*01	GGCATAGCAGCTGGTACTACTACAATAGAAA	6	35.6	NA	37	36.9	0 (0.0%)	0	NA	9.00E-04	0.005
29	IGHJ6*02	IGHV1-3*02	TTCGGGG	3	50.9	0.09	70.4	65.5	0 (0.0%)	0	0	0	0
29	IGHJ4*02	IGHV3-30*03	TAATAACCAGTCAAAATATCGTAATCCGTCCCCCA	12	12.3	0.1	4.4	3	1 (0.01%)	0.001	0	0	0
29	IGHJ5*02	IGHV1-8*01	GGGGAGGTTCCAGTTATACCCAATCCGGGGATCGGGCAGCTGGTACTACTACCCCGAC	43	8.4	0.3	6.4	8.1	16 (0.2%)	0.02	0	0.01	0.007
30	IGHJ5*02	IGHV4-34*01	CCCCCTCGTATATAGCATACACCATTAGTACAATAAAGGCGTGAGGATCTCCC	30	93.1	5	96.3	90	18 (2.9%)	2.8	0.07	2.1	6.9
31	IGHJ5*02	IGHV3-11*01	CCTTCCTTTAATAACTCCCCGAACCATAGTAATGC	12	38.8	0	NA	1	0 (0.0%)	0	0	NA	0
31	IGHJ4*02	IGHV3-11*01	CCCACACGGTATTTTGCAACTCCCCGAACCTCTCCTATCTTTGCCCCA	37	38.1	0	NA	0	4 (0.4%)	0.05	0	NA	0
32	IGHJ4*02	IGHV1-17*01	AGTAGCAGCTACCACCACTACCCTAGGAGAA	NA	49.9	0.2	50.7	50.5	11 (1.4%)	0.006	0.09	0.02	0.02
32	IGHJ4*02	IGHV3-47*01	AGTAGCAGCTACCACCACTACAATATCCTCATC	6	49.1	0.2	49.2	49.4	9 (1.2%)	0.005	0	0.02	0.02
33	IGHJ3*02	IGHV3-41*01	GAACATAACTCCCCGAACTCCCC	NA	66.6	0.4	99.7	98.9	3 (0.3%)	0.001	0	0	0

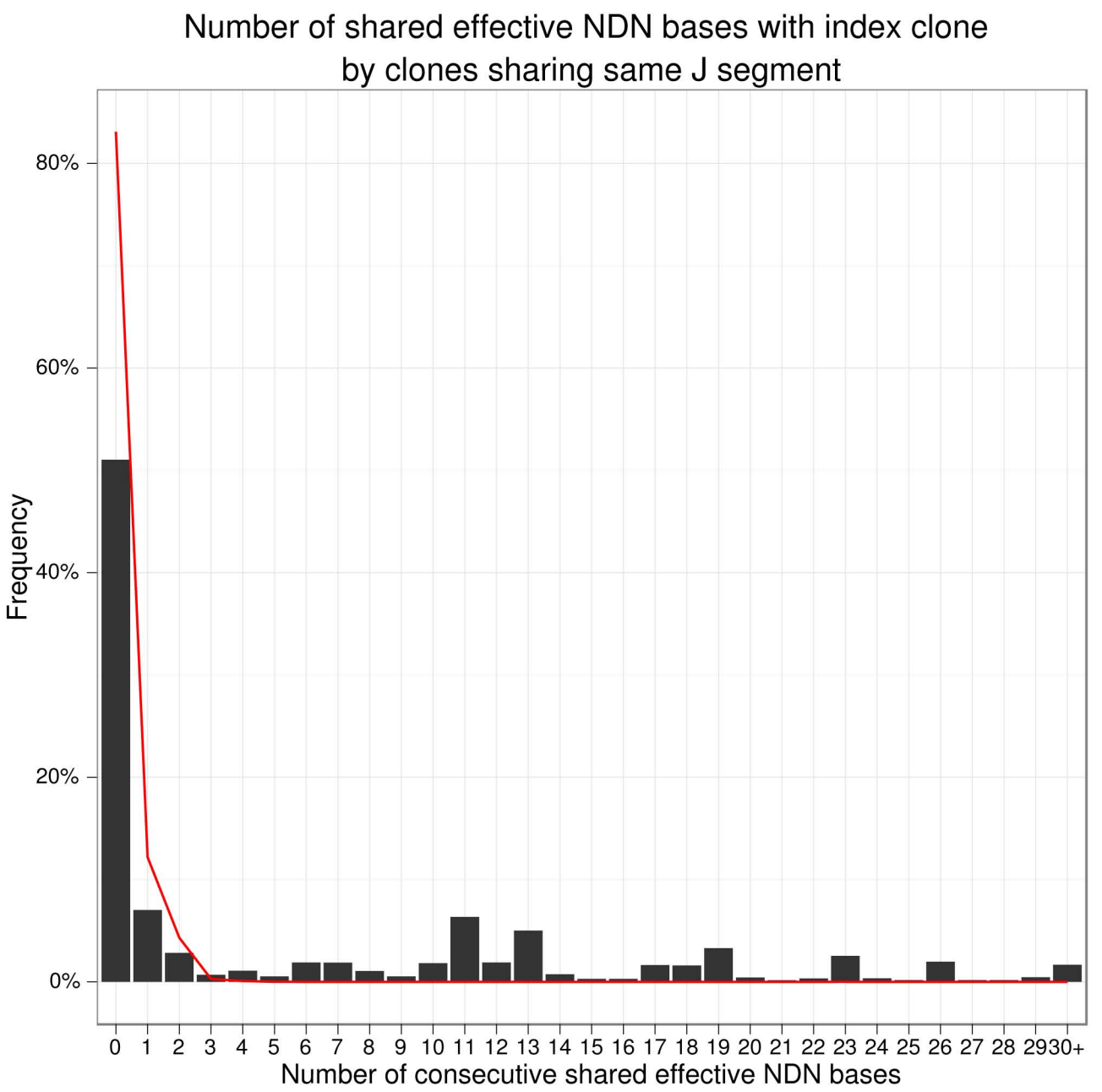
34	IGHJ5*02	IGHV1-2*02	CGGTCGAGG	NA	85.4	6.9	86	85.2	0 (0.0%)	0	0	0	0.05
34	IGHJ4*02	IGHV1-46*01	ACCGTGCCCCTTTATCCCCAGTTGCCTCTCG	21	14.1	0.6	13.9	14.7	0 (0.0%)	0	0.02	0	0
35	IGHJ6*02	IGHV3-23*01	GCTCCCACTAAGTCAGGAA	11	48.7	0.8	56.8	53.7	1020 (24.4%)	1.6	0.02	1.5	1.1
35	IGHJ3*02	IGHV3-13*03	CGAGCTGCTAGATTGGGTC	11	14.7	0.2	10.1	9.9	129 (3.1%)	0.1	0	0.2	0.1
36	IGHJ4*02	IGHV1-3*02	CATGTCCCTCCCCAGTTTAGGGGGAATG	21	22.4	0	95.6	98	0 (0.0%)	0	0	0	0
38	IGHJ5*02	IGHV3-30*18	GTAGTTCCAGCCCCG	7	65.5	3.2	60.1	58.7	154 (13.4%)	0.3	0	0.2	0.09
38	IGHJ6*02	IGHV3-13*01	GGGATCCGGGCGGACCCCATTTTTCGGCGATGGCT	NA	31.5	1.5	37.4	39.5	0 (0.0%)	0	0	0	7.00E-04
39	IGHJ4*02	IGHV4-34*01	CTGCTCCAT	3	99	NA	99.8	99.3	0 (0.0%)	0	NA	0	0
40	IGHJ4*02	IGHV3-13*01	GCTCGCCAGCTGCTGCTATACC	7	84.9	1.8	89.4	89.7	24 (0.3%)	0.02	0	0.02	0.03
42	IGHJ6*02	IGHV6-1*01	CAGCTGGTACTACTACCCGGACCCCAC	13	56	1.2	58.5	55.6	4 (3.4%)	0.001	0	6.00E-04	0
42	IGHJ5*02	IGHV2-5*10	ACCCGGGGACGTACCAGCTGCTGCTATAGGGGGGGTCTGT	24	43.2	0.9	40.8	43.7	1 (0.8%)	1.00E-04	0.1	0	0
43	IGHJ4*02	IGHV3-74*01	TGGGGGGTATTGACCAGTCAAAATATCGTAAGAAGCTCGCGGGA	22	58.1	NA	52.1	49.3	2 (8.7%)	0.02	NA	0.004	0.1
43	IGHJ6*03	IGHV3-13*01	GGGGACAGAGGGCCACTCCAAAAATCGTAATACCTCGGTCGA	23	40.8	NA	46.5	48.9	0 (0.0%)	0	NA	0	0.006
44	IGHJ5*02	IGHV3-23*03	CCTTAGCGA	4	73.9	0.8	70.2	60.4	0 (0.0%)	0	0	0	0
44	IGHJ4*02	IGHV3-23*03	AGAGGGGTAAGTAATAACCAGTCAAAATAGTA	16	25.2	0.6	29.8	39.5	3 (0.8%)	3.00E-04	0.03	0	5.00E-04
45	IGHJ4*02	IGHV3-11*01	CCCAAGGGATAGCAGTCACCACCACTCACACGGCTT	18	71.7	0.6	62.6	2.6	5 (0.3%)	0.01	0	0.02	0
45	IGHJ6*02	IGHV3-21*01	AGCCCCA	4	8.2	0	16.5	0.1	0 (0.0%)	0	0	0	0
46	IGHJ6*02	IGHV1-69*01	TAATAACCACTCCAAAAATTACCTTACCCCCCGACGC	NA	94.6	2.1	97.8	97.5	0 (0.0%)	0	0	0	0
47	IGHJ4*02	IGHV3-64*02	CCCTATACCACTACTATCATAGTAATACCCA	11	98.5	1.5	99.7	99.8	60 (5.1%)	0.2	0	0.2	0.04
48	IGHJ6*02	IGHV3-13*01	CGAATGTGTACCAGCTGCTGCTATA	9	29.1	0.5	29.9	30.8	0 (0.0%)	0	0	0	0
48	IGHJ4*02	IGHV1-3*02	TAACCACTACTATCATAGTAATACACCAA	7	24.3	1	24	24.1	3 (3.6%)	0.004	0	0.01	0.003
48	IGHJ4*02	IGHV3-11*04	TCGTACCCACCCTGGACGAGCTGCTATACTCGGA	18	23.9	1.3	23.7	22.3	7 (8.4%)	0.003	0	0.003	8.00E-04
48	IGHJ5*02	IGHV3-7*01	ACCATCGCAGCTATTTGAAGA	13	22.5	0.5	22.2	22.6	20 (24.1%)	0.02	0	0.02	0.02
49	IGHJ4*02	IGHV3-33*01	CCTCCCAAAGTACCAGCCACTGCGAGA	15	99	0.7	99.7	99.2	4 (0.6%)	7.00E-04	0	2.00E-04	0.003
50	IGHJ3*02	IGHV1-3*02	GCCCTATCCCCGTAGAGGC	10	74.6	0.8	85.6	44.5	2632 (40.3%)	7	0.1	9.1	5.3
50	IGHJ3*02	IGHV2-5*10	GCCCTATCCCCGTAGAGGCCTCCCCCAGGGTCTGT	26	6.9	0.06	0.3	44.3	2224 (34.0%)	4.3	0.07	5.6	3.8
51	IGHJ6*03	IGHV3-33*01	ATACGGTATAGCTACGGAGTACCCGA	20	5.8	0	39.5	39.8	4 (0.5%)	5	0	1.2	12.4

Table S4. Upstream and downstream V_H replacement in evolved clonotypes.

Patient ID	Clone	Number of evolved clonotypes with V segment 3' compared to index clone V segment	Number of evolved clonotypes with V segment 5' compared to index clone V segment	Number of evolved clonotypes where V segment cannot be compared to index clone V segment
3	1	0	15	1
5	1	1	23	1
5	2	0	0	613
5	3	0	0	30
6	1	0	3	0
6	2	0	22	0
6	3	0	1	0
6	4	0	2	0
6	5	0	853	4
7	1	1	24	0
8	1	257	20	2
8	2	0	0	22
9	1	0	42	0
9	2	0	0	0
10	1	0	1	1
10	2	0	0	12
10	3	2	9	4
12	1	0	6	2
12	2	2	2	0
14	1	9	13	0
15	1	317	1803	3
15	2	287	3642	3
15	3	287	3642	3
15	4	287	3642	3
15	5	317	1803	3
16	1	0	2	0
16	2	0	1	1
17	1	0	10	1

17	2	0	2	0
17	3	0	0	0
18	1	0	0	366
18	2	1	139	3
18	3	264	68	3
19	1	0	0	0
20	1	0	0	89
20	2	14	320	37
20	3	0	0	63
21	1	0	0	0
22	1	3	32	0
23	1	48	2404	23
23	2	35	87	3
23	3	149	2285	23
25	1	0	7	0
26	1	0	31	6
26	2	0	159	0
27	1	0	0	0
27	2	0	22	0
28	1	6	2	0
28	2	0	0	0
28	3	0	9	0
29	1	0	0	0
29	2	0	1	0
29	3	0	16	0
30	1	6	12	0
31	1	0	4	0
31	2	0	0	0
32	1	0	9	0
32	2	2	11	0
33	1	1	2	0
34	1	0	0	0
34	2	0	0	0

35	1	17	916	87
35	2	2	123	4
36	1	0	0	0
38	1	2	76	76
38	2	0	0	0
39	1	0	0	0
40	1	2	22	0
42	1	0	1	0
42	2	0	4	0
43	1	2	0	0
43	2	0	0	0
44	1	0	0	0
44	2	0	3	0
45	1	0	5	0
45	2	0	0	0
46	1	0	0	0
47	1	10	50	0
48	1	0	0	0
48	2	0	7	0
48	3	0	20	0
48	4	0	3	0
49	1	0	4	0
50	1	11	2265	356
50	2	154	1714	356
51	1	3	0	1


Number of shared effective NDN bases with index clone
by clones sharing same J segment
Frequency
80%
60%
40%
20%
0%
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30+
Number of consecutive shared effective NDN bases