Infection as a cause of childhood leukemia: virus detection employing whole genome sequencing

Christoph Bartenhagen,[1] Ute Fischer,[2*] Klaus Korn,[3] Stefan M. Pfister,[4,5,6] Michael Gombert,[2] Cai Chen,[2,7] Vera Okpanyi,[2] Julia Hauer,[2] Anna Rinaldi,[8] Jean-Pierre Bourquin,[8] Cornelia Eckert,[9] Jianda Hu,[7] Armin Ensser,[3] Martin Dugas[1*] and Arndt Borkhardt[2*]*

[1]*Institute of Medical Informatics, University of Münster, Germany;* [2]*Department of Pediatric Oncology, Hematology and Clinical Immunology, University Children's Hospital, Medical Faculty, Heinrich-Heine-University, Düsseldorf, Germany;* [3]*Institute for Clinical and Molecular Virology, University Hospital, Friedrich Alexander University of Erlangen-Nuremberg, Erlangen, Germany;* [4]*Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), Heidelberg, Germany;* [5]*Department of Pediatric Oncology, Hematology & Immunology, Heidelberg University Hospital, Germany;* [6]*German Cancer Consortium (DKTK), Heidelberg, Germany;* [7]*Fujian Institute of Hematology, Fujian Medical University Union Hospital, Fuzhou, China;* [8]*Pediatric Oncology, University Children's Hospital Zurich, Switzerland and* [9]*Department of Pediatrics, Division of Oncology and Hematology, Charité Berlin, Germany*

Correspondence: ute.fischer@med.uni-duesseldorf.de
doi:10.3324/haematol.2016.155382

Supplemental

Supplemental Material and Methods

Patients

Genomic DNA was isolated from bone marrow obtained from 14 pediatric ALL patients (*ETV6-RUNX1* positive, n=7; high hyperdiploid, n=7). Presence of the *ETV6-RUNX1* fusion gene was confirmed by FISH analysis and RT-PCR, high hyperdiploidy by cytogenetic karyotype analysis. In addition, NGS analysis confirmed the respective ALL subtype. Bone marrow was aspirated at diagnosis, remission and/or relapse. Blast percentage was >70% in the leukemia samples. Mononuclear cells were derived by Ficoll density centrifugation from bone marrow, DNA was isolated using standard protocols and sequencing was carried out as described (1). The local ethics committee approved the research and written informed consent was given by all parents.

The datasets of the analyzed high hyperdiploid patients were previously used for genomic studies (1, 2). None of the published analyses is reproduced in this letter. The data sets have been reanalyzed in the present manuscript with a different scope. Genotype data will be deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/), which is hosted by the EBI, under accession number EGAS00001000670 (high hyperdiploid ALL data sets) and EGAS00001000684 (*ETV6-RUNX1* positive ALL data sets).

Library preparation and paired-end sequencing

Genomic DNA was isolated employing the AllPrep DNA Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. Genomic DNA was sheared into fragments using Covaris Adaptive Focused Acoustics (Covaris, Woburn, MA). Automatic ligation of sequencer-specific adapters (Illumina, San Diego, CA) to the DNA fragments

was carried out employing a SPRIworks Fragment Library Systems I (Beckmann Coulter, Krefeld, Germany). PCR amplification generated sequencing libraries of 350 to 400 bp fragments purified by gel extraction. DNA concentrations were determined employing High Sensitivity DNA Chips (Agilent, Waldbronn, Germany). Paired-end sequencing was performed as recommended by the manufacturer (on a Genome Analyzer IIx or a HiSeq2000, both from Illumina), respectively. Sequencing statistics are provided in Supplemental Table 1.

Data analysis

Datasets

1. Viral reference sequences

In total, 25.525 viral genomes, as deposited in the Genome Information Broker for Viruses (GIB-V) (3), were screened for presence and integration. The dataset contained the genomes of DNA and RNA including several strains and clones of the same organism.

2. Repeats in viral sequences

Before running the pipeline, the viral sequences were screened for repeats and regions of low complexity with RepeatMasker (4). The coordinates were used for alignment filtering as described below.

3. Sequence homology between viral and human sequences

All viral sequences were further compared to the human reference genome (GRCh37.55) with MUMmer (4), which reported regions of high sequence homology between 30 and 9.463bp. Their coordinates were used for alignment filtering as well.

Bioinformatic pipeline

1. Alignment against human and viral reference genomes

Viral sequence detection starts with the subtraction of all human sequences from the data sets. All reads are aligned to the reference genome, the remaining unmapped reads are candidates for sequences of non-human, viral origin and are extracted from the alignment for further processing. How this first subtraction step is performed is up to the user. We chose BWA (5) in paired-end mode with default parameter settings to map all patient datasets against the human reference genome (GRCh37.55). BWA is a very fast alignment tool, but one may replace it with any other short read aligner.

Next, our pipeline discards potential viral sequences with a similarity of 85% or higher between the viral origin and the human genome. Hence, all previously unmapped reads were aligned in single-end mode once more against the human genome with higher sensitivity and a mapping of minimum 85% of the read with Bowtie2 (6). The scoring and penalty parameter settings were set as follows:

bowtie2 --very-sensitive --mp 1,1 --rdg 2,1 --rfg 2,1 --score-min L,0,-0.15

For this part of the pipeline, we preferred Bowtie2 due to its various parameters, which allow detailed control of multiple alignments (important for the mapping against the virus database) and mapping rates (important for subtraction of human sequences).

The effect of skipping the whole subtraction step i.e. both previous alignments and starting directly with an alignment against the viral genome sequences is shown in Supplemental Table 4. It demonstrates the importance of a sensitive subtraction of human sequences to avoid false-positive virus detections.

All reads, which were unmapped after the BWA and Bowtie2 alignment, were then aligned against all 25.525 viral sequences simultaneously with Bowtie2.

Since the viral reference contains some highly similar genomes (e.g. clones and strains of

the same organism), multiple alignments (up to 1,500) to more than one viral reference sequence were allowed and reported for further analysis (k-mode in Bowtie 2).

bowtie2 -k 1500 -x gib_v_database

A test run with BLAT instead of Bowtie2 for the viral reference alignment is shown in Supplemental Table 5.

2. Alignment filtering

Multiple alignments against the same viral sequence and duplicate reads, i.e. alignments of more than one read to the same position, were excluded except for one using Samtools (7). Alignments against viral sequences were then filtered by overlaps with repeat sequences and homologies with human sequences (coordinates were previously computed, as described above). All viruses with at least two alignments were then exported for manual inspection and validation.

3. Detection of viral integration

All paired-end reads with one alignment on the human genome and one alignment on viral sequences are candidates for viral integration. For that purpose, for every viral alignment, the human alignment was searched for the corresponding partner read. All reads the spanning human and a viral genome were exported for manual inspection and validation.

Simulation of viral sequences and integration

The reference genomes of ten viruses were selected for simulation: Simian virus 40 (NCBI Nucleotide accession nr. EF579658), Merkel cell polyomavirus (EU375803), Human herpesvirus 5 (FJ527563), Human herpesvirus 4 (AJ507799), Human immunodeficiency virus 1 (AJ291719), Human T-lymphotropic virus 1 (AY563953), Human adenovirus 1

(AF534906), Human papillomavirus 16 (K02718), Human parvovirus 4 (AY622943) and Human herpesvirus 3 (AJ871403). All virus genomes were randomly mutated with frequencies 0%, 5%, 10%, 15%, 20% and 30%. Positions could not be drawn twice, unannotated regions were excluded and it was assured that a nucleotide was not substituted by the same one. For all the 60 (mutated) genomes, the following two scenarios were simulated:

1. Viral sequences

This scenario corresponds to the presence of the viral sequences in the DNA without (evidence of) integration (Supplemental Figure 2). A total of 1,000 non-overlapping 50 bp paired-end reads (i.e. 2,000 sequences) were drawn at random positions from each (mutated) viral genome.

The reads were processed according to our pipeline (including alignment against the human reference genome), and those reads were counted, which could be mapped back against a viral genome of the same type in the GIB-V database.

2. Viral integration

This second scenario simulates the integration of a viral genome into the host, i.e. a junction between host and viral DNA within the host's genome, which can be detected by a spanning paired-end read across the integration site. For every mutated viral genome, 20 integration sites were picked randomly from chromosome 1 of the human reference genome, giving 40 breakpoints for each virus. Between 80 and 100% of the mutated viral genome was inserted as a whole at each site.

Paired-end reads with 50bp length, 250bp insert-size (50bp standard deviation) and a linear increasing per base error rate were simulated from the modified chromosome 1 using dwgsim (8) with 5x sequence depth. The reads were processed by our pipeline and

the sequence coverage and the recall of integration sites was calculated for every virus and mutation frequency.

Theoretical probability of virus detection

The probability for obtaining at least one read from a viral sequence of given length (shown in Fig. 1A) was computed as

$$P(X \geq 1) = 1 - P(X \leq 0) = 1 - B(0 \mid p, n).$$

Given a binomial distribution with n experiments (i.e. the number of reads; viral origin: yes/no) and a probability $p = \frac{v - r}{h + v - r}$ of sequencing a read of length $r = 50\text{bp}$ from a viral sequence of length v within the human genome of length $h = 3.095.677.412\,bp$ (for hg19).

Supplemental References

1. Chen C, Bartenhagen C, Gombert M, Okpanyi V, Binder V, Rottgers S, Bradtke J, Teigler-Schlegel A, Harbott J, Ginzel S, Thiele R, Fischer U, Dugas M, Hu J, Borkhardt A. Next-Generation-Sequencing-Based Risk Stratification and Identification of New Genes Involved in Structural and Sequence Variations in Near Haploid Lymphoblastic Leukemia. Genes, Chrom & Cancer 2013; 52: 564–579.
2. Chen C, Bartenhagen C, Gombert M, Okpanyi V, Binder V, Röttgers S, Bradtke J, Teigler-Schlegel A, Harbott J, Ginzel S, Thiele R, Husemann P, Krell PF, Borkhardt A, Dugas M, Hu J, Fischer U. Next-generation-sequencing of recurrent childhood high hyperdiploid acute lymphoblastic leukemia reveals mutations typically associated with high risk patients. Leuk Res. 2015;39(9):990-1001.
3. Hirahata, M. et al. Genome Information Broker for Viruses (GIB-V): database for comparative analysis of virus genomes. Nuc Acids Res 2007; 35:D339–D342.
4. Smit, AFA. Hubley, R. & Green, P. RepeatMasker Open-3.0.

http://www.repeatmasker.org (1996-2010)

5. Kurtz, S. et al. Versatile and open software for comparing large genomes. Genome Biol 2004;5: R12.
6. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009; 25:1754–1760.
7. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012; 9:357-359.
8. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25:2078–2079.
9. Whole Genome Simulation. http://sourceforge.net/apps/mediawiki/dnaa/index.php?title=Whole_Genome_Simulation

Supplemental Tables

Supplemental Table 1: Genome size and nucleotide accession numbers for the 10 selected viruses used for simulation experiments.

Virus	Genome size (kb)	NCBI nucleotide accession number
SV40	5.2	AF579658
Merkel cell polyomavirus	5.3	EU375803
ADV1	36.0	AF534906
HPV-16	7.9	K02718.1
VZV (HHV3)	124.9	AJ871403.1
HIV	9.7	AJ291719
Parvovirus 4	5.2	AY622943
HTLV-1	8.3	AY563953
EBV (HHV4)	171.8	AJ507799
CMV (HHV5)	235.6	FJ527563

Supplemental Table 2: Sequencing and mapping statistics for whole-genome sequencing of 14 patients with B-precursor ALL (HeH, high hyperdiploid; age, age at diagnosis).

No.	Subtype	Age	Time point	Total sequencing reads [10^6]	Mapped sequencing reads (to human reference)			Unmapped sequencing reads	
					Reads [%]	Coverage [%]	Coverage [fold]	[%]	total [10^6]
1	*ETV6-RUNX1+*	3	Diagnosis	370	97	92	5.9	3	11
			Remission	483	97	92	7.7	3	15
			Relapse	497	97	92	7.9	3	15
2	*ETV6-RUNX1+*	2	Diagnosis	492	97	91	7.8	3	15
			Remission	483	97	91	7.7	3	15
			Relapse	522	97	91	8.4	3	16
3	*ETV6-RUNX1+*	4	Diagnosis	499	98	91	8.0	2	10
			Remission	741	97	92	11.8	3	22
			Relapse	588	96	91	9.3	4	24
4	*ETV6-RUNX1+*	2	Diagnosis	298	78	83	2.7	22	66
			Remission	129	90	60	1.3	10	13
5	*ETV6-RUNX1+*	3	Diagnosis	715	84	91	9.7	16	114
			Remission	439	96	90	6.9	4	18
6	*ETV6-RUNX1+*	3	Diagnosis	570	88	91	8.2	12	68
			Remission	401	90	90	5.9	10	40
			Relapse	499	93	90	7.5	7	35
7	*ETV6-RUNX1+*	5	Diagnosis	367	93	90	5.6	7	26
			Remission	336	96	90	5.3	4	13
			Relapse	419	95	90	6.5	5	21
8	HeH	3	Diagnosis	149	86	68	1.5	14	21
			Remission	135	71	59	1.1	29	39
			Relapse	126	89	64	1.3	11	14
9	HeH	3	Diagnosis	252	90	83	2.6	10	25
			Remission	259	88	84	2.6	12	31
			Relapse	319	88	87	3.3	12	38
10	HeH	3	Diagnosis	208	87	80	2.1	13	27
			Remission	170	87	75	1.7	13	22
11	HeH	3	Diagnosis	746	74	91	8.8	26	194
			Remission	777	63	92	7.8	37	288
			Relapse	930	80	92	11.9	20	186
12	HeH	3	Remission	563	83	91	7.6	17	96
			Relapse	550	88	92	7.8	12	66
13	HeH	3	Remission	483	88	92	7.0	12	58
			Relapse	545	86	92	7.7	14	76
14	HeH	3	Diagnosis	402	95	92	6.3	5	20
			Remission	372	94	92	5.7	6	22
Mean values (±SD)				440 (±196)	89 (±8)	87 (±9)	6.1 (±3.0)	11 (±8)	49 (±60)

Supplemental Table 3: Sequencing and mapping statistics for whole-genome sequencing of 14 age- and gender-matched controls (ICGC PedBrain cohort).

No.	Control	Total sequencing reads [10^6]	Mapped sequencing reads (to human reference)		Unmapped sequencing reads	
			Reads [%]	Coverage [fold]	[%]	total [10^6]
1	PA3	1.473	95	48	5	73
2	PA10	806	95	26	5	40
3	PA25	841	95	27	5	42
4	PA27	1.242	94	40	6	74
5	PA42	1.050	93	34	7	73
6	PA48	1.203	95	39	5	60
7	PA62	981	95	32	5	49
8	PA71	1.169	94	38	6	70
9	PA85	985	95	32	5	49
10	PA87	1.076	94	35	6	64
11	PA91	1.056	90	34	10	105
12	PA135	1.230	94	40	6	73
13	PA144	1.342	95	44	5	67
14	PA159	1.490	95	49	5	74
Mean values (±SD)		1.139 (±209)	94 (±1)	37 (±7)	6 (±1)	65 (±17)

Supplemental Table 4: Virus detection in *ETV6-RUNX1* positive ALL.

Virus	Patient 1			Patient 2			Patient 3			Patient 4		Patient 5		Patient 6			Patient 7		
	D	Rm	Rl	D	Rm	Rl	D	Rm	Rl	D	Rm	D	Rm	D	Rm	Rl	D	Rm	Rl
BacteriophagebIL170											2, 72								
BacteriophagebIL311											2, 72								
Bacteriophagesk1											2, 72								
BovinepolyomavirusDNA	2, 102																		
Emilianiahuxleyivirus86isolateEhV86				2,58															
EnterobacteriaphageP7																	6, 306		
EpsteinBarrvirusartifactualjoin				2, 102					3, 69								2, 69		
GQ979703AF152407AY365422AY576278FJ010786GQ25											2, 72								
HIV1														2, 102			2, 102		
HIV1cloneES116fromUSA																	2, 102		
HIV1cloneES120fromUSA																	2, 102		
HIV1cloneES424fromUSA																	2, 102		
HIV1clonepIIIBfromUSA														2, 102			2, 102		
HIV1E9fromtheUSA																	2, 102		
HIV1isolate01BR047fromBrazil																	2, 102		
HIV1isolate04KMH5fromSouthKorea																	2, 102		
HIV1isolate101208fromUSA																	2, 102		
HIV1isolate101211fromUSA																	2, 102		
HIV1isolate508283clonepbf25fromU																	2, 102		
HIV1isolate508286clonepbf26fromU																	2, 102		
HIV1isolate508294clonepbf28fromU																	2, 102		
HIV1isolate515596clonenp8fromUSA																	2, 102		
HIV1isolateCNHN24subtypeBThaifrom																	2, 102		
HIV1isolateCTL016fromDenmark																	2, 102		
HIV1isolateCTL043fromDenmark																	2, 102		
HIV1isolateCu43fromCuba																	2, 102		
HIV1isolateL8157fromUSA																	2, 102		
HIV1isolateMNcloneMNTQfromtheUSA																	2, 102		
HIV1isolatePS1038Day0fromAustralia																	2, 102		
HIV1isolatePS1038Day174fromAustral																	2, 102		
HIV1isolateWCM32P0896fromUSA																	2, 102		
HIV1patientWClPRsample1985clone4																	2, 102		
HIV1patientWClPRsample1985clone5																	2, 102		
HIV1patientWClPRsample1985clone52																	2, 102		
HIV1patientWClPRsample1985clone54																	2, 102		
HIV1patientWClPRsample1990clone12																	2, 102		
HIV1patientWClPRsample1990clone31																	2, 102		
HIV1patientWClPRsample1990clone32																	2, 102		
HIV1strainESX24252fromSpain																	2, 102		
HIV1strainH434clone8A3fromNether																	2, 102		
HIV1strainH434clone8F4fromNether																	2, 102		
Humanherpesvirus4completewildtypeg				2, 102					3, 69								2, 69		
Humanherpesvirus4strainAG876				2, 102															
Humanherpesvirus4strainGD1				2, 102															
Humanherpesvirus6		4, 204																	
Humanherpesvirus6BDNAstrain		4, 204																	
Humanherpesvirus7strainRK		5, 216	2, 102					13, 510									4, 115	5, 205	
HumanimmunodeficiencyvirusHIV1cloneA																	2, 102		
Humanimmunodeficiencyvirustype1														2, 102			2, 102		
Humanimmunodeficiencyvirustype1isolat																	2, 102		
Humanimmunodeficiencyvirustype1isolat																	2, 102		
Humanimmunodeficiencyvirustype1isolat																	2, 102		

Humanimmunodeficiencyvirustype1isolat				2, 102	2, 102
HumanTcellleukaemiatypeIII					2, 102
Lactococcuslactisphagejj50		2, 72			
Lactococcusphage712		2, 72			
LactococcusphagebIBB29		3, 108			
LactococcusphageCB13		3, 108			
LactococcusphageCB19		2, 72			
LactococcusphageCB20		2, 72			
LactococcusphageSL4		2, 72			
PropionibacteriumphagePA6	4, 204	2, 72		4, 204	
Saimiriineherpesvirus1strainMV54	2, 102				
SimianimmunodeficiencyvirusgenomicRN					2, 102
Torquetenominivirus6DNAisolate		2, 72			
Torquetenovirus10DNAisolate		241, 381			
Torquetenovirus12DNAisolate		214, 382			
Torquetenovirus15DNAisolate		68, 206			
Torquetenovirus16DNAisolate		143, 289			
Torquetenovirus19DNAisolate		279, 447			
Torquetenovirus26DNAisolate		18, 54			
Torquetenovirus27DNAisolate		39, 171			
Torquetenovirus28DNAisolate		92, 153			
Torquetenovirus3strainHEL3		167, 323			
Torquetenovirus4DNAisolate		20, 87			
Torquetenovirus6isolateKAV		196, 301			
Torquetenovirus7isolatePMV		165, 282			
Torquetenovirus8DNAgenotype22		310, 423			
Torquetenovirusclone817		296, 427			
TorquetenovirusDNAgenotype23		376, 452			
TorquetenovirusDNAisolate		106, 187			
TorquetenovirusDNAisolate		93, 178			
TorquetenovirusDNAisolate		149, 249			
TorquetenovirusDNAisolate		86, 186			
TorquetenovirusDNAisolate		68, 159			
TorquetenovirusDNAisolate		86, 177		4, 169	
TorquetenovirusDNAisolate		87, 179			
TorquetenovirusDNAisolate		246, 423			
TorquetenovirusDNAisolate		127, 313			
Torquetenovirusisolate2h		230, 348			
Torquetenovirusisolate3h		433, 604			
Torquetenovirusisolatetth10		104, 291			
Torquetenovirusisolatetth11g2		96, 288			
Torquetenovirusisolatetth13		142, 252			
Torquetenovirusisolatetth14		159, 314			
Torquetenovirusisolatetth16		180, 309			
Torquetenovirusisolatetth17		181, 345			
Torquetenovirusisolatetth18		113, 297			
Torquetenovirusisolatetth19		115, 250			
Torquetenovirusisolatetth20		90, 293			
Torquetenovirusisolatetth21		113, 297			
Torquetenovirusisolatetth22g4		161, 276			
Torquetenovirusisolatetth23		137, 225			
Torquetenovirusisolatetth25		181, 345			
Torquetenovirusisolatetth26		173, 299			
Torquetenovirusisolatetth27		180, 345			
Torquetenovirusisolatetth29		159, 314			
Torquetenovirusisolatetth3		165, 334			

Torquetenovirusisolatetth31			181, 345	
Torquetenovirusisolatetth4			172, 268	
Torquetenovirusisolatetth5			159, 314	
Torquetenovirusisolatetth6			113, 297	
Torquetenovirusisolatetth7			131, 243	
Torquetenovirusisolatetth8			161, 276	
Torquetenovirusisolatetth9			169, 337	
TorquetenovirusstrainSIA109			164, 1007	
TTvirusgeneforORF1andORF2isolate			333, 464	
TTvirusgenotype1aDNA			213, 339	
TTvirusgenotype1orf2/5orf2/4orf1g			213, 339	
TTvirusisolate			332, 429	
TTvirusisolate			352, 427	
TTvirusisolateGH1			247, 318	
TTvirusisolateJA1			115, 203	
TTvirusisolateJA20			431, 434	
TTvirusisolateJA4			186, 288	
TTvirusisolateJA9			203, 334	
TTvirusisolateT3PB			193, 319	
TTvirusisolateTUPB		2, 66	136, 317	
TTvirusisolateTWHORF2andORF1genes			146, 302	
TTvirusisolateUS32			176, 316	
TTvirusisolateUS35			179, 319	
TTvirusPolishisolateP/1C1			294, 467	
TTvirussle1931			159, 256	
TTvirussle1932			160, 256	
TTvirussle1957			277, 352	
TTvirussle2057			204, 361	
TTvirussle2058			204, 361	
TTvirussle2061			204, 361	
TTvirussle2065			204, 361	
TTvirussle2072			166, 339	
TTvirusstrainBDH1			134, 314	
TTvirusstrainTTVCHN1			174, 338	
TTvirusTTVCHN2			65, 133	
V01555J02070K01729K01730V01554X00498X00499	2, 102	3, 69		2, 69

The table presents the viruses detected in *ETV6-RUNX1* positive ALL patients. D, diagnosis sample; Rm, remission sample; Rl. relapse sample. Sample-Columns show alignment and coverage against viral genome as number of reads and coverage in bp (“2, 102” i.e. 2 reads, 102 bp coverage). Similarity to other viral genomes: At least 50% of the reads mapped to this viral genome. Filtering steps: 1. Alignment of unmapped reads against the human genome using BOWTIE2: maximum sensitivity and at least 85% sequence similarity. 2. Removal of multiple alignments against the same viral genome. 3. Removal of alignment duplicates (i.e. same position on the same viral genome). 4. Removal of alignments which overlap with repeat regions on the viral genome (previously detected with the program RepeatMasker). 5. Removal of alignments in regions with high sequence similarity to the human genome (previously detected with the program Mummer). 6. Removal of viral genomes with less than 2 alignments.

Viruses detected due to similarity to PhiX (detected because of PhiX DNA spike-in in the whole genome sequencing procedure), or similarity to cloning vectors (e.g. HIV gag, pol, env sequences), or viruse not infecting humans were not considered in Table 2 and Figure 1C.

Supplemental Table 5: Virus detection in high hyperdiploid ALL.

Virus	Pat. 8			Pat. 9			Pat. 10			Pat. 11			Pat. 12		Pat. 13		Pat. 14		
	D	Rm	Rl	D	Rm	Rl	D	Rm	D	Rm	Rl	Rm	Rl	Rm	Rl	D	Rm
AY543070AJ001191AJ585756AY339622K00165K00172				2, 49															
AY548170AH009871AY005330AY005331AY005332AY00														2, 102					
B19virusisolateJ35													4, 204						
B19virusisolateNAN													4, 204						
BacteriophageT5strainATCC11303B5				2, 49															
BacteriophageT5strainst0deletionmut				2, 49															
Enterobacteria													2, 102						
EnterobacteriaphageDE3													2, 102						
EnterobacteriaphageP7			2, 72			2, 72			2, 102										
EpsteinBarrvirusartifactualjoin							8, 288	15, 540							2, 51		2, 62		
ErythrovirusB19strainHV													4, 204						
GQ221974AY156040AY169795AY325311AY446863AY44					3, 108														
GU937742AY156044AY169798AY446860AY446871AY44					2, 72														
HumancytomegalovirusstrainAD169					2, 72														
Humanherpesv					4, 144														
Humanherpesvirus3														2, 102					
Humanherpesvirus3DNAstrain														2, 102					
Humanherpesvirus3DNAstrain														2, 102					
Humanherpesvirus3isolateHHV3M2DR														2, 102					
Humanherpesvirus3strain03500														2, 102					
Humanherpesvirus3strain11														2, 102					
Humanherpesvirus3strain22														2, 102					
Humanherpesvirus3strain32passage22														2, 102					
Humanherpesvirus3strain32passage5														2, 102					
Humanherpesvirus3strain32passage72														2, 102					
Humanherpesvirus3strain36														2, 102					
Humanherpesvirus3strain49														2, 102					
Humanherpesvirus3strain8														2, 102					
Humanherpesvirus3strainBC														2, 102					
Humanherpesvirus3strainCA123														2, 102					
Humanherpesvirus3strainKel														2, 102					
Humanherpesvirus3strainNH293														2, 102					
Humanherpesvirus3strainSD														2, 102					
Humanherpesvirus3strainSVETA														2, 102					
Humanherpesvirus3strainVarilRix														2, 102					
Humanherpesvirus3strainVariVax														2, 102					
Humanherpesvirus3														2, 102					
Humanherpesvirus4completewildty							8,	15,							2,		2,		

peg		288	540		51		62		
Humanherpesvirus4strainAG876		8, 288	12, 432						
Humanherpesvirus4strainGD1		7, 252	14, 504						
Humanherpesvirus5strainAD169substra	2, 72								
Humanherpesvirus5strainAF1	2, 72								
Humanherpesvirus5strainHAN13	2, 72								
Humanherpesvirus5strainHAN20	2, 72								
Humanherpesvirus5strainU11	2, 72								
Humanherpesvirus5strainU8	2, 72								
Humanherpesvirus5strainVR1814	2, 72								
Humanherpesvirus6				2, 102			2, 102		
Humanherpesvirus6BDNAstrain				2, 102			3, 153		
Humanherpesvirus6							2, 102		
Humanherpesvirus7strainRK		2, 72	2, 72		4, 186	9, 338	7, 236	15, 374	
HumanparvovirusB19isolateKU1				4, 204					
SaimiriineherpesvirusstrainMV54					2, 102				
Smallanellovirus1		8, 116	8, 105					6, 121	
Smallanellovirus2		11, 165	11, 154					7, 126	
Torquetenomidivirus1DNAisolate		17, 324	15, 273					7, 126	
Torquetenomidivirus2DNAisolate		8, 241	9, 191					7, 126	
TorquetenomidivirusDNAisolate		12, 222 bp	14, 262 bp					9, 177	
TorquetenomidivirusDNAisolate		11, 218	11, 162					7, 126	
TorquetenomidivirusDNAisolate		14, 272	11, 193					8, 126	
TorquetenomidivirusDNAisolate		14, 332	14, 245					9, 177	
TorquetenomidivirusDNAisolate		12, 247	13, 234					8, 126	
TorquetenomidivirusDNAisolate		5, 159	8, 204					2, 50	
TorquetenomidivirusDNAisolate		9, 176	10, 190					7, 121	
TorquetenomidivirusDNAisolate		14, 303	14, 252					7, 121	
TorquetenomidivirusDNAisolate		9, 176	11, 246					3, 86	
TorquetenomidivirusDNAisolate		14, 266	14, 270					8, 126	
TorquetenomidivirusDNAisolate		16, 349	14, 241					6, 177	
TorquetenomidivirusDNAisolate		10, 273	14, 277					3, 102	
TorquetenomidivirusDNAisolate		10, 201	12, 242					2, 51	
TorquetenomidivirusDNAisolate		14, 298	15, 252					7, 126	
TorquetenomidivirusDNAisolate		10, 222	11, 190					8, 126	
TorquetenomidivirusDNAisolate		10, 201	13, 226					3, 102	
TorquetenovirusDNAisolate	13, 448							35, 1083	
TorquetenovirusstrainSIA109	11, 359							3, 118	
V01555J02070K01729K01730V01554X00498X00499		8, 288	13, 468		2, 51		2, 62		

The table presents the viruses detected in high hyperdiploid ALL patients. Presented as in Supplemental Table 6.

Supplemental Table 6: Virus detection in age- and gender-matched controls

Virus	PA 10	PA 135	PA 144	PA 159	PA 25	PA 27	PA 3	PA 42	PA 48	PA 62	PA 71	PA 85	PA 87	PA 91
Abelsonmurineleukemiavirus												34, 1580		19, 1240
Amphotropicmurineleukemiavirusstrain												44, 309		
Autographacalifornicanucleopolyhedroviru	14, 1182	30, 1882	27, 1669	34, 1999	10, 815	26, 1674	25, 1665	30, 1663	23, 1494	21, 1723	27, 1736	5, 501		30, 1480
Bacteriophage933W			2, 202	5, 470	2, 159	11, 1022	4, 404	4, 278	5, 505	8, 619	3, 303			6, 549
Bacteriophagef1	5, 299	12, 304	7, 440	11, 333	4, 207	9, 498	13, 275	23, 332	32, 333	22, 145	12, 140			7, 293
BacteriophageHK022			3, 303				3, 303	3, 302	4, 340				2, 202	
BacteriophageHK620			2, 202					2, 201	4, 340				2, 202	
BacteriophageHK97		3, 235	5, 359	5, 505			3, 303	5, 504	7, 643					
BacteriophageIf1	5, 353	15, 584	13, 536	27, 613	3, 266	9, 515	11, 516	18, 607	17, 524	11, 491	10, 450	4, 320		19, 547
BacteriophageP22								5, 405	3, 239					
BacteriophageP22pbi								5, 405	4, 340					
BacteriophagephiK	59, 165	78, 173	76, 188	76, 186	54, 164	76, 179	75, 172	76, 183	78, 189	83, 195	76, 185	65, 179	60, 165	76, 179
BacteriophageRB32									2, 103					
BacteriophageS13	9518, 5352	9640, 5352	9643, 5352	9703, 5352	9534, 5352	9630, 5352	9625, 5352	9653, 5352	9654, 5352	9667, 5352	9623, 5352	9497, 5352	9544, 5352	9670, 5352
BacteriophageS13circularDNA	9619, 5352	9753, 5352	9739, 5352	9792, 5352	9635, 5352	9753, 5352	9723, 5352	9750, 5352	9770, 5352	9759, 5352	9726, 5352	9611, 5352	9648, 5352	9764, 5352
BacteriophageSPBc2														8, 808
Bombyxmandarinanucleopolyhedrovirusis	4, 400	13, 793	9, 739	15, 855	3, 303	6, 598	10, 727	7, 630	7, 601	6, 539	9, 549			12, 869
Bombyxmorinuclearpolyhedrosisvirusiso	4, 400	12, 692	8, 736	14, 878	3, 303	6, 598	11, 828	6, 598	7, 601	6, 539	10, 603			10, 763
CASBREmurineleukemiavirusviralgenom												12, 151		2, 130
Cauliflowermosaicvirusgenome														2, 140
Cauliflowermosaicvirusgenome														2, 140
CauliflowermosaicvirusisolateBBC														2, 140
CauliflowermosaicvirusisolateNY8153														2, 140
Citrusexocortisviroid												38, 166	39, 145	19, 153
Citrusexocortisviroidclone2/5												208, 381	284, 377	74, 374
ColiphageNC51	9974, 5385	10048, 5385	10055, 5385	10083, 5385	9984, 5385	10027, 5385	10029, 5385	10028, 5385	10057, 5385	10062, 5385	10025, 5385	9957, 5385	9989, 5385	10054, 5385
ColiphagephiX174isolateSC2	10311, 5352	10312, 5352	10311, 5352	10312, 5352	10311, 5352	10311, 5352	10311, 5352	10312, 5352	10312, 5352	10312, 5352	10313, 5352	10311, 5352	10311, 5352	10311, 5352
ColiphageWA10	9238, 5352	9474, 5352	9520, 5352	9552, 5352	9271, 5352	9400, 5352	9457, 5352	9476, 5352	9473, 5352	9509, 5352	9403, 5352	9185, 5352	9290, 5352	9459, 5352
CucumbermosaicvirussegmentRNA1		2, 103						7, 112			3, 109			
DG75Murineleukemiavirus												503, 3620	11, 553	57, 2532
Enterobacteria	13, 1034	70, 5720	50, 4425	76, 6902	10, 890	43, 3847	38, 3727	48, 4242	73, 6790	62, 4545	43, 3687	8, 702	6, 606	53, 4968
Enterobacteriaphage2851	5, 393	20, 647	19, 817	27, 1309	10, 777	29, 1561	8, 461	21, 1104	28, 1137	16, 1105	18, 1145	6, 547		18, 644
Enterobacteriaphagealpha3	52, 160	63, 160	60, 163	65, 162	48, 132	58, 158	64, 165	63, 166	66, 165	67, 163	64, 184	50, 161	60, 161	62, 164
EnterobacteriaphageAR1DNA								2, 156	2, 103					
EnterobacteriaphageCUS3		2, 156		2, 202			3, 303		2, 202					
EnterobacteriaphageDE3	59, 1593	96, 4984	71, 4452	105, 5954	40, 1179	63, 3688	76, 3349	110, 3292	75, 5474	73, 4107	84, 3938	11, 923	18, 1217	97, 4644
Enterobacteriaphagef1DNAisolate	5, 299	12, 304	7, 440	11, 333	4, 207	9, 498	13, 275	23, 332	32, 333	22, 145	12, 140			7, 293
Enterobacteriaphagef1DNAisolate	5, 299	12, 336	8, 445	12, 333	4, 207	9, 498	10, 270	22, 338	32, 341	18, 134	10, 132			7, 293
Enterobacteriaphagef1isolateF1ances	5, 299	12, 304	7, 440	11, 333	4, 207	9, 498	13, 275	23, 332	32, 333	22, 145	12, 140			7, 293
EnterobacteriaphageG4isolateAnc	18, 156	38, 155	39, 159	40, 157	16, 120	35, 158	37, 157	38, 155	41, 157	39, 156	27, 156	24, 157	26, 156	40, 158
EnterobacteriaphageG4isolateG4anc	18,	38,	39,	40,	16,	35,	37,	38,	41,	39,	27,	24,	26,	40,

es	156	155	159	157	120	158	157	155	157	156	156	157	156	158
EnterobacteriaphageID2Moscow/ID/2001	40, 154	55, 156	53, 161	56, 159	39, 157	53, 155	52, 156	52, 154	55, 164	63, 160	57, 169	41, 157	41, 158	59, 164
EnterobacteriaphageMin27				3, 283	2, 159	7, 618	2, 202	3, 177	3, 303	5, 446	3, 303			3, 246
EnterobacteriaphageMS2isolateST4														171, 3456
EnterobacteriaphageP1mod749	9, 552	34, 771	26, 699	36, 760	15, 642	32, 686	18, 696	23, 693	36, 701	14, 626	10, 503	42, 773	20, 921	46, 1431
EnterobacteriaphageP22DNAstrain								5, 405	4, 340					
EnterobacteriaphageP7	33, 1136	75, 1151	50, 1169	89, 1149	46, 1148	105, 1175	72, 1063	97, 1188	109, 1154	74, 1149	67, 1159	10, 624	17, 745	107, 1126
EnterobacteriaphagephiX174isolate3F9	10313, 5352	10313, 5352	10312, 5352	10313, 5352	10312, 5352	10312, 5352	10312, 5352	10312, 5352	10312, 5352	10313, 5352	10312, 5352	10312, 5352	10312, 5352	10312, 5352
EnterobacteriaphageRB14								2, 156	2, 103					
EnterobacteriaphageSf6	6, 340	9, 556	19, 859	27, 1110	2, 186	9, 636	11, 943	17, 1203	9, 615	11, 420	8, 491	3, 303	4, 404	16, 809
EnterobacteriaphageST104DNA							2, 202	5, 280						
EnterobacteriaphageSt1	39, 163	55, 172	62, 198	65, 186	41, 196	62, 179	60, 195	60, 200	64, 196	65, 195	61, 180	49, 178	46, 197	63, 179
EnterobacteriaphageT4								2, 156	2, 103					
EnterobacteriaphageT4T								2, 156	2, 103					
EnterobacteriaphageVT2Sakaiproviral				2, 182		7, 618		4, 278	2, 202	5, 446	4, 404			
EnterobacteriaphageYYZ2008	6, 552	12, 920	6, 524	26, 1765	6, 566	14, 1274	20, 1356	12, 828	20, 1293	24, 1789	8, 808			24, 1508
EpsteinBarrvirusartifactualjoin		21, 2090										2, 202	9, 909	64, 5196
FBRmurineosteosarcomacompletepr												92, 1521	2, 202	50, 860
FriendmurineleukemiavirusFB29												42, 286	2, 132	3, 303
Friendmurineleukemiavirusgenomic RNAcl												36, 274		3, 303
Friendmurineleukemiavirus												28, 274		2, 202
Friendspleenfocusformingvirusg												14, 533		13, 479
Friendspleenfocusformingvirus												132, 1378	3, 154	16, 766
GenomeofphageG4	4, 142	7, 147	7, 119	8, 123	4, 116	4, 145	6, 115	5, 146	11, 146	12, 146	6, 147	3, 102	3, 107	5, 107
HIV1										25, 1564				
HIV1clonepIIIBfromUSA										22, 1407				
HIV1E9fromtheUSA										23, 1451				
HIV1strainNLXJDC6441X2fromChina										10, 654				
Humanadenovirus6strainTonsil99prot	2, 202	2, 202		2, 202		92, 4780		3, 302	2, 202					4, 403
HumanadenovirusCserotype5						84, 4465		3, 302						2, 201
HumanadenovirusJJS2010strain16700	2, 202	2, 202		2, 202		86, 4737		3, 302	2, 202					4, 403
Humanadenovirustype1subgroupC	2, 202	2, 202		2, 202		91, 4738		2, 201	2, 202	2, 201				3, 302
Humanadenovirustype5strainNHRCAd5F						84, 4465		3, 302						2, 201
HumancytomegalovirusstrainAD169										3, 231				2, 202
Humanherpesv										2, 202				2, 202
Humanherpesvirus3							4, 404							
Humanherpesvirus3							4, 404							
Humanherpesvirus4completewildtypeg		21, 2090										2, 202	9, 909	64, 5185
Humanherpesvirus4strainAG876		17, 1686bp											7, 707bp	55, 4519bp
Humanherpesvirus5strainHAN38										3, 231				2, 202
Humanherpesvirus6								2, 202						
Humanherpesvirus6BDNAstrain								2, 202						
Humanherpesvirus7strainRK	82, 1091	152, 1594	142, 1241	101, 1355	58, 1187	128, 1582	108, 2085	77, 988	73, 808	111, 1084	70, 1037	78, 860	71, 1250	107, 1325

Humanimmunodeficiencyvirus1DNA clone										7, 608				
Humanimmunodeficiencyvirus1proviral										7, 697				
HumanimmunodeficiencyvirusHIV1cloneA										3, 303				
HumanimmunodeficiencyvirusHIV1cloneA										9, 675				
Humanimmunodeficiencyvirustype1isolat										25, 1564				
Humanimmunodeficiencyvirustype1										25, 1564				
HumanTcellleukaemiatypeIII										23, 1452				
HypososterdidymatorvirussegmentSHF														9, 446
Mastadenovirush5gene						84, 4465		3, 302						2, 201
Moloneymurineleukemiavirus												14, 231		2, 102
Moloneymurineleukemiavirusneuropatho												9, 126		
Moloneymurineleukem												14, 231		2, 102
Mousemammarytumorviruscompleteprovi												43, 2245		
MuLV												192, 1816	5, 232	21, 1015
MurineleukemiavirusGraffiGV12DNA												109, 720		4, 402
MurineleukemiavirusGraffiGV14DNA												147, 793	3, 176	8, 404
MurineleukemiavirusisolateNeRV												70, 962		12, 721
MurineleukemiavirusMCF1233												127, 2563	3, 129	34, 1593
MurineleukemiavirusSL33												43, 715	3, 129	10, 593
MurineleukemiavirusstrainBM5eco												89, 1232		8, 582
MurineleukemiavirusstrainSRS196												10, 227		3, 261
Murineosteosarcomavirus												93, 1527	2, 202	50, 806
PhageBP4795	11, 730	26, 864	25, 1225	37, 1706	15, 1018	44, 2341	11, 755	29, 1363	47, 1694	23, 1466	30, 1540	8, 671		34, 1428
PhagecdtIDNA		2, 202		3, 303					3, 226	2, 129				
PhageM13genome	5, 299	12, 336	8, 445	12, 333	4, 207	9, 498	12, 279	23, 338	33, 341	21, 145	11, 133			7, 293
phageMS2genome														175, 3356bp
Plutellaxylostellamultiplenucleopolyh	13, 1081	27, 1615	25, 1467	32, 1807	8, 613	25, 1573	24, 1564	29, 1562	21, 1292	20, 1622	25, 1634	4, 400		29, 1379
Rachiplusiaoumultiplenucleopolyhedrov	4, 239	8, 551	7, 470	9, 488	3, 301	6, 370	6, 555	6, 352	6, 519	4, 364	6, 418			9, 315
Rauschermurineleukemiavirus												40, 387		3, 189
Retroviridae												242, 2850	6, 257	37, 1659
Saimiriineherpesvirus1strainMV54		2, 202		2, 202										
Salmonellaphagec341							3, 303	2, 201	4, 340	2, 126				2, 202
SalmonellaphageSE1								5, 280						
Salmonellatyphimuriumbacteriophage ES1							3, 303			2, 126				
Shigellaflexneribacterio														2, 202
Shrimpwhitespotsyndromevirus	14, 904	15, 1036	24, 1203	42, 1220	10, 693	45, 1149	17, 961	24, 1038	35, 1204	20, 1071	11, 840	8, 578	9, 639	54, 1257
SimianimmunodeficiencyvirusgenomicRN										12, 609				
Smallanellovirus1		8, 141												
Smallanellovirus2		17, 182												
Stx1convertingphageDNA				3, 283		7, 618	3, 303	4, 278	3, 303	5, 446	4, 404			2, 145
Stx2convertingphage1717	8, 697	20, 748	43, 2585	43, 2285	13, 1112	47, 2695	23, 1548	39, 2472	40, 1965	24, 1841	29, 1991	8, 778	8, 641	38, 1827

Stx2convertingphage86DNA				3, 283	2, 159	3, 303	4, 404	2, 202	2, 202	2, 143				3, 246
Stx2convertingphageIDNA			2, 202	5, 470	2, 159	11, 1022	4, 404	4, 278	5, 505	8, 619	3, 303			6, 549
Stx2convertingphageIIDNA	6, 494	20, 647	18, 789	26, 1212	10, 777	29, 1561	10, 663	19, 902	29, 1238	16, 1105	19, 1145	6, 547		20, 789
Torquetenomidivirus1DNAisolate		38, 497												
Torquetenominivirus1DNAisolate											2, 103			
Torquetenominivirus5DNAisolate		2, 169			4, 111						2, 103			
Torquetenominivirus7DNAisolate		24, 178					2, 139							
Torquetenovirus16DNAisolate					2, 202									4, 404
TorquetenovirusDNAisolate														5, 505
TorquetenovirusDNAisolate											2, 154			4, 404
Torquetenovirusisolatetth19		5, 482												
TTvirusisolateJA1												2, 193		
TTvirusisolateTUPB					3, 303									5, 506
TTVlikeminivirusDNAisolate											2, 103			
VacciniavirusGLV1h68	66, 2060	66, 2432	47, 3002	92, 3816	43, 1541	49, 3162	64, 2319	96, 2565	61, 3527	96, 3989	70, 2800	9, 909	9, 762	37, 2558
WoodchuckhepatitisBvirus										6, 448				
XenotropicMuLVrelatedvirus22Rv1comp												414, 3688	8, 532	49, 2542

The table presents the viruses detected in blood samples from age- and gender-matched controls whole genome sequenced in the ICGC PedBrain project. Presented as in Supplemental Table 6.

Supplemental Table 7: Virus detection in 1000 genomes data sets.

Virus	HG00100	HG00106	HG00103	HG00117	HG00116	NA12878	NA18507
BacteriophageS13						2, 202	375, 4915
BacteriophageS13circularDNA						2, 202	382, 4927
ColiphageID1							220, 3860
ColiphageID22							281, 4129
ColiphageID32							8, 200
ColiphageID34						2, 202	259, 4436
ColiphageID45						2, 202	343, 4770
ColiphageNC11						2, 202	293, 4321
ColiphageNC16						2, 202	316, 4595
ColiphageNC1						2, 202	255, 3935
ColiphageNC28							3, 115
ColiphageNC29							3, 115
ColiphageNC35							5, 158
ColiphageNC37						2, 202	315, 4595
ColiphageNC3							8, 200
ColiphageNC41						2, 202	319, 4632
ColiphageNC51						2, 202	403, 5198
ColiphageNC56						2, 202	298, 4564
ColiphageNC5						2, 202	326, 4641
ColiphageNC7						2, 202	239, 3877
ColiphagephiX174isolateAnc						2, 202	437, 5258
ColiphagephiX174isolateC						2, 202	437, 5258
ColiphagephiX174isolateCG						2, 202	432, 5231
ColiphagephiX174isolateCGGhi						2, 202	432, 5231
ColiphagephiX174isolateCGGhiC						2, 202	425, 5231
ColiphagephiX174isolateCGGhiMhi						2, 202	423, 5231
ColiphagephiX174isolateCGGlo						2, 202	423, 5181
ColiphagephiX174isolateCGGloC						2, 202	414, 5178
ColiphagephiX174isolateCGGloMhi						2, 202	421, 5173
ColiphagephiX174isolateCM						2, 202	437, 5258
ColiphagephiX174isolateCMMhi						2, 202	424, 5226
ColiphagephiX174isolateCMMhiC						2, 202	421, 5199
ColiphagephiX174isolateCMMhiMhi						2, 202	423, 5227
ColiphagephiX174isolateCMMlo						2, 202	437, 5258
ColiphagephiX174isolateCMMloC						2, 202	429, 5231
ColiphagephiX174isolateCMMloMhi						2, 202	426, 5258
ColiphagephiX174isolateS1						2, 202	436, 5258
ColiphagephiX174isolateSC1						2, 202	433, 5231
ColiphagephiX174isolateSC2						2, 202	436, 5258
ColiphagephiX174isolateSCS1						2, 202	431, 5231
ColiphagephiX174isolateSCS2						2, 202	435, 5258
ColiphagephiX174isolateSCSC1						2, 202	428, 5227
ColiphagephiX174isolateSCSC2						2, 202	434, 5260
ColiphageWA10						2, 202	337, 4816
ColiphageWA11						2, 202	217, 3791
ColiphageWA13							6, 200
ColiphageWA45							8, 200
ColiphageWA4						2, 202	207, 3598
EnterobacteriaphageDE3							2, 200
EnterobacteriaphagephiX174isolate100						2, 202	434, 5258
EnterobacteriaphagephiX174isolate10A						2, 202	432, 5258
EnterobacteriaphagephiX174isolate10B						2, 202	426, 5258
EnterobacteriaphagephiX174isolate10C						2, 202	435, 5258
EnterobacteriaphagephiX174isolate10D						2, 202	427, 5258

EnterobacteriaphagephiX174isolate10E						2, 202	425, 5258
EnterobacteriaphagephiX174isolate10F						2, 202	426, 5258
EnterobacteriaphagephiX174isolate250						2, 202	425, 5258
EnterobacteriaphagephiX174isolate30A						2, 202	423, 5253
EnterobacteriaphagephiX174isolate30B						2, 202	437, 5258
EnterobacteriaphagephiX174isolate30C						2, 202	426, 5258
EnterobacteriaphagephiX174isolate3A9						2, 202	435, 5258
EnterobacteriaphagephiX174isolate3B9						2, 202	436, 5258
EnterobacteriaphagephiX174isolate3C9						2, 202	420, 5245
EnterobacteriaphagephiX174isolate3D9						2, 202	426, 5258
EnterobacteriaphagephiX174isolate3E9						2, 202	431, 5258
EnterobacteriaphagephiX174isolate3F9						2, 202	433, 5258
EnterobacteriaphagephiX174isolateAP1						2, 202	439, 5258
EnterobacteriaphagephiX174isolateDEL						2, 202	437, 5258
EnterobacteriaphagephiX174isolateJAC						2, 202	437, 5258
EnterobacteriaphagephiX174isolatePhi						2, 202	437, 5258
EnterobacteriaphagephiX174strainalph						2, 202	423, 5202
EnterobacteriaphagephiX174strainbeta						2, 202	425, 5221
EnterobacteriaphagephiX174straingamm						2, 202	437, 5258
EpsteinBarrvirusartifactualjoin	33, 3368	41, 4265	29, 2790	1682, 89013	4886, 148271	202028, 169979	166543, 174668
HepatitisCvirusgenomicRNAisolate							138, 1381
HepatitisCvirussubtype1aclonepHCV							494, 2154
HepatitisCvirussubtype1agenomicRNA							59, 1160
HepatitisCvirussubtype1ais							178, 1876
HepatitisCvirussubtype1aisolate03							163, 1401
HepatitisCvirussubtype1aisolate03P							20, 752
HepatitisCvirussubtype1aisolatecol							179, 1443
HepatitisCvirussubtype1aisolateDN0							173, 1629
HepatitisCvirussubtype1aisolateDN1							175, 1570
HepatitisCvirussubtype1aisolateDN2							47, 1262
HepatitisCvirussubtype1aisolateHCV							187, 2000
HepatitisCvirussubtype1aisolateTN1							135, 1419
HepatitisCvirussubtype1aisolateTN6							175, 1763
HepatitisCvirussubtype1aisolateV02							166, 1435
HepatitisCvirussubtype1aisolateV03							145, 1366
HepatitisCvirussubtype1a							507, 2166
HepatitisCvirussubtype1apolyprotein							202, 1765
HepatitisCvirussubtype1astrainHCT							140, 1543
Humanadenovirus6strainTonsil99prot	6, 112	4, 112					
HumanadenovirusCserotype5	6, 112	4, 112					
Humanadenovirustype1subgroupC	6, 112	4, 112					
Humanadenovirustype5strainNHRCAd5F	6, 112	4, 112					
Humanherpesvirus4completewildtypeg	33, 3368	41, 4265	29, 2790	1681, 86363	4872, 139079	197601, 157953	164060, 163421
Humanherpesvirus4strainAG876	27, 2788	34, 3588	23, 2279	1511, 78495	4444, 129281	178233, 149747	145380, 156239
Humanherpesvirus4strainGD1	30, 3044	37, 3889	26, 2490	1591, 82811	4643, 135124	188680, 156181	152460, 161797
Humanherpesvirus7strainRK	3, 222	2, 216		2, 152		13, 593	176, 1193
Mastadenovirush5gene	6, 112	4, 112					
Pestivirustype1							6, 238

The table presents the viruses detected in whole genome sequencing data from the 1000 genomes project (1000genomes.org). Presented as in Supplemental Table 6. For simulations HG00100, HG00106, HG00103, HG00117 and HG00116 were used.

Supplemental Table 8: Aligning directly to viral genomes. Exemplary test results for patient 1.

	Diagnosis		Relapse		Remission	
Viral genome	**PE**	**Cov.**	**PE**	**Cov.**	**PE**	**Cov.**
Atelineherpesvirus3_AF083424	4	55 bp	2	102 bp	4	140 bp
Aviansarcomavirusproviralcsrcgeneco_L21974	4	83 bp	2	102 bp	3	65 bp
Bovineherpesvirustype11_AJ004801	6	190 bp	5	166 bp	5	207 bp
Bovineviraldiarrheavirus1NADL_M31182	8	131 bp	15	148 bp	7	100 bp
Ectocarp_AF204951AF204952AF210454U95206X76296			2	59 bp		
Emilianiahuxleyivirus86isolateEhV86_AJ890364	6	189 bp	3	106 bp	3	153 bp
Equidherpesvirus2_U20824	2	102 bp			2	102 bp
GardnerRasheedfelinesarcomaviru_X00255K01487	2	102 bp	5	120 bp	5	115 bp
HepatitisCvirussubtype1aisolateHCV_EU155319			3	52 bp		
Humanherpesvirus6_AF157706L13162L14772L16947					2	102 bp
Humanherpesvirus6BDNAstrain_AB021506	2	102 bp			4	205 bp
Humanherpesvirus7strainRK_AF037218	515	1353 bp	528	1422 bp	569	1500 bp
Humanherpesvirus8strainGK18_AF148805					2	52 bp
Pestivirusgiraffe1H138_AF144617	10	156 bp	16	156 bp	8	104 bp
Pestivirustype1_AF268278	8	131 bp	15	148 bp	7	100 bp
Pestivirustype1cytopathicgenomicRNA_U86599	2	54 bp				
RoussarcomavirusSchmidtRuppinDgeno_D10652	4	83 bp			3	65 bp
RoussarcomavirusstrainSchmidtRuppin_AF052428	4	83 bp			3	65 bp
Saimiriineherpesvirus1strainMV54_HM625781	5	165 bp	7	271 bp	3	152 bp
Simiansarcomavirusprovira_V01201J02394J02397	3	111 bp	6	96 bp	6	139 bp

To test whether the sensitivity of virus detection could potentially be increased by starting with the alignment to viral sequences instead of using the sequences that were not aligned to the human genome we altered the virus detecting workflow. This measure resulted in a greater number of detected virus species. However, detailed inspection showed that all of these surplus virus species were falsely detected due to sequence similarities between human and viral sequences. In addition, many of the detected virus species do not infect humans. This procedure did in fact not increase the sensitivity compared to our original workflow. PE, number of paired end reads; Cov., covered viral sequence in bp.

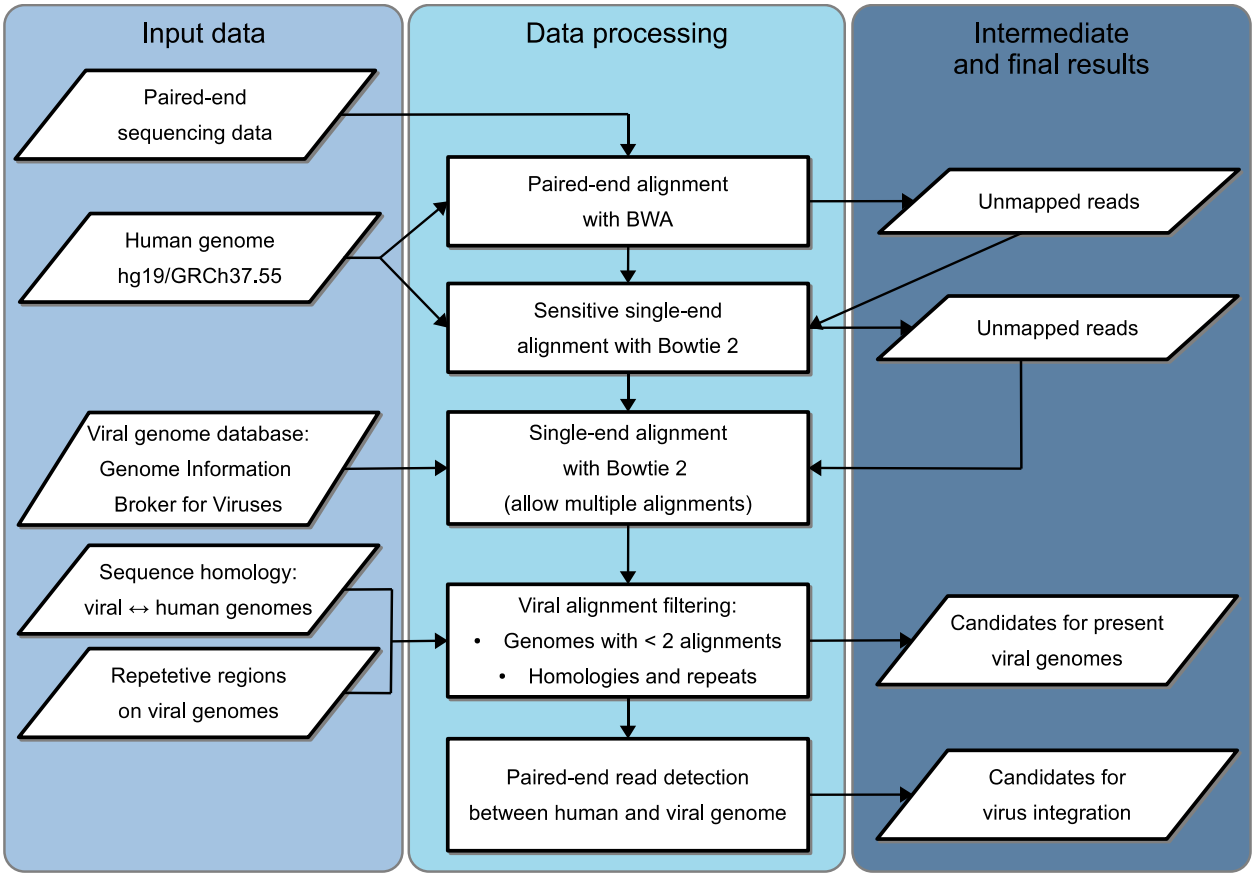
Supplemental Table 9: Alignment using BLAT. Exemplary test results for patient 1.

	Diagnosis		Relapse		Remission	
Viral genome	**PE**	**Cov.**	**PE**	**Cov.**	**PE**	**Cov.**
Amsactamooreientomopoxvirus_AF250284	3	135 bp				
Bacteriophage69_AY954951	2	102 bp				
Bacteriophage85_AY954953	2	102 bp				
Bovineherpesvirustype11_AJ004801	4	87 bp			3	50 bp
BovinepolyomavirusDNA_D13942D00755M74843	2	101 bp				
Cercopithecineherpesvirus16strainX31_DQ149153	4	188 bp	4	133 bp	3	79 bp
Cercopithecineherpesvirus1strainE249_AF533768					3	123 bp
CotesiacongregatavirussegmentCircle1_AJ632314	5	1351 bp	8	921 bp	6	919 bp
CotesiacongregatavirussegmentCircle1_AJ632315			3	295 bp		
CotesiacongregatavirussegmentCircle2_AJ632324					2	429 bp
CotesiacongregatavirussegmentCircle2_AJ632326	2	46 bp	2	48 bp		
Cryptoph_AY229987AY096241AY096242X77048X79569			2	112 bp		
Culexnigripalpusbaculovirus_AF403738	13	111 bp	10	114 bp	11	107 bp
Cyprinidherpesvirus3DNAstrain_AP008984	7	363 bp	5	270 bp	4	468 bp
Emilianiahuxleyivirus86isolateEhV86_AJ890364	25	1527 bp	39	2369 bp	28	1888 bp
EpsteinBarrvirusartifactualjoin_M80517M75989	3	67 bp	7	138 bp	2	43 bp
Equidherpesvirus2_U20824	4	165 bp	9	390 bp	14	776 bp
Gallidherpesvirus2_AF147806			2	273 bp	5	727 bp
Gallidherpesvirus2serotype1isolate_AF243438	2	1060 bp	5	1061 bp	5	1330 bp
Gallidherpesvirus2strainCVI988_DQ530348					5	320 bp
GlyptafumiferanaeichnovirusDNAsegmen_AB289987	3	126 bp	5	142 bp	8	139 bp
Glyptapantelesflavicoxisbracovirusseg_EU001284	3	139 bp			2	188 bp
HepatitisCvirussubtype1aisolateHCV_EU155215	2	32 bp			3	32 bp
HepatitisCvirussubtype1aisolateHCV_EU155275			2	31 bp		
HepatitisCvirussubtype1aisolateHCV_EU155299					2	102 bp
HepatitisCvirussubtype1aisolateHCV_EU155319	65	462 bp	68	394 bp	69	425 bp
Honeysuckleyellowveinbeta[Japan_AB236326	2	48 bp	2	48 bp	6	171 bp
Honeysuckleyellowveinmosaicvirusass_AJ543430					2	54 bp
Humanherpesvirus4completewildtypeg_AJ507799	3	67 bp	6	358 bp	5	557 bp
Humanherpesvirus6_AF157706L13162L14772L16947	6	104 bp	7	103 bp	7	269 bp
Humanherpesvirus6_X83413					8	286 bp
Humanherpesvirus6BDNAstrain_AB021506	12	245 bp	21	382 bp	15	584 bp
Humanherpesvirus7strainRK_AF037218	21	1391 bp	20	779 bp	44	1474 bp
Humanherpesvirus8strainGK18_AF148805	3	302 bp				
ImpatiensnecroticspotvirusisolateHD_GU112504	2	36 bp				
Jaagsiektesheepretrovirus_DQ838494			3	88 bp		
Microplitisdemolitorbracovirussegment_AY875687	3	44 bp			2	387 bp
Molluscumcontagiosumvirussubtype1_U60315					2	314 bp
NSMglycoproteinprecursor{Msegment}[to_S48091	2	71 bp				
ParameciumbursariaChlorellavirusFR48_DQ890022			3	67 bp		
PhlebovirusspCoAr171616segmentS_EF201836	2	36 bp				
PropionibacteriumphagePA6_DQ431235	2	102 bp				
RicestripevirussegmentRNA3isolateY_FM242704					2	211 bp
Saimiriineherpesvirus1strainMV54_HM625781			5	537 bp	6	116 bp
Staphylococcusaureusphagephi11_AF424781	2	102 bp				
StaphylococcusphagephiETA2DNAstrain_AP008953	2	102 bp				
StaphylococcusphagephiMR25DNA_AB370205	2	102 bp				
TanapoxvirusisolateTPVRoC_EF420157	2	553 bp				
Torquetenovirus12DNAisolate_AB064605	3	60 bp	2	58 bp	3	104 bp
Torquetenovirus19DNAisolate_AB025946			2	65 bp	2	41 bp
Torquetenovirus27DNAisolate_AB064595	3	61 bp	3	63 bp		

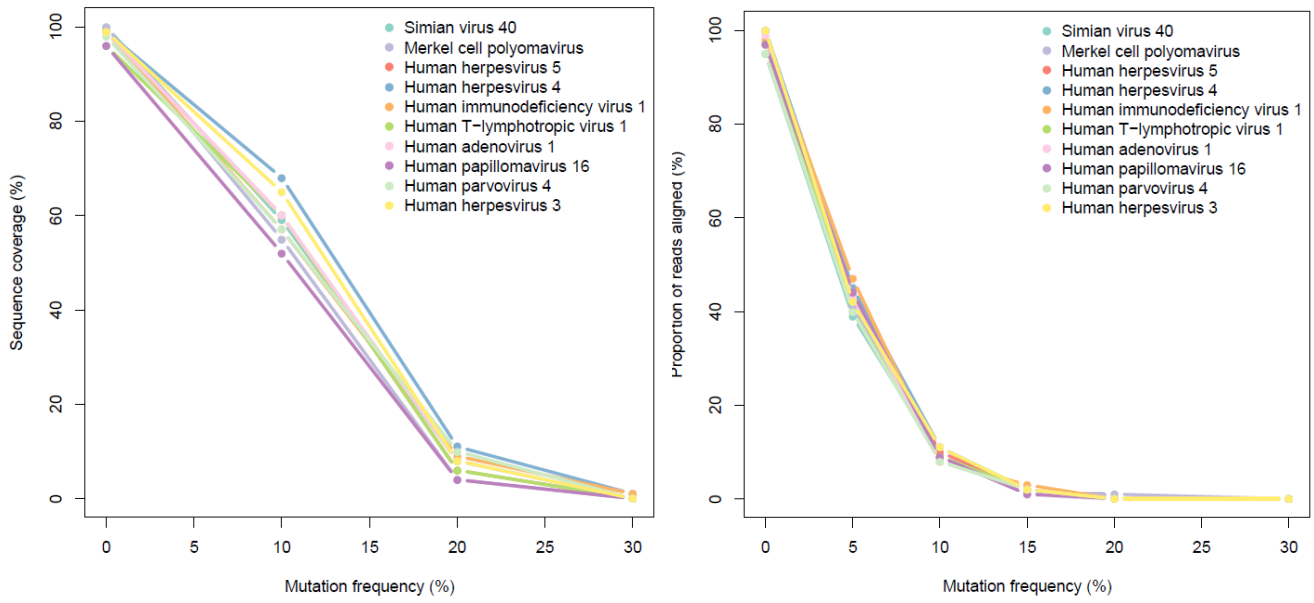
Torquetenovirus4DNAisolate_AB041957					2	76 bp
TorquetenovirusDNAisolate_AB064600	3	61 bp	3	65 bp	2	61 bp
TorquetenovirusDNAisolate_AB064602	3	60 bp	3	63 bp	2	60 bp
TorquetenovirusDNAisolate_AB064604	2	61 bp	2	61 bp	3	112 bp
Torquetenovirusisolatetth20_AJ620216	2	59 bp	2	60 bp	3	61 bp
Torquetenovirusisolatetth6_AJ620212	2	57 bp			2	57 bp
TTvirusisolate_AB038619					3	72 bp
TTvirusisolate_AB038620			3	71 bp	3	66 bp
V01555J02070K01729K01730V01554X00498X00499	3	67 bp	7	138 bp	2	43 bp
Venezuelanequineencephalitisvirusstr_AF075258			4	66 bp		

We tested BLAT ("Blast Like Alignment Tool", an optimized version of BLAST comparable to BLASTn: Kent, W. J. BLAT---The BLAST-Like Alignment Tool. Genome Research 12, 656–664 (2002)) as an alternative alignment tool for the alignment against the virus database. Using this tool we did not detect viral integrations, but more alignments against viruses than with our described pipeline. However, most of the virus sequences identified belonged to viruses not infecting humans. Regarding the alignment, it was obvious that BLAT mapped short fragments allowing for partial alignments and large gaps of more than 100 bp. The aligned reads were mainly comprised of repetitive sequence such as polyA, -T, -G, or -C or repeated k-mers. In addition, due to their repetitive nature, the reads had low base and sequence qualities (mean <20 for all bases, minimum 2 Phred-scaled) (Supplemental Fig. 3); usually values between 30 and 40 are realistic and expected for high base calling certainty). BLAT or BLAST don’t take base qualities into account, in contrast to BWA or BOWTIE2. The obtained viral sequences were seemingly falsely aligned by BLAT in a sporadic, stochastic fashion. Careful analysis of the mapped reads demonstrated that false positive detections were caused by mis-alignment of highly repetitive sequences. This procedure did not increase the sensitivity compared to our original workflow. PE, number of paired end reads; Cov., covered viral sequence in bp.

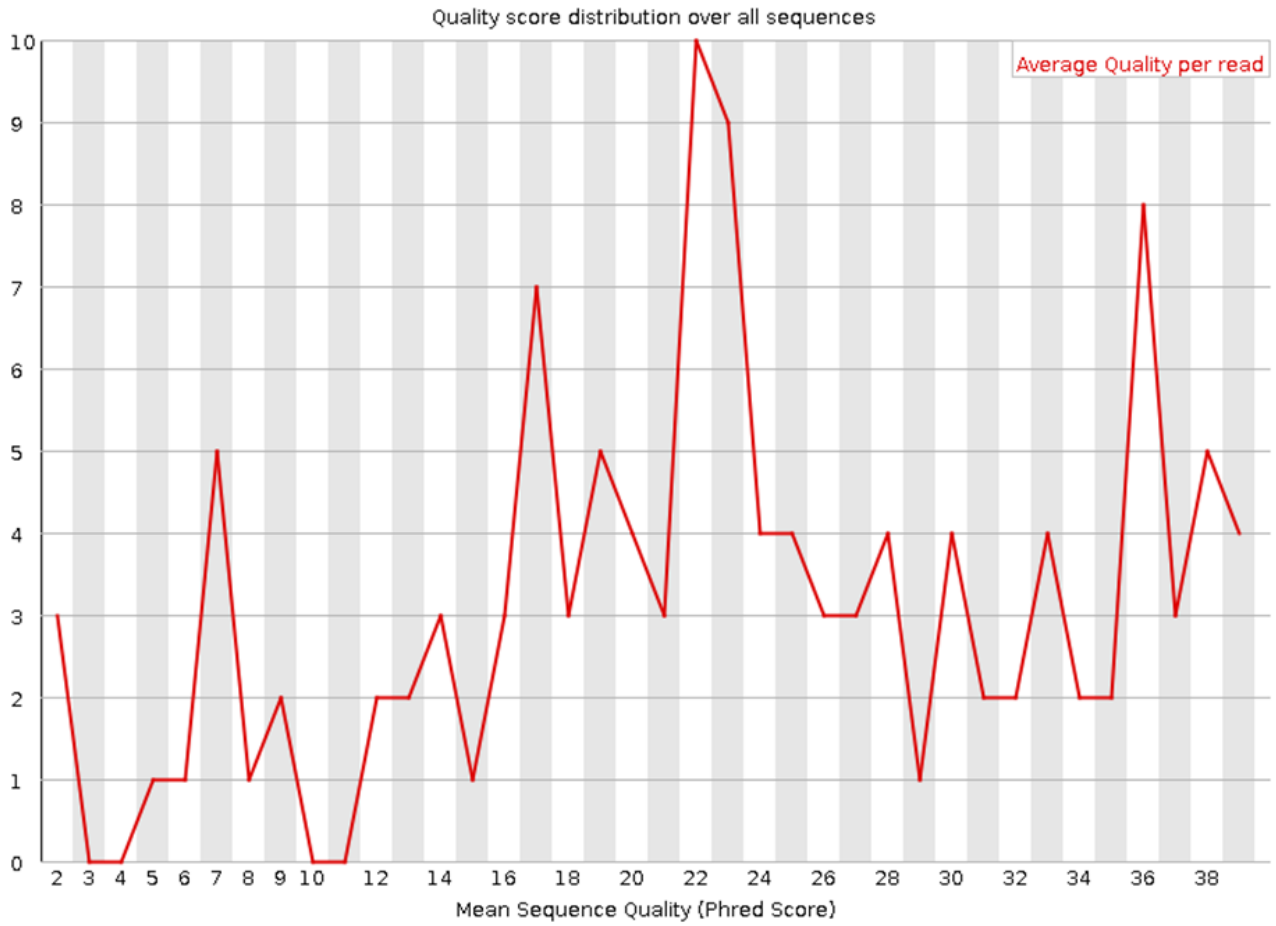
Supplemental Figure



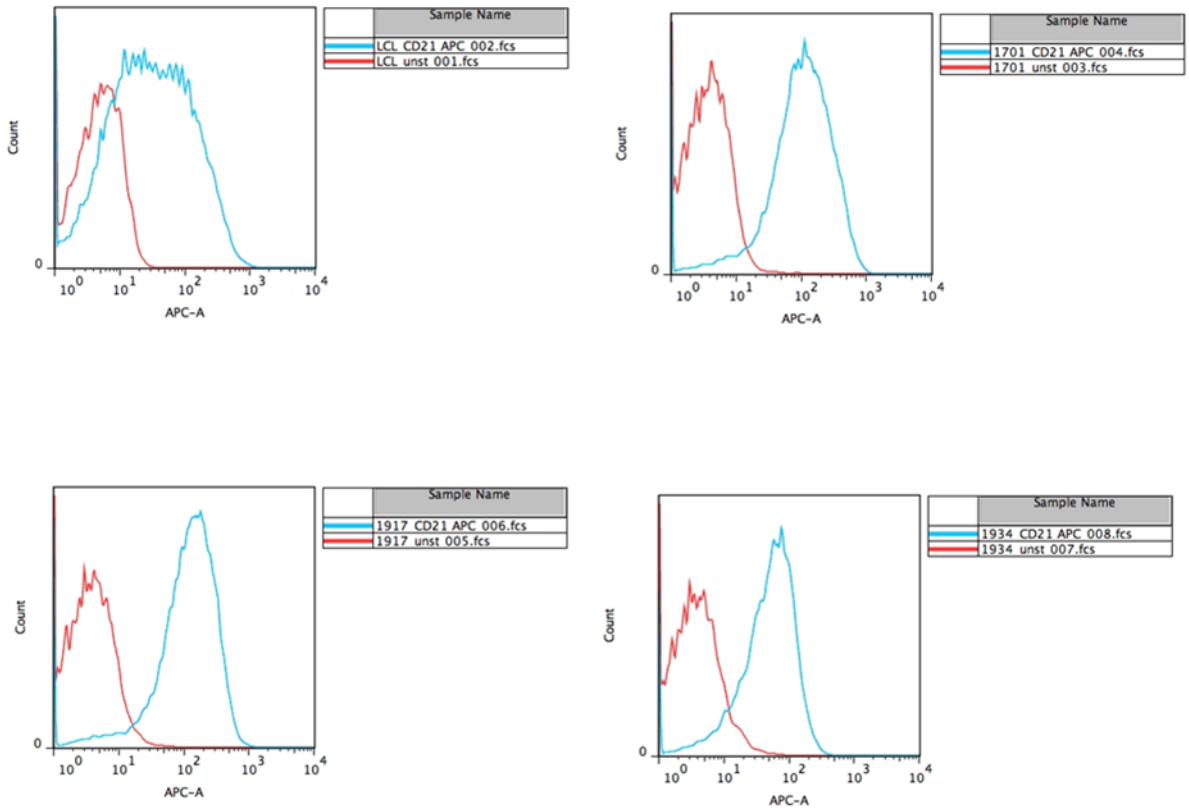
Supplemental Figure 1: **Bioinformatic detection of integrated viruses in whole-genome sequencing data sets.** Paired-end sequences were aligned against the human reference genome (GRCh37.55) employing BWA (Li & Durbin, 2009). The alignment was repeated with all previously unmapped reads with a higher sensitivity and a minimum mapping concordance of 85% using Bowtie2 (Langmead & Salzberg, 2012). The remaining unmapped reads were then aligned against the viral reference genomes (Genome Information Broker for Viruses database (Hirahata *et al*, 2007)) with Bowtie2 allowing multiple matches against more than one genome. The virus alignment was filtered once more for repeat sequences and regions of high homology with the human reference genome. Paired-end reads spanning the human and a viral genome were considered for breakpoint detection of viral integration. Viruses with at least two single-end alignments or one spanning read were chosen for manual inspection and validation. In total, 25.525 viral genomes, as deposited in the Genome Information Broker for Viruses were screened for presence and integration.



Supplemental Figure 2: **Bioinformatic detection of viruses in whole-genome sequencing data sets. (A)** Coverage of integrated viral genome sequence with increasing viral mutation rate. Ten common viruses were selected for simulation experiments (Supplementary Table 3). The percentage of covered integrated viral genome sequences decreases with an increase in the mutation rate of the annotated virus sequences. For each virus, 20 integration sites were randomly distributed across the human chromosome 1. The respective wildtype virus sequences correspond to 0% mutation rate. Depicted are the results for simulated paired end reads of 50 bp length and 5 fold sequencing depth. The reads were processed by the virus detection pipeline and the sequence coverage of integration sites was calculated for every virus and every mutation frequency. **(B)** Modeling the detection of viral sequences in the DNA without integration or evidence of integration. A total of 1,000 non-overlapping 50 bp paired-end reads (i.e. 2,000 sequences) were drawn at random positions from each mutated viral genome. The reads were processed according to the virus detection pipeline (including alignment against the human reference genome), and those reads were counted, which could be mapped back against a viral genome of the same type in the GIB-V database. The proportion of sequencing reads that can be aligned decreases with increasing viral mutation rate.



Supplemental Figure 3: Quality report (FastQC, version 0.11.3) of virus reads identified by BLAT alignment shows low sequence quality (Phred scores). Sequence quality averages around a Phred score of 20 (1% error rate). Phred scores of 3-10 equate to an error rate of 50% or 10%, respectively.



Supplemental Figure 4: CD21 is expressed in pre B-ALL cells. Flow cytometric measurement of cell surface expression of CD21 employing a specific anti-CD21 antibody conjugated with fluorescent APC (BD Biosciences, Heidelberg, Germany). Primary patient pre B-ALL cells (“LCL”) and three different human patient derived pre B-ALL xenografts grown in NSG mice (“1793, 1917, 1914”) were stained with anti-CD21-APC (blue histogram) or left unstained (red histogram).