www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 10 November 2014
Accepted: 07 May 2015
Published: 13 July 2015

Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data

Michael Forster[1,*], Silke Szymczak[1,*,†], David Ellinghaus[1], Georg Hemmrich[1], Malte Rühlemann[1], Lars Kraemer[1], Sören Mucha[1], Lars Wienbrandt[2] & Martin Stanulla[3], UFO Sequencing Consortium within the I-BFM Study Group[4], Andre Franke[1]

Several pathogenic viruses such as hepatitis B and human immunodeficiency viruses may integrate into the host genome. These virus/host integrations are detectable using paired-end next generation sequencing. However, the low number of expected true virus integrations may be difficult to distinguish from the noise of many false positive candidates. Here, we propose a novel filtering approach that increases specificity without compromising sensitivity for virus/host chimera detection. Our detection pipeline termed Vy-PER (Virus integration detection bY Paired End Reads) outperforms existing similar tools in speed and accuracy. We analysed whole genome data from childhood acute lymphoblastic leukemia (ALL), which is characterised by genomic rearrangements and usually associated with radiation exposure. This analysis was motivated by the recently reported virus integrations at genomic rearrangement sites and association with chromosomal instability in liver cancer. However, as expected, our analysis of 20 tumour and matched germline genomes from ALL patients finds no significant evidence for integrations by known viruses. Nevertheless, our method eliminates 12,800 false positives per genome (80× coverage) and only our method detects singleton human-phiX174-chimeras caused by optical errors of the Illumina HiSeq platform. This high accuracy is useful for detecting low virus integration levels as well as non-integrated viruses.

Two primary drivers for developing virus integration detection methods are the fields of disease therapy and disease etiology. In gene therapy and immunotherapy studies, a major consideration is the non-integration[1] or at least the safe integration of a vector's payload into the host genome[2–4]. In disease etiology, prominent examples of integrations into the host genome are the retroviruses human T-lymphotropic virus (HTLV) in adult T-cell leukemia[5] and human immunodeficiency virus (HIV) in acquired immune deficiency syndrome (AIDS). Recent studies reported that integration of HIV at specific genomic locations leads to clonal expansion of virus-infected cells – slowing viral decay under combination antiretroviral therapy (cART) – and to cancer initiation[6,7]. Other etiologically important viruses that may integrate using different methods are hepatitis B virus (HBV) in liver cancer and human papillomavirus (HPV) in cervical, anal, oropharynx and other cancers[8–10]. Epstein-Barr virus (EBV) is associated with Burkitt's lymphoma[11] and is routinely used to immortalise cell lines[12,13], but has also been

[1]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Schleswig-Holstein, D-24105 Kiel, Germany. [2]Department of Computer Science, Christian-Albrechts-University of Kiel, Schleswig-Holstein, D-24118 Kiel, Germany. [3]Department of Pediatric Haematology and Oncology, Hannover Medical School, Lower Saxony, D-30625 Hannover, Germany. [4]See the Author Contributions for details. [*]These authors contributed equally to this work. [†]Current address: Institute of Medical Informatics and Statistics, University Medical Center Schleswig-Holstein, Campus Kiel, Schleswig-Holstein, D-24105 Kiel, Germany. Correspondence and requests for materials should be addressed to M.F. (email: m.forster@ikmb.uni-kiel.de)

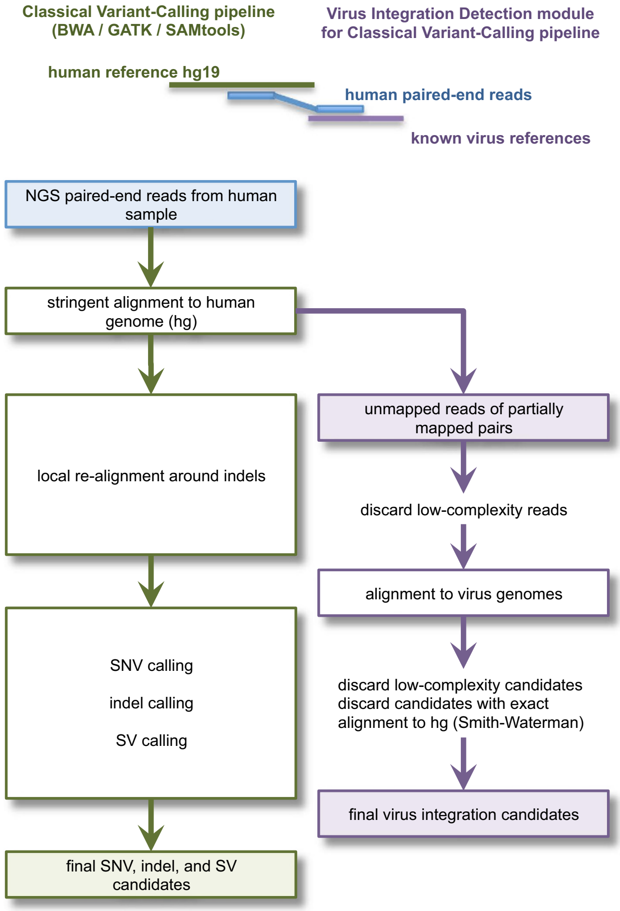
Classical Variant-Calling pipeline
(BWA / GATK / SAMtools)
Virus Integration Detection module
for Classical Variant-Calling pipeline
human reference hg19
human paired-end reads
known virus references
NGS paired-end reads from human sample
stringent alignment to human genome (hg)
local re-alignment around indels
SNV calling
indel calling
SV calling
final SNV, indel, and SV candidates
unmapped reads of partially mapped pairs
discard low-complexity reads
alignment to virus genomes
discard low-complexity candidates
discard candidates with exact
alignment to hg (Smith-Waterman)
final virus integration candidates

Figure 1. Detection of virus integrations into the human genome using paired-end sequencing. After the computationally expensive classical alignment of all paired-end reads to the human genome, the pipeline splits into classical variant calling and virus integration detection.

reported to integrate into the host genome at very low frequencies[14–16]. Regardless of whether viruses are integrated into the host genome or not, one causal mechanism for cancer development is binding of virus proteins to the tumour suppressor *p53*, thereby inhibiting apoptosis. This mechanism is exploited for example by HBV[17], HPV[18], herpes simplex type 1[19], measles[20], or simian virus type 40[21]. Another causal mechanism was hypothesised when HBV integration throughout the human genome was detected in liver cancer[22] and associated with genomic instability, putatively caused by fusion transcripts. A third causal mechanism was more recently suggested when HBV integration sites were found to recurrently cluster near genomic rearrangement sites and were therefore associated with chromosomal instability (chromothripsis)[23–25]. This discovery led to systematic and large-scale virus integration analyses of The Cancer Genome Atlas data[11,26], and of data in many ongoing cancer studies including the childhood acute lymphoblastic leukemia (ALL) deep sequencing pilot study initiated by the German Federal Office for Radiation Protection (http://goo.gl/q7SaUZ). Childhood ALL is mostly classifiable by fusion genes[27,28] and is usually associated with exposure to radiation[29,30], like other childhood cancers[31]. The exact causes of ALL remain unknown. To our knowledge, there have been no previous studies on virus integrations in childhood ALL.

Virus integrations can be detected using Illumina paired-end sequencing (Fig. 1), if one read maps to the human genome and its paired end to a virus genome[23,24]. We analysed our childhood leukemia data accordingly using own scripts, as the data were too large for publicly available pipelines even on our high performance Linux cluster (928 cores with the hardware limit of 125 GB RAM per job). In addition, the pipelines that we tested were quite slow and did not detect virus integrations in benchmarking runs of our leukemia data. Our initial scripts identified a surprisingly large number of virus candidates. On closer inspection we noticed that many virus candidates were simply long stretches of unspecific short tandem repeats (STRs) or homopolymers (Table 1). These unspecific sequence stretches turned out to be identical in many virus genomes as well as the human genome. We therefore developed filters to eliminate false positives efficiently without compromising the sensitivity to detect true chimeric human/virus sequences. One solution to eliminate nearly all false positives would be to filter low-complexity reads with DUST[32], as implemented for example in BLAST[33] searches, or with PRINSEQ[34]. To avoid

Virus candidate sequence	Probable virus candidate	More likely explanation
TTAGGGTTAGGGCTAGGGCTAGGGCTAGGGCTAGGGCTAGGGCTAGGGCTAGGGCT	Cyprinid herpesvirus 3	STR (human telomere)
CTCTCTCTCTCTCTCTCACACACACACACACACACACACACACACACACACACAC	Ictalurid herpesvirus 1	STRs
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTA	Caviid herpesvirus 2	mainly homopolymer
TATATATATATATATATATATATATATTTTTTTTTTTTTTTTTTTTTTTT	Cotesia congregata bracovirus	STR and homopolymer
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTANCCCTAACCCTA	Human herpesvirus 6A	STR (human telomere)
TAACCCTAACCCTAACCCTAACCCTAGCCCTAACCCTAACCCTAACCCTA	Human herpesvirus 7	STR (human telomere)

Table 1. Typical false positive virus candidates before final filtering.

removing potentially informative reads (see Supplementary Results online), we prefer to filter reads by only considering the length of its longest STR. A second solution to eliminate false positives would be to cluster the virus/human chimera candidates within a genomic window as implemented in the VirusSeq pipeline[26], but this method by definition loses the sensitivity to detect low-frequency events represented by singleton or low-frequency chimeras. A third method would be to obtain supporting information on potential viruses from unmapped read-pairs, as implemented in the VirusSeq and VirusFinder pipelines[35], which is limited to the co-occurrence of a virus, not its integration into the host genome. Our aim for high sensitivity was motivated by the knowledge that for example HIV or HTLV has a tropism for T-lymphocytes with CD4 receptors[36,37], leaving most other cell types uninfected, and that viruses such as EBV or HTLV integrate at seemingly random sites and are difficult to detect unless a major clonal expansion takes place within the cell population[14–16,37]. With respect to the efficient use of computer resources, the computationally expensive alignment of all read-pairs to the human genome should ideally be performed just once, i.e. in the context of a classical BWA-based next generation sequencing (NGS) pipeline (Fig. 1). The resulting specific and efficient analysis pipeline would allow us to routinely scan all of our future genome, transcriptome, and targeted next generation sequencing data for integrations of known viruses before further wet lab tests or computationally more expensive analyses are performed on selected samples.

Material and Methods

Whole genome sequencing of leukemia samples. In 2012 the German Federal Office for Radiation Protection initiated a multi-center deep sequencing pilot study on childhood ALL (Supplementary consortium data online). We sequenced the whole genomes of paired tumour and control samples from ten pediatric patients according to the Illumina TruSeq paired-end sequencing protocol. The tumours had characteristic genomic rearrangements detected by PCR[38]. Samples were taken at diagnosis in different hospitals. The tumour cell content was between 60% and 95%. Each tumour sample was sequenced to a minimal coverage of 80× using an average of eight HiSeq 2000 lanes. As in nearly all current leukemia studies, the control samples were collected from the same patients when they were in remission, defined by a tumour cell content (MRD, minimal residual disease) less than 0.01%. Each control sample was sequenced to a minimal coverage of 40× using an average of four lanes.

The paired tumour and control samples were collected within the International BFM-Study Group (I-BFM-SG) from Austria, France, and Germany. The patients were enrolled by the AIEOP-BFM study group (Austria, Germany) and the FRALLE study group (France). Informed consent for the use of spare specimens for research was obtained from study individuals, parents or legal guardians. The research project reported here was approved by the Ethics Committee of the Medical Faculty, Christian-Albrechts-University, Kiel, Germany. All methods were carried out in accordance with the approved guidelines. The biological samples and sequence data are not consented for access outside of our consortium.

Whole genome sequencing protocol. Illumina TruSeq 2 × 101 bp paired end sequencing was performed on Illumina HiSeq 2000 and HiSeq 2500 instruments (Illumina, San Diego, CA). In brief, we used 1 μg of DNA per sample, which is fragmented by sonication to lengths of approximately between 250 and 650 bps before end-repair and ligation of Illumina sequencing adapters and molecular barcodes. The resulting libraries are each attached to random positions on a glass carrier plate (flowcell), and sequenced: Put simply, in the "first run", the positions of all clusters on the glass plate are optically identified, and the first end of each library is sequenced to a length of 101 bps. In the second run, the library is sequenced from the second end to a length of 101 bps. Importantly, the position of each library on the flowcell is optically identified by the instrument and in rare cases a library position may be identified differently in the second run, leading to a chimeric sequencing artefact consisting of two different libraries.

Bioinformatic availability and implementation. All Vy-PER scripts (**V**irus integration detection b**Y** **P**aired **E**nd **R**eads) are freely available at http://www.ikmb.uni-kiel.de/vy-per. The downloads include example scripts for a Linux cluster, human genome hg19 reference, NCBI virus genome references and benchmark data. The Vy-PER scripts are implemented in Python 2.7 with pysam v0.6, R, the Bioconductor[39] packages quantsmooth[40] and RColorBrewer, and bash scripts. Further prerequisites are BLAT[41], BWA[42], Phobos 3.3.12 (Christoph Mayer, http://www.rub.de/spezzoo/cm/cm_phobos.htm), and SAMtools[43]. For final filtering we used a Smith-Waterman implementation on the FPGA-based hardware RIVYERA S6-LX150 from SciEngines (www.sciengines.com), but for general users, Vy-PER also provides the option of using BLAT on a Linux cluster.

The bioinformatic workflow is summarised in Fig. 1 and described in detail in the Supplementary Methods online. In brief, the first step is the classical alignment of the human sequencing reads to the human genome prior to classical NGS variant calling, for example using BWA and SAMtools. This alignment is generally performed with stringent settings to keep run-time low and to map reads with high confidence for subsequent variant calling. However, a percentage of human reads will remain as "unmapped reads" that could not be aligned to the human reference. After this first step, the pipeline splits into two sub-pipelines that can be run in parallel if a computer cluster is available: (i) classical variant-calling, and (ii) virus integration detection. In the virus detection sub-pipeline, the first step is to extract paired-ends for which one read aligned to the human genome and the other end did not. The second step is to discard low-complexity reads with a long short tandem repeat or homopolymer. The third step is to test remaining reads for viral origin by alignment to known virus genomes. Some reads may map full-length or partially to a virus genome. In the fourth step, these candidate virus sequences are tested for low complexity, as a partial-length mapping result to a virus genome may consist only of a short tandem repeat or homopolymer (Table 1). In the final step, the remaining candidate sequences are tested for human origin using exact alignment to the human genome. This is computationally extremely expensive and can therefore usually be performed for only a few hundred sequences on a normal Linux computer cluster, hence the preceding multi-step filtering/testing process. Two sub-steps are used to speed up this exact alignment to the human genome: Each candidate is exactly aligned to its own small reference sequence window around the end of the read-pair that mapped to the human genome. Those candidates failing the first, local alignments are consecutively exactly aligned against the entire human genome. For exact alignment, BLAT is used with special settings (Supplementary Methods online), and for optional speed increase we use the Smith-Waterman implementation on FPGA hardware.

Bioinformatic pipeline comparison. We here also compare the virus integration detection by Vy-PER with several recently published pipelines using their default settings or recommendations: the VirusSeq pipeline[26], the VirusFinder pipeline v.2.0[35], and the ViralFusionSeq pipeline revision1289[44]. As the VirusFinder pipeline does not include virus reference sequences on its project homepage, we followed the authors' recommendations and downloaded the RINS reference sequences[45]. Out of interest, we also used VirusSeq's gibVirus reference sequences (http://gib-v.genes.nig.ac.jp/) for VirusFinder. The benchmark timings were obtained by running each pipeline on a single 16-core linux computer node with 120 GB RAM, except for Vy-PER which only required a single core and less than 10 GB RAM; all file input/output was performed via a 10 Gbit/s network to a central EMC Isilon X200 storage cluster. We included the Amazon EC2 cloud-based instance of SURPI[46] for detecting the presence of viruses without the ability to detect their integration. We tested SURPI v1.0.18 (July 10, 2014) using the EC2 Amazon Machine Image ami-a28f4cca on an i2.4xlarge instance in Amazon's US-East region, as described in the SURPI protocol.

Example data for the bioinformatic pipeline comparison. Download links to all four example data sets are available at http://www.ikmb.uni-kiel.de/vy-per.

NA12878V is a synthetic data set that we constructed by extracting 10,000 read-pairs from a 1000 Genomes Project[47] Illumina paired-end whole genome sequencing BAM file for individual NA12878[48] and adding 11,205 read-pairs from multiple individuals. The 11,205 read-pairs consist of 9,832 human/human pairs and 1,373 human/virus chimeras: The human/virus chimeras comprise 1,334 enterobacteria phage phiX174 chimeras, 2 enterobacteria phage M13 chimeras, and 37 human herpes virus type 3 (HHV-3) chimeras. The 9,832 human/human read-pairs include sequence stretches with similarities to virus sequence stretches, in order to test whether the benchmarked pipelines incorrectly detect herpes viruses (false positives). For patient confidentiality reasons, the original FASTQ sequence files and sequences in analysis or results files are removed from the publicly downloadable data. The paired-ends were aligned to hg19 using BWA and SAMtools. Then the Vy-PER pipeline was run as described.

L526401A liver cancer sample (RNA). As our patient data cannot be publicly released, we here demonstrate the detection of known HBV integrations into liver cancer cell genomes employing publicly available Illumina paired-end (2 × 50 bp) transcriptome data[26] with known HBV integrations. Before Vy-PER was run, the 53,060,622 paired-ends were aligned to hg19 using BWA and SAMtools, which required 3 h 2 minutes on a 16-core node. Then the Vy-PER pipeline was run as described.

198T and *268T* comprise publicly available "cleaned" Illumina paired-end (2 × 90 bp) whole genome tumour data with known HBV integrations, from two liver cancer patients in a study of 88 liver cancer patients[23]. The original raw sequences do not seem to be publicly available. We selected subsets of *198T* and *268T* sequencing data to test the sensitivity of our method and of other approaches at ultra low virus integration content. The subsets consist of 82,708,061 (198T) and 82,450,511 (268T) read-pairs, respectively. Before the Vy-PER run, alignment to hg19 was performed using BWA and SAMtools, which

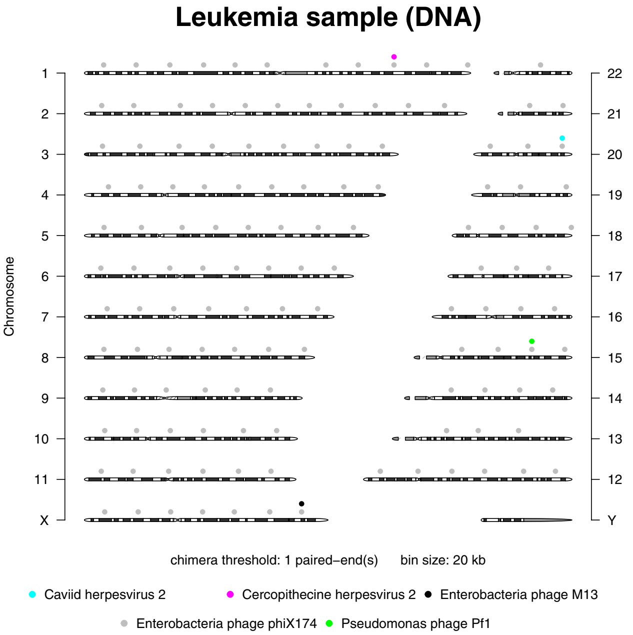

Figure 2. Vy-PER ideogram summary plot. Negative example without true virus integrations: Patient genome sequenced with 40× coverage and analysed with highest sensitivity, eliminating 6400 false positives and leaving three unsupported singletons (pseudomonas phage, caviid herpes, cercopithecine herpes) which we manually eliminated as alignment artefacts. PhiX and M13 originated from the 1% Illumina sequencing library control spike-in. The plot shows integration sites down to single chimeras.

required 3 hours 22 minutes (198T), and 3 hours 45 minutes (268T), respectively. Then the Vy-PER pipeline was run as described.

Results

No evidence of somatic virus integration in childhood acute lymphoblastic leukemia samples with genomic rearrangements. Using our Vy-PER method, we searched for virus integrations in 10 tumour samples that were sequenced to a minimal coverage of 80×, and in 10 matched normal samples from the same patients that were sequenced to a minimal coverage of 40×. Before final filtering (see Supplementary Methods online), we detected on average 1600 virus candidates per lane (12,800 per 80× genome). On manual inspection, most candidates appeared to be false positives. Table 1 shows a representative snapshot of false positive virus candidates with intriguingly plausible viruses, e.g. the common herpes viruses. However, their candidate sequences are non-specific STR-rich or homopolymer-rich sequences. It can be argued that some herpes types are known to integrate into human telomeres[49], but on the other hand, some herpes virus genomes simply have homologies to the human telomeres[50]. When we aligned a number of full-length HiSeq reads from herpes candidates to the expected human genome sequence window using an interactive DIALIGN-based tool[51], we found that these reads were in fact from the human genome. Our alignment exercise suggested that a more accurate and more sensitive alignment of the virus candidates to the host genome would be an effective filter to eliminate false positives. Therefore, we implemented the final filtering of candidates (Fig. 1). The STR-based filtering of virus candidate sequences and exact alignment of remaining sequences to the expected human genome window cut the number of false positive virus candidates to only 12 candidates per 80× genome, on average. The final exact alignment to the whole human genome eliminated most of these remaining few sequences. In the end, no true virus/human chimeras were detected (Fig. 2, Table 2) even at the lowest reporting threshold of 1 paired-end in 20 M bps. The only reported candidates were sparsely distributed phiX174 chimeras, a few singletons, and a cluster of three enterobacteria phage M13 chimeras in the remission sample of patient 8 but not in the tumour. The phiX chimeras are probably optical positioning errors of the HiSeq platform when the sequencing of the second read of a paired-end run is performed. PhiX libraries are technical controls for base calling on the Illumina platforms. They are spiked into the completed sample libraries as a fraction of about 1% of the total libraries. The phiX libraries should not normally ligate with sample libraries, as both libraries are blunt-ended. The fraction of phiX chimeras in the ALL samples comprised about 0.0001% of the total paired-ends, or about 160 chimeras per lane. Enterobacteria phage M13 is used for bacterial cloning, as in the commercial production of phiX. The M13 singletons and the small M13 cluster that we detected are a trace contaminant of the Illumina phiX174 libraries. This suggests that our STR-filtering and host-genome re-alignment strategy is very

Weighted candidates	Unambiguously aligned reads	Virus	ID
1655.3	1655	Enterobacteria phage phiX174	NC_001422.1
1.0	1	Pseudomonas phage Pf1	NC_001331.1
1.0	1	Caviid herpesvirus 2	NC_011587.1
1.0	1	Enterobacteria phage M13	NC_003287.2
1.0	1	Cercopithecine herpesvirus 2	NC_006560.1

Table 2. Top viruses in a leukemia patient sample after final filtering, low stringency (threshold: 1 chimera). Weighted candidates: For reads that align ambiguously to two or more viruses, we consider the top 3 viruses, assigning the highest weight to the first virus and the lowest weight (e.g. 0.3) to the last virus; the weight for an unambiguous alignment is 1.0.

	Leukemia	NA12878V	L526401A	198T	268T
Data type	WGS	WGS	RNA-Seq	WGS	WGS
HiSeq lanes	1.0	$\approx$ 0.0001	$\approx$ 0.33	$\approx$ 0.5	$\approx$ 0.5
Read length	2 × 101 bp	2 × mixed	2 × 50 bp	2 × 90 bp	2 × 90 bp
Read pairs	2.1×10^8	2×10^4	5.3×10^7	8.2×10^7	8.2×10^7
SURPI on EC2					
fast	–	14.8 h	–	–	–
comprehensive	–	24.5 h	–	–	–
EC2 cost	–	$ 250.00	–	–	–
ViralFusionSeq	19.1 h	3 mins	4 h	12.3 h	12.4 h
VirusFinder					
RINS virus db	14.5 h	2.7 h	9.8 h	21.5 h	15.4 h
gibVirus db	12.8 h	2.7 h	13.7 h	20.1 h	12.4 h
VirusSeq	195 h	17 mins	14.7 h	57.6 h	58.3 h
Vy-PER	**1 h**	**7 mins**	**21 mins**	**19 mins**	**19 mins**

Table 3. Wall clock run time comparison between different bioinformatic pipelines for five examples. WGS whole genome sequencing, RNA-Seq whole transcriptome sequencing, EC2 Amazon elastic cloud computing. Wall clock times obtained on a 16 core computer, except for Vy-PER which only needed a single core of a 16 core computer and $<$ 1 minute on the connected FPGA computer. The average number of WGS read pairs per HiSeq lane in our leukemia project was 1.65×10^8.

promising for the elimination of human sequence reads and the preservation of virus/host chimeras, even down to singletons.

Bioinformatic pipeline comparison. *Childhood acute lymphoblastic leukemia whole genome data from one HiSeq lane.* To compare run times between different pipelines (Table 3), we selected one HiSeq 2000 lane of data from one childhood ALL sample and ran each pipeline on an identical 16-core computer, except for Vy-PER which only requires a single core. One lane is approximately equivalent to a 10× covered genome. Analysis of the full genomic data was not feasible due to the computational duration and the unavailability of sufficient computing resources. After the classical BWA/SAMtools alignment to hg19, which required 13.5 hours, Vy-PER completed within 1 hour. VirusFinder required only 12.75 hours because it did not detect any viruses, but each detected virus would roughly add another 13 hours due to the renewed alignment of all reads to a combined hg19/virus reference sequence. ViralFusionSeq required 19 hours, the greater part of which is due to BWA-SW alignment which can be slower than BWA alignment. VirusSeq ran for 195 hours, mainly due to its use of the Mosaik aligner[52], and failed in its final step. We could not test the Amazon EC2 cloud based instance of SURPI on this data set, due to severe performance issues. With respect to their detection results, VirusFinder, ViralFusionSeq and VirusSeq did not report human/virus chimeras in the leukemia data set (Table 4), ignoring the M13 and phiX seen by Vy-PER. VirusSeq incorrectly reported non-integrated carp herpesvirus sequences (Table 5). VirusFinder with the recommended RINS database incorrectly reported human herpesvirus type 5 (HHV-5) which we analysed and found to be the harmless sequencing control library component M13. With the alternative gibVirus database, VirusFinder reported the non-integrated virus J02482M

Example (viruses)		ViralFusionSeq	VirusFinder (RINS)	VirusFinder (gibVirus)	VirusSeq	Vy-PER
Leukemia B2265L8 (phiX)	TP	–	–	–	–	phiX
	FP	–	–	–	–	–
NA12878V (HHV3, M13, phiX)	TP	–	–	–	–	HHV3, M13, phiX
	FP	–	–	–	–	CHV2 (singleton)
L52640A (HBV, phiX)	TP	HBV	–	–	HBV	HBV, phiX
	FP	–	–	–	–	–
198T (HBV)	TP	–	–	–	–	HBV
	FP	–	–	–	–	–
268T (HBV)	TP	HBV	–	–	–	HBV
	FP	–	–	–	–	–

Table 4. human/virus chimera detection comparison between different bioinformatic pipelines for five examples. RINS VirusFinder-recommended database, gibVirus alternative database, TP true positive, FP false positive. phiX enterobacteria phage phiX174, HHV3 human herpesvirus 3, M13 enterobacteria phage M13, CHV2 caviid herpesvirus 2, HBV hepatitis B virus.

Example (viruses)		SURPI on EC2	Viral Fusion Seq	VirusFinder (RINS)	VirusFinder (gibVirus)	VirusSeq	Vy-PER
Leukemia (phiX, M13)	TP	not run	–	phiX	J0 (phiX)	–	phiX
	FP	not run	–	HHV5	DE3, P7	carp herpesvirus	–
NA12878V (HHV3, M13, phiX)	TP	HHV3, phiX	–	HHV3, phiX	J0 (phiX), HHV3	–	HHV3, M13, phiX
	FP	α3, f1, G4, phiK	–	–	S13	–	CHV2 (singleton)
L52640A (HBV, phiX)	TP	not run	HBV	HBV, phiX	HBV, phiX	HBV	HBV, phiX
	FP	not run	–	–	–	–	–
198T (HBV)	TP	not run	–	HBV	HBV	–	HBV
	FP	not run	–	–	–	carp herpesvirus	–
268T (HBV)	TP	not run	HBV	HBV	HBV	–	HBV
	FP	not run	–	–	–	carp herpesvirus	–

Table 5. Virus detection comparison between different bioinformatic pipelines for five examples, regardless of whether virus integration into the host genome was detected. EC2 Amazon elastic cloud computing, RINS VirusFinder-recommended database, gibVirus alternative database, TP true positive, FP false positive. phiX enterobacteria phage phiX174, M13 enterobacteria phage M13, J0 J02482M10348M 10379M10714M10749M10750M10866, HHV5 human herpesvirus 5, DE3 enterobacteria phage DE3, P7 enterobacteria phage P7, HHV3 human herpesvirus 3, α3 enterobacteria phage alpha3, f1 enterobacteria phage f1, G4 enterobacteria phage G4, phiK enterobacteria phage phiK, S13 bacteriophage S13, CHV2 caviid herpesvirus 2, HBV hepatitis B virus.

10348M10379M10714M10749M10750M10866 (Table 5) and its *de-novo* assembled contig which our subsequent manual BLAST search identified as phiX174.

NA12878V synthetic genome data. The known human/virus chimeras were reported by Vy-PER but no other pipelines: Vy-PER reported 24 × human herpesvirus 3 (HHV-3), 1185 × phiX, and 2 ×

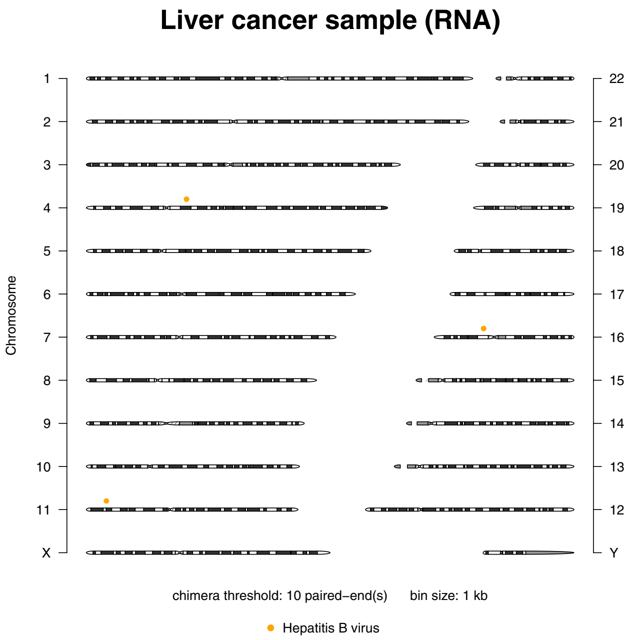

Figure 3. Vy-PER ideogram summary plot. Positive example with true virus integrations: Publicly available whole transcriptome liver cancer data analysed with default sensitivity, showing HBV candidate loci on chromosomes 4, 11 and 16. The plot only shows integrations supported by 10 or more chimeras.

Chr	Start	End	Candidates	Virus
4	63647816	63648816	16.0	Hepatitis B virus
4	63651319	63652319	10.0	Hepatitis B Virus
11	12711328	12712328	71.0	Hepatitis B Virus
16	31413359	31414359	44.0	Hepatitis B Virus
16	31414755	31415755	18.0	Hepatitis B Virus
16	31418770	31419770	10.0	Hepatitis B Virus

Table 6. Virus candidate loci in liver cancer sample after final filtering, high stringency (threshold: 10 supporting paired-ends)

M13 chimeras where the human reads were aligned with a mapping quality of at least 20. HHV-3 and phiX were reported as non-integrated viruses by SURPI and VirusFinder, but not by ViralFusionSeq or VirusSeq (Table 5). The known M13 phage was detected by no other pipeline. The herpes-like decoy sequences in this example data did not trigger any false positive calls by any pipeline except for Vy-PER, which reported a single caviid herpesvirus chimera. SURPI and VirusFinder (gibVirus database) reported some false positive non-integrated enterobacteria phages. It should be noted that SURPI removes low-complexity reads as the first step, possibly over-aggressively, as true reads are removed which can be aligned by BWA (Supplementary Results online).

L526401A liver cancer sample (RNA). The known HBV integration loci were reported by Chen and colleagues[26] and we reproduced these loci using their VirusSeq pipeline (Supplementary Table S1 online). All reported HBV loci were also detected by Vy-PER. Figure 3 displays the virus candidate loci in the genome that were detected by Vy-PER with the stringent default setting of at least ten supporting chimeras, i.e. at least ten virus/host paired-ends are required to support a virus integration locus, and Table 6 lists the virus candidate loci in 1000 bp bins. Note that there are three integration loci on chromosome 16, one more than reported by VirusSeq. However, there is no split read to support this last locus, only the respective paired read that was aligned to hg19. In transcriptome data, a paired-end library may conceivably span two or more exons, which could lead to an additional candidate locus that is 4000 bp distant from the true integration locus. Table 7 shows the number of virus candidates reported by Vy-PER if singletons are enabled, and Fig. 4 shows the corresponding candidate loci. The VirusSeq pipeline includes gene annotation of the integration loci. However, the annotation is occasionally misleading, here for the integration locus chr4:63647816-63648816 which is 1.5 megabases distant from the 3' end of the *TECRL* gene. This locus is nevertheless annotated as "*TECRL*/3-prime", which the VirusSeq authors have also

Weighted candidates	Unambiguously aligned reads	Virus	ID
181.0	181.0	Hepatitis B virus	NC_003977.1
7.0	7.0	Enterobacteria phage phiX174	NC_001422.1

Table 7. Top viruses in liver cancer (RNA-Seq) after final filtering, low stringency (threshold: 1 chimera).

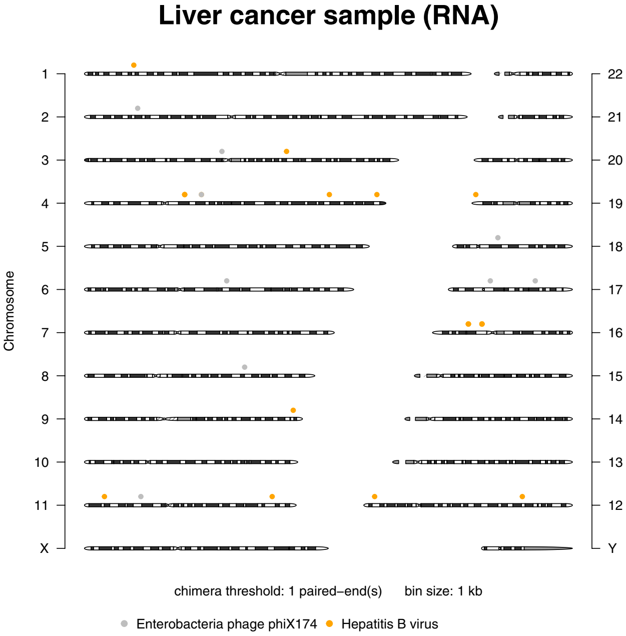

Figure 4. Vy-PER ideogram summary plot. HBV integration loci into the liver cancer genome detected at low stringency (threshold: 1 chimera), also showing detected phiX singletons.

copied into their publication[26]. The actual locus that VirusSeq computed is correct, but it is located in a gene desert. The nearest gene, *LPHN3*, is less than half the distance to *TECRL*. The ViralFusionSeq pipeline reported 10 clipped fusion sequences of which most are duplicates, but did not report their integration loci (Supplementary Table S2 online). Confusingly, the 100 bp fusion sequences reported by ViralFusionSeq are a tandem repeat of identical 50 bp reads. A BLAT search for the integration locus did not place any of these fusion reads correctly, due to their short length of only 50 nucleotides and the unavailable paired-end information. Only one such read aligned to the correct locus on chromosome 11, but the alternative locus on chromosome 20 was ranked higher in the BLAT search. The VirusFinder pipeline did not report any virus integrations at all (Table 4 and Supplementary Table S1 online). Finally, no other pipeline reported the phiX chimeras seen by Vy-PER, but VirusFinder reported non-integrated phiX sequences.

198T liver cancer genome data. The liver cancer tissue sample harbours seven known human/HBV breakpoints[23]. We tested the sensitivity of all pipelines using a low-coverage subset of the whole genome data (Table 3). The only pipeline that was able to detect HBV in this data subset was Vy-PER. The virus integration event is detected by merely a single human/HBV chimera, but its locus is nevertheless validated by the known data[23]. There are no phiX chimeras in this data set, because the publicly deposited data has been "cleaned". Looking at false positives, VirusSeq again incorrectly reported a non-integrated carp herpesvirus (Table 5).

268T liver cancer genome data. The liver cancer tissue sample harbours four known human/HBV breakpoints[23]. Only Vy-PER and ViralFusionSeq detected HBV integration in the low-coverage data subset of the full whole genome data. Vy-PER reported a single human/HBV chimera with a confidently mapped human read and a split human/virus read, which is validated by the known data[23]. ViralFusionSeq reported seven human/HBV fusion reads but did not resolve the coordinates (Supplementary Table S2 online). Manual BLAT and BLAST searches mapped four fusion reads to the same locus that was detected by Vy-PER, and the remaining three fusion reads to two further known loci (Supplementary Results online). We manually identified one HBV insertion to have microhomologies of four shared nucleotides between virus and host at the first end of the insertion (detected by Vy-PER), and of five shared nucleotides at the other end (detected by our manual searches for the fusion reads reported by ViralFusionSeq) (Supplementary Results online). The third insertion locus even showed a microhomology of 13 nucleotides. With respect to false positives, VirusSeq again incorrectly reported a non-integrated carp herpesvirus (Table 5).

Discussion

Virus integrations into the host genome of cancer tissue can be detected using paired-end whole genome or transcriptome sequencing data. However, the low number of expected true virus integrations may be difficult to distinguish from the noise of many false positive candidates if aiming to increase sensitivity. For example, some viruses target only a fraction of cell types, such as the human herpesvirus-6 (HHV-6) which has a tropism for CD4+ T-lymphocytes[53]. We therefore analysed false positive virus integration candidates and developed accurate filtering methods and efficient Python and R scripts that can be integrated into any BWA-based next-generation sequencing analysis pipeline. The key strategies to eliminate false positives are to (a) remove reads consisting mainly of unspecific STRs or homopolymers, (b) remove virus candidates with sequences consisting mainly of unspecific STRs or homopolymers, (c) make sure that virus candidates do not align to the host genome, using highly accurate and sensitive alignment tools such as BLAT with its most sensitive setting or optionally Smith-Waterman, and (d) test for supporting candidates, as singletons may be artefacts. Finally, phiX and M13 are Illumina's sequencing platform control spike-ins and not true virus infections. The optional FPGA-based Smith-Waterman alignments to the complete human genome are useful if FPGA hardware is available. Even without this optional stage, only quite a small number of false positives remain after using BLAT to align virus candidates to a reduced human genome sequence window of e.g. length 500 bps. In a clinical setting, it is not a practical option to use BLAT with its most sensitive settings to align candidates to the entire human genome, as this may take hours or days, but for pure research the run time is less important. In this context it is worth noting that BLAST uses DUST-based masking and does not align some human reads (Supplementary Results online), making BLAT or Smith-Waterman more suitable for this specific task.

The results shown here demonstrate that Vy-PER is sensitive, specific and efficient enough to be used as a first pipeline for routinely scanning all NGS data sets for virus integrations. The efficiency of Vy-PER is explained by its filtering design for speedy processing of genomic data that can be furthermore parallelised for multi-lane genomes (map-reduce approach), and – when integrated into a BWA based pipeline – by avoiding the duplication of any expensive whole genome alignment computations. Once viruses of concern have been found in a sample, a more detailed analysis with ViralFusionSeq, VirusFinder and VirusSeq can then be performed. These latter pipelines are still in their early days and fairly complicated to install and understand, but each has its own merits. All of them also perform a search for viruses that did not integrate into the host genome. VirusSeq annotates the integration loci with gene information and computes breakpoints. VirusFinder is supposed to do the same, but our tested version (version 2) did not compute any integrations. ViralFusionSeq reports only the human/virus fusion reads but not the integration locus. Whereas Vy-PER was designed with concise clinical reporting in mind, the other pipelines are more detailed and do not provide a summary graphic or table. Vy-PER uses the NCBI virus genomes which are sufficient for uniquely pinpointing a virus at species level, and the summary tables show whether some or all virus reads were unambiguously assigned to a single virus species (Table 2). Similarly, for example, TIPP (Taxon Identification and Phylogenetic Profiling, http://www.cs.utexas.edu/users/phylo/research/projects.html) dumbs down its level of detail as necessary, from strain to species or even to higher levels in the phylogeny, when ambiguous reads are encountered. However, the tested other pipelines currently still report viral strain information, which can give users a false sense of security, especially as there are vast homologies between the strains of a species, and we are unsure how correct these virus databases actually are. For example, the less experienced user may well be excited by the carp herpesvirus incorrectly reported in three out of five example data sets by one pipeline, or the human herpesvirus type 5 (instead of M13) incorrectly reported by another pipeline. No doubt each pipeline was developed for a specific type of virus and cancer data set, so it is probably worthwhile for a researcher to run a suspicious data set, i.e. one in which Vy-PER detected a virus, through each of these pipelines and to keep the potential problems in mind. It is for example remarkable that ViralFusionSeq detected viruses in only two of the five example data sets, but nevertheless detected three of the four fusion sequences in the low-coverage data subset of tumour sample 268T. Vy-PER only detected one of these fusions to be a virus integration event into the human host, because it searches only for chimeric paired-ends and not also for fusion sequences. Nevertheless, the scan for chimeras by Vy-PER was more sensitive than the approaches in other pipelines, correctly detecting the known chimeras in all five examples.

Surprisingly, Vy-PER is also able to detect non-integrated viruses. Taking the example of phiX (1% Illumina sequencing control library spike-in) and M13 (trace contamination of phiX library), the known presence of these non-integrated viruses is detected as chimeric virus/human read-pair artefacts. In our whole genome leukemia sequencing study, on average 160 phiX/human chimeras were detected per HiSeq lane (0.0001% of read-pairs). The chimeric sequencing artefacts can be interpreted as randomly sampled non-integrated virus sequences. Indeed, on the related topic of bacterial identification from shotgun sequencing, it was recently suggested[54] that only 100 randomly sampled 100 bp Illumina reads from pure bacterial DNA could be sufficient to identify the bacterial strain with over 90% specificity. For detecting non-integrated pathogens, the sensitivity is better if all paired-end reads that could not be aligned to the human genome are scanned. The scan can be followed by our approach to eliminate false positives, which is not limited to chimeras or to viruses.

As a matter of speculation, future studies may find that the concept of a human-virus breakpoint at a single base pair is not appropriate for all viruses. On the one hand it is known that retroviruses integrate at a distinct breakpoint when an integrase binds to its preferred nucleotide motifs[55] which are then duplicated and flank the integrated viral sequence[56]. However, other viruses, specifically HBV, may integrate using the non homologous end joining (NHEJ) pathway[57,58] which implies a shared microhomology between virus and host at the integration site, as observed in our manual inspections of the integration sites in data set 268T. The NHEJ mechanism suggests that the HBV integration into the human genome may not be the driver event for genomic instability, but that the HBV integration may have occurred as a passenger event after double-stranded breakage of the human chromosome.

In the treatment of HIV infections, cART may be unable to completely eliminate HIV if replication competent proviruses are integrated into long lived or proliferating CD4+ cells. Of particular interest are the integration loci which are preferentially located in genes that promote cell proliferation[6,7]. Using Vy-PER, the integration loci in clonally expanded cell populations would be detectable either using normal paired-end shotgun sequencing or even better, using targeted sequencing like linker-mediated polymerase chain reaction[6] or integration site loop amplification[7]. To speed up the bioinformatic analysis, the full virus genome reference could be replaced by just the HIV genome references.

Finally, the wealth of incidental metagenomic findings in human NGS data is highly exciting for potential pathogen detection, even when only microscopically low read numbers are available that our approach can detect and assign to a virus genome.

References

1. Bertino, P. *et al.* Fowlpox-based survivin vaccination for malignant mesothelioma therapy. *Int. J. Cancer* **133,** 612–623 (2013).
2. Baum, C., Schambach, A., Bohne, J. & Galla, M. Retrovirus vectors: toward the plentivirus? *Mol. Ther.* **13,** 1050–1063 (2006).
3. Soliman, H. Developing an effective breast cancer vaccine. *Cancer Control* **17,** 183–190 (2010).
4. Bertino, P. *et al.* Vaccination with a piggyBac plasmid with transgene integration potential leads to sustained antigen expression and CD8(+) T cell responses. *Vaccine* **32,** 1670–1677 (2014).
5. Watanabe, T. Current status of HTLV-1 infection. *Int. J. Hematol.* **94,** 430–434 (2011).
6. Maldarelli, F. *et al.* HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345,** 179–783 (2014).
7. Wagner, T. A. *et al.* HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345,** 570–573 (2014).
8. Kurdgelashvili, G. *et al.* Incidence of potentially human papillomavirus-related neoplasms in the United States, 1978 to 2007. *Cancer* **119,** 2291–2299 (2013).
9. Bishop, J. A. *et al.* Human papillomavirus-related carcinomas of the sinonasal tract. *Am. J. Surg. Pathol.* **37,** 185–192 (2013).
10. Hillbertz, N. S., Hirsch, J.-M., Jalouli, J., Jalouli, M. M. & Sand, L. Viral and molecular aspects of oral cancer. *Anticancer Res.* **32,** 4201–4212 (2012).
11. Tang, K.-W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M. & Larsson, E. The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat. Commun.* **4,** 1–9 (2013).
12. Miller, G. Immortalization of human lymphocytes by Epstein-Barr virus. *Yale J. Biol. Med.* **55,** 305–310 (1982).
13. Tosato, G. *et al.* Epstein-Barr virus immortalization of normal cells of B cell lineage with nonproductive, rearranged immunoglobulin genes. *J. Immunol.* **137,** 2037–2042 (1986).
14. Takakuwa, T. *et al.* Integration of Epstein-Barr virus into chromosome 6q15 of Burkitt lymphoma cell line (Raji) induces loss of BACH2 expression. *Am. J. Pathol.* **164,** 967–974 (2004).
15. Reisinger, J., Rumpler, S., Lion, T. & Ambros, P. F. Visualization of episomal and integrated Epstein-Barr virus DNA by fiber fluorescence *in situ* hybridization. *Int. J. Cancer* **118,** 1603–1608 (2006).
16. Morissette, G. & Flamand, L. Herpesviruses and chromosomal integration. *J. Virol.* **84,** 12100–12109 (2010).
17. Wang, X. W. *et al.* Hepatitis B virus X protein inhibits p53 sequence-specific DNA binding, transcriptional activity, and association with transcription factor ERCC3. *PNAS* **91,** 2230–2234 (1994).
18. Lechner, M. S. *et al.* Human papillomavirus E6 proteins bind p53 *in vivo* and abrogate p53-mediated repression of transcription. *EMBO J.* **11,** 3045–3052 (1992).
19. Boutell, C. & Everett, R. D. The herpes simplex virus type 1 (HSV-1) regulatory protein ICP0 interacts with and Ubiquitinates p53. *J. Biol. Chem.* **278,** 36596–36602 (2003).
20. Cruz, C. D. *et al.* Measles virus V protein inhibits p53 family member p73. *J. Virol.* **80,** 5644–5650 (2006).
21. Dobbelstein, M. & Roth, J. The large T antigen of simian virus 40 binds and inactivates p53 but not p73. *J. Gen. Virol.* **79,** 3079–3083 (1998).
22. Ferber, M. J. *et al.* Integrations of the hepatitis B virus (HBV) and human papillomavirus (HPV) into the human telomerase reverse transcriptase (hTERT) gene in liver and cervical cancers. *Oncogene* **22,** 3813–3820 (2003).
23. Sung, W.-K. *et al.* Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat. Genet.* **44,** 765–769 (2012).
24. Jiang, Z. *et al.* The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome Res.* **22,** 593–601 (2012).
25. Zhang, Z. Genomic landscape of liver cancer. *Nat. Genet.* **44,** 1075–1077 (2012).
26. Chen, Y. *et al.* VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics* **29,** 266–267 (2013).
27. Pui, C.-H., Robison, L. L. & Look, A. T. Acute lymphoblastic leukaemia. *Lancet* **371,** 1030–1043 (2008).
28. Pui, C.-H., Carroll, W. L., Meshinchi, S. & Arceci, R. J. Biology, risk stratification, and therapy of pediatric acute leukemias: an update. *J. Clin. Oncol.* **29,** 551–565 (2011).
29. Ziegelberger, G. *et al.* Childhood leukemia–risk factors and the need for an interdisciplinary research agenda. *Prog. Biophys. Mol. Biol.* **107,** 312–314 (2011).
30. Eden, T. Aetiology of childhood leukaemia. *Cancer Treat. Rev.* **36,** 286–297 (2010).
31. Pearce, M. S. *et al.* Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. *Lancet* **380,** 499–505 (2012).
32. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* **13,** 1028–1040 (2006).

33. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
34. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27,** 863–864 (2011).
35. Wang, Q., Jia, P. & Zhao, Z. VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PLoS One* **8,** e64465 (2013).
36. Goff, S. P. Host factors exploited by retroviruses. *Nat. Rev. Microbiol.* **5,** 253–263 (2007).
37. Cook, L. B., Rowan, A. G., Melamed, A., Taylor, G. P. & Bangham, C. R. M. HTLV-1-infected T cells contain a single integrated provirus in natural infection. *Blood* **120,** 3488–3490 (2012).
38. Van der Velden, V. H. J. *et al.* Detection of minimal residual disease in hematologic malignancies by real-time quantitative PCR: principles, approaches, and laboratory aspects. *Leukemia* **17,** 1013–1034 (2003).
39. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).
40. Eilers, P. H. C. & de Menezes, R. X. Quantile smoothing of array CGH data. *Bioinformatics* **21,** 1146–1153 (2005).
41. Kent, W. J. BLAT---The BLAST-Like Alignment Tool. *Genome Res.* **12,** 656–664 (2002).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
43. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
44. Li, J.-W. *et al.* ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. *Bioinformatics* **29,** 649–651 (2013).
45. Bhaduri, A., Qu, K., Lee, C. S., Ungewickell, A. & Khavari, P. A. Rapid identification of non-human sequences in high-throughput sequencing datasets. *Bioinformatics* **28,** 1174–1175 (2012).
46. Naccache, S. N. *et al.* A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res.* **24,** 1180–1192 (2014).
47. Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
48. Forster, M. *et al.* From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software. *Nucleic Acids Res.* **41,** e16 (2013).
49. Arbuckle, J. H. *et al.* The latent human herpesvirus-6A genome specifically integrates in telomeres of human chromosomes *in vivo* and *in vitro*. *PNAS* **107,** 5563–5568 (2010).
50. Gompels, U. A. *et al.* The DNA sequence of human herpesvirus-6: structure, coding content, and genome evolution. *Virology* **209,** 29–51 (1995).
51. Morgenstern, B., Frech, K., Dress, a & Werner, T. DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics* **14,** 290–294 (1998).
52. Lee, W.-P. *et al.* MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. *PLoS One* **9,** e90581 (2014).
53. Takahashi, K. *et al.* Predominant CD4 T-lymphocyte tropism of human herpesvirus 6-related virus. *J. Virol.* **63,** 3161–3163 (1989).
54. Gautier, L. & Lund, O. Low-bandwidth and non-compute intensive remote identification of microbes from raw sequencing reads. *PLoS One* **8,** e83784 (2013).
55. Serrao, E. *et al.* Integrase residues that determine nucleotide preferences at sites of HIV-1 integration: implications for the mechanism of target DNA binding. *Nucleic Acids Res.* **42,** 5164–5176 (2014).
56. Krishnan, L. & Engelman, A. Retroviral integrase proteins and HIV-1 DNA integration. *J. Biol. Chem.* **287,** 40858–40866 (2012).
57. Bill, C. A. & Summers, J. Genomic DNA double-strand breaks are targets for hepadnaviral DNA integration. *PNAS* **101,** 11135–11140 (2004).
58. Nguyen, D. H., Ludgate, L. & Hu, J. Hepatitis B virus-cell interactions and pathogenesis. *J. Cell. Physiol.* **216,** 289–294 (2008).

Acknowledgments

Eike Zell administrated the childhood acute lymphoblastic leukemia sequencing consortium. Markus Schilhabel, Melanie Friskovec, Catharina von der Lancken, and Melanie Schlapkohl performed whole genome next-generation sequencing. Dr. Eva Ellinghaus helped to organise the project and edited the manuscript. Jan Christian Kässens provided informatics support to FPGA users. Dr. Adam Grundhoff advised on the differences between retroviral and HBV integration. We gratefully acknowledge Vy-PER presentation feedback at the HiTSeq 2014 conference. This work was supported by the German Federal Office for Radiation Protection (BfS); the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (grant number 01ZX1306A, sysINFLAME); the Deutsche Forschungsgemeinschaft (DFG) Clusters of Excellence 'Inflammation at Interfaces'; and the EU Seventh Framework Programme [FP7/2007-2013, grant number 262055, ESGI].

Author Contributions

M.S. and A.F. designed and coordinated the study; M.S. and members of the UFO Sequencing Consortium within the I-BFM Study Group were involved in leukemia patient recruitment and provided the leukemia samples; M.F., S.S. and L.K. wrote the Vy-PER pipeline; L.W. and D.E. designed, coordinated and supervised the code development on the FPGA hardware; M.F., S.S., G.H., M.R., S.M. and L.W. performed bioinformatic analyses and interpreted the data; M.F., S.S., D.E., G.H., M.R., S.M. and A.F. wrote the manuscript; M.F., S.S. and G.H. created Fig. 1, S.S. created Figs 2–4; M.F. created tables 1,2,6,7, M.F. and M.R. created tables 3 and S1, M.F., S.S. and M.R. created Tables 4,5, M.R. created Table S2. Consortia: UFO Sequencing Consortium within the I-BFM Study Group. Department of Pediatric Oncology, Hematology and Clinical Immunology, Heinrich Heine University, Düsseldorf, Germany: Arndt Borkhardt, Ute Fischer; Department of Oncology, University Children's Hospital Zurich, Zurich, Switzerland: Jean-Pierre Bourquin, Anna Rinaldi; Pediatric Hematology and Oncology, Charité University Hospital, Berlin, Germany: Cornelia Eckert; Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Germany: Michael Forster, Andre Franke; Federal Office for Radiation Protection, Oberschleißheim, Germany: Sabine Hornhardt; European Molecular Biology Laboratory, Genome

Biology, Heidelberg, Germany: Jan O. Korbel, Stephanie Sungalee; Max Planck Institute for Molecular Genetics, Berlin, Germany: Hans Lehrach, Thomas Risch, Hans-Jörg Warnatz, Marie-Laure Yaspo; Pediatric Hematology and Oncology, Hannover Medical School, Hannover, Germany: Martin Stanulla.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Forster, M. *et al.* Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data. *Sci. Rep.* **5**, 11534; doi: 10.1038/srep11534 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/