Method

Using reference-free compressed data structures to analyze sequencing reads from thousands of human genomes

Dirk D. Dolle,[1,6] Zhicheng Liu,[1,2,6] Matthew Cotten,[1] Jared T. Simpson,[3,4] Zamin Iqbal,[5] Richard Durbin,[1] Shane A. McCarthy,[1] and Thomas M. Keane[1,2]

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom; [2]European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, United Kingdom; [3]Ontario Institute for Cancer Research, Toronto, Ontario M5G 0A3, Canada; [4]Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3G4, Canada; [5]Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, United Kingdom

We are rapidly approaching the point where we have sequenced millions of human genomes. There is a pressing need for new data structures to store raw sequencing data and efficient algorithms for population scale analysis. Current reference-based data formats do not fully exploit the redundancy in population sequencing nor take advantage of shared genetic variation. In recent years, the Burrows–Wheeler transform (BWT) and FM-index have been widely employed as a full-text searchable index for read alignment and de novo assembly. We introduce the concept of a population BWT and use it to store and index the sequencing reads of 2705 samples from the 1000 Genomes Project. A key feature is that, as more genomes are added, identical read sequences are increasingly observed, and compression becomes more efficient. We assess the support in the 1000 Genomes read data for every base position of two human reference assembly versions, identifying that 3.2 Mbp with population support was lost in the transition from GRCh37 with 13.7 Mbp added to GRCh38. We show that the vast majority of variant alleles can be uniquely described by overlapping 31-mers and show how rapid and accurate SNP and indel genotyping can be carried out across the genomes in the population BWT. We use the population BWT to carry out nonreference queries to search for the presence of all known viral genomes and discover human T-lymphotropic virus 1 integrations in six samples in a recognized epidemiological distribution.

[Supplemental material is available for this article.]

Recent years have seen the number of whole human genomes sequenced continue to increase dramatically through large-scale population and medical sequencing projects such as the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015), UK10K (The UK10K Consortium 2015), and GoNL (The Genome of the Netherlands Consortium 2014). The scale-up of human population sequencing has enabled us to detect sequence variants down to extremely low minor allele frequencies (The 1000 Genomes Project Consortium 2015), explore variation in ancient human lineages and isolated populations (Raghavan et al. 2015), and use genomics to discover rare disease-causing mutations (Katsanis and Katsanis 2013). Current predictions estimate that we will have sequenced 1M human genomes in the near future (Stephens et al. 2015), which will present formidable informatics scaling challenges.

The sequencing data produced by current high-throughput sequencing technologies consists of paired reads on the order of 100 bp, along with their base qualities, with the vast majority of aligned data currently stored in the SAM/BAM format (Li et al. 2009). The SAM/BAM format, originally developed by the 1000 Genomes Project (1000GP), requires on the order of one byte per base pair, with the vast majority of the space being taken by the base qualities (Hsi-Yang Fritz et al. 2011). Recently, the CRAM format has been proposed (Hsi-Yang Fritz et al. 2011) and adopted by the Global Alliance for Genomics and Health consortium (https://genomicsandhealth.org/) to provide a more sustainable foundation for exploring strategies for sequencing read data compression, such as controlled loss of base qualities, a strategy that can result in more accurate genotyping (Yu et al. 2015; Ochoa et al. 2016). One key innovation of the CRAM format is to only store the differences in individual sequencing reads relative to the reference genome. Furthermore, when one considers that the vast majority of variants per individual are shared among multiple individuals (The 1000 Genomes Project Consortium 2015), there is also significant duplication of nonreference sequences.

In parallel, there is increasing interest in methods for rapid searching of large collections of sequencing reads from many individuals. Iqbal et al. (2012) developed the Cortex assembler for representing sequencing reads from multiple samples using colored de Bruijn graphs for genome assembly and reference-free variant identification (Iqbal et al. 2012). Applications that were presented include variant calling from a single high-coverage genome,

[6]**These authors contributed equally to this work.**
Corresponding authors: rd@sanger.ac.uk, sm15@sanger.ac.uk, tk2@ebi.ac.uk
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.211748.116.

27:300–309 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/17; www.genome.org

detection of novel sequence from a population not present in the reference, and genotyping of simple and complex variants highly divergent from the reference. However, the implementation only scaled to around 10 human genomes on standard hardware, orders of magnitude lower than what is required. Recently, Bloom filters in the form of sequence bloom trees (SBTs) were used to build highly compressed partial text indexes given a large set of input sequences and demonstrate rapid sequence searches with low memory requirements (Bloom 1970; Solomon and Kingsford 2016). An SBT structure was constructed from 2652 RNA-seq experiments, requiring 200 GB. In recent years, the Burrows–Wheeler transform (BWT) and FM-index have been widely employed to build full-text indexes for read alignment (Langmead et al. 2009; Li and Durbin 2009), read-error correction (Li 2015), and de novo genome assembly (Simpson and Durbin 2011). The key features of using a BWT structure to index sequencing reads are that it is inherently reference-free, full-text, compressed, and coupled with the FM-index, enabling rapid sequence searches of arbitrary *k*-mer sizes across the entire set of sequences without rebuilding the index for different values of *k*.

Results

Data processing and BWT construction

Figure 1 gives an overview of the data processing strategy. We begin with the whole-genome low coverage and exome sequencing reads from the final phase of the 1000GP (2705 individuals over 26 populations) consisting of ~87 Tbp and 922 billion reads (The 1000 Genomes Project Consortium 2015). We used a combination of examining the base qualities for each read and querying the sequences against a preconstructed Cortex graph (Iqbal et al. 2012) to carry out error correction and removal of poor quality reads (see Methods). This resulted in a set of 734 billion unchanged reads, 85 billion corrected reads, and 103 billion reads that could not be corrected and were discarded. We took advantage of the reference strand labeling in the Cortex de Bruijn graph (obtained by labeling nodes during a traversal of the reference sequence) to reverse complement read sequences with a clear reverse strand orientation with respect to the reference genome (see Methods). We normalized the read lengths so that we could identify completely identical read sequences and only store one copy of the sequence in the BWT by trimming to 73 or 100 bp, depending on whether the original read sequence was >100 bp. For each resulting read, we used a key-value pair database (RocksDB, http://rocksdb.org/) to record the read groups, number of corrected bases, and number of bases >Q20 using the read sequence as the key. We next sorted the 53 billion sequence keys reverse lexicographically and constructed the BWT structure. The 53 billion unique sequences (keys) produced an average of 15.45 reads for each key. In Figure 2, we benchmarked the total size of the BWT using both the uncorrected and corrected reads with increasing numbers of individuals (using the reads from a 5-Mbp region on Chr 20). The plot shows that using the uncorrected reads, the BWT continues to linearly increase in size, independent of the sort order. For the error-corrected reads, BWTs produced from reverse lexicographic sorting order (RLO) were an order of magnitude smaller in size than lexicographic order (LO). The BWT of a collection of strings is the series of characters preceding a suffix in the lexicographically ordered set of all possible suffixes extracted from these strings. Arranging the strings in the collection in reverse lexicographical order prior to BWT construction assures that identical characters preceding the same suffix are grouped together and hence can be better compressed by methods like run length encoding (Cox et al. 2012). The effect of error correction of the reads can be observed with the total BWT around two orders of magnitude larger with uncorrected reads. With error correction and RLO sorting, the total size of the BWT begins to plateau from ~1500 to 2000

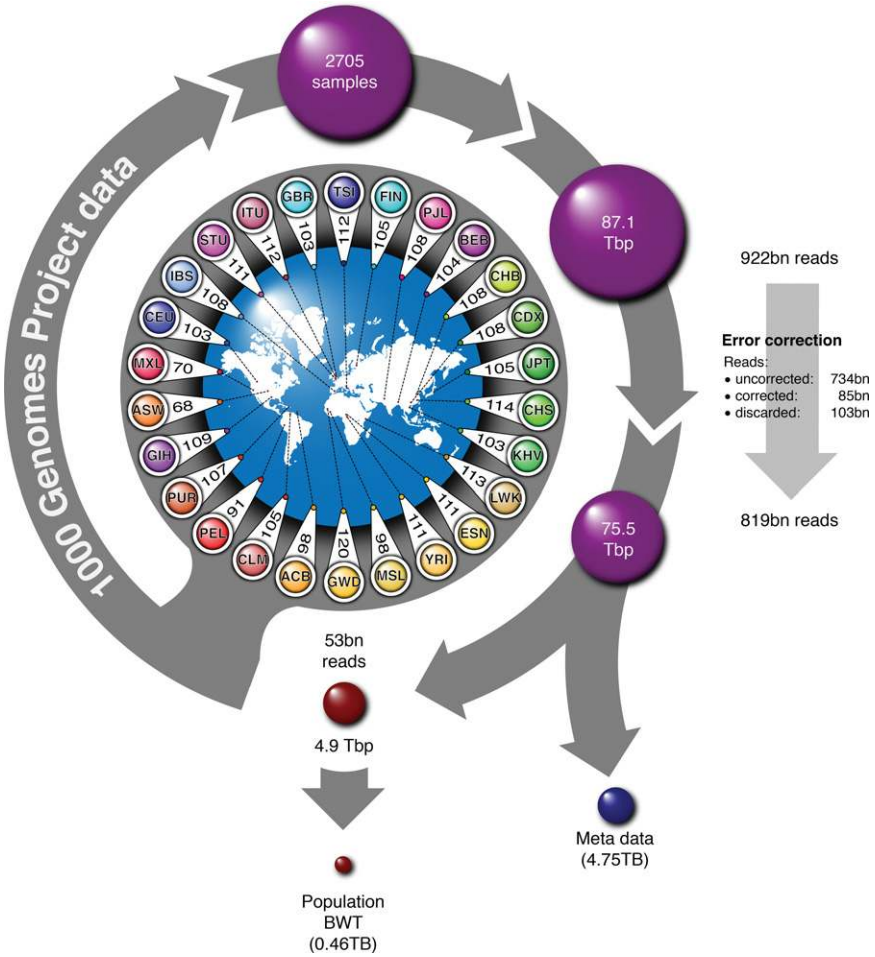


Figure 1. Sequencing reads from 2705 individuals (low-coverage whole-genome and exome sequencing) from 26 populations comprising a total of 922 billion reads (87.1 Tbp) used for the 1000GP population BWT. Reads were first error-corrected using a Cortex graph (Iqbal et al. 2012). The error-corrected reads were then trimmed to either 100 or 73 bp, unique sequences identified on the forward strand, quality values discarded, and the metadata stored in a separate database. This resulted in 4.9 Tbp consisting of 53 billion nonredundant reads.

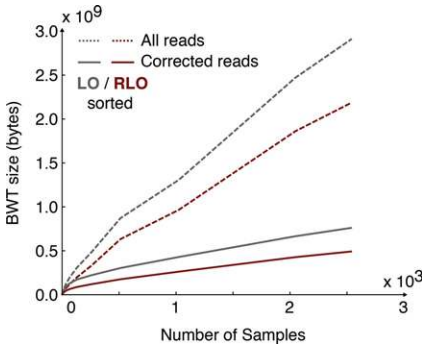
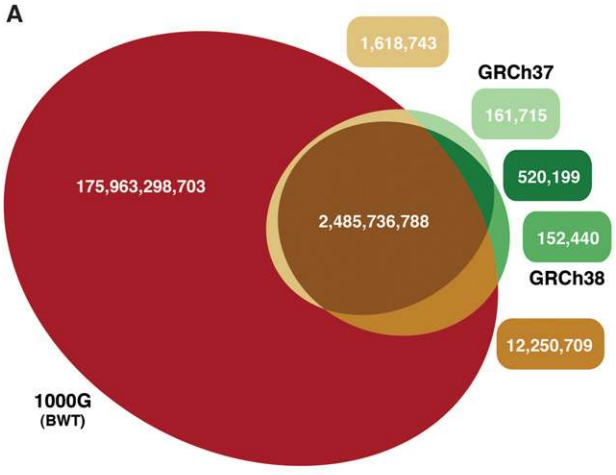


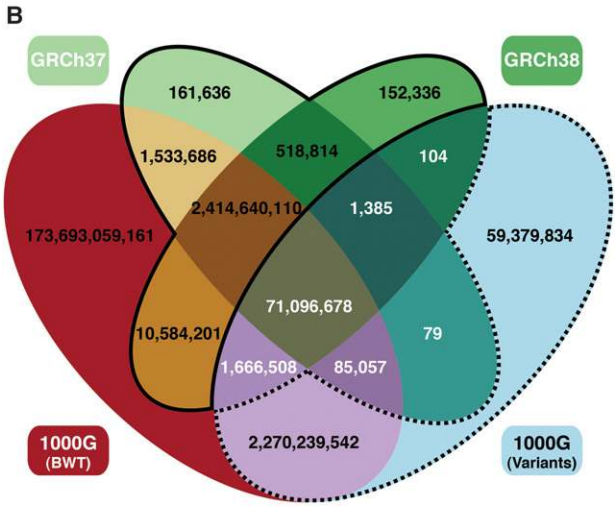
Figure 2. Sequences were sorted by reverse lexicographic order to build the population BWT. Different sorting orders were tested for their effect on the BWT size using the 1000GP reads aligned to a 5-Mbp region.

genomes. The final size of the BWT for the entire data set was 464 GB (split over 16 smaller BWTs based on read prefix in order to load into system memory over multiple servers) (Supplemental Table S1), and the corresponding RocksDB metadata database was 4.75 TB (0.09 bytes per bp). The resulting population BWT server can be queried for exact matches to arbitrary length *k*-mer sequences and return either the count of matching read sequences, the matching read sequences, or the matching read sequences with sample metadata (Supplemental Fig. S1). We benchmarked the query completion time for 100,000 *k*-mer queries for the different types of server responses and *k* values (Supplemental Table S2), finding that, for smaller values of *k*, returning matching read counts was the fastest primarily due to the network time required for transferring large quantities of read sequences. At larger values of *k*, where less matching reads are found, the difference between requesting read counts and full matching read sequences is considerably reduced. In the remainder of the paper, we call the resulting population BWT the 1000GP BWT, or where unambiguous, just the BWT.

Population support for human reference assemblies and variation

We first used the population BWT to assess the direct support in the 1000GP read data for every base of two recent versions of the human reference assembly (GRCh37 and GRCh38) and support for the SNP and short indel variants called by the 1000GP. We extracted all forward strand 31-mers contained in both reference assemblies and queried the population BWT for reads matching these 31-mers. Finally, we also generated all 31-mers contained in the reads stored in the BWT Read Server. The vast majority of reference 31-mers (Fig. 3A) are supported by the 1000GP BWT (99.97%) and mostly shared between both assemblies (99.41%), with 0.07% of GRCh37 31-mers lost from the change from GRCh37 to GRCh38, with 0.49% gained in GRCh38. One estimate of the completeness of the 1000GP BWT is to calculate the proportion of 31-mers derived from high-depth Illumina sequencing reads of a sample that are already in the BWT. We tested one sample contained in the BWT (NA12878) and another not part of the BWT (NA12877), and in both cases, we found that only 0.95%–1.1% of 31-mers are not found in the BWT (Supplemental Table S6). We further queried the 1000GP BWT for all 31-mers generated by the SNP and indel variants found by the 1000GP (The 1000 Genomes Project Consortium 2015). Figure 3B shows the intersections of these four 31-mer sets. Considering the reference genomes





C

Supported		Intergenic	Intron	Exon	Coding	Total (kb)
Shared	GRCh37	1,157,992	1,310,169	79,673	33,584	2,581,417
	GRCh38	1,141,507	1,320,005	86,426	33,484	2,581,422
	(CTM)	(< 1)				
Exclusive	GRCh37	1,768	1,196	142	59	3,165
	GRCh38	12,072	1,459	92	54	13,677
	(CTM)	(2,824)				

Not supported		Intergenic	Intron	Exon	Coding	Total (kb)
Shared	GRCh37	549	640	42	4	1,235
	GRCh38	541	646	45	4	1,235
Exclusive	GRCh37	174	163	16	5	358
	GRCh38	313	30	2	1	346
	(CTM)	(242)				

Figure 3. (*A*) 31-mer intersection of two human reference assemblies (GRCh37 and GRCh38) and the 1000GP population BWT. (*B*) 31-mer intersection of two human reference assemblies, 1000GP population BWT, and all 31-mers generated from the 1000GP phase 3 SNP and indel variants (The 1000 Genomes Project Consortium 2015). 31-mers shared between reference sets and variant set (white numbers) make up for ~3% of each data set and almost all (99.998%) are supported by the 1000GP population BWT. (*C*) A breakdown of the regions on the two human assemblies with and without 1000GP population BWT support that are shared or exclusive to either genome build (all numbers are kbp), in four functional categories. (CTM) Centromeric sequence.

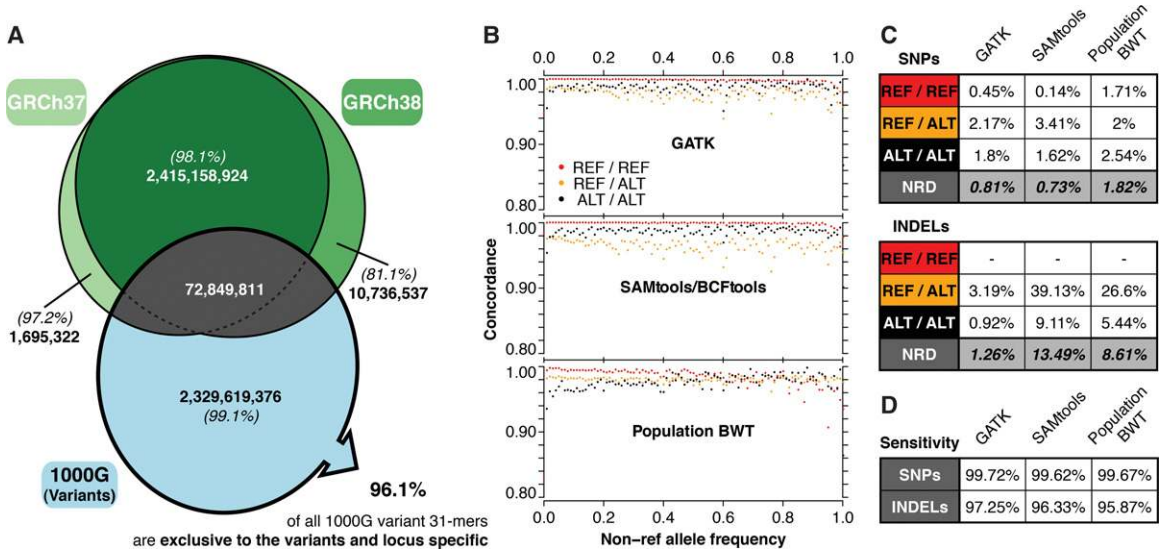


Figure 4. (*A*) Intersection of the human reference assembly 31-mers and the 1000GP SNP and indel variant 31-mers. The percentages in parentheses give the proportion of these 31-mers that are locus-specific (no other combination of variants in either the same or a different locus in the GRCh37 assembly generates the identical 31-mer). Of all 31-mers generated based on 1000GP variants, 96.1% are locus-specific and exclusive to the variants set, with 91.8% containing a single alternative allele. (*B*) SNP genotyping of the 1000GP samples at Illumina Omni chip exome-only sites by 31-mer querying of the BWT compared to single sample calling with GATK HaplotypeCaller (v3.5) and SAMtools (v1.1). Dots indicate genotype concordance for variants at different allele frequencies. (*C*) Genotype discordance rates for SNPs (Omni exome-only: 80,973 sites, all samples) and indels (Genome in a Bottle [Zook et al. 2016] exome in NA12878: 654 sites). (*D*) Sensitivity of each method expressed as the fraction of total genotypes for which a genotype call was made.

and the 1000GP variant 31-mers, the vast majority of 31-mers were either reference (solid black outline) or variant specific (dotted outline) and supported by the 1000GP BWT (overlap with the red ellipse). Figure 3C shows the amount of sequence gained or lost over four functional categories based on the GENCODE human genome annotation (Harrow et al. 2012). When the 31-mers are converted into reference genome regions, 3.1 Mbp (1.6M 31-mers, 0.07%) of sequence that has population BWT support was lost in the transition from GRCh37 to GRCh38, but roughly 7.5 times (13.6 Mbp) more was gained (10.5M 31-mers, 0.49%). We examined the read coverage for the regions in GRCh37 that do not contain 31-mer support from the 1000GP BWT. The vast majority of these regions are 50–60 bp (Supplemental Fig. S2), with >70% (89% and 73.7% for GRCh37 exclusive regions and those shared with GRCh38, respectively) overlapping at least one variant. Sixty-five percent and 39% (for GRCh37 exclusive regions and those shared with GRCh38, respectively) overlap a locus for which GRCh37 contains the minor allele or an error. Interestingly, the majority of the unsupported GRCh38 sequence is located in the new synthetic centromeric regions (CTM; 242 kbp) (Fig. 3C), although 2.8 Mbp of the new centromere is supported. The amount of coding sequence without population support in GRCh37 consists of 9.2 kbp in 203 protein-coding genes and 4.6 kbp in 123 genes for GRCh38 (Supplemental Table S4), reflecting the flipping of bases to the major allele. Interestingly, there were 12 protein-coding genes that contain unsupported 31-mers only found in GRCh38.

Reference-free population genotyping

Figure 3B shows that the majority (97%) of 31-mers derived from the 1000GP variation catalog are distinct from the reference genome. Furthermore, we determined that 99% of these 1000GP variant 31-mers not found in the references are locus-specific (no other combination of variants in either the same or a different locus in the GRCh37 assembly generates the identical 31-mer) (Fig. 4A). The 31-mers shared between the reference genomes and the variants are likely to be in regions containing repeats longer than 31 bp, which were still callable in the 1000GP by using the untrimmed longer reads or read pair information. Informed by this analysis, we developed a simple SNP and indel genotyping strategy based on querying the population BWT for *k*-mer sequences to test for read support of the reference and alternative allele for every individual. We tile across each genotyping site with overlapping *k*-mers upstream of and downstream from the site and query the population BWT for exact matching reads. We assign a genotype to each sample by recording how many of the reads from the sample match best to the reference or alternative allele (see Methods).

For SNPs, we benchmarked the approach using the Illumina Infinium BeadChip Omni2.5-8 genotypes in the 1000GP exome regions as a truth set. We initially evaluated the effect different values of *k* have on the population BWT genotyping accuracy using all Chromosome 20 sites, finding that $k = 34$ produced the lowest nonreference discordance (Supplemental Table S3). We genotyped all of the Omni chip positions in the 1000GP exome regions with single sample calling using GATK HaplotypeCaller, SAMtools/BCFtools, and the 1000GP population BWT. Figure 4B shows that the population BWT genotyping compares favorably to GATK and SAMtools across the allele frequency spectrum. The overall nonreference discordance rate is slightly higher for the population BWT genotyping (1.82%) compared to the GATK (0.81%) and SAMtools (0.73%). For heterozygous SNPs, the population BWT approach is more accurate than the two reference-based callers (discordance rate of 2% vs. 2.17% for GATK, and 3.41% for SAMtools). When we stratify the sites by the

number of flanking variants identified by the 1000GP, the proportion of correctly genotyped sites is reduced for all methods (Supplemental Fig. S9). The proportion of sites genotyped was >99% for all three approaches (Fig. 4D). A runtime comparison for genotyping NA12878 with SAMtools and GATK compared to the BWT is given in Supplemental Table S5. Although the BWT took five times longer, it completed genotyping for all of the samples, as every 31-mer query returns matching reads for all samples and therefore, over all samples, is many times faster.

We developed a similar approach for indel genotyping by testing reference and alternative alleles by dense *k*-mer tiling across the indel site (see Methods), querying the population BWT with the resulting *k*-mers and assigning a genotype to each individual based on the matching reads returned (see Methods). For indels, we use the Genome in a Bottle (GIAB) consortium gold standard indel genotypes for NA12878 for evaluation (Zook et al. 2016). Initially, we tested the effect different values of *k* have on genotyping accuracy using Chromosome 20 sites, determining that $k = 25$ produced the most accurate genotypes (Supplemental Table S3). We genotyped the indels with GATK HaplotypeCaller, SAMtools/BCFtools, and the 1000GP population BWT (see Methods). The indel genotyping accuracy varied widely between the callers. GATK produced the lowest nonreference discordance (1.26%), followed by the 1000GP population BWT (8.61%) and SAMtools (13.49%) (Fig. 4C). This is not so surprising since the GIAB indel calls are largely derived from GATK genotypes, and there is often poor overlap between indel discovery tools (Narzisi et al. 2014).

Nonreference queries

As the population BWT is a full-text index of the read sequences, irrespective of whether they align to the reference genome or not, it enables rapid testing of hypothesis-driven queries. We sought to assess the proportion of sequences of viral origin contained in the 1000GP reads. An earlier study using 150 individuals from the 1000GP found evidence for 0.13% of reads coming from nonhuman DNA (Tae et al. 2014). To expand this to the full set of samples, we downloaded 257,943 viral sequences from the CoreNucleotide division of GenBank and used the Kraken classifier (Wood and Salzberg 2014) to define a set of 102.6M virus-specific 31-mers (see Methods; Fig. 5A). We queried the 1000GP population BWT with these 31-mers initially for read counts (to remove very highly abundant low complexity sequences), then returned matching read sequences, and finally queried the metadata database for sample information. The population BWT queries were run in under 2 d, with the sample metadata retrieval taking 7 d (see Methods). The most prevalent source of nonhuman sequences is the herpesviruses, including Epstein-Barr virus, used in the creation of the lymphoblastoid cell lines (LCLs) that were the DNA source for many of the 1000GP samples. The distribution of the number of EBV matching reads largely follows the documented DNA source in the 1000GP (Fig. 5B), with a few notable exceptions which are likely misclassified as being from blood. The DNA that is recorded as being of unknown origin appears to be almost entirely from LCLs, having a similar distribution of EBV reads as the documented LCL-derived samples. Of the viruses identified (excluding

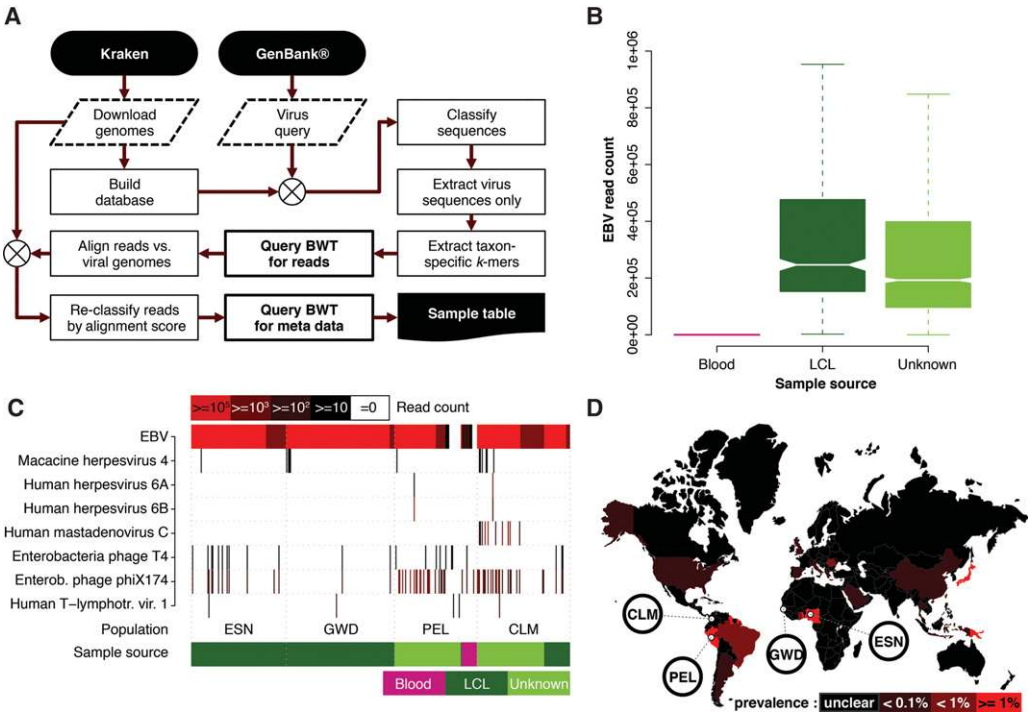


Figure 5. (*A*) Reference genomes (Human, bacteria, plasmids, and viruses) were downloaded using Kraken's (Wood and Salzberg 2014) built-in routines and a Kraken database generated. GenBank was queried for all virus sequences and the resulting sequence set classified using Kraken to identify taxon-specific 31-mers which were used to query the population BWT for matching reads. Retrieved read sequences were reclassified by alignment to the viral genomes stored in the Kraken database. Finally, sample metadata were retrieved for the final read set. (*B*) Notched boxplot showing the distribution of human herpesviruses (including EBV) read counts stratified by documented DNA source. Nonoverlapping notches indicate a significant difference of the medians at the 5% level. (*C*) The populations for which at least one sample contains >10 HTLV-1 reads (black bars) and other virus taxa with >99 reads (red bars) in at least one sample are shown (for all populations, see Supplemental Figs. S3–S8). (*D*) World map showing HTLV-1 prevalence in different countries, with 1000GP populations that show signal for this virus highlighted.

EBV), 69 occur in at least one sample at >10 reads and 14 at >100 reads (Supplemental Figs. S2–S7).

Figure 5C gives the species source for the most frequently found sequences per individual for four particular population groups (see Supplemental Figs. S3–S8 for all populations). Enterobacteria phage phiX174, the Illumina library spike-in sequence, is also prevalent across all of the populations at greater than or equal to 100 reads in 605 samples. Adenovirus C is reported to be present in almost all populations worldwide (Garnett et al. 2002); however, our analysis shows that it is almost completely absent in several populations (e.g., Gambian in Western Divisions in the Gambia, and Esan in Nigeria) (Fig. 5). The absence of adenovirus in some groups and high levels in other groups suggests differences in the sample preparation (Adenovirus C is often used as a recombinant vector for cell culture reagents) (Luo et al. 2007) or differences in adenovirus in these populations.

One interesting finding is the presence of Human T-lymphotropic virus 1 (HTLV-1) reads found in six individuals (Fig. 5C; Table 2, below). HLTV-1 can integrate into the genome and is known to have infected human populations for thousands of years, with the virus being transferred from mother to child, through sexual contact, or through contaminated blood products (Derse et al. 2007; Verdonck et al. 2007). The known epidemiological distribution spans areas of southern Japan (Satake et al. 2012), sub-Saharan Africa, the Caribbean, and South America, where >1% of the general population is infected (Verdonck et al. 2007). For the most part, carriers remain asymptomatic, but HTLV-1 infection has been associated with exceptionally severe diseases, such as adult T-cell leukemia/lymphoma (Takatsuki 2005) and an inflammatory disease of the central nervous system called HTLV-1-associated myelopathy/tropical spastic paraparesis (Gessain et al. 1985; Lezin et al. 2005). The 31-mer querying method identified six samples from five populations with potential HTLV-1 integrations. For those individuals, we also aligned the entire original read set to a reference genome containing GRCh38 and a HTLV-1 consensus sequence, confirming the presence of HTLV-1 in these genomes and slightly increasing the HTLV-1 read support for each sample (Table 2, below). The populations in which we detected HTLV-1 presence largely follow the known epidemiological distribution with HTLV-1 positive samples from Africa and South American populations and were sequenced at six different centers. We did not observe HTLV-1 in any Japanese samples (reported HTLV-1 prevalence of 0.66% and 1.02% [Satake et al. 2012]), although Japan has had HTLV-1 population screening in place since 1986 (Inaba et al. 1989).

Discussion

In this paper, we show how BWT indexes can be used for efficient compression and indexing of large collections of sequencing reads from thousands of individuals. Unlike traditional reference-based alignment approaches, the population BWT has a sublinear growth as more individuals are included in the structure. However, this is dependent on the sequencing data having a low error rate so that the majority of new sequences observed in each individual represent true genetic variation. One of the main difficulties of using the 1000GP data with this approach is that most individuals were sequenced to low coverage (7–8×). For error correction, we used a Cortex de Bruijn graph that was built from these reads and was error-cleaned by removing tips (short contigs unconnected at one end) and unitigs that were at low frequency in all populations. The fraction of error-corrected reads was quite low (9.2%), since our error-correction strategy was deliberately conservative as we wanted to avoid removing true genetic variation. It is still notable that the resulting population BWT contains over 35 times more 31-mers than are present in the reference genomes and the SNP and indel variants (Fig. 3A). It has been suggested that existing variation catalogs fail to account for 35%–68% of some types of structural variation and 25% of short indels (Gordon et al. 2016). Therefore, unaccounted genetic variants, variants located in inaccessible regions of the genome, and nonhuman sequences could contribute to these novel 31-mers. Our virus sequence analysis only accounted for 102 M 31-mers in the population BWT; therefore, it is more likely that these novel 31-mers are due to remaining errors in the sequencing reads. One could perform more stringent error correction to reduce the sequencing errors at a cost of removing true low frequency variants. More recent approaches to per-sample read-error correction are most effective with comparatively high sequencing depth (30–50×) per sample (Simpson and Durbin 2011; Li 2015). Therefore, we envisage that, as the cost of human sequencing continues to decrease and higher depth sequencing becomes the norm, the population BWT could be an efficient storage medium for indexing large collections of human samples.

The most significant storage saving in this approach comes from discarding the base qualities after base error correction is carried out. It remains an open question as to what proportion of base qualities need to be retained for accurate variant discovery and genotyping, with increasing evidence showing that discarding or quantile binning of base qualities does not have a detrimental effect (Yu et al. 2015; Ochoa et al. 2016). However, many applications of next-generation sequencing (e.g., clinical sequencing) rely on highly accurate identification of novel rare variants. One alternative approach to completely discarding base qualities could be a controlled loss of base qualities. For example, there could be an iterative process of population BWT construction, where genomes are continually added. Initially, with few genomes, the majority of the sequencing reads will contain novel *k*-mers, and as more genomes are added, we will observe the same *k*-mers in multiple individuals across the population. One could envisage an approach where base qualities are only maintained for reads that support novel *k*-mers, with these *k*-mers being constantly queried against the BWT for increasing population read support, with the goal of eventually discarding these base qualities as increasing support is observed in the population. One could employ the BEETL-Fastq BWT-based data structure to create a side structure of compressed and searchable indexes of read sequences including base qualities (Janin et al. 2014).

One of the limitations of this approach is that this implementation of a population BWT does not maintain read pair information. In our SNP and indel genotyping, read pair information could be incorporated into the genotyping strategy to derive more accurate genotypes. Read pairs would also be particularly useful for structural variant discovery and genotyping, as most existing structural variation detection algorithms use a combination of split reads and read pairs for supporting evidence (Keane et al. 2014; Layer et al. 2014). In our virus analysis, efficient retrieval of read pairs would enable more rapid localization of the HTLV-1 viral integrations by avoiding the need to realign the full original read set. For SNP and indel discovery, retrieval of read pairs would enable local haplotype assembly and phasing of discovered variants, which could aid correct alignment of highly variable loci into a variation graph (Church et al. 2015), especially with library technologies that conserve long-range phase information (Putnam et al. 2016; Zheng et al. 2016). Recording read pair information

could significantly increase the amount of metadata required from just basic sample level information to knowledge of every unique read pair combination or sets of reads from the same molecule. The ability to store and efficiently retrieve all of the read pairs of a sample could enable the use of the population BWT as a highly compressed, searchable, and scalable archival format for sequencing data.

One of the benefits of choosing the BWT and FM-index as the underlying data structure is that the construction process does not constrain the length of possible *k*-mer queries. In de Bruijn-based approaches such as Cortex (Iqbal et al. 2012) and SBT (Solomon and Kingsford 2016), the *k*-mer must be fixed at the time of index construction. In the SNP and indel genotyping, the genotyping accuracy varied depending on the *k*-mer. The length of the *k*-mer used to assess an individual site can be affected by the number of mutations in the local region, where smaller, more densely sampled *k*-mers could potentially produce more accurate genotypes in regions of high mutation rates. Using dynamic *k*-mer queries for genotyping and the incorporation of read pair information are potential avenues for further improving genotyping accuracy.

Using whole-genome sequencing reads to classify reads into taxonomic groups has become the basis for metagenomic analysis (Gilbert and Dupont 2011). We used a metagenomics *k*-mer classification approach to detect evidence for nonhuman sequences in the 1000GP reads. Several studies have cautioned against overinterpretation of unexpected sequences found in sequencing reads due to the possibility of laboratory kit or reagent contamination (Lusk 2014; Salter et al. 2014). For these reasons, our finding of evidence for low levels of HTLV-1 in several 1000GP samples should be treated with caution. On the one hand, the epidemiological distribution of the samples found to contain HTLV-1 fits the known pattern, we can localize many of the putative integrations using read pairs (Table 2, below), and the samples were sequenced at multiple different centers. However, we cannot rule out the possibility of kit or reagent contamination without further laboratory validation of the results.

Methods

Sequencing data

The sequencing reads were downloaded in FASTQ format from the 1000GP ftp site and correspond to the phase 3 sequencing data freeze (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/20130502.phase3.analysis.sequence.index) consisting of 2574 in total and 2535 of these with both low coverage whole-genome and exome sequencing (The 1000 Genomes Project Consortium 2015).

Error correction

Read-error correction was carried out using the Cortex software (Iqbal et al. 2012) (https://github.com/iqbal-lab/cortex). Briefly, Cortex is a de novo De Bruijn graph assembler that allows simultaneous assembly of multiple samples and variants to be called without reliance on mapping of reads to a reference genome. We used the Cortex graph that contains a merge of all of the populations in the 1000GP (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130718_phase3_samples_cortex_graphs/phase.all_pops.ctx) (see Supplemental Methods of The 1000 Genomes Project Consortium 2015). The Cortex graph was loaded into memory, and then the reference genome (GRC37) was parsed, annotating each *k*-mer with the direction in which it was seen (forward, reverse, or both). If a read was <73 bp in length or contained any character other than ACGT, it was discarded. If all of the base qualities for a read were greater than or equal to Q20, the read was kept without correction. Correction was seeded by finding a 31-mer of $Q > 20$ bases and extending greedily by shifting one base at a time. On shifting and meeting a $Q < 20$ base, if there was precisely one single-base correction of a $Q < 20$ base which changed a *k*-mer absent from the Cortex graph to a *k*-mer present in the Cortex graph, this change was made. If all of the *k*-mers in a read were annotated consistently with the read coming from the reverse strand of the reference genome (i.e., either unannotated, or annotated as being seen in the reverse strand of the reference), the corrected read was reverse-complemented and printed in the forward direction; otherwise, it was printed in the same orientation as the input data. This was done purely to improve compression in the BWT. Finally, read sequences were trimmed to two reads lengths: 73 and 100 bp. If a corrected read was >100 bp, then it was trimmed to 100 bp; if a read was between 73 and 100 bp in length, it was trimmed to 73 bp; and if a read was <73 bp, it was discarded. For base error correction, we used a modified version of error_correction.c from Cortex (https://github.com/wtsi-svi/cortex@fc26874).

Read deduplication and metadata

The error correction process output the read sequence (in forward orientation), the read name, the number of corrected bases during, and the number of low quality (<Q20) bases. Corrected read sequences were sorted in reverse lexicographic order, with duplicates removed. For each unique read sequence in the final read set, we stored the read groups (2 bytes), the number of corrected bases (1 byte), and the number of low-quality bases (1 byte). This information was stored in a RocksDB (v2.6), with the unique read sequences as keys.

BWT and FM-index construction

The reads were split into 16 partitions based on the last 2 bp in the read sequence (see Supplemental Table S1), with the reads for each partition sorted in reverse lexicographic order. Then, we used SGA v0.10.13 (Simpson and Durbin 2011) to construct the BWT string for each read collection. SGA outputs BWT strings in run-length encoding (RLE), with each byte representing a continuous run of the same character. The first three bits of a byte encode the five different characters (i.e., ACGT$). The last five bits of the same byte encode the number of the runs for that character up to the length of 31. The cumulative size of the run-length encoded BWTs on disk was 464GB.

The Burrows–Wheeler transform renders an important property, Last-to-First (column) mapping, i.e., the *i*th occurrence of character X in the last column corresponds to the *i*th occurrence of X in the first column. The FM-index (Ferragina and Manzini 2000), based on BWT and LF mapping, allows for fast query of a pattern and locates every occurrence of the searched pattern. We built an index structure based on the run-length encoded BWT string. With such index, we were able to search for a *k*-mer, extend to the full read from matched location, and get the full read sequence in linear time. The implementation of this index can be found at https://github.com/wtsi-svi/ReadServer.

System setup

We set up a server to allow fast query of a given *k*-mer and return information about the number of matched reads, the matched read sequences, and for each matched read, the list of samples that the read was derived from. To achieve high-throughput and fast response, we created a message queue-based application server that sends *k*-mer sequence requests to the 16 BWTs across four

physical servers (Supplemental Fig. S1). Each machine has four applications, and each application has a BWT partition and its associated index structure loaded in memory (total memory required: 561GB). The hardware of these four machines is varied. One machine has 32 (logical) cores with 256GB RAM. The other three machines have 20 (logical) cores and 188GB RAM. All machines run on the Ubuntu 12.04LTS system.

Reference genome analysis

We generated 31-mer sets for GRCh37 and GRCh38 by extracting 31-mers starting on every position (forward strand) in both assemblies for all autosomes and gonosomes. Any 31-mers containing IUPAC ambiguity codes were discarded. The 31-mers were queried against the population BWT to check for support in the 1000GP read set (forward and reverse orientation). The population BWT 31-mers (used in Fig. 3A,B) were generated from the final corrected set of read sequences.

To measure the completeness of the 1000GP BWT, we downloaded high-coverage sequencing reads for NA12877 (ERR194146) and NA12878 (ERR194147) from the Illumina Platinum genomes. Reads were error-corrected with BFC (-s 3g -t 32 FASTQ_IN> FASTQ_OUT) (Li 2015). In the resulting reads, NA12878 had a higher rate of remaining unique 31-mers (0.34) than NA12877 (0.31). We extracted a random 31-mer from a randomly selected read to a total coverage of one 31-mer for every tenth read. The same read could give rise to more than one 31-mer, and independently drawn 31-mers could be exact duplicates (same read, same position in the read). Thirty-one-mers containing N's were discarded, the remaining 31-mers randomly split into 10 independent sets, and the sets per sample queried against the 1000GP BWT. Finally, 31-mers were then binned based on the number of read matches (Supplemental Table S6).

1000 Genomes variant 31-mers

For each individual in 1000GP, we created a maternal and paternal genome by substituting the phased variants and generated 31-mers that overlap with every nonreference position. We excluded unphased, nondiploid (except gonosomal hemizygous), or conflicting variants (e.g., SNPs in regions which are also called as being deleted on the same chromosome copy), variants for which the exact coordinates could not be determined, reference alleles where an individual chromosome copy was contradictory (e.g., a region genotyped as reference for a deletion that also contains another nonreference variant), and filtered reference alleles that collided with each other by discarding all downstream reference loci within the overlapping region. In total, 0.16% of the variants were excluded. For each of the resulting haploblocks, every contained 31-mer was generated and queried against the population BWT.

SNP genotyping

We used the 1000GP Illumina Omni chip data produced at the Broad institute (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20131122_broad_omni/Omni25_genotypes_2141_samples.b37.v2.vcf.gz) for the list of gold standard SNP genotypes. There were 1668 samples in the Omni chip genotypes that were included in the phase 3 1000GP freeze. Genotyping was carried out with the population BWT by generating a reference and alternate allele using 99 bp of flanking sequence for each site. We tiled each allele sequence with 34-mers with a step of 10 bp. We queried the population BWT with the 34-mers and carried out a local Smith-Waterman alignment (match +1, mismatch penalty −4, gap open penalty −6, gap extension penalty −1) of the returned reads onto the reference and alternate allele, excluding divergent hits (if only mismatches, then allow maximum of three mismatches, otherwise allow a maximum of one indel and eight points penalty). Using the number of reads supporting the reference or alternative alleles, we assigned genotypes according to Table 1. Finally, we output a new VCF file with the population BWT-determined genotypes.

Indel genotyping

We downloaded a recent version of the Genome in a Bottle (GIAB) NA12878 variant set (v2.18, ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv2.18/NISTIntegratedCalls_14datasets_131103_allcall_UGHapMerge_HetHomVarPASS_VQSR-v2.18_all_nouncert_excludesimplerep_excludesegdups_excludedecoy_excludeRepSeqSTRs_noCNVs.vcf.gz) and filtered it to include only monoallelic indels. We then used SAMtools/BCFtools v1.1 (samtools mpileup -gut DP,DV,DP4,SP,DPR,INFO/DPR -EQ 0 -p -b [NA12878_BAM_FOFN] -f [GRCh37_REF] -l [NA12878_VARIANTS_BED] | bcftools call -mf GQ,GP -O z) and GATK v3.5 (java7 -jar -Xmx28G GenomeAnalysisTK.jar -R [GRCh37_REF] -I [NA12878_BAM_LIST] -L [NA12878_VARIANT_INTERVALS_BED] -T HaplotypeCaller -stand_call_conf 4 –genotyping_mode GENOTYPE_GIVEN_ALLELES –output_mode EMIT_ALL_SITES –alleles [NA12878_VARIANTS_VCF]) (intervals are the indel start and end position padded by 150 bp) to call and/or genotype those variants based on NA12878 low-coverage and exome sequencing data. The SAMtools/BCFtools calls were subsequently left normalized as we did not specify the alleles during the genotyping stage. The GATK results did not require this step. For population BWT genotyping, we generated a reference and alternate sequence for each indel by adding 100 bp of flanking sequence to either the reference or alternate allele. We then generated 25-mers (1-bp step) from these sequences and queried the population BWT for matching reads. Any 25-mer with >100,000 matches or with a homopolymer length >14 bp were excluded. We further generated flanking sequences between 100 and 200 bp upstream of and downstream from each variant. If a 25-mer from these regions was found in any of the reads returned from the BWT, we considered that as evidence that the corresponding read is either pointing away from the variant or too far away to overlap it, and hence discarded the read. All remaining reads were collapsed into a nonredundant read set and aligned against both the reference and alternative alleles using exonerate v2.2.0 (–model ungapped –dnawordlen 25 –percent 90 –bestn 1). Alignment hits were subsequently filtered for reads that reach at least 2 nt from the flank into the variant locus. Each valid read was then assigned to either the reference or alternative allele based on the highest alignment score, with reads with equal scores being discarded. Finally, the full sample metadata were retrieved and the indels genotyped per sample using the read count thresholds in Table 1.

Table 1. The population BWT SNP genotype assignment scheme

Constraint	Genotype
$N = 0, M = 0$	./.
$N > 0, M = 0$	0/0
$N = 0, N > 0$	1/1
$(N/M) < 0.125$	1/1
$(N/M) > 8$	0/0
$0.125 < (N/M) < 8$	0/1

(N) Number of reference supporting reads, (M) number of alternative allele supporting reads, (.) unknown genotype, (0) reference allele, (1) first alternative allele.

Table 2. The population and sample identifiers for the individuals found to contain evidence for HTLV-1 viral integrations

Population	Sample	Center	Read counts		Unique integrations
			BWT	Alignment	
ACB	HG02325	Broad Institute	1	2	n/a
CLM	HG01357	Illumina, BCM	224	282	8
ESN	HG03370	WUGSC, WTSI	40	92	2
GWD	HG02675	Broad Institute	104	123	3
PEL	HG01918	BGI	19	32	n/a
PEL	HG01917	BGI	21	34	2

The read count columns indicate the number of viral reads by searching the population BWT and by realignment of all of the sequencing reads from the individuals to a GRCh38 + HTLV-1 reference genome.

Viral genome analysis

We downloaded 257,943 viral sequences from the CoreNucleotide division (http://www.ncbi.nlm.nih.gov/nuccore/ on 20/02/2015) of GenBank (search string: "((((((txid10239[Organism]) AND 2000[SLEN]:300000[SLEN])))))) NOT patent"). We generated the virus taxon-specific 31-mers using Kraken v0.10.5-beta (Wood and Salzberg 2014) by generating a database containing fully assembled virus ("kraken-build –download-library viruses"), bacteria ("kraken-build –download-library bacteria"), and plasmid ("kraken-build –download-library plasmids"), and the GRCh38 human reference assembly ("kraken-build –download-library human") (built on 16/03/2015). We used this Kraken database to classify the virus sequences downloaded from GenBank. Of the 257,943 input sequences, 244,656 (94.8%) could be classified, 243,123 (94.3%) as viruses covering 4093 of 5808 virus taxa (70.5%). From the Kraken output files, we extracted 102,655,127 taxon-specific 31-mers. We queried the population BWT with these 31-mers, returning counts for the number of matching reads (query time 2d3h48′26″ CPU time, 2d16h5′7″ wall clock time using 80 threads). Of the 102.6 M 31-mers, 435,799 from 886 taxa had matches in the population BWT. Of these,1369 31-mers match very large numbers of reads (>100,000), indicating that these contain little information and match repetitive or low-complexity sequences, and so they were discarded. We subsequently did full read sequence retrieval queries for the remaining 434,430 31-mers (0d5h2′19″ CPU time, 0d5h9′22″ wall clock time, using 10 threads). All reads returned from the population BWT were collapsed into a nonredundant set of sequences per taxon ID resulting in a final size of 113,193,726 reads.

Although we can be sure that each read contains at least one taxon-specific 31-mer, this could be due to one or more sequencing errors in the 31-mer. Therefore, we reclassified the reads by short read alignment to the genome sequences using Smalt v0.7.5.1 (http://www.sanger.ac.uk/science/tools/smalt-0), which enabled us to examine the relative alignment score of matches to assess the classification. Based on the alignment results, we chose a threshold of 75% of the maximum alignment score per read and included only reads that exceeded this threshold when aligning to a virus genome while staying below for any other kind of target sequence (human, bacteria, or plasmid). Each read fulfilling these criteria was then assigned to the virus to which it aligned. In case of equal best matches to different virus genomes, one was chosen at random. Using this filter, 107,234,569 reads (94.7%) could be assigned to a virus covering 289 virus taxa. To assign samples, we queried the population BWT metadata database for sample information per read (total run time was 7d17h25′32″ CPU time, 8d0h42′13″ wall clock time).

For the samples found to contain HTLV-1, we downloaded the original FASTQ files from the 1000GP ftp site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/) and aligned all of the reads using BWA MEM v0.7.12 to a reference genome containing GRCh38 + HTLV-1. Table 2 gives the relative read counts for reads found to contain HTLV-1 from the BWT queries and alignment of the reads (no minimum mapping quality or length threshold for hits).

Software availability

The collection of software used to build the population BWT server is available on Github (https://github.com/wtsi-svi/ReadServer) and in Supplemental File S1.

Acknowledgments

This work was supported by the Wellcome Trust. We thank John Marshall (Wellcome Trust Sanger Institute) for providing technical help and support for several aspects of this project.

Author contributions: J.T.S., S.A.McC., R.D., Z.I., and T.M.K. initiated the ideas for the study. D.D.D., Z.L., S.A.McC., and T.M.K. carried out the software development and computational analysis. M.C. provided advice for the virus genome analysis. D.D.D. and T.M.K. drafted the manuscript. All authors read and approved the final manuscript.

References

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526:** 68–74.

Bloom BH. 1970. Space/time trade-offs in hash coding with allowable errors. *Commun ACM* **13:** 422–426.

Church DM, Schneider VA, Steinberg KM, Schatz MC, Quinlan AR, Chin C-S, Kitts PA, Aken B, Marth GT, Hoffman MM, et al. 2015. Extending reference assembly models. *Genome Biol* **16:** 13.

Cox AJ, Bauer MJ, Jakobi T, Rosone G. 2012. Large-scale compression of genomic sequence databases with the Burrows–Wheeler transform. *Bioinformatics* **28:** 1415–1419.

Derse D, Crise B, Li Y, Princler G, Lum N, Stewart C, McGrath CF, Hughes SH, Munroe DJ, Wu X. 2007. Human T-cell leukemia virus type 1 integration target sites in the human genome: comparison with those of other retroviruses. *J Virol* **81:** 6731–6741.

Ferragina P, Manzini G. 2000. Opportunistic data structures with applications. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science, FOCS '00*, p. 390. IEEE Computer Society, Washington, DC.

Garnett CT, Erdman D, Xu W, Gooding LR. 2002. Prevalence and quantitation of species C adenovirus DNA in human mucosal lymphocytes. *J Virol* **76:** 10608–10616.

The Genome of the Netherlands Consortium. 2014. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46:** 818–825.

Gessain A, Barin F, Vernant JC, Gout O, Maurs L, Calender A, de Thé G. 1985. Antibodies to human T-lymphotropic virus type-I in patients with tropical spastic paraparesis. *Lancet* **2:** 407–410.

Gilbert JA, Dupont CL. 2011. Microbial metagenomics: beyond the genome. *Annu Rev Mar Sci* **3:** 347–371.
Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW, et al. 2016. Long-read sequence assembly of the gorilla genome. *Science* **352:** aae0344.
Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22:** 1760–1774.
Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E. 2011. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res* **21:** 734–740.
Inaba S, Sato H, Okochi K, Fukada K, Takakura F, Tokunaga K, Kiyokawa H, Maeda Y. 1989. Prevention of transmission of human T-lymphotropic virus type 1 (HTLV-1) through transfusion, by donor screening with antibody to the virus. One-year experience. *Transfusion* **29:** 7–11.
Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. 2012. *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet* **44:** 226–232.
Janin L, Schulz-Trieglaff O, Cox AJ. 2014. BEETL-fastq: a searchable compressed archive for DNA reads. *Bioinformatics* **30:** 2796–2801.
Katsanis SH, Katsanis N. 2013. Molecular genetic testing and the future of clinical genomics. *Nat Rev Genet* **14:** 415–426.
Keane TM, Wong K, Adams DJ, Flint J, Reymond A, Yalcin B. 2014. Identification of structural variation in mouse genomes. *Front Genet* **5:** 192.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25.
Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15:** R84.
Lezin A, Olindo S, Olière S, Varrin-Doyer M, Marlin R, Cabre P, Smadja D, Cesaire R. 2005. Human T lymphotropic virus type I (HTLV-I) proviral load in cerebrospinal fluid: a new criterion for the diagnosis of HTLV-I–associated myelopathy/tropical spastic paraparesis? *J Infect Dis* **191:** 1830–1834.
Li H. 2015. BFC: correcting Illumina sequencing errors. *Bioinformatics* **31:** 2885–2887.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.
Luo J, Deng Z-L, Luo X, Tang N, Song W-X, Chen J, Sharff KA, Luu HH, Haydon RC, Kinzler KW, et al. 2007. A protocol for rapid generation of recombinant adenoviruses using the AdEasy system. *Nat Protoc* **2:** 1236–1247.
Lusk RW. 2014. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS ONE* **9:** e110808.
Narzisi G, O'Rawe JA, Iossifov I, Fang H, Lee Y, Wang Z, Wu Y, Lyon GJ, Wigler M, Schatz MC. 2014. Accurate *de novo* and transmitted indel detection in exome-capture data using microassembly. *Nat Methods* **11:** 1033–1036.
Ochoa I, Hernaez M, Goldfeder R, Weissman T, Ashley E. 2016. Effect of lossy compression of quality scores on variant calling. *Brief Bioinform* bbw011.
Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* **26:** 342–350.
Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Ávila-Arcos MC, Malaspinas A-S, et al. 2015. Population Genetics. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349:** aab3884.
Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* **12:** 87.
Satake M, Yamaguchi K, Tadokoro K. 2012. Current prevalence of HTLV-1 in Japan as determined by screening of blood donors. *J Med Virol* **84:** 327–335.
Simpson JT, Durbin R. 2011. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* **22:** 549–556.
Solomon B, Kingsford C. 2016. Fast search of thousands of short-read sequencing experiments. *Nat Biotechnol* **34:** 300–302.
Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. 2015. Big data: astronomical or genomical? *PLoS Biol* **13:** e1002195.
Tae H, Karunasena E, Bavarva JH, McIver LJ, Garner HR. 2014. Large scale comparison of non-human sequences in human sequencing data. *Genomics* **104:** 453–458.
Takatsuki K. 2005. Discovery of adult T-cell leukemia. *Retrovirology* **2:** 16.
The UK10K Consortium. 2015. The UK10K project identifies rare variants in health and disease. *Nature* **526:** 82–90.
Verdonck K, González E, Van Dooren S, Vandamme A-M, Vanham G, Gotuzzo E. 2007. Human T-lymphotropic virus 1: recent knowledge about an ancient infection. *Lancet Infect Dis* **7:** 266–281.
Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15:** R46.
Yu YW, Yorukoglu D, Peng J, Berger B. 2015. Quality score compression improves genotyping accuracy. *Nat Biotechnol* **33:** 240–243.
Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, et al. 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34:** 303–311.
Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3:** 160025.

Received June 22, 2016; accepted in revised form December 14, 2016.

Using reference-free compressed data structures to analyze sequencing reads from thousands of human genomes

Dirk D. Dolle, Zhicheng Liu, Matthew Cotten, et al.

Genome Res. 2017 27: 300-309 originally published online December 16, 2016
Access the most recent version at doi:10.1101/gr.211748.116