Supplement for: "Building a Pan-genome Reference for a Population"

Ngan Nguyen, Glenn Hickey, Daniel Zerbino, Brian Raney, Dent Earl, Joel Armstrong, W. James Kent, David Haussler, Benedict Paten

Center for Biomolecular Science and Engineering, University of California Santa Cruz, CA, USA

1 Availability

The Cactus alignment source is in one convenient distribution at: `https://github.com/glennhickey/progressiveCactus`. It includes the source code for this project, as well as the HAL source code, which contains the hal2AssemblyHub.py script, for generating comparative assembly hubs that can be viewed using the UCSC browser. The MHC assemblies and the comparative assembly hub is at: `http://hgwdev.cse.ucsc.edu/~benedict/MHCBrowserForRecomb`.

2 Transitive Sequence Graph

Supp. Figure 1 shows an illustration of a transitive sequence graph.

3 Sequence Assemblies

The input samples include 16 human MHC haploid assemblies. In each case sets of scaffolds were obtained as described below and then converted into sets of contigs. This was done by splitting the scaffolds at scaffold gaps, defined as contiguous subsequences of 10 or more 'N' or 'n' characters, resulting in the removal of the scaffold gaps and the replacement of previously contiguous scaffolds with multiple separate contigs.

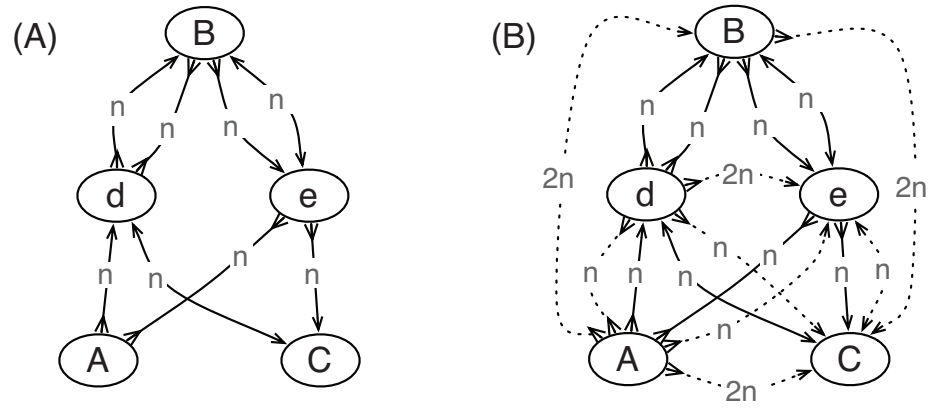

Fig. 1. An illustration of a transitive sequence graph. (A) A sequence graph of n sequences of "A d B e C" and n sequences of "A -e B -d C". (B) A transitive sequence graph of the same sequences in (A).

Of the assemblies, 8 were the GRCh37 haplotypes [1, 2]. These sequences (chr6:28477754-33448354, chr6_apd_hap1, chr6_cox_hap2, chr6_dbb_hap3, chr6_mann_hap4, chr6_mcf_hap5, chr6_qbl_hap6, chr6_ssto_hap7) were obtained from the UCSC genome browser GRCh37/hg19 database.

One assembly was the Venter MHC sequence (chr6:28284180-33170530), which was extracted by mapping the Venter assembly (September 2007 release, `ftp://ftp.jcvi.org/pub/data/huref/`) to the GRCh37 MHC loci. The mapping was done using LASTZ (`http://www.bx.psu.edu/~rsharris/lastz/`, version 1.02.00), with the minimum identity set to 97% and default parameters otherwise.

Two other assemblies came from the 'African' (NA18057) [3] and the 'Asian' (Yh1) [4] genomes. For these two genomes, we obtained the scaffolds from the BGI de novo assemblies (see `ftp://public.genomics.org.cn/BGI/yanhuang/Genomeassembly/african2.scafSeq.closure.gz` and `ftp://public.genomics.org.cn/BGI/yanhuang/Genomeassembly/asm_yanh.scafSeq.closure.gz`, respectively) and extracted out the sequences that mapped to the GRCh37 MHC loci as done for the Venter assembly.

The last 5 assemblies were from the 1000 Genomes pilot project trios, including NA12878, NA12892, NA19238, NA19239 and NA19240. The NA12878 MHC assembly was extracted

from the recently de novo assembled NA12878 genome [5] (see http://www.ncbi.nlm.nih.gov/nuccore?term=GL582980:GL586310), again by mapping the scaffolds as described above.

The other 4 assemblies were made using the Velvet de novo assembly program [6], version 1.1.06. For each assembly, we only used the Illumina reads http://www.illumina.com/systems/hiseq_2000.ilmn that were mapped to the GRCh37 MHC main region, using the 1000 Genomes project alignments (downloaded from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/data/NAxxxxx/alignment/, where NAxxxxx is replaced with the sample name, e.g. NA12892). The Velvet parameters used were as follows: *kmer 25, exp_cov auto*, *ins_length* and *ins_length_sd* obtained using the perl script *observed-insert-length.pl*, which was included in the Velvet package (http://www.ebi.ac.uk/~zerbino/velvet/Manual.pdf).

In addition, we also added the MHC sequence of the chimpanzee reference assembly as an outgroup. This sequence was extracted from the UCSC genome browser chimpanzee (assembly panTro3) by mapping the GRCh37 MHC main region to this assembly using the 'Covert' function of the UCSC browser.

4 Creating Human Haplotype Alignments

To create the alignments (homology relation) we use an adapted version of the Cactus alignment program [7]. The Cactus program's CAF algorithm starts by using the LASTZ pairwise alignment program (http://www.bx.psu.edu/~rsharris/lastz/, version 1.02.00) to generate a set of pairwise alignments between all the input sequences. It then filters these pairwise alignments to create a consistent multiple sequence alignment (MSA).

In the adapted version of Cactus used for this work the following parameters are passed to LASTZ: *–hspthresh=1800 –identity=X*, where $X = 95 = \lfloor 100 - \frac{300}{4}(1 - e^{-d\beta\frac{4}{3}}) \rfloor$ is the maximum likelihood identity expected by the Jukes Cantor model [8], given a liberal estimate of the maximum evolutionary distance between the human and the Chimpanzee outgroup

of $d = 0.015$ [9], and a conservative factor of $\beta = 3$ to allow for regional accelerations in the substitution rate.

4.1 MSA Post Processing

Having constructed an initial MSA and the pan-genome reference P. Ref. for the samples above we "trimmed" the alignment so that the PGF sample, which is a single contig, was present in the first and last columns of P. Ref. This trimming resulted in a MSA and P. Ref. that could then be fairly compared with PGF, as no sequence included in the MSA mapped to before or after the interval defined by PGF.

The trimming was achieved by recomputing the MSA and P. Ref. with exactly the same parameters as in the initial run, but with suffixes and prefixes of contigs that mapped to columns in P. Ref. that preceded the first column containing positions from PGF or proceeded the last column containing positions of PGF removed.

5 Sample Composition

To understand the contribution of each sample to P. Ref. we analyze the alignment of each sample in the MSA. In this work each sample can be considered a set of contigs. A contig and the bases it contains are *covered* by the alignment if one or more bases in the contig is recurrent, and therefore included in P. Ref. Supp. Figure 2 shows the total length of covered contigs for each sample, the number of recurrent bases (equivalent with the number of bases aligning to P. Ref.) and the number of bases in each sample that align to bases in PGF.

There are substantial differences in the numbers of covered bases between the samples. These differences are largely explained by the use of different sequencing technologies: the Venter and the haploid GRCh37 samples were generated using Sanger sequencing and have larger average numbers of covered bases (avg. 4,162,770 bp) than the remaining samples (avg. 3,168,268 bp), which all used Illumina short read (avg. 47bp in this study) sequencing technology [3]. However, there are two clear exceptions to this pattern. The Sanger sequenced

APD sample has noticeably fewer such bases (2,320,747 total bp). Conversely, the Illumina sequenced NA12878 sample (4,192,579 total bp) has similar coverage with the Sanger sequenced samples, probably because it had higher sequencing coverage and a greater variety of paired end libraries than the other Illumina samples [5].

For a sample, the difference between the number of covered bases and the number of recurrent bases is the number of non-recurrent bases in the MSA. These are bases which are either part of relatively rare segregating polymorphisms, or erroneous due to mis-alignment or mis-assembly. Encouragingly, the Chimpanzee sample has by far the largest number of non-recurrent bases in the MSA, a total of 185,330 bp (3.31% of covered bp); in contrast the average human sample has only 19,362 such bases (0.52% of covered bp). Summing across the human samples, a total of 309,896 bp are non recurrent (5.54% of all columns and non-recurrent bp).

6 Identifying SNVs and Indels in the MSA

Given the MSA and a designated reference sample, generally either P. Ref. or PGF, we made variation predictions for each of the other input samples. This was achieved by analysing columns and subgraphs of the MSA. We now describe the prediction of SNVs, indels and nonlinear breakpoints.

6.1 SNVs

For each column containing a position from an input sample and a position from a chosen reference we predict a SNV for the sample with respect to the reference if the oriented bases of the two positions differ. The set of SNVs for a given input sample and reference is then the set of all such SNVs.

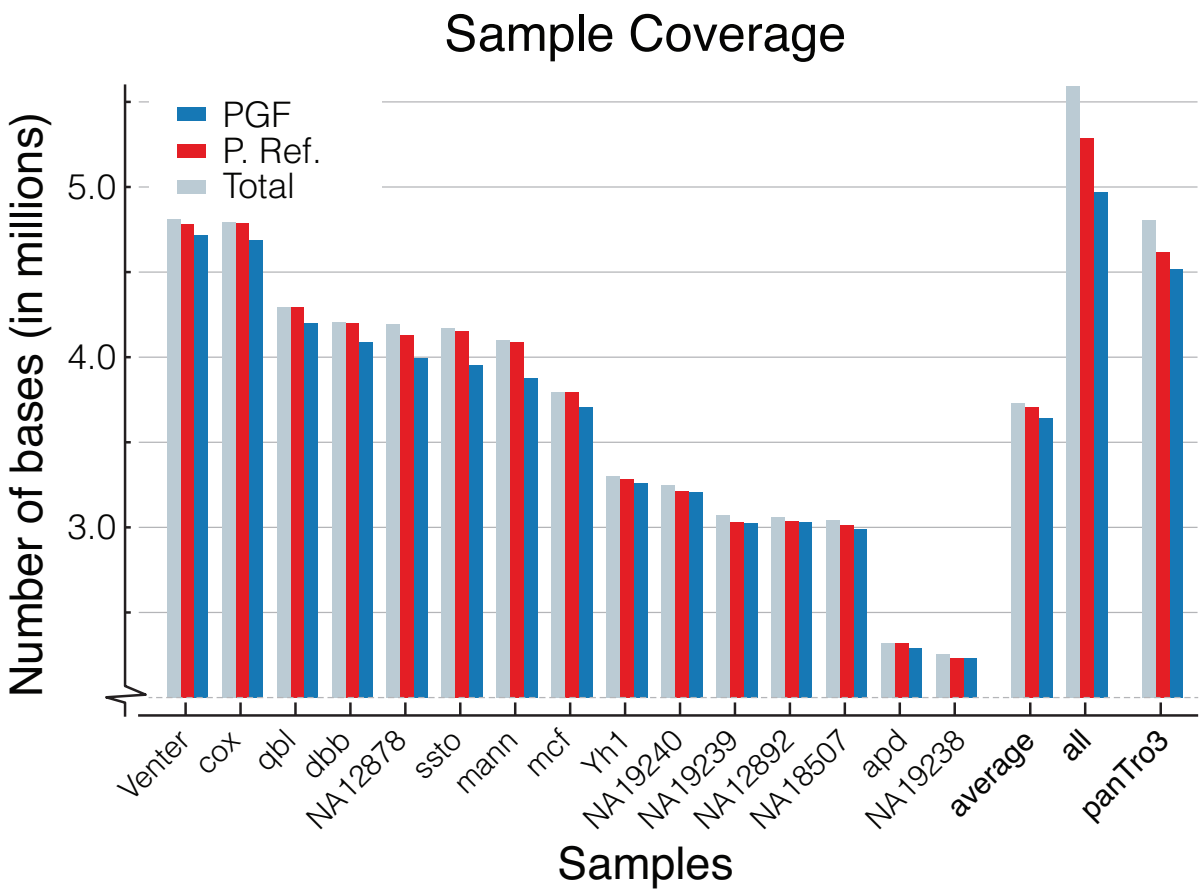

Fig. 2. P. Ref. contains ∼6% of recurrent bases that are not represented in PGF, and Sanger sequenced samples have higher coverage than Illumina sequenced samples. The columns show the number of bases from each sample mapped to the Cactus MSA, to P. Ref. and to PGF. The 'average' category gives the average over all human samples. The 'all' category considers all columns and unaligned bases in all the human samples, i.e. as 1 base per homology set.

6.2 Indels

Let $G = (V, E)$ be the bidirected graph for our MSA constructed such that $\alpha = 1$. Let $G' = (V' \subset V, E' \subset E)$ be the subgraph of G containing only nodes and direct adjacency edges representing a chosen reference sample R and input sample T. Due to the synteny partitioning process, any column in G' represents at most one position from each of R and T, while because $\alpha = 1$ every unaligned position is also represented by a node.

Let C be an $M, 2$ cycle in G'. C has one positive node, A, and one negative node, B, and both must contain positions from R and T. C can be subdivided into two paths P_1 and P_2

that both include A and B, but are otherwise disjoint. Let x_i and x_j be a pair of positions such that $[x_i] = A$, $[x_j] = B$ and P_1 represents $x_i <_S x_j$. Similarly, let y_k and y_l be a pair of positions such that $[y_k] = A$, $[y_l] = B$ and P_2 represents $y_k <_S y_l$. Without loss of generality assume that x is the sequence in R and y is the sequence in T. If $j - i > 1$ then we count a deletion in T with respect to R of length $j - i$. Similarly if $l - k > 1$ then we count an insertion in T with respect to R of length $l - k$.

7 dbSNP/1000 Genomes Comparisons

We compared our SNV and short ($\leq$ 10 bp) indel predictions made with respect to PGF to the intersection of the dbSNP database (build 134) [10] and the 1000 Genomes project (1KGP) data (release 20110521) [11].

The dbSNP data for the GRCh37 MHC was obtained from the UCSC Genome Browser 'snp134' table (assembly GRCh37/hg19) for the region chr6:28477754-33448354. The SNVs included were all records classified as 'single' or 'MNP' (Multiple Nucleotide Polymorphism). We considered each base of the MNPs as equivalent to one SNV. The short insertions include records classified as 'insertion' or 'in-del' with length $\leq$ 10 bases. Similarly, the short deletions include records classified as 'deletion' or 'in-del' with length $\leq$ 10 bp.

The 1KGP data was obtained from `ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/`, again including all records within the region chr6:28477754-33448354.

A variant predicted by the Cactus MSA and present in the intersection of the dbSNP and the 1KGP datasets was called *true positive*.

7.1 A Comparison of the MSA's Variation Predictions to the dbSNP/1KGP Data Confirms the High Accuracy of the MSA

In overview, the MSA made 56,080 distinct SNV predictions relative to PGF, of which 42,584 (76%) were confirmed by dbSNP/1KGP. This accounts for 28% of all SNVs currently in the

dbSNP/1KGP data. Given that there were only 15 samples other than PGF used in this study, observing 28% of the population variation is significant.

One important set of predicted SNVs are those that are present in P. Ref, as these reflect differences between the pan-genome reference sequence and PGF. Approximately 97% of such SNVs are contained in the dbSNP/1KGP data, leaving 264 total possible "false positives" (false positive with respect to the dbSNP/1KGP). The majority of these (91% or 241 SNVs) occurred in bases that were labeled as either being repetitive or proximal to a breakpoint in one or more of the samples. Repetitive regions and breakpoint vicinities are challenging cases and often result in multiple equivalent solutions in alignment. Therefore, it is expected to observe disagreements between the MSA and the dbSNP/1KGP data in these regions. A careful manual analysis of the other 23 non-repetitive, non-breakpoint-proximal cases of the 264 false positives revealed supporting evidence for 21 of them (Supp. Subsection 7.2). Overall, the results confirm the SNVs between P.Ref. and PGF and validate P.Ref.'s quality.

We see a similar picture with short indels to that with SNVs, but a generally higher level of disagreement between the MSA predictions and the dbSNP/1KGP data. Overall, the MSA made 22,360 indel predictions of which 14,575 (65%) were confirmed, accounting for 34% of all short indels currently in the dbSNP/1KGP data.

False Positive SNVs Likely Resulted From the Assembly Quality of the Illumina Sequenced Samples and Reduced by the Recurrence Requirement

With 76% of the total predicted SNVs confirmed by dbSNP, there were 24% of false positives. A large proportion of these false positives came from the Illumina sequenced samples (Supp. Figure 3). On average, the Sanger sequenced samples had a 98% true positive rate in comparison to only a 78% rate in the Illumnia sequenced ones.

To further investigate these false positive SNVs we subdivided the MSA predictions into categories based upon: First, their presence outside of a sequence of PGF annotated as repetitive, and therefore hard to correctly assign homology to; second, their distance

from a breakpoint within the MSA, which might result in misalignment, and third, whether they were recurrent, i.e. predicted by multiple samples and therefore were less likely to be erroneous (Supp. Figure 3).

Being outside of repetitive sequence and more than 5 bp from a breakpoint results in a small positive increase in the average true positive rate (2.7% and 3.1% increase, respectively); combined this effect is even stronger (4.9% increase). SNVs within repeats and near breakpoints are therefore likely to be genuine candidates for misalignment and consequent false SNV prediction.

Being recurrent had a small effect on the Sanger samples (avg. 0.6% increase in true positive rate), but a huge effect on the Illumina sequenced samples (avg. 17% increase in true positive rate). Looking at only recurrent SNVs, the overall true positive rate is 95% and is similar between Sanger (avg. 98% per sample) and Illumina sequenced samples (avg. 96% per sample). Looking only at recurrent SNVs also only leads to a 13% average reduction in the total number of SNVs called, a much smaller reduction than that of looking at SNVs only in non-repeat regions (54%) or SNVs not proximal to a breakpoint (19%). The high accuracy in the SNV predictions of the Sanger sequenced samples, together with the significant improvement in accuracy of the Illunima sequenced samples' SNV predictions when the "recurrent" condition was required, suggest that most of the false positives may be attributed to sample's assembly quality (sequencing or assembling errors) and not alignment errors.

Assessing false negative rates is harder as SNV and indel calls for the individual GRCh37 haplotypes were not available. However, given the high true positive rate, I estimate the false negative rate for each of these samples by assuming the MSA predictions are correct and using the total number of previously reported SNV predictions [1]. Given this caveat, in the haploid (Sanger) samples I see an average false negative rate of 2% per sample. In the diploid samples I see an average false negative rate of 59.5% per sample, which is reasonable given that, as mentioned, I expect to miss half of all their variants.

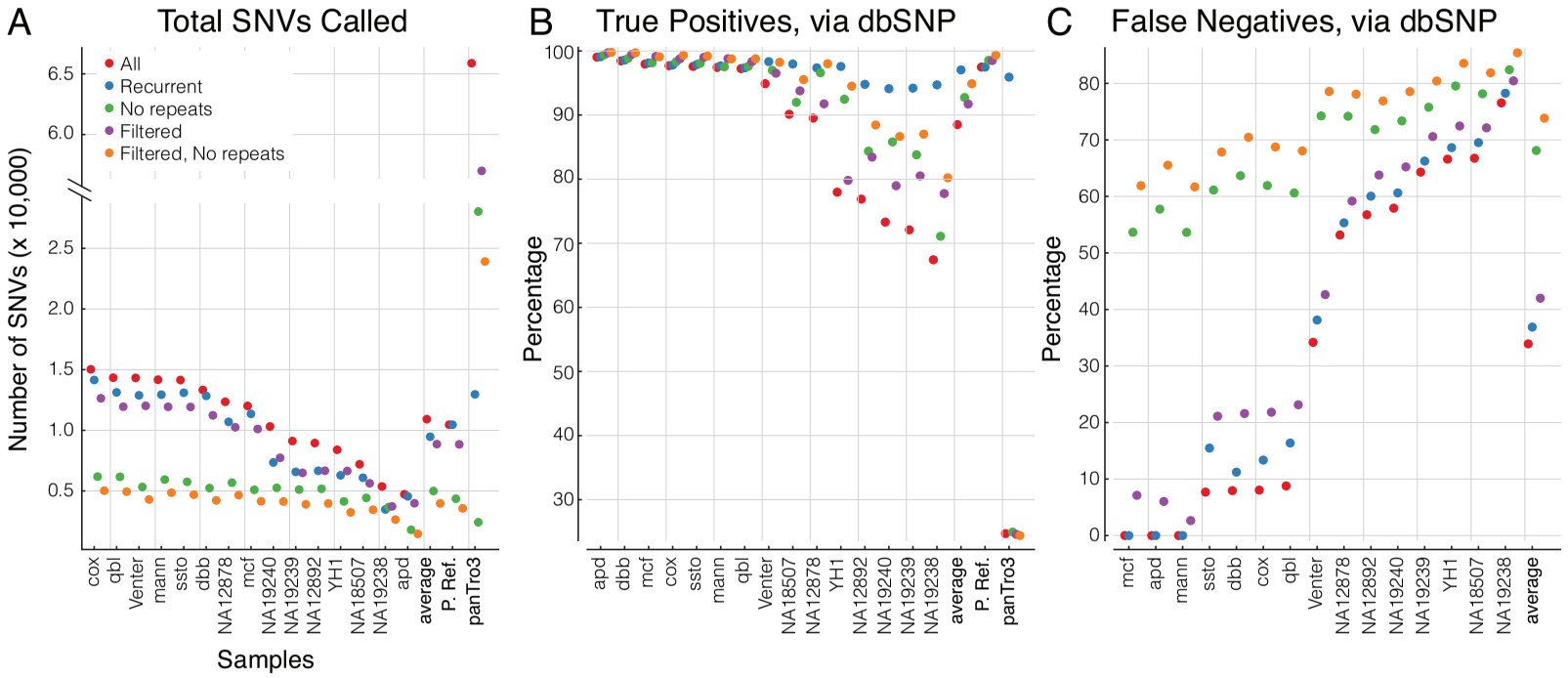

Fig. 3. A detailed comparison of the SNVs predicted by the Cactus MSA to those in dbSNP shows high accuracy in variation predictions of the Sanger sequenced samples and in recurrent variation predictions of the Illumina sequenced samples. (A) The total number of SNVs predicted by the MSA with respect to PGF. (B) The proportion of SNVs predicted by the MSA with respect to PGF already present in dbSNP. (C) The proportion of (previously reported) SNVs for a given sample in dbSNP not predicted by the MSA. Key gives categories of SNVs predicted by the MSA; 'All': All SNVs, 'Recurrent': All SNVs present in at least two samples, 'No repeats': Excluding SNVs at bases labeled repetitive in PGF, 'Filtered': Excluding SNVs within columns that are within 5 bp of a breakpoint in any sequence, 'Filtered, No Repeats': Intersection of previous two categories.

7.2 Manual Analysis of False Positive SNVs

We manually analysed P. Ref. false positives (not in dbSNP or the 1KGP data) SNVs with respect to PGF using the UCSC Genome Browser [12] and the Browser's unpublished MULTIZ [13] multiple sequence alignment of the GRCh37 haplotypes.

We separated the SNVs into five categories: 'confirmed', 'dbSNP bug', 'alignment disagreement', 'recurrent', and 'single'. SNVs were labelled 'confirmed' if they were observed in the MULTIZ MSA. SNVs that were not in dbSNP build 134 due to a bug in dbSNP were labelled 'dbSNP bug'; we reported this bug and it has now been fixed in dbSNP build 135. SNVs were labelled 'alignment disagreement' when the MULTIZ alignment opted for indels and the Cactus MSA opted for substitutions. SNVs were 'recurrent' if there was no evidence from the MULTIZ MSA (often due to missing data or non-aligned sequences), but the SNVs were present in two or more samples from the Cactus MSA. SNVs were 'single' if there was no evidence from the MULTIZ MSA and the SNVs were present in only one sample.

Chrom	Start	P.Ref. Allele	PGF Allele	Annotation
chr6	30463203	T	G	C
chr6	30463204	T	G	C
chr6	32535163	C	G	C
chr6	32535165	G	C	C
chr6	32535237	A	G	C
chr6	32536032	A	T	C
chr6	32536033	C	A	C
chr6	32536034	C	T	C
chr6	32536962	C	T	B
chr6	32547908	G	A	ID
chr6	32548727	C	G	S
chr6	32550935	G	A	B
chr6	32551306	G	A	R
chr6	32551307	A	G	R
chr6	32551852	C	G	S
chr6	32553317	A	C	B
chr6	32570019	T	C	C
chr6	32633098	A	C	B
chr6	32633101	T	C	B
chr6	32633906	T	C	B
chr6	32689731	G	C	C
chr6	32689732	C	T	ID
chr6	32689733	A	T	ID

Table 1. A manual analysis of P.Ref. non-repetitive and filtered SNVs with respect to PGF that were not in dbSNP or 1000 Genome Project data. These SNVs were neither within the repetitive regions (non-repetitive) nor proximal to a breakpoint (filtered). 'Chrom', 'Start': location of each SNV relative to the positive strand of GRCh37. Annotation: 'C': SNVs were confirmed by an independent MULTIZ multiple sequence alignment (see Supplementary Section 7.2). 'B': a bug in dbSNP build 134 that had been fixed in build 135, SNVs were indeed in dbSNP. 'ID': disaggreement between Cactus MSA and other alignments, in which Cactus MSA called substituitions while other alignments called indels. 'R': SNVs were not confirmed by MULTIZ MSA but recurrent within the input samples. 'S': SNVs were not confirmed by MULTIZ MSA and not reccurent (single).

8 Gene Mapping

To assess how genes mapped to P. Ref. in comparison to PGF, we aligned RefSeq [14]) transcripts and Genbank RNAs [15] to the GRCh37 assembly (excluding the alternative

loci) and to a hybrid GRCh37/P. Ref. assembly, which was the GRCh37 assembly with the MHC region replaced by the P. Ref. sequence. The alignments were done using Blat [16], version 34x10. For each sequence the best alignments were chosen using the *pslCDnaFilter* program available in the UCSC Genome Browswer source code. Alignments with less than 95% base identity were discarded. In addition, two different coverage filterings, 90% and 95%, were applied, which respectively required that the alignments covered at least 90% and 95% of the RNA's/transcript's bases to be kept. RefSeq and Genbank RNAs were obtained from the UCSC Genome Browser tables, *refGene* and *all_mrna*, respectively, GRCh37 assembly.

HLA-DRB pseudogenes genes that were not in the RefSeq database (HLA-DRB2, HLA-DRB7, HLA-DRB8, HLA-DRB9) were mapped to P. Ref. using Blat with identity $\geq 95\%$. Sequences for these genes were obtained from the UCSC Genome Browser, using the predicted mRNA sequences generated by Ensembl (HLA-DRB2: ENST00000419200, HLA-DRB7:ENST00000422566, HLA-DRB8:ENST00000436297, HLA-DRB9: ENST00000449413).

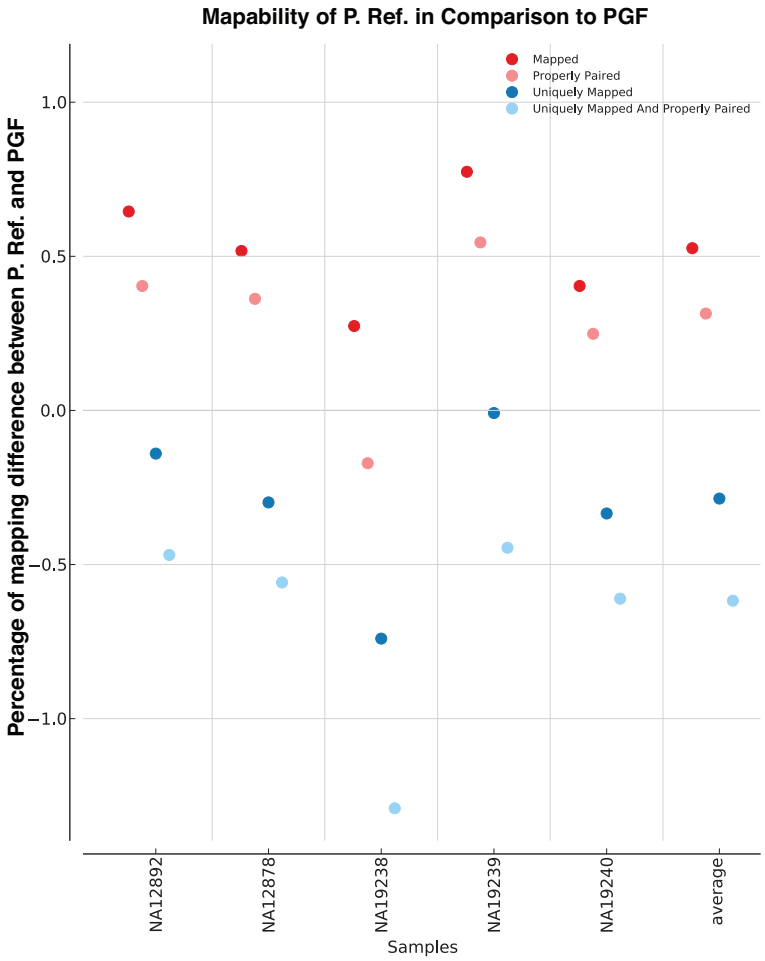

Fig. 4. A UCSC Browser display [12] of the MHC HLA-DRB hypervariable region in a prototype P. Ref. MHC reference browser. Self Align tracks: Alignment of the region against itself with a 90% and 95% minimum identity threshold. It demonstrates that much of the region is homologous to itself at 90% identity, but very little at 95%, which is substantially below the threshold required in the MSA to create homology. Gene tracks: Genes identified by alignment and using RefSeq annotations (see Supplementary Section 8). Snake tracks: Subsequences of contiguous bases aligned to the reference are shown as rectangles. SNVs with respect to the reference are coloured red, otherwise bases are coloured light blue. The lines connecting the rectangles show adjacencies between the bases. In addition to genes that are present in PGF (known genes HLA-DRB5, HLA-DRB1 and pseudogenes HLA-DRB9, HLA-DRB6), P. Ref. also contains genes that are recurrent in the input samples (HLA-DRB3, HLA-DRB4 and pseudogenes HLA-DRB2, HLA-DRB7, HLA-DRB8). The MSA shows clearly the relationship of the samples in the region, e.g COX and QBL have the same DRB group and are grouped together. Lines coloured orange indicate adjacencies that contain unaligned bases only present in one sample.

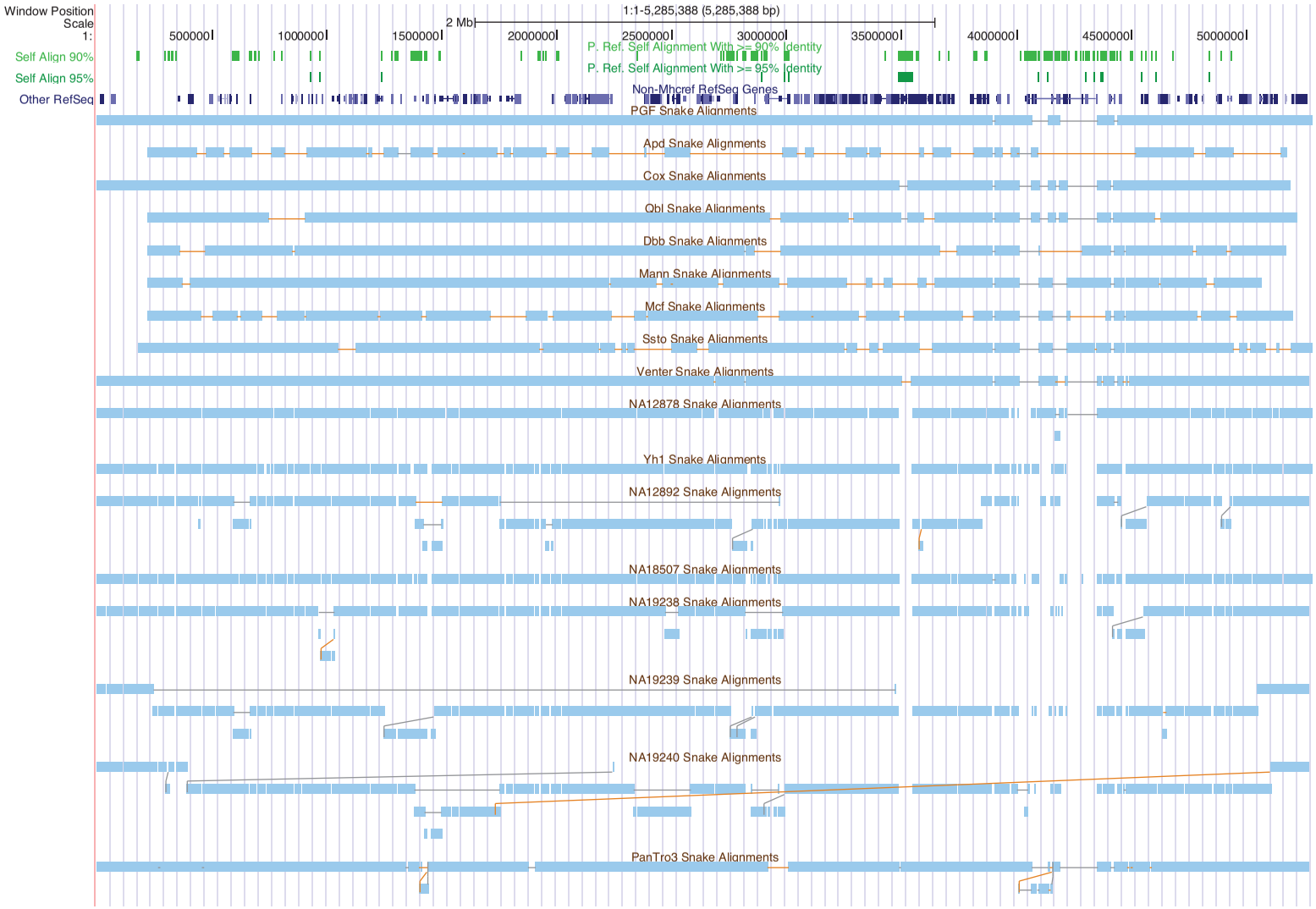

Fig. 5. A UCSC Browser screenshot showing a prototype P. Ref. MHC reference browser. The figure is arranged similarly to Figure 4

	90%		95%	
	P. Ref.	PGF	P. Ref.	PGF
Genbank RNA	3209	2986	3095	2881
RefSeq transcript	371	368	370	367
RefSeq genes all tx mapped	213	210	212	209
RefSeq genes ≥ 1 tx mapped	213	211	212	210

Table 2. Statistics on RNAs and RefSeq transcripts mapping to either references, GRCh37 or P. Ref. Columns: '90%': RNAs must have at least 90% bases aligned to the reference, '95%': RNAs must have at least 95% bases aligned to the reference, 'P. Ref. ': the reference is P. Ref., 'GRCh37': the reference is GRCh37 MHC main locus. Rows: 'Genbank RNA': number of Genbank RNAs mapped best to the appropriate reference with the appropriate base coverage. 'RefSeq transcript': similar to 'Genbank RNAs' but for RefSeq transcripts instead of Genbank RNAs. 'RefSeq genes all tx mapped': number of RefSeq genes that have all the transcripts mapped best to the appropriate reference with the appropriate base coverage. 'RefSeq genes ≥ 1 tx mapped': number of RefSeq genes that have at least one transcript mapped best to the appropriate reference with the appropriate base coverage.

9 Short Read Mapping

To test P. Ref. as a target for mapping experiments, we constructed versions of P. Ref. excluding a held out 1000 Genomes sample and then compared mappings made with BWA of the held out sample to PGF and the held out P. Ref. The mappings include Illumina reads that were mapped to PGF and Illumina unmapped reads (see Supplementary Section 3 for the data source). Unpaired and paired reads were mapped using *bwa samse* (*-n 10000*) and *bwa sampe* (*-n 10000 -N 10000 -a 1000*), respectively. The *bwa* version was 0.5.9-r16.

Sample	MD Reads	Total MD bases	% Repeats	SNV Rate	dbSNP ESR	bcftools ESR
NA12878	52,731	354,575	74.08	0.0144	2.6	3.2
NA12892	30,542	264,912	73.63	0.0100	2.7	2.8
NA19238	31,335	249,473	71.34	0.0088	3.0	2.7
NA19239	40,308	266,904	71.77	0.0161	2.8	3.5
NA19240	50,928	327,527	72.71	0.0128	2.7	2.9
average	41,168	292,678	72.70	0.0124	2.7	3.1

Table 3. An analysis of PGF mapping discordant reads show that these reads map mostly to repetitive regions that have an enrich in SNVs called by dbSNP/1KGP. 'MD Reads': Total PGF mapping discordant reads. 'Total MD bases': Total mapping discordant bases in PGF. '% Repeats': Proportion of mapping discordant bases in PGF classified as repetitive. 'SNV Rate': Number of SNVs predicted by dbSNP/1KGP per mapping discordant base in PGF. 'dbSNP ESR': dbSNP 'Enriched SNV Ratio', ratio of mapping discordant SNV rate (previous column) over overall SNV rate predicted by dbSNP/1KGP. 'bcftools ESR': bcftools 'Enriched SNV Ratio', ratio of mapping discordant SNV rate over overall SNV rate predicted by bcftools.

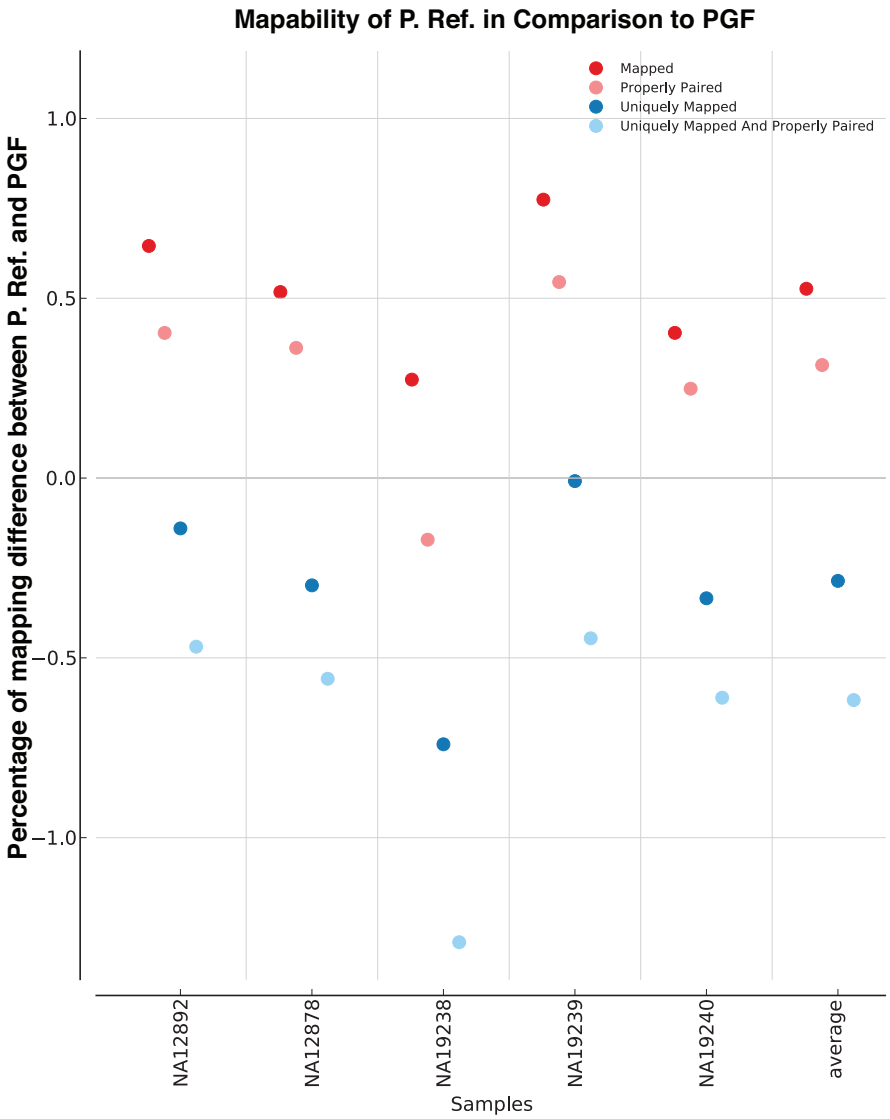

Fig. 6. A comparison of short read mapping to P. Ref. and to PGF shows that P. Ref. has slightly more mapped and properly paired reads and slightly less uniquely mapped reads. The y-axis is the ratio of the difference between the number of reads of a sample mapping to a P. Ref. constructed without the sample in question and the number of reads of the sample mapping to PGF over the number of reads of the sample mapping to PGF. 'Mapped': For all reads. 'Properly Paired': For paired reads (see methods for definition of 'proper pairing'). 'Uniquely Mapped': As 'Mapped', but ignoring reads that map to multiple region equally well. 'Uniquely Mapped and Properly Paired': As 'Properly Paired', but ignoring reads that map to multiple region equally well.

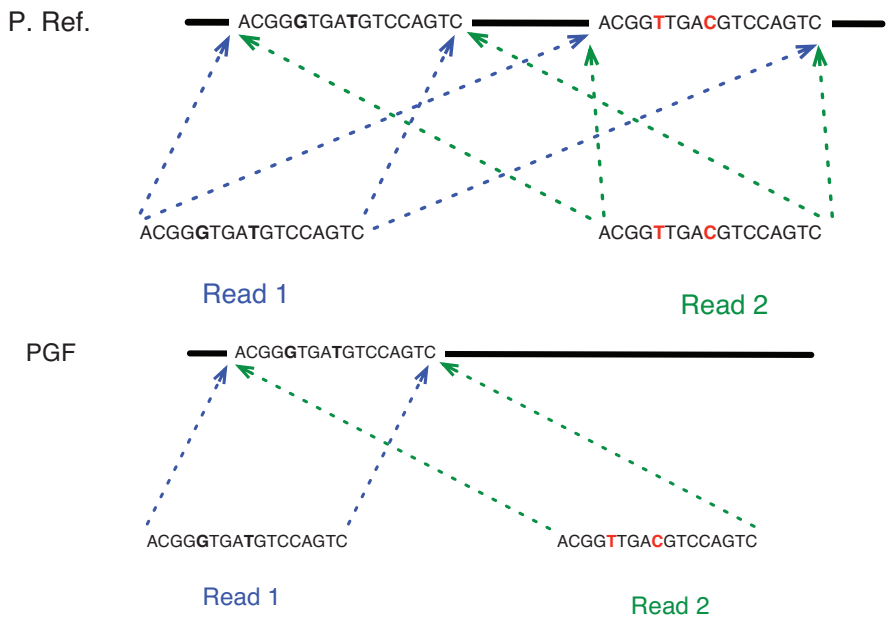

Fig. 7. An example scenario of reads mapping to a paralog when the true ortholog is missing. Here, P. Ref. contains the orthologous sequence of Read 2 and therefore results in non-unique mapping. PGF does not have orthologous sequence of Read 2 and results in the mis-mapping of Read 2 to the orthologous sequence of Read 1, resulting in unique-mapping, but higher SNPs.

References

[1] Horton, R., Gibson, R., Coggill, P., Miretti, M., Allcock, R.J., Almeida, J., Forbes, S., Gilbert, J.G.R., Halls, K., Harrow, J.L., Hart, E., Howe, K., Jackson, D.K., Palmer, S., Roberts, A.N., Sims, S., Stewart, C.A., Traherne, J.A., Trevanion, S., Wilming, L., Rogers, J., de Jong, P.J., Elliott, J.F., Sawcer, S., Todd, J.A., Trowsdale, J., Beck, S.: Variation analysis and gene annotation of eight mhc haplotypes: the mhc haplotype project. Immunogenetics **60**(1) (Jan 2008) 1–18

[2] Church, D.M., Schneider, V.A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.C., Agarwala, R., McLaren, W.M., Ritchie, G.R.S., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., Matthews, L., Whitehead, S., Chow, W., Torrance, J., Dunn, M., Harden, G., Threadgold, G., Wood, J., Collins, J., Heath, P., Griffiths, G., Pelan, S., Grafham, D., Eichler, E.E., Weinstock, G., Mardis, E.R., Wilson, R.K., Howe, K., Flicek, P., Hubbard, T.: Modernizing reference genome assemblies. PLoS Biol **9**(7) (Jul 2011) e1001091

[3] Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Cheetham, R.K., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M.J., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M.D., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Catenazzi, M.C.E., Chang, S., Cooley, R.N., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fajardo, K.V.F., Furey, W.S., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Jones, T.A.H., Kang, G.D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ng, B.L., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Pike, A.C., Pinkard, D.C., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Rodriguez, A.C., Roe, P.M., Rogers, J., Bacigalupo, M.C.R., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Sohna, J.E.S., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley,

G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R., Smith, A.J.: Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456**(7218) (Nov 2008) 53–9

[4] Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., Hu, Y., Yang, Z., Zheng, H., Hellmann, I., Inouye, M., Pool, J., Yi, X., Zhao, J., Duan, J., Zhou, Y., Qin, J., Ma, L., Li, G., Yang, Z., Zhang, G., Yang, B., Yu, C., Liang, F., Li, W., Li, S., Li, D., Ni, P., Ruan, J., Li, Q., Zhu, H., Liu, D., Lu, Z., Li, N., Guo, G., Zhang, J., Ye, J., Fang, L., Hao, Q., Chen, Q., Liang, Y., Su, Y., San, A., Ping, C., Yang, S., Chen, F., Li, L., Zhou, K., Zheng, H., Ren, Y., Yang, L., Gao, Y., Yang, G., Li, Z., Feng, X., Kristiansen, K., Wong, G.K.S., Nielsen, R., Durbin, R., Bolund, L., Zhang, X., Li, S., Yang, H., Wang, J.: The diploid genome sequence of an asian individual. Nature **456**(7218) (Nov 2008) 60–5

[5] Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S., Jaffe, D.B.: High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA **108**(4) (Jan 2011) 1513–8

[6] Zerbino, D.R., Birney, E.: Velvet: algorithms for de novo short read assembly using de bruijn graphs. Genome Res **18**(5) (May 2008) 821–9

[7] Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., Haussler, D.: Cactus: Algorithms for genome multiple sequence alignment. Genome Res **21**(9) (Sep 2011) 1512–28

[8] Cantor, T.H., Cantor, C.R.: Evolution of protein molecules. Mammalian protein metabolism **1** (1969) 22–123

[9] Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S., Reich, D.: Genetic evidence for complex speciation of humans and chimpanzees. Nature **441**(7097) (Jun 2006) 1103–8

[10] Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Sirotkin, K.: dbsnp: the ncbi database of genetic variation. Nucleic Acids Res **29**(1) (Jan 2001) 308–11

[11] 1000-Genomes-Project-Consortium: A map of human genome variation from population-scale sequencing. Nature **467**(7319) (Oct 2010) 1061–73

[12] Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T.R., Giardine, B.M., Harte, R.A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R.M., Learned, K., Li, C.H., Meyer, L.R., Pohl, A., Raney, B.J., Rosenbloom, K.R., Smith, K.E., Haussler, D., Kent, W.J.: The ucsc genome browser database: update 2011. Nucleic Acids Res **39**(Database issue) (Jan 2011) D876–82

[13] Miller, W., Rosenbloom, K., Hardison, R.C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D.C., Baertsch, R., Blankenberg, D., Pond, S.L.K., Nekrutenko, A., Giardine, B., Harris, R.S., Tyekucheva, S., Diekhans, M., Pringle, T.H., Murphy, W.J., Lesk, A., Weinstock, G.M., Lindblad-Toh, K., Gibbs, R.A., Lander, E.S., Siepel, A., Haussler, D., Kent, W.J.: 28-way vertebrate alignment and conservation track in the ucsc genome browser. Genome Res **17**(12) (Dec 2007) 1797–808

[14] Pruitt, K.D., Tatusova, T., Brown, G.R., Maglott, D.R.: Ncbi reference sequences (refseq): current status, new features and genome annotation policy. Nucleic Acids Res **40**(1) (Jan 2012) D130–5

[15] Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J., Sayers, E.W.: Genbank. Nucleic Acids Res **40**(1) (Jan 2012) D48–53

[16] Kent, W.J.: Blat–the blast-like alignment tool. Genome Res **12**(4) (Apr 2002) 656–64