64,976 whole genome haplotypes from the Haplotype Reference Consortium and efficient algorithms to use them

Richard M. Durbin[1], Warren Kretzschmar[2], Shane A. McCarthy[1], Sayantan Das[3], Petr Daněček[1], Christian Fuchsberger[3],Olivier Delaneau[4], Hyun Min-Kang[3], Gonçalo Abecasis[3], Jonathan Marchini[2] on behalf of the Haplotype Reference Consortium

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK
[3]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109 [4]Département de Génétique et Développement, Faculté de Médecine, Université de Genéve, Geneva, Switzerland

Introduction

Genotype imputation infers missing genotypes in samples from a reference panel, so can fill in full sequences from genome-wide association study (GWAS) data. It is central to modern genetic association studies, supporting meta-analysis and increasing coverage and power. The Haplotype Reference Consortium (HRC) has combined whole genome sequence data from **32,488 samples** from 20 studies to build an imputation reference panel at **39,235,157 single nucleotide variant sites** with allele frequencies down to 0.01% (1/10,000). We provide internet-accessible servers for imputation from whole genome genotype data, and show that we can impute at the same accuracy sites with up to an order of magnitude lower allele frequency than 1000 Genomes Phase 3.

We also describe `pbwt` [1], a new extremely rapid and scalable imputation software using the positional Burrows-Wheeler transform (PBWT).

The Haplotype Reference Consortium Release 1

Contributing datasets

This first release of the Haplotype Reference Consortium reference panel (HRC.r1) combines data from 20 large whole-genome sequencing projects. The majority of these cohorts consist of samples of European origin making the panel particularly suitable for imputing into samples of European ancestry.

Cohort	Samples	Avg. coverage	Nationality
UK10K	3,715	6.5×	UK
Sardinia	3,445	4×	Italy
IBD	4,478	4×/2×	UK
GoT2D	2,710	4× + Exome	US
BRIDGES	2,487	6-12×	US
1000 Genomes Phase 3	2,495	4× + Exome	Various
GoNL	748	12×	Netherlands
AMD	3,222	4×	US
HUNT	1,023	4×	US
SiSu + Kuusamo	1,918	4×	Finland
INGI-FVG	250	4-10×	Italy
INGI-Val Borbera	225	6×	Italy
MCTFR	1,325	10×	US
HELIC	247	4×/1×	Greece
ORCADES	398	4×	UK, Orkney
inCHIANTI	676	7×	UK
GECCO	1,131	4-6×	US
GPC	697	30×	US
ProjectMinE	935	45×	Netherlands
NEPTUNE	403	4×	US
	32,488		

Table 1: Cohorts contributing to the HRC.r1 reference panel.

Data Processing

- → Take initial calls from each cohort.
- → Select calls with overall minor allele count (MAC)≥5 and carry out cross cohort QC/consistency, removing some sites and samples (see Figure 1).
- → Obtain genotype likelihoods (GLs) at these sites from the original BAMs using SAMtools.
- → Recall haplotypes from GLs using a modified version of SNPtools [2] that restricts the parental haplotype search based on the original haplotypes.
- → Finally re-phase with SHAPEIT and remove sites with MAC<5 in new call set.

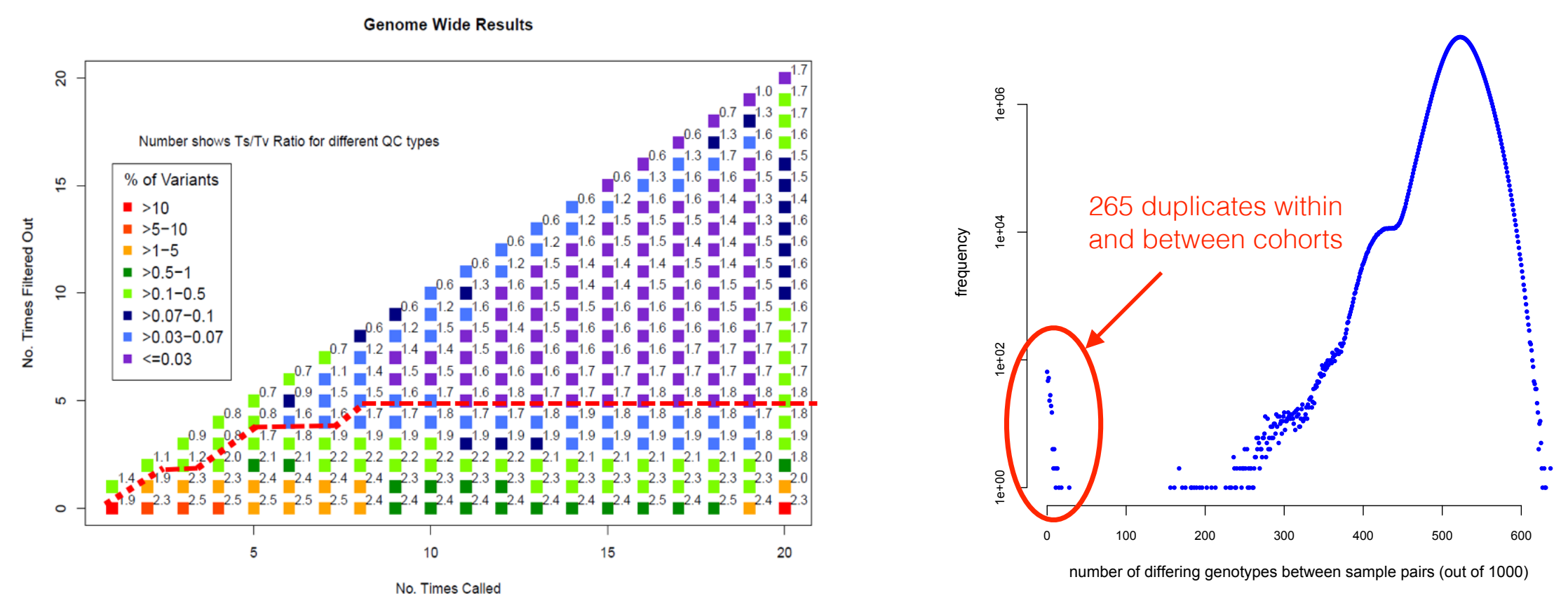


Figure 1: Site and sample filtering. Left: One site filter was to remove sites filtered out by 5 or more of the original cohorts (fewer if originally called in fewer); Right: we removed 265 duplicate samples.

Improved imputation performance over existing panels

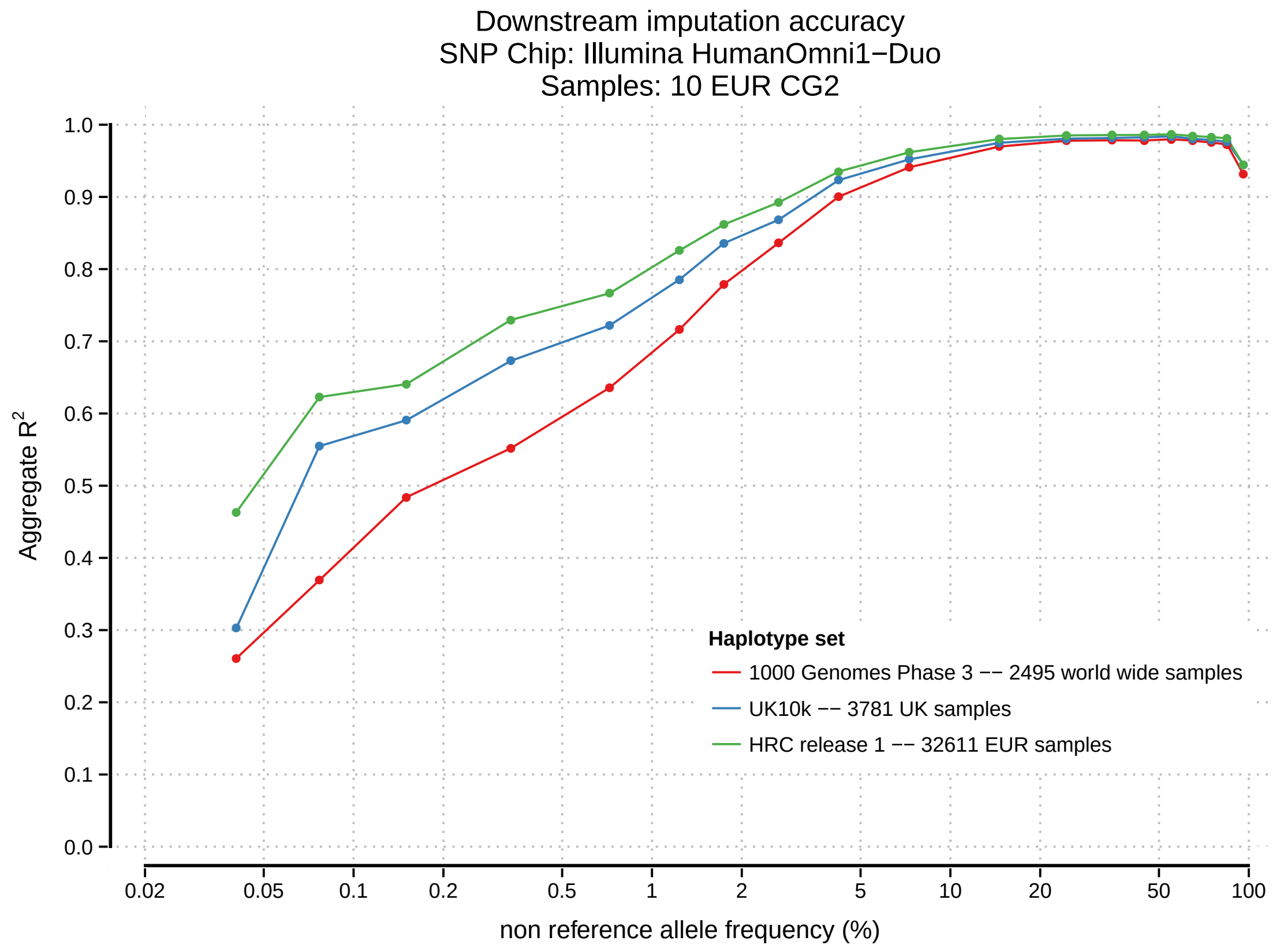


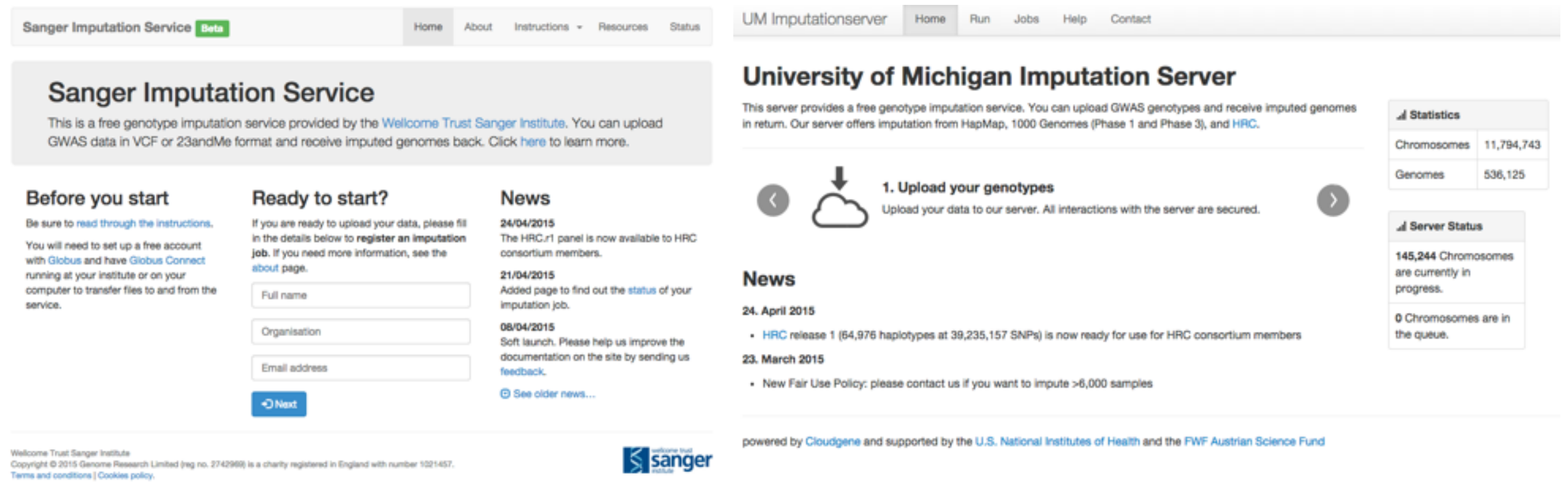
Figure 2: Imputation performance on various panels. Pseudo-GWAS genotypes created by selecting genotypes at Illumina HumanOmni1Duo chip sites from chromosome 20 of 10 CEU deep sequenced samples. Imputation evaluated by comparing high coverage genotype calls not in the selected chip sites with those imputed from the various panels using IMPUTE2.

- → HRC.r1 provides substantial increase over 1000 Genomes Phase 3 imputation (5-10× lower MAF).
- → Improvement over UK10K and other large single study panels while also greatly increasing the density of imputable sites (∼22M for UK10K to ∼40M for HRC.r1).

Using the resource

Making the full reference panel publicly available is not possible at the moment due to restrictions on data access, so we are making the panel available to users via **public imputation servers** at the Sanger Institute and at the University of Michigan. Users upload phased or unphased GWAS data and imputed genotypes and dosages are returned.

The imputation engine used at the Sanger Institute is `pbwt` (see final section of this poster), while at the University of Michigan it is `minimac3` [3]. Both servers optionally pre-phase GWAS data before imputation using SHAPEIT2 [4].



Sanger Imputation Service

This is a free genotype imputation service provided by the Wellcome Trust Sanger Institute. You can upload GWAS data in VCF or 23andMe format and receive imputed genomes back. Click here to learn more.

Before you start

Be sure to read through the instructions.

You will need to set up a free account with Globus and have Globus Connect running at your institute or on your computer to transfer files to and from the service.

Ready to start?

If you are ready to upload your data, please fill in the details below to register an imputation job. If you need more information, see the about page.

News

24/04/2015
The HRC.r1 panel is now available to HRC consortium members.

21/04/2015
Added page to find out the status of your imputation job.

08/04/2015
Soft launch. Please help us improve the documentation on the site by sending us feedback.

See older news...

Wellcome Trust Sanger Institute
Copyright © 2015 Genome Research Limited (reg no. 2742969) is a charity registered in England with number 1021457.
Terms and conditions | Cookies policy

University of Michigan Imputation Server

This server provides a free genotype imputation service. You can upload GWAS genotypes and receive imputed genomes in return. Our server offers imputation from HapMap, 1000 Genomes (Phase 1 and Phase 3), and HRC.

1. Upload your genotypes
Upload your data to our server. All interactions with the server are secured.

News

24. April 2015
- HRC release 1 (64,976 haplotypes at 39,235,157 SNPs) is now ready for use for HRC consortium members

23. March 2015
- New Fair Use Policy: please contact us if you want to impute >6,000 samples

powered by Cloudgene and supported by the U.S. National Institutes of Health and the FWF Austrian Science Fund

Statistics

Chromosomes	11,794,743
Genomes	536,125

Server Status

145,244 Chromosomes are currently in progress.

0 Chromosomes are in the queue.

Figure 3: Imputation servers. Left: the Sanger Institute server (`imputation.sanger.ac.uk`) uses `pbwt`. Right: The University of Michigan server (`imputationserver.sph.umich.edu`) uses `minimac3`.

The HRC.r1 panel is currently available via the imputation servers to HRC contributors and will be available openly to all on July 1st.

HRC Future Plans

Chromosome X The currently available panel only contains the autosomes. We are currently working on adding in chromosome X.

Availability in the EGA We intend to deposit a subset of data in EGA with managed access for imputation only.

GRCh38 We are planning to provide a GRCh38 version by liftover in the first instance.

Release 2 Later in 2015 we will begin building release 2 with more data representing more populations outside of Europe including Africa, Asia and the Americas. If you have a cohort with sequencing data that you would like to see incorporated, please get in touch.

Positional Burrows-Wheeler Transform (PBWT) and imputation

The positional Burrows-Wheeler transform (PBWT) provides a compressed representation of a haplotype panel in which variant calls are implicitly available in sorted order at each site, sorted based on the preceding sequence [1]. The entire PBWT data for HRC release 1 is 3.9GB at 0.0015 bytes per genotype, compared to 70GB for the gzipped `.hap` files used by IMPUTE2.

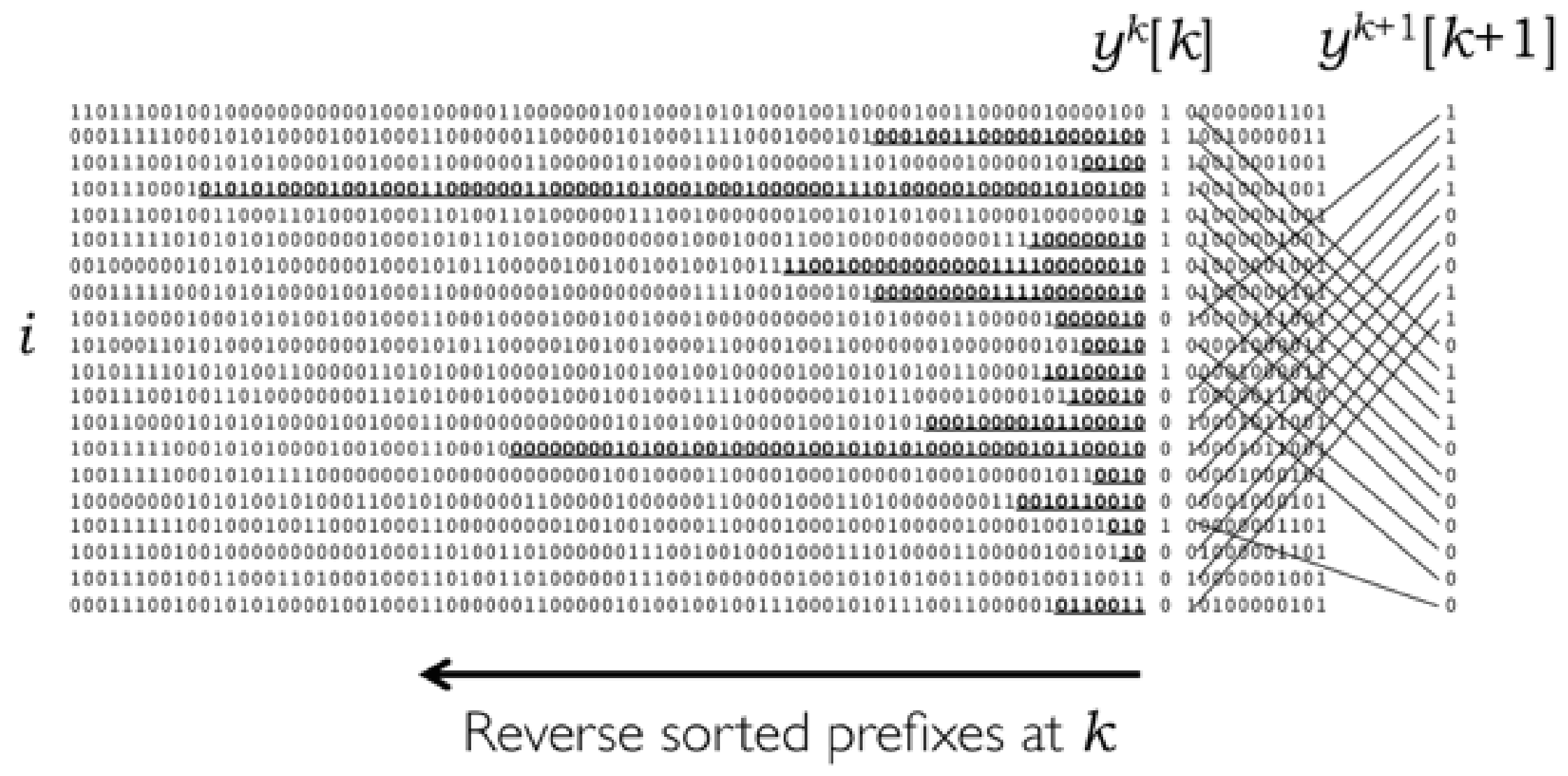


Figure 4: Procedure for updating the sort order and deriving the PBWT.

This structure also supports very fast exact matching of new sequences to find optimal local matches in the panel, in a manner equivalent to Burrows-Wheeler transform based matching in read aligners such as `bwa`. We have developed a fast imputation algorithm by first finding, for each phased input haplotype, locally maximal matches to the observed data, then imputing missing values by taking a weighted sum of the values on the local matches. This is implemented in the `pbwt` package and can be used as `pbwt -readVcfGT INPUT.vcf.gz -referenceImpute PANEL -writeVcfGz OUTPUT.vcf.gz`.

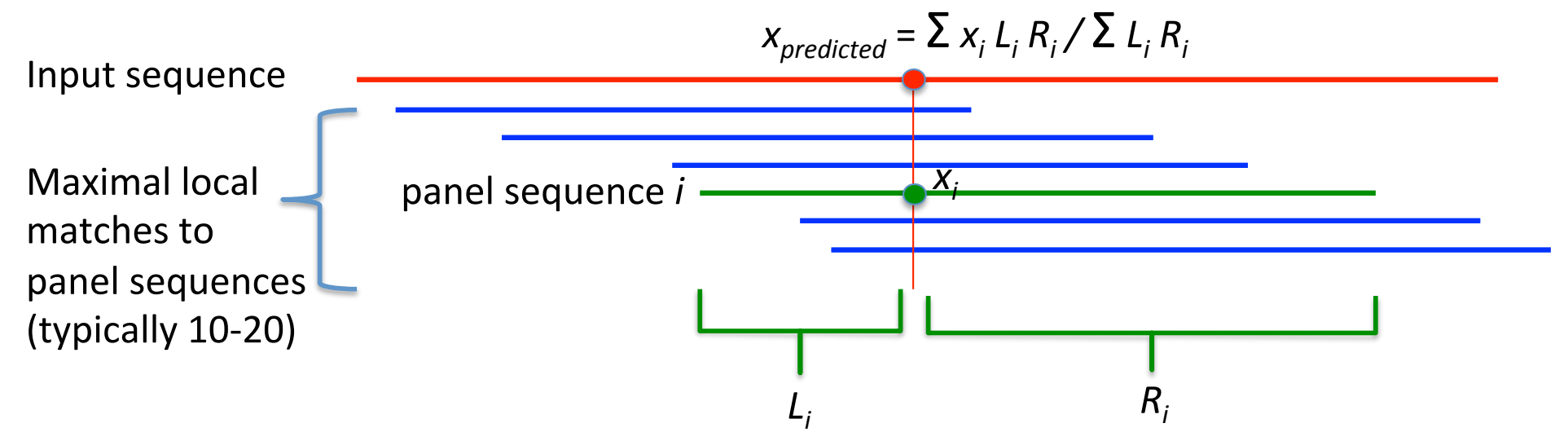
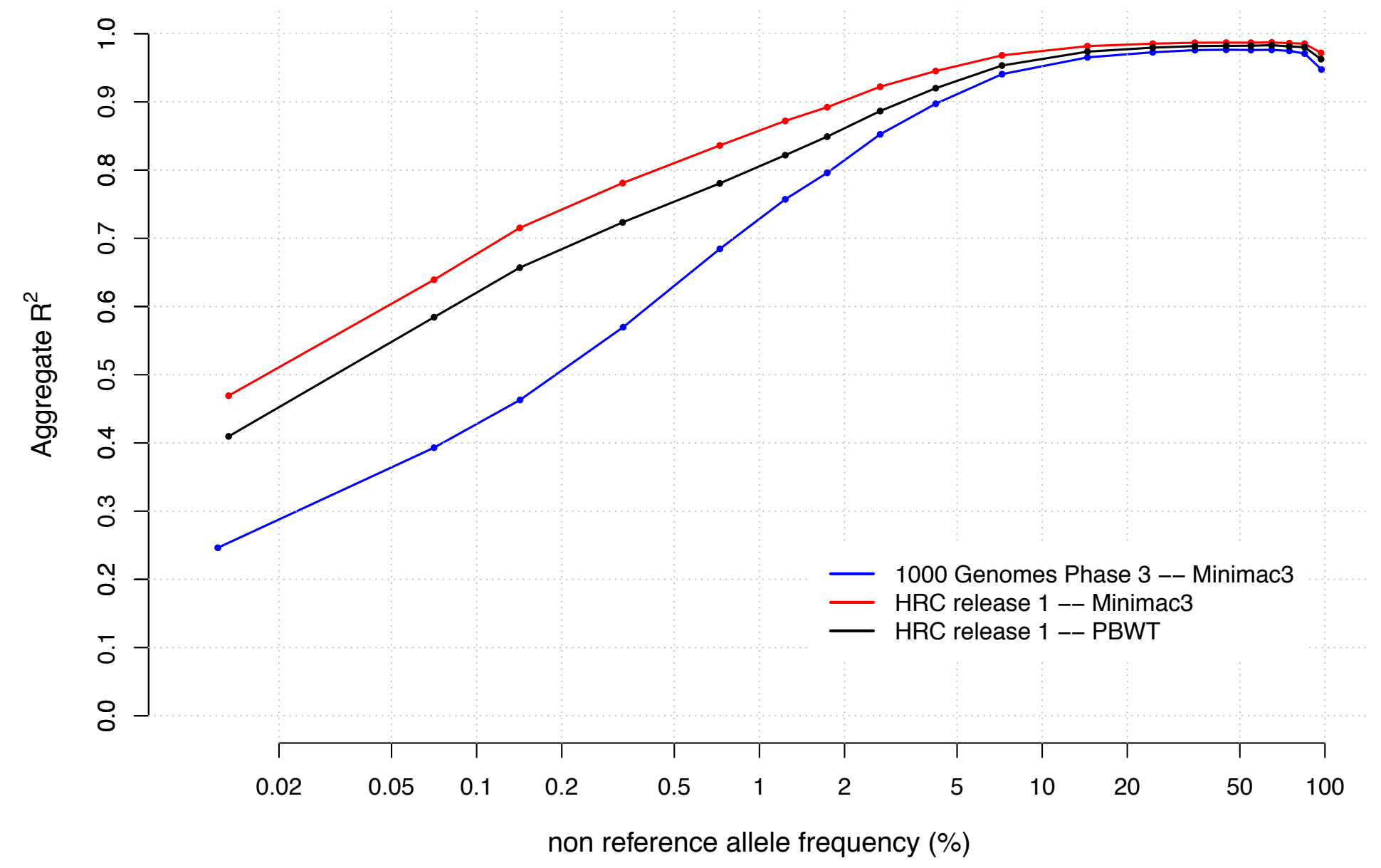

$x_{predicted} = \Sigma\, x_i\, L_i\, R_i \,/\, \Sigma\, L_i\, R_i$
Input sequence
Maximal local matches to panel sequences (typically 10-20)
panel sequence i
x_i
L_i
R_i

Figure 5: Imputation based on local maximal haplotype matching.

This implementation can impute 1000 samples against HRC release 1 in on average 134 CPU seconds per sample, with 1.63GB memory requirement for chromosome 1 (imputation is split by chromosome), with accuracy a little lower than Minimac and IMPUTE2.



References

[1] Richard Durbin. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics*, 30(9):1266–1272, 2014.

[2] Yi Wang, James Lu, Jin Yu, Richard A Gibbs, and Fuli Yu. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome Research*, 23(5):833–842, 2013.

[3] Christian Fuchsberger, Gonçalo R Abecasis, and David A Hinds. minimac2: faster genotype imputation. *Bioinformatics*, page btu704, 2014.

[4] Olivier Delaneau, Jean-Francois Zagury, and Jonathan Marchini. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, 10(1):5–6, 2013.

Haplotype Reference Consortium	`www.haplotype-reference-consortium.org`
Sanger imputation server	`imputation.sanger.ac.uk`
UMich imputation server	`imputationserver.sph.umich.edu`
PBWT	`github.com/richarddurbin/pbwt`
Contact	`richard.durbin@sanger.ac.uk`