A Suite Of Programs For Pre- And Post-imputation Data Checking

N.W. Rayner[1,2,3], N. Robertson[1,2], Anubha Mahajan[1] M.I. McCarthy[1,2,4]

1) WTCHG, University Oxford, Oxford, United Kingdom; 2) OCDEM, University Oxford, United Kingdom; 3) Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom ;4) NIHR Oxford Biomedical Research Centre, Churchill Hospital, Oxford, United Kingdom

Contact: wrayner@well.ox.ac.uk

Introduction

- The availability of the public imputation servers at the Wellcome Trust Sanger Institute and the University of Michigan has greatly reduced the complexity associated with performing an imputation run.
- Despite this, thorough quality control prior to imputation is still vital to ensure the data are correctly aligned to the reference genome.
- Post-imputation, verifying the run has completed successfully can be tricky to do from log files.
- To simplify these processes we have developed a suite of programs in Perl and Java that check and summarise the data at both the pre- and post-imputation stages.

Pre-imputation Checking

- The pre-imputation checking program compares a Plink format marker (.bim) and frequency (.freq) file to the selected reference panel.
- Currently 1000G phase 3 and HRC r1 and r1.1 reference panels are supported.
- The program checks all variants in the .bim file that match the reference for a variety of potential issues (*Table 1*) using both location information as well as SNP names.

Issue Checked	*Removed or updated?*
Chromosome and position	Updated
Alleles	Removed
Allele frequency (AF)	Removed if AF difference > 0.2*
Strand	Updated
Palindromic SNPs	Removed if MAF >0.4
Reference allele	Updated
Variant naming	Updated**

Table 1: Potential issues checked for in .bim file. *AF difference used can be set by the user, or removed entirely. **Variant names are not automatically updated although a list of naming differences is produced.

- Summary statistics for the run are written to a log file (*Figure 1*) and for each potentially updatable issue a list of variants affected is produced.
- Variants not matching the reference or failing one of the above cut offs are listed for removal.

```
Options Set:
Reference Panel:              HRC
Bim filename:                 PI.qcPlus.bim
Reference filename:           HRC.r1-1.GRCh37.wgs.mac5.sites.tab.gz
Allele frequencies filename: PI.qcPlus.frequency.frq
Allele frequency threshold:  0.2

Matching to HRC

Position Matches
 ID matches HRC 680362
 ID Doesn't match HRC 48119
 Total Position Matches 728481
ID Match
 Different position to HRC 1953
No Match to HRC 1661
Skipped (X, XY, Y, MT) 0
Total in bim file 732119
Total processed 732095

Indels (ignored in r1) 1

SNPs not changed 210061
SNPs to change ref alt 516520
Strand ok 726545
Total Strand ok 726581

Strand to change 56
Total checked 730434
Total checked Strand 726601
Total removed for allele Frequency diff > 0.2 605
Palindromic SNPs with Freq > 0.4 2076

Non Matching alleles 1757
ID and allele mismatching 1005; where HRC is . 935
Duplicates removed 23
```

Figure 1: Log file from the pre-imputation checking.

- A shell script of Plink commands is created to automate the use of these variant lists within Plink to update or remove variants.

Post-Imputation Checking

- The post-imputation checking program takes information from the output of the current set of imputation programs or servers and produces a range of charts and tables based on information score, alternate AF, MAF and position.
- The resulting plots allow a quick visual assessment of the quality of an imputation run (*Figures 2-7*).

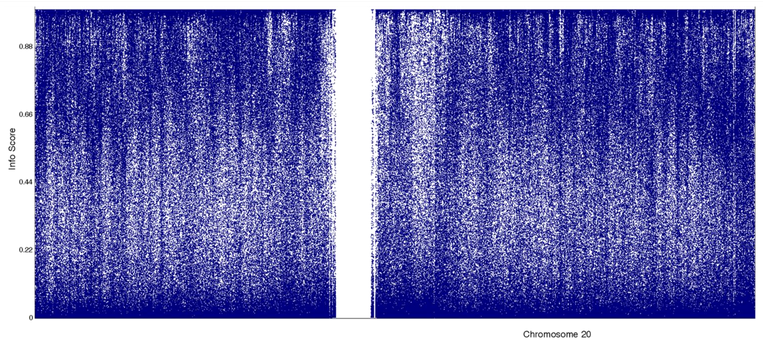


Figure 2: Information score plotted per chromosome. Plots are coloured red if there are 2 sections of >1MB without variants, indicating possible imputation failures.

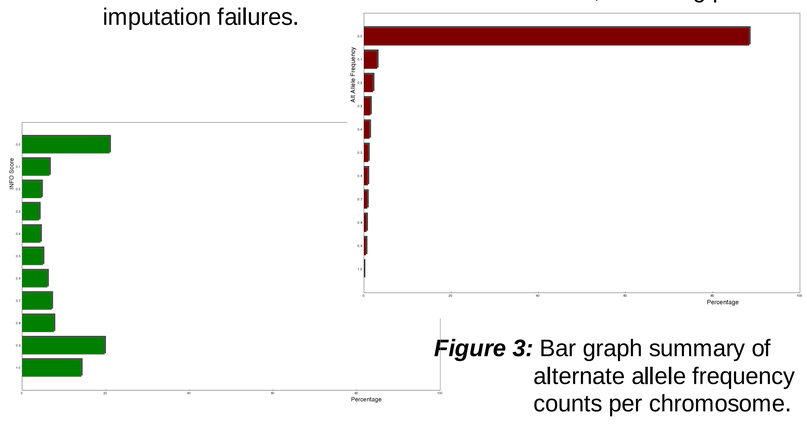


Figure 3: Bar graph summary of alternate allele frequency counts per chromosome.

Figure 4: Bar graph summary of Information score counts per chromosome.

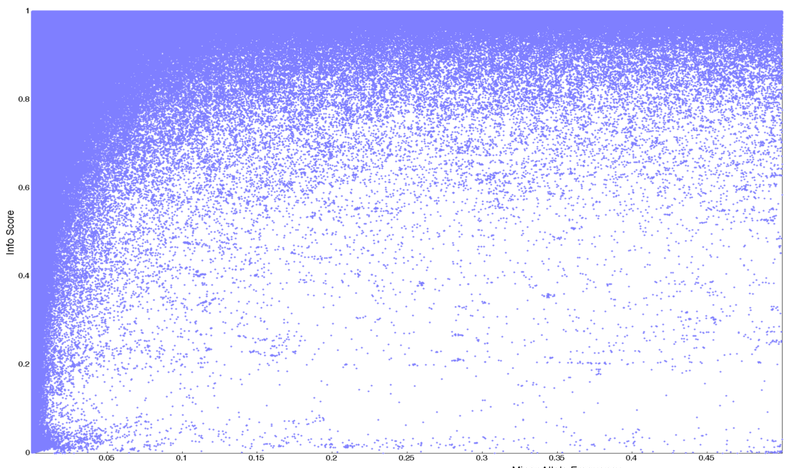


Figure 5: Plot of information score vs MAF across the chromosome.

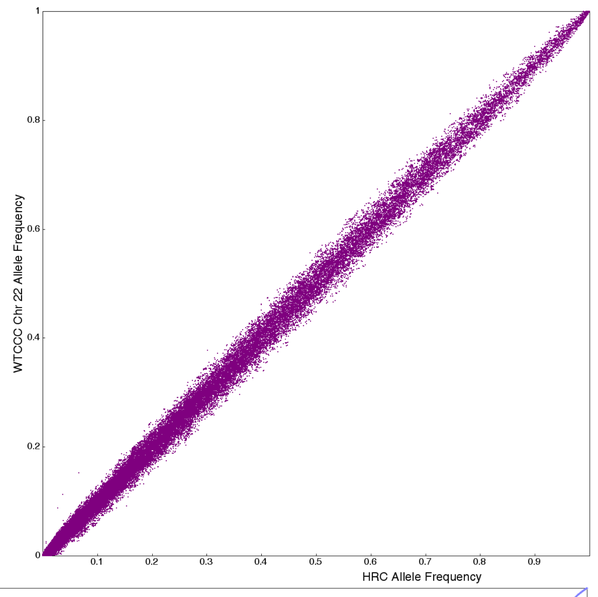


Figure 6: Plot of the allele frequency comparison to the reference (HRC in this case).

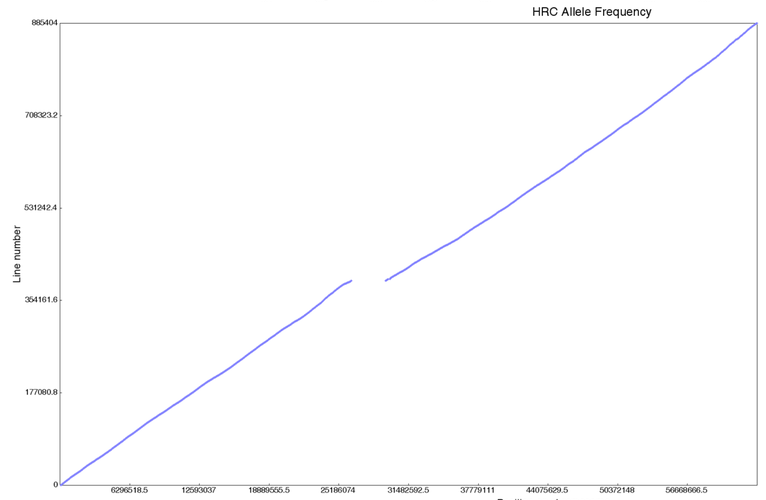
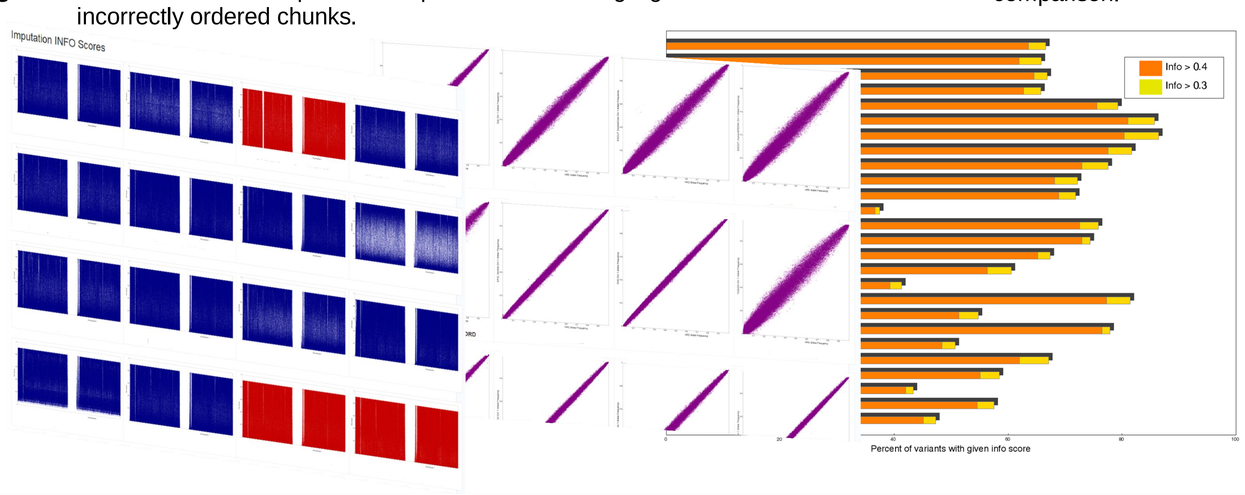


Figure 7: Plot of chromosome position vs. position in file to highlight failed or incorrectly ordered chunks.



Visualisation

- As the output for a single imputation run consists of 132 images and 70 tables for ease of visualisation these are collated into a single html file (*Figure 8*) viewable in any current web browser.
- The program can also be run to summarise and compare multiple imputed data sets (*Figure 9*).

Availability

- All the programs and instructions are available at:

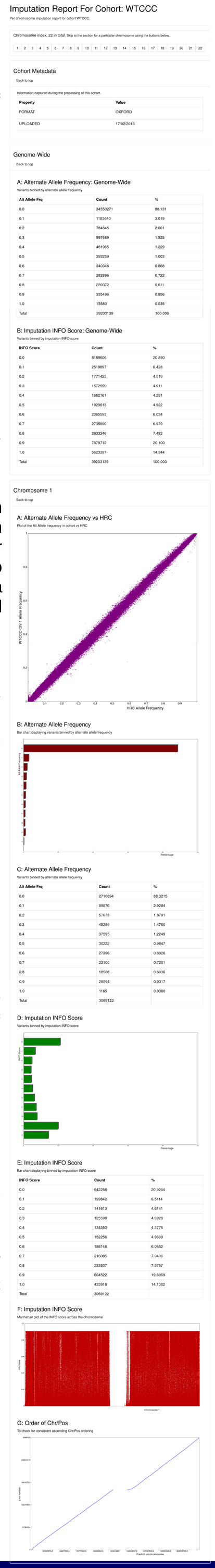
  www.well.ox.ac.uk/~wrayner/tools

Future Work

- The post-imputation program is reliant on external libraries for chart plotting, we plan to offer the program via a web server to avoid installation issues.

Summary

- These programs simplify the checking of large data sets and have been used extensively in single cohort studies, as well as for large consortia such as DIAMANTE where >35 genome-wide data sets were prepared using our pre-imputation program and >25 were checked post-imputation.
- Using the post-imputation program we were able to easily identify a wide range of imputation issues such as AF differences and missing or duplicated regions.



Imputation Report For Cohort: WTCCC

Per chromosome imputation report for cohort WTCCC.

Chromosome index, 22 in total. Skip to the section for a particular chromosome using the buttons below.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

Cohort Metadata

Back to top

Information captured during the processing of this cohort.

Property	Value
FORMAT	OXFORD
UPLOADED	17/02/2016

Genome-Wide

Back to top

A: Alternate Allele Frequency: Genome-Wide

Variants binned by alternate allele frequency

Alt Allele Frq	Count	%
0.0	34550271	88.131
0.1	1183640	3.019
0.2	784645	2.001
0.3	597669	1.525
0.4	481965	1.229
0.5	393259	1.003
0.6	340346	0.868
0.7	282896	0.722
0.8	239372	0.611
0.9	335496	0.856
1.0	13580	0.035
Total	39203139	100.000

B: Imputation INFO Score: Genome-Wide

Variants binned by imputation INFO score

INFO Score	Count	%
0.0	8189606	20.890
0.1	2519897	6.428
0.2	1771425	4.519
0.3	1572599	4.011
0.4	1682161	4.291
0.5	1929613	4.922
0.6	2365593	6.034
0.7	2735890	6.979
0.8	2933246	7.482
0.9	7879712	20.100
1.0	5623397	14.344
Total	39203139	100.000

Chromosome 1

Back to top

A: Alternate Allele Frequency vs HRC

Plot of the Alt Allele frequency in cohort vs HRC

WTCCC Chr 1 Allele Frequency

HRC Allele Frequency

B: Alternate Allele Frequency

Bar chart displaying variants binned by alternate allele frequency

C: Alternate Allele Frequency

Variants binned by alternate allele frequency

Alt Allele Frq	Count	%
0.0	2710694	88.3215
0.1	89876	2.9284
0.2	57673	1.8791
0.3	45299	1.4760
0.4	37595	1.2249
0.5	30222	0.9847
0.6	27396	0.8926
0.7	22100	0.7201
0.8	18508	0.6030
0.9	28594	0.9317
1.0	1165	0.0380
Total	3069122	

D: Imputation INFO Score

Variants binned by imputation INFO score

E: Imputation INFO Score

Bar chart displaying binned by imputation INFO score

INFO Score	Count	%
0.0	642258	20.9264
0.1	199842	6.5114
0.2	141613	4.6141
0.3	125590	4.0920
0.4	134353	4.3776
0.5	152256	4.9609
0.6	186148	6.0652
0.7	216085	7.0406
0.8	232537	7.5767
0.9	604522	19.6969
1.0	433918	14.1382
Total	3069122	

F: Imputation INFO Score

Manhattan plot of the INFO score across the chromosome

Chromosome 1

G: Order of Chr/Pos

To check for consistent ascending Chr/Pos ordering

Line number

Position on chromosome

Figure 8: *(right)* Single html page containing tables, embedded images and the navigation bar.

Figure 9: *(below)* Plots of multiple cohort summaries to enable quick comparison.