Development and Use Of a Pipeline to Generate Strand and Position Information for Common Genotyping Chips

N.W. Rayner[1,2], M.I. McCarthy[1,2,3]

1) WTCHG, University Oxford, Oxford, United Kingdom; 2) OCDEM, University Oxford, United Kingdom; 3) NIHR Oxford Biomedical Research Centre, Churchill Hospital, Oxford, United Kingdom

Contact: wrayner@well.ox.ac.uk

Introduction

- Increasing numbers of consortia are being formed to analyse data sets from multiple sources using meta-analysis.
- A key issue to be dealt with prior to any meta-analysis is ensuring all the data sets are reported to the same genomic build and strand.
- Imputation can align strand for most SNPs, however A/T and G/C SNPs having a MAF close to 50% or SNPs that do not appear in the reference panel can be missed.
- Standardising genome build in a meta-analysis can also be problematic, as data may have been generated on previous genome builds.
- Options such as the UCSC liftOver tool do exist to remap the data however these are unable to provide information on the current strand orientation.
- To address these issues an automated pipeline to map the SNPs from the different genotyping platforms to the genome has been created, along with tools to use the data to determine the current strand and remap genotype data files.

Pipeline

- The core of the pipeline is a local installation of a BLAT server, which maps the flanking or probe sequences taken from the chip annotation files to the genome.
- This is wrapped up in a set of Perl programs which automate the running of the server and the processing of the input and output files (**Figure 1**).
- The flanking or probe sequences in fasta format are run against the relevant genome in BLAT.
- BLAT output provides the start position for the sequence and its orientation to the genome.
- This is processed automatically to extract the position of the SNP in the sequence and the orientation of SNP to the reference genome.
- The final output is given as a set of three files:
  1. Chromosome, position and relative orientation of each SNP to the genome build used (.strand file).
  2. SNPs not mapping to the genome above the specified threshold (.miss file).
  3. SNPs that show mapping to more than one position on the genome (.multiple file).

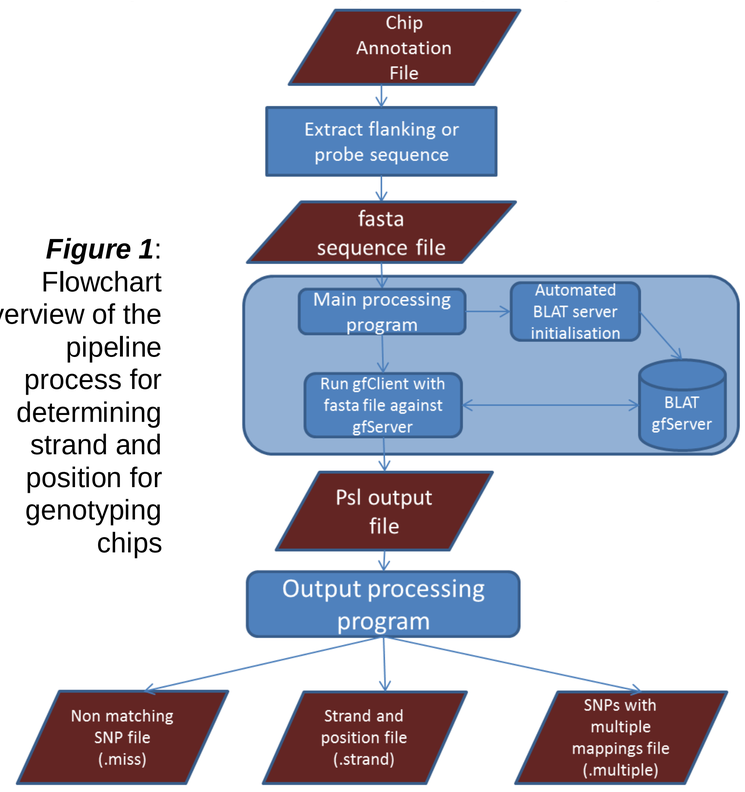

Figure 1: Flowchart overview of the pipeline process for determining strand and position for genotyping chips

- For Illumina chips the input data is aligned with respect to the TOP strand. There are a number of different orientations possible from GenomeStudio and the other genotype calling algorithms e.g. Illumina and source strand.
- Alternate orientation files have been generated for a number of chips and can easily be generated on request for other chips.
- Exporting and storing data from Illumina chips on the TOP strand will allow updating to latest build and forward strand by simply adding the relevant strand and position file (**Figure 2**).

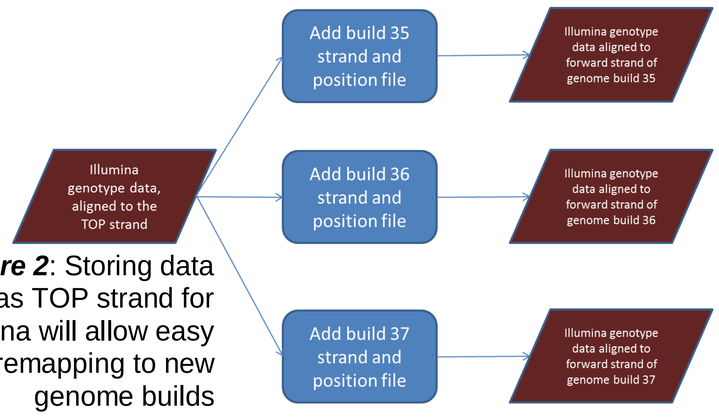

Figure 2: Storing data as TOP strand for Illumina will allow easy remapping to new genome builds

Validation

Comparison with HapMap

- As a validation, the HumanOmniExpress12v1 genotyping chip was mapped to NCBI build 36 using the pipeline and the positions compared to the HapMap.
- Of 733,202 loci present on the chip 689,924 have a matching rs ID in the HapMap.
- Of the 689,924 overlapping SNPs 99.7% had an identical position and chromosome (**Figure 3**)

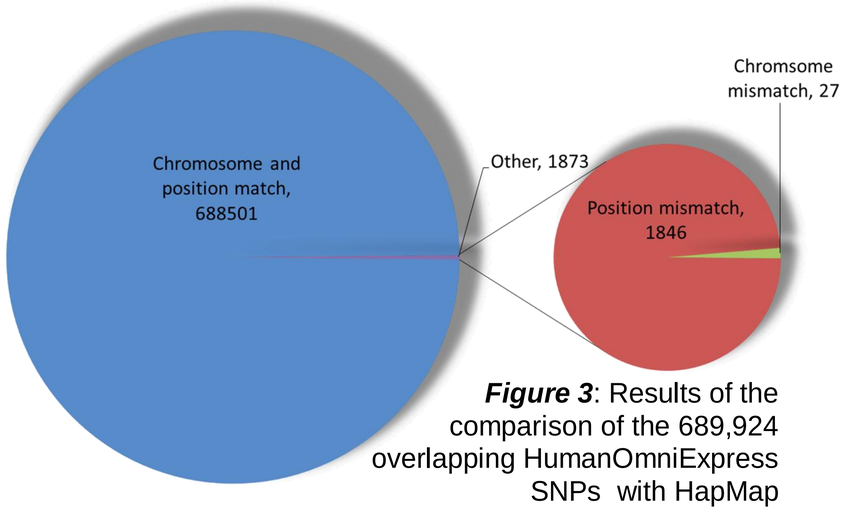

Figure 3: Results of the comparison of the 689,924 overlapping HumanOmniExpress SNPs with HapMap

Comparison with Liftover

- Positions for the HumanOmni2.5M-4v1 on genome build 36 were remapped to build 37 using the UCSC liftOver tool and the strand and position files (**Table 1*).***

	Genome Build		
	36	37	% Remapped
LiftOver	2,450,000	2,448,782	99.95
Pipeline	2,450,000	2,449,626	99.98

Table 1: Showing the number of SNPs successfully remapped from genome build 36 to 37 using both LiftOver and the strand and position pipeline

- The build 37 positions from liftOver were compared to the build 37 positions derived from the strand and position pipeline (**Table 2**).

	n	Percentage of total overlap
Total of build 37 SNP overlap	2,448,671	100%
Chromosome and position identical	2,445,556	99.87%
Different chromosome	331	0.014%
Different position	2,784	0.11%

Table 2: Showing the total number of SNPs overlapping between build 37 liftOver and the strand and position remapped HumanOmni2.5M along with the total number of chromosome and position differences in the overlapping set

Remapping genome build

- The files can also, as described above, be used to remap data to new genome builds.
- The data can be used in a number of different settings, e.g. Plink to update ped/map data, or via a script to update gen or analysis summary files.
- As part of the remapping a program has been developed to determine the current strand of an Illumina genotype data set, giving a known starting point for updating and remapping (**Figure 4**)

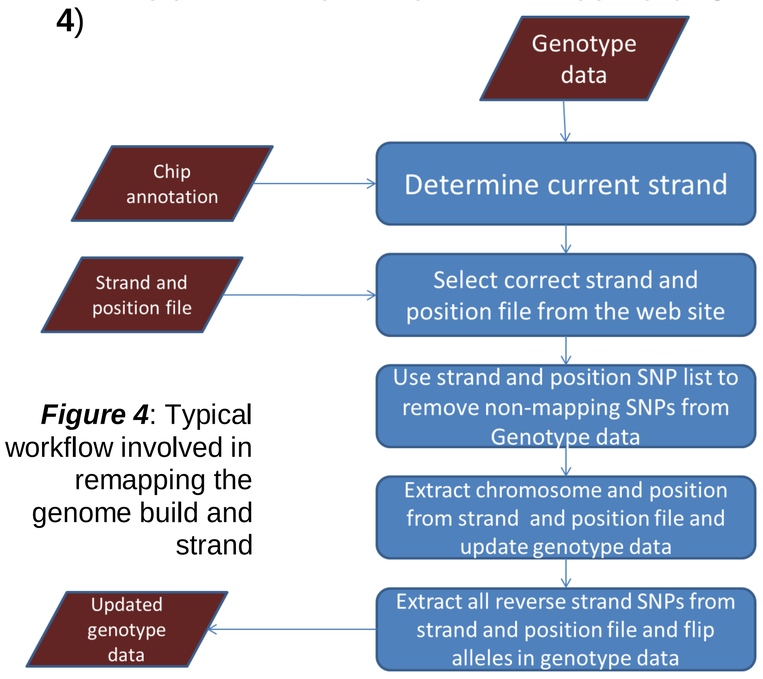

Figure 4: Typical workflow involved in remapping the genome build and strand

Usage to date

- The strand and position files have been utilised by a number of consortia such as MAGIC, GIANT, trans-ethnic fine mapping and DIAGRAM in a variety of different ways.
- The DIAGRAM metabochip replication effort, consisting of 26 cohorts comprising ~85,000 samples all utilising the Illumina Cardio-Metabochip genotyping platform, have successfully used these files to combine directly genotyped data on 3,425 SNPs to follow up top signals from the earlier DIAGRAM meta-analyses.

Web Site

- In addition to those used by the consortia other common genotyping chips are being run through the pipeline for all genome builds >35 (**Table 3**)
- The resulting files are placed on a web site where they are freely available for download.
  http://www.well.ox.ac.uk/~wrayner/strand/

Chip	NCBI Genome Build		
	35	36	37
HumanHap300v1		✓	
HumanHap300v2		✓	✓
HumanCNV370v1	✓	✓	✓
HumanHap550-2v3	✓	✓	✓
Human610Quadv1		✓	✓
HumanHap650Yv3		✓	✓
Human660W-Quad		✓	✓
HumanOmniExpress12v1		✓	✓
Human1Mv1_C		✓	
Human1M-Duov3		✓	✓
HumanOmni1M-Quad_v1		✓	✓
Humn1.2MDuoCustom_v1		✓	✓
HumanOmni2.5-4v1		✓	✓
HumanOmni2.5-8v1		✓	✓
CardioMetabochip_11395247		✓	✓
Affymetrix 500K		✓	✓
Affymetrix 6.0			✓

Table 3: List of SNP genotyping chips and the build that have been processed and are currently on the web site

- The tools to check existing strand and update gen files will also be made available via the web site.
- The web page will be updated to keep pace with new chips and genome builds.

Summary

- The pipeline produces high quality mappings showing high concordance to other methods and data sets, and by using the pre-packaged files for the chips should prove to be faster and easier to use.