**TECHNICAL NOTES:**

When analizing SNP array data (i. e. Infinium Global Screening Array v2.0) there are same challenges to be resolved:

1) the large number of different file formats used by the genetics community,

2) the ambiguous A/T and G/C single nucleotide polymorphisms (SNPs) for which the strand is not obvious,

3) check for errors in manifest file, cluster file, strand designations, build reference

For many statistical analyses, such as meta-analyses of GWAS and genotype imputation, it is vital that the datasets to be used will be corrected always having in mind the points resumed in this paragraph.

**DNA strand designations**

* ILMN Strand, a.k.a. Design Strand: The strand used by Illumina to design probes based on thermodynamic stability and locus specificity according to NCBI BLAST. For this reason, it can differ from the Customer/ Source strand.
* Forward/Reverse (Fwd/Rev) Strand: Used by dbSNP, Fwd/Rev designations can change with NCBI Genome Build updates, so Genome Build must be specified when reporting Fwd/Rev strands. 1. For SNPs in standard array products, Fwd strand = Source strand, and originates from dbSNP. 2. For custom array product SNPs without rsid’s, the customer can identify the Source strand as Fwd or Rev, based on their own criteria. Illumina custom product files use the customer’s Fwd/Rev designations. Note: The Fwd strand, as identified in Illumina standard product files, should not be confused with Plus (+) strand, which HapMap interchangeably calls the “forward strand.”
* Plus/Minus (+/-) Strand: The standard designation for all eukaryotic organisms used by HapMapand 1000 Genomes Project. The 5′ end of the (+) strand is at the tip of the short arm (p arm) of the chromosome and the 5′ end of the (-) strand is at the tip of the long arm (q arm). (+/-) designations can change with NCBI Genome Build updates, so Genome Build must be specified when reporting (+/-) strands.
* Source Strand: Same as the Customer strand. The strand submitted to the Illumina designer for probe design. 1. For standard SNPs, it is the Fwd strand as reported in the source database (i.e., dbSNP). 2. Custom content can be reported as rsid’s or as the DNA sequences or chromosomal regions, depending on the format submitted by the customer.
* Top/Bottom (Top/Bot) Strand:Top/Bot nomenclature was developed by Illumina using sequence-based context to assign strand designations that does not change regardless of database or genome assembly used. (e.g., depending on the NCBI Genome Build referenced, strand and allele designations can change). Top/Bot is not directly related to Fwd/Rev or (+/-).Top/Bot strand is determined by examining the SNP and the surrounding DNA sequence and it only applies to SNPs with two possible alleles. See the Top/Bot A/B Allele bulletin for more details.

**Basic Considerations**

Genotype data can be coded on either the forward genomic strand or the reverse genomic strand (e.g. a SNP coded T/G on the forward strand would be coded A/C on the reverse strand). The strand used to store the genotypes is not always the same within a dataset (i.e. the same strand may not be used for all variants) or between the different datasets to be aligned (i.e. the same strand may not be used for a variant present in both datasets); these differences can be intentional or accidental.

To complicate matters, most of the common file formats do not define the strand used. For some types of SNPs, it is fairly straightforward to detect and correct the strand differences. For example, a T/G SNP is non-ambiguous as its complement on the other strand is A/C. However, G/C and T/A variants are ambiguous or cryptic as their complementary alleles are C/G and A/T, respectively. This ambiguity means it is more difficult to detect and resolve strand issues for these SNPs.

Of course, it is possible to simply exclude all ambiguous variants, however, modern genotyping chips often contain many A/T and G/C SNPs; the ImmunoChip has 25,740 such SNPs (1.7% of all SNPs), the ExomeChip 244,771 (11.9%) and the Omni5-quad 144.578 (3.4%). Simply excluding these variants will limit the power of a GWAS meta-analysis where the A/T or G/C variant is the causal variant or is in higher LD to the causal variant. In the case of imputation it has also been shown that more input genotypes yield imputed genotypes of higher quality [[4](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4307387/#CR4)], so if it is possible to include the A/T and G/C variants, this is more desirable. In the cases where the strand of the genotypes is known, there are many solutions to easily correct the strands of one dataset or to simply state explicitly the strand used, for example as is possible in IMPUTE2 [[5](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4307387/#CR5)] or METAL [[6](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4307387/#CR6)]. In practice, however, this information is not always available or trustworthy.

One solution to the problem of unknown strands is to compare the minor allele between two datasets. However, use of the minor allele is not ideal as it can differ between datasets and populations, especially for common variants. PLINK [[7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4307387/#CR7)] employs a more powerful approach to detect strand inconsistencies between cases and controls. However, this method requires many manual steps, re-coding of phenotypes before and after the actual alignment, manual alignment of the non-ambiguous SNPs and merging the data into one dataset, and finally a script needs to be written to parse the alignment results from PLINK to determine the actual alignment. When using PLINK, it is not possible to align genotypes with posterior probabilities.

**Will Rayner Tools:**

Will Rayner of the Wellcome Trust Center for Human Genetics has provided a series of useful scripts that can fix the alignment of the strands.

**STRAND issue**

* The designation of strand is an extremely important issue. The Illumina strand TOP, or front Illumina strand, IS NOT the same as the front GRCh37 strand.
* All genotypes must be in the forward strand of GRCh37. An easy way to get this is to export genotypes from GenomeStudio using TOP allele annotations (typical GenomeStudio output), which we ask you to update to the advance chain of compilation 37 using the scripts provided by Will Rayner in Sanger. A description of Illumina's TOP/BOT scheme is here. Will's instructions for use, including scripts, are available here.
* //www.well.ox.ac.uk/~wrayner/strand/

Will Rayner has also created thread files for common genotyping chips in a variety of genetic constructs. These were created using BLAT, to map neighboring sequences (obtained from annotation files) to the reference human genome.

In some situations, the position in Plink files does not match that of the string file, when the difference is greater than 10bp we remove these variants because we can not be sure in the string. Will also gives a file that lists variants that have more than 1 high quality match (>90%) with the genome. It might be a good idea to remove these variants, since it means that these probes attach to different parts of the genomes.

It is important to have the version of the.strand file corresponding to the genotype matrix used, and choose the version corresponding to the build 37 (to match the chip reference panel). Note that Will Rayner's files assume that the alleles are in the TOP chain. If you use the option to delete variants that had more than one quality match with the genome, you need to add the corresponding file.multiply.

**CHANGE STRAND CODIFICATION: FROM “TOP/BOTTOM” TO “FORWARD/REVERSE”**

The easiest way to process the data, is to:

1. export them from GenomeStudio as a plink data set and ensuring you are using the TOP strand: .PED + .MAP + SAMPLESHEET
2. you can double check this using the chipendium tool

<http://mccarthy.well.ox.ac.uk/static/software/chipendium/>  which will give the chip and strand). You have to upload .BIM file or .txt containing rsID’s to check the manifest and the strand designation to “TOP”.

1. You have to download the Strand and Position Files from here: [GSA-24v2-0\_A1-b37-strand.zip](https://www.well.ox.ac.uk/~wrayner/strand/GSA-24v2-0_A1-b37-strand.zip). The data for each chip and genome build combination are freely downloadable. Each zip file contains three files, these are:

.strand file

.miss file

.multiple file

*.strand file This contains all the variants where the match to the relevant genomic sequence >90%. The strand file contains six columns, SNP id, chromosome, position, %match to genome, strand and TOP alleles. The SNP ids used are those from the Illumina annotation file and so are not necessarily the latest ones for that position from dbSNP. The alleles listed are the Illumina TOP alleles, if you are in any doubt whether your data file can be used with these strand files a check of your non A/T G/C SNPs alleles vs the strand file should confirm this for you. If there are differences then it is likely your genotype file has been created by exporting on a different strand, such as ILMN or SOURCE in this case please check the links in the menu above to those pages. If that is still unsuccessful then if you can provide a list of the SNP ids and their alleles on the chip (a plink .bim file is ideal for this) it is likely a strand file can be created for you.*

*.miss file The .miss file gives the ids of the SNPs that did not reach the required threshold for mapping to the genome, the position and strand of the best match are given.*

*.multiple file The .multiple file contains SNPs that had more than 1 high quality match (>90%) to the genome, in this instance the better match is taken for the .strand file, the number of these is reported, this is first numeric column in the file. If there are two matches of the same quality then one is chosen at random and the number of identical matches is reported, this is the second numeric column in the file.*

1. Updating the strand and position. Once you have the TOP strand data file you can then use the update\_build.sh script on the web site, with the corresponding strand file, this will set both the genome build as well as the strand (this strategy can be used to update data to any genome build as the TOP strand is not genome build specific).

A script developed by Neil Robertson for updating the chromosome, position and strand of binary ped files using these strand and position files can be downloaded here:

update\_build.sh <bed-file-stem> <strand-file> <output-file-stem>

Usage is:

* <bed-file-stem> is the name of your binary ped set minus the .bed, .bim or .fam extension
* <strand-file> is appropriate strand file for you chip and current strand orientation (TOP, SOURCE, ILMN)
* <output-file-stem> is the name of the new output file to create again minus the .bed, .bim or .fam extension

1. After running the update\_build.sh the data file will be forward strand on the genome build of choice and ready for QC and analysis.
2. If you are using imputation and have the data on genome build 37 then you can use the pre imputation checking tool (under Tools GitHub iCMLab) to verify that the data are correctly aligned with the reference panel for imputation, if you have done strand updates this is just a double check but can be useful to do to verify that everything has worked correctly.