Health and Retirement Study:
Information for dbGaP users on annotation issues in the Illumina HumanOmni2.5-4v1_D manifest

August 2014

Sarah Nelson[1*], Wei Zhao[2], Jennifer Smith[2*], Jessica Faul[2]

1- University of Washington Genetics Coordinating Center, Seattle, WA
2- Health and Retirement Study, University of Michigan, Ann Arbor, MI

Table of Contents

*Correspondence may be addressed to sarahcn@uw.edu and smjenn@umich.edu

Background

This document describes annotation issues in the Illumina HumanOmni2.5-4v1_D manifest, which affected genotype and imputation data for the Health and Retirement Study 2006/2008 samples (HRS1-2). The relevant dbGaP accession numbers are phg000207.v1 (HumanOmni2.5 array genotypes) and phg000264.v (imputation, 1000 Genomes Project reference). Here we briefly describe the issues and provide guidance to dbGaP users on how to account for them in the current datasets. The University of Washington Genetics Coordinating Center (UWGCC) will be providing updated datasets to post onto dbGaP; however, until the updated datasets are posted, users will need to implement the recommended changes below.

For more information on the annotation issues, including how they were discovered and investigated, please see the appendix at the end of this document.

Primary Issue: Reversed A/B alleles

For 18,763 strand ambiguous SNPs (i.e., with A/T or C/G alleles), the A/B allele designation is reversed in the Illumina manifest. According to Illumina technical support, this error was introduced when redesigning the manifest from version B to version D, but was corrected in subsequent versions (e.g., version H). This error was propagated into any dataset or annotation that relied on the mapping of A/B alleles to design, TOP, forward, or plus(+) strand alleles*.

* For more on allele naming conventions and strand designations in Illumina genotyping, see:

Nelson, S. C., Doheny, K. F., Laurie, C. C., & Mirel, D. B. (2012). Is “forward” the same as “plus”?...and other adventures in SNP allele nomenclature. *Trends Genet*, *28*(8), 361–363.

Secondary Issue: Incorrect RefStrand designation

In identifying the A/B allele reversal issue noted above, the UWGCC further identified likely errors in Illumina’s RefStrand designations. RefStrand indicates the orientation of the probe sequences (“design strand”) with respect to the human genome reference sequence: either “+” or “-”. The UWGCC preformed BLAT searches on the “TopGenomicSeq” entries in the Illumina manifest and identified 608 SNPs where the BLAT result was in conflict with RefStrand.

This error was propagated into any dataset or annotation that relied on the mapping of A/B alleles to plus(+) strand alleles – namely the 1000 Genomes Project imputation.

Table of Issues

Users should refer to the accompanying Excel file “HRS1-2_HumanOmni2.5-4v1_D_flaggedSNPs.xlsx” to identify which SNPs are affected by which issues(s) above. In this table of issues, “probes.alleles.disc=TRUE” flags the SNPs with the primary issue, reversed A/B alleles. The secondary issue, incorrect RefStrand designation, is flagged by “blat.strand.disc=TRUE.” See the data dictionary in the first tab of this Excel file for a full definition of all the fields in this table of issues.

Actions to address issues in dbGaP dataset components

Below is a list of actions users can take to correct the annotation issues in HRS1-2.

1) PLINK dataset

Affected files

CIDR_HRS_Top_subject_level (.bed, .bim, .fam)

CIDR_HRS_Top_subject_level_filtered (.bed, .bim, .fam)

How to identify affected SNPs

"probes.alleles.disc=TRUE"

Action needed

Use PLINK "--flip" command at all affected SNPs

2) Allele mapping file

Affected files

SNP_allelemap.csv

How to identify affected SNPs

"probes.alleles.disc=TRUE" and "blat.strand.disc=TRUE"

Action needed

Where "probes.alleles.disc=TRUE" and "blat.strand.disc!=TRUE", entries in the alle.design, alle.top, alle.fwd, and alle.plus fields should be swapped between alle.AB=A and alle.B=B. That is, for the alle.AB=A row, update alle.design, alle.top, alle.fwd, and alle.plus with values from alleAB=B row. Similarly for the alle.AB=B row, update alle.design, alle.top, alle.fwd, and alle.plus with values from alleAB=A row. Another way to think of this is as a flipping alle.design, alle.top, alle.fwd, and alle.plus, according to the following table:

Old allele	New allele
A	T
C	G
G	C
T	A

Where "probes.alleles.disc=TRUE" and "blat.strand.disc=TRUE", entries in the alle.design, alle.top, and alle.fwd fields should be swapped between alle.AB=A and alle.AB=B. *The alle.plus field does not need to be changed.*

Where "probes.alleles.disc!=TRUE" and "blat.strand.disc=TRUE", the alle.plus field should be swapped between alle.AB=A and alle.AB=B.

3) Imputation

Affected files

All imputed genotype probabilities files, e.g., “HRS_chr#.gprobs”

How to identify affected SNPs

“problem.type2.SNP=TRUE”

Action needed

Update the alleles at all affected SNPs (“problem.type2.SNP=TRUE”), which can be implemented with the GTOOL program’s “--strand” command (http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html#Orient). Create a strand file with all the “problem.type2.SNP=TRUE” SNPs and indicate “-” as the strand. Then run GTOOL as follows:

```
gtool -O --g HRS_chr#.gprobs --strand problem.type2.SNPs.strand --og
HRS_strandFix_chr#.gprobs --log strandFix_chr#.log
```

Note that while this corrects the plus(+) strand alleles in the .gprobs files for these “problem.type2.SNP=TRUE” SNPs, the imputation was nevertheless run with these SNP not aligned to the plus(+) strand. The appendix includes an investigation of what effect these SNPs had on the imputation accuracy at neighboring imputed variants. While no systematic or severe hit in accuracy was observed, the possibility of these SNPs introducing some error into the imputation cannot be ruled out.

Recommendation to use HRS1-3

dbGaP users are encouraged to use the more recent, combined dataset HRS1-3 dataset when it becomes available on dbGaP. The primary issue of reversed A/B alleles is not present in HRS1-3.While HRS1-2 was genotyped on version D of the HumanOmni2.5 manifest, HRS3 was genotyped on the more recent version H. When the UWGCC created the combined HRS1-3 dataset, any SNPs with cross-dataset duplicate discordances were removed, which effectively removed all of the SNPs with reversed A/B allele designations in HRS1-2. Furthermore, incorrect RefStrand designations are expected to be less of an issue in the HRS1-3 imputation, as the UWGCC found the RefStrand assignments in version H of the array to be more consistent with BLAT results compared to version D. Thus version H was used to align study SNPs to the plus(+) strand prior to the HRS1-3 imputation.

Appendix

Additional information on the investigations undertaken by HRS analysts and the UWGCC to identify and explore the issues in the HumanOmni2.5-4v1_D manifest

Identified Issues

1) Discrepancy between RefStrand designation with BLAT

First, we checked strand annotation in the Illumina array manifest (HumanOmni2.5-4v1_D.csv). We did a BLAT search of TopGenomicSeq for all chr1-22,X SNPs and found top BLAT matches at ~99.7% of array probes (at matching chr and position). Then we checked BLAT matches against the Illumina array manifest "RefStrand" and "IlmnStrand" columns (which give +/- and TOP/BOT for design alleles, respectively), which should follow the rules below:

a. Where IlmnStrand=TOP, BLAT strand should=RefStrand
b. Where IlmnStrand=BOT, BLAT strand should!=RefStrand

We found discrepancies for 608 SNPs. Luckily, for strand unambiguous SNPs where Illumina strand annotation was wrong and we didn't initially have alleles on the + strand going into imputation, IMPUTE2 flipped the study alleles to align with reference panel strand. Indeed, the log records show that IMPUTE2 did this for 400 SNPs. However, this would still be a problem for strand ambiguous SNPs because the imputation software would not be able to recognize the problem and would carry over the wrong information into the imputation. Table 1 is the breakdown of SNP type for the 608 SNPs (with strand ambiguous types bolded), and only 25 of these are strand ambiguous (**A/T or C/G** alleles). Among them, 14 SNPs were imputation basis SNPs in HRS1-2 imputation. Please note that this issue itself is not of major concern as it is incidental and happens to every manifest.

Table 1. The breakdown of SNP type for the 608 SNPs with inconstant BLAT search results

SNP type	A/C	A/G	**A/T**	**C/G**	**G/C**	**T/A**	T/C	T/G
Number	60	215	**4**	**9**	**7**	**5**	228	80

2) Discrepancy between A/B alleles and ProbeSeqs

Next we looked at the design of the ProbesSeq in the HumanOmni2.5-4v1_D manifest. All **strand ambiguous SNPs** are Inifinium I assays (two bead types) and thus have both an AlleleA_ProbeSeq and AlleleB_ProbeSeq as shown in the following Figure. *Note: strand unambiguous SNPs are all single bead type, Infinium II assay*.

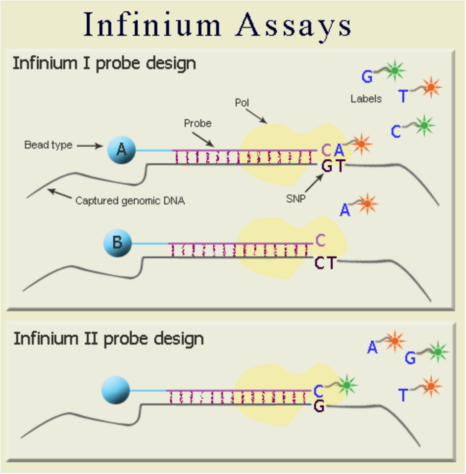


Image: http://www.ncbi.nlm.nih.gov/genome/probe/doc/DistrIllumina.shtml

Table 2. ProbeSeq design for SNP kgp7085589 in different Omni2.5 array manifests

array	Ilmn Strand	SNP	AlleleA_ProbeSeq	AlleleB_ProbeSeq
HumanOmni2.5-4v1_D	TOP	[C/G]	TTGCATGCTGCGACCAGGGATCCTCTCTCCATGTCTGTCGCTGGCCTTGG	TTGCATGCTGCGACCAGGGATCCTCTCTCCATGTCTGTCGCTGGCCTTGC
HumanOmni2.5-4v1_H	TOP	[C/G]	TTGCATGCTGCGACCAGGGATCCTCTCTCCATGTCTGTCGCTGGCCTTGC	TTGCATGCTGCGACCAGGGATCCTCTCTCCATGTCTGTCGCTGGCCTTGG
HumanOmni2.5-8v1_C	TOP	[C/G]	TTGCATGCTGCGACCAGGGATCCTCTCTCCATGTCTGTCGCTGGCCTTGC	TTGCATGCTGCGACCAGGGATCCTCTCTCCATGTCTGTCGCTGGCCTTGG

According to the design, we would expect the last nucleotide of the Allele A probe to be the design allele A, and the same is true for allele B. However, we found that there are a total of 18,763 strand ambiguous SNPs whose last nucleotides of the ProbeSeqs were inconsistent with A/B design alleles. Table 2 lists an example SNP kgp7085589 (red highlighting). As shown in the table, the manifest 4v1_D does NOT follow the rule whereas the manifests 4v1_H and 8v1_C are correct. We communicated this with Illumina technical support team and they confirmed that it was a mistake in the manifest D and was corrected in the following manifest. Because all of those SNPs are strand ambiguous, the problem was not fixed by imputation software. Among them, 13,961 were imputation basis SNPs in HRS1-2.

Influence on imputation

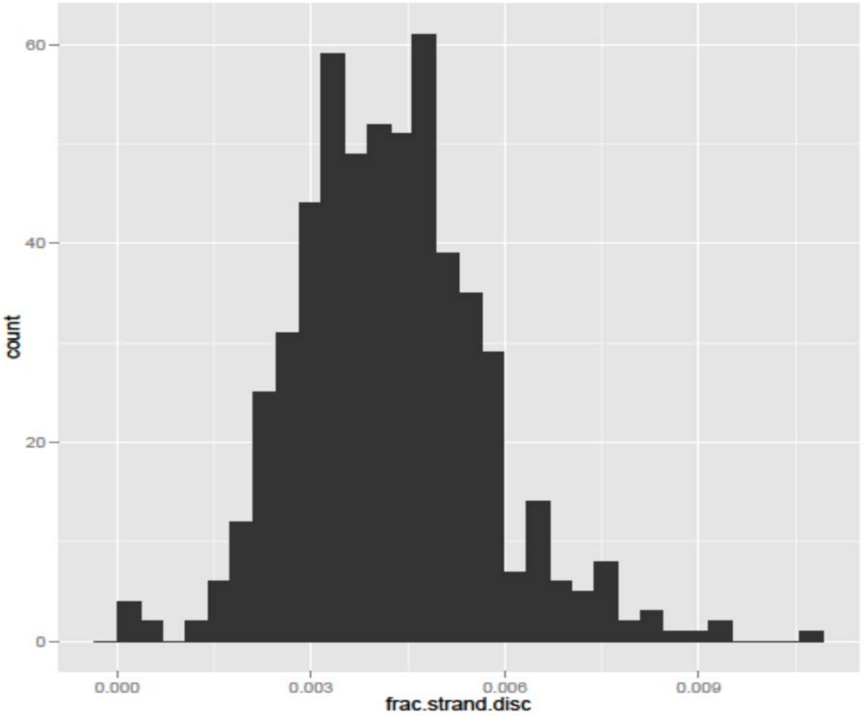
To examine how these problematic SNPs affected 1000 Genomes imputation, we examined:

- How strand misalignments are distributed across imputation segments
- How strand misalignments affect imputation accuracy at nearby imputed SNPs – i.e., are the imputed frequencies of nearby SNPs also reversed?

Note: This analysis was performed before we had our final list of strand ambiguous SNPs (before we knew the exact cause of the strand flip problem). For this analysis, SNPs with "apparent strand misalignment" were defined as those that had concord_type0<0.4 and info_type>0.8 from IMPUTE2 internal masked SNP testing. This captured the vast majority of truly misaligned strand ambiguous SNPs.

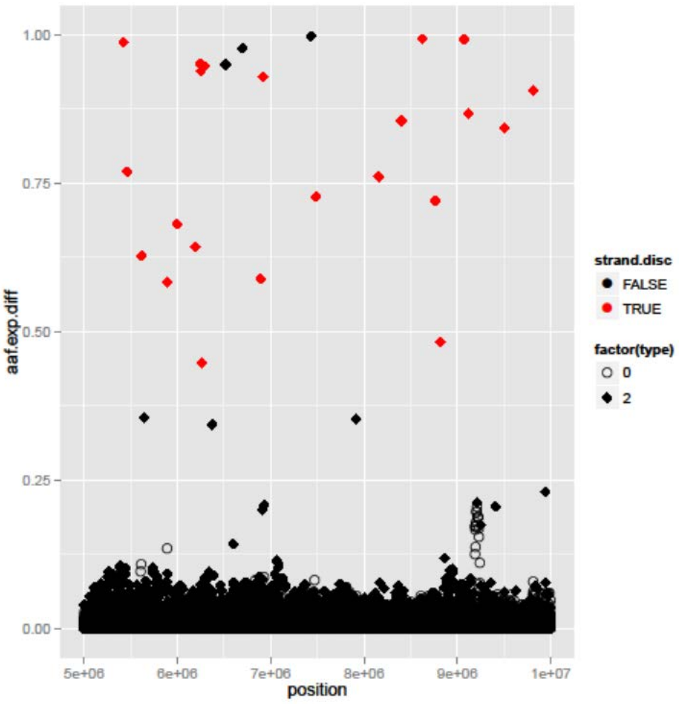
1) Distribution of type 2 SNPs (directly genotyped SNPs) with apparent strand misalignments:

We found that the type 2 SNPs with apparent strand misalignments were quite evenly distributed across imputation segments: range 0 – 92 per segment, mean 16. Below is a histogram of the fraction of type 2 SNPs per segment with apparent strand misalignments – the max observation is 1% of type 2 SNPs with strand issues per segment.



2) Strand misalignments don't appear to strongly affect imputation accuracy at nearby imputed SNPs:

We randomly chose one imputation segment with an average fraction of type 2 SNPs with strand misalignments (0.4-0.5%) and compared the allele frequency for the imputed SNPs in the imputation output to the 1000 Genomes reference samples. The segment we chose is segment 2 (from 5-10MB) on chromosome 6. It has 37,500 type 0 (imputation target) SNPs and 4,742 type 2 (imputation basis) SNPs. Among them, there are 22 type 2 SNPs with apparent strand discrepancies. The following plot shows the absolute value difference between imputed alternate allele frequency (AAF) across all HRS samples ("exp_freq_a1" in imputed metrics files) and the EUR AAF (position as X axis, difference in allele frequency on Y axis). For imputed SNPs, this plot is limited to a subset where EUR and AFR AAF were within 0.2 (yielding 34,006 type 0 SNPs in this plot).

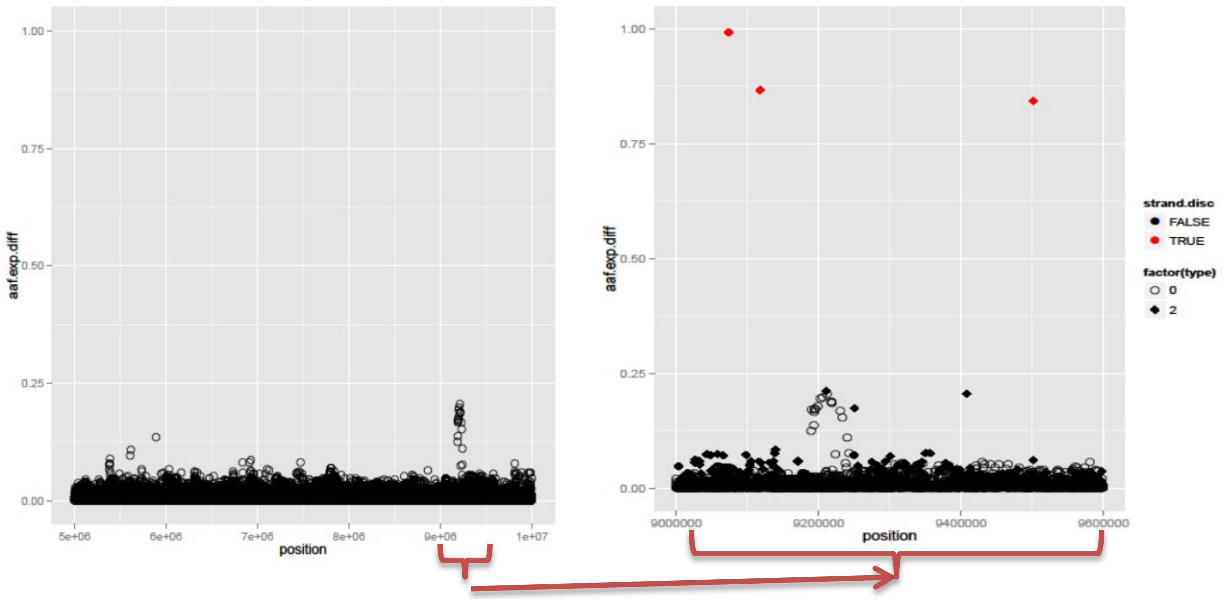


Type 0 (imputed) are open circles; type 2 (observed) are closed diamonds. Type 2 are flagged as strand discrepant (concord_type0<0.4 & info_type0>0.8) are red.

You can see imputed SNPs have imputed AAF comparable to EUR AAF.

The three black diamonds with high aaf.exp.diff are truly strand misaligned, but just not colored red because of the masked SNP metrics thresholds that we used to define strand discrepant for this specific analysis.

...honing in on only imputed (open circles) from above:



As seen in the left most plot of type 0 SNPs only, most imputed AAF are within 0.1 of EUR AAF. There is, however, one spike of SNPs ~9.4-9.6MB: 13 imputed SNPs with aaf.exp.diff>0.15. There are 3 strand discrepant type 2 SNPs in this interval (see zoomed in plot to right, in which type 2 SNPs have been added back in).

Conclusion: With evidence from this imputation segment (selected at random from imputation segments with an average fraction of apparently strand discrepant type 2 SNPs), it does NOT appear that strand discrepant type 2 SNPs are negatively affecting imputed SNPs in any severe or systematic way. That is, imputed AAF in study samples is comparable to reference AAF; it does not look like type 0 variants are imputed to the wrong strand *for most samples*. The spike of imputed SNPs ~9.4-9.6MB with a > 0.15 difference in AAF suggest that strand discrepant type 2 SNPs could cause imputation errors in *some samples* – but not enough to change the overall AAF in the imputed sample set.