Deelen et al. BMC Research Notes 2014, 7:901
http://www.biomedcentral.com/1756-0500/7/901

BMC Research Notes

TECHNICAL NOTE Open Access

Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration

Patrick Deelen[1,2*†], Marc Jan Bonder[2†], K Joeri van der Velde[1,2], Harm-Jan Westra[2], Erwin Winder[1,2], Dennis Hendriksen[1,2], Lude Franke[2] and Morris A Swertz[1,2*]

Abstract

Background: To gain statistical power or to allow fine mapping, researchers typically want to pool data before meta-analyses or genotype imputation. However, the necessary harmonization of genetic datasets is currently error-prone because of many different file formats and lack of clarity about which genomic strand is used as reference.

Findings: Genotype Harmonizer (GH) is a command-line tool to harmonize genetic datasets by automatically solving issues concerning genomic strand and file format. GH solves the unknown strand issue by aligning ambiguous A/T and G/C SNPs to a specified reference, using linkage disequilibrium patterns without prior knowledge of the used strands. GH supports many common GWAS/NGS genotype formats including PLINK, binary PLINK, VCF, SHAPEIT2 & Oxford GEN. GH is implemented in Java and a large part of the functionality can also be used as Java 'Genotype-IO' API. All software is open source under license LGPLv3 and available from www.molgenis.org/systemsgenetics.

Conclusions: GH can be used to harmonize genetic datasets across different file formats and can be easily integrated as a step in routine meta-analysis and imputation pipelines.

Keywords: GWAS, Imputation, Meta-analysis, Linkage disequilibrium

Background

Genome-wide association studies (GWAS) increasingly require the integration of multiple genetic data sets to reach sufficient resolution and statistical power, either by imputing missing genotypes or by pooling datasets for a meta-analysis. However, there are two major challenges to be resolved: 1) the large number of different file formats used by the genetics community, and 2) the ambiguous A/T and G/C single nucleotide polymorphisms (SNPs) for which the strand is not obvious. For many statistical analyses, such as meta-analyses of GWAS [1] and genotype imputation [2], it is vital that the datasets to be used are aligned to the same genomic strand.

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 [3] 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

* Correspondence: patrickdeelen@gmail.com; m.a.swertz@gmail.com
†Equal contributors
[1]University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, the Netherlands
[2]University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands

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], 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] or METAL [6]. 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] 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.

Implementation

Here, we present Genotype Harmonizer (GH): a new command-line tool to automate genotype data harmonization. GH can read commonly used file formats (PLINK, binary PLINK, VCF, SHAPEIT2 & Oxford GEN) and align a study dataset to a specified reference without any prior knowledge of the strand used. After alignment, GH writes data back to a chosen format (PLINK, binary PLINK, SHAPEIT2 or Oxford GEN). All handling of the genotype data and loading genotypes from the different formats is implemented in our Genotype IO library, which also allows integration of the harmonization tools into other software. GH consists of 25,000 lines of code with a high unit test coverage of over 60% at conditional level and continuous build testing. GH is written in Java and has been tested under Linux, Windows, and OS-X. All source code is available at www.github.com/molgenis/systemsgenetics.

GH implements a fully automated method that assigns the strand of ambiguous SNPs by selecting nearby non-ambiguous SNPs that are in linkage disequilibrium (LD) in both the study data and the reference data. GH correlates the estimated haplotype frequencies between the study data and the reference data. If GH finds more negative correlations than positive ones in haplotype frequencies, the ambiguous SNP is swapped to the other strand. When GH is unable to align a SNP (e.g. because of a lack of surrounding SNPs), this ambiguous SNP is excluded from the set. It is possible to prevent exclusion of variants that could not be aligned using LD, GH can optionally perform alignment using the minor allele for variants that have a minor allele frequency below a specified value.

Findings

Usage in an imputation workflow

We advise applying GH to pre-phased data before imputation. When pre-phasing using SHAPEIT2 [8] and imputing using IMPUTE2, GH can read the SHAPEIT2 output directly and can write aligned results in the same format for direct use by IMPUTE2 (Figure 1). Performing the alignment after the pre-phasing step ensures that pre-phasing does not need to be repeated when imputing using a different reference set or a newer version of a reference set. GH can also update the variant identifiers of the study data to match the reference set identifiers using the --update-id option. An example command is:

```
GenotypeHarmonizer.sh --input shapei-
t2Output --ref
refInVcf --output targetPath --update-id
```

Usage to harmonize GWAS data

GH can also be used in merging or meta-analysis of different GWAS datasets (Figure 1). One of the datasets can be used as a reference and the other datasets can be aligned to it, or all the cohorts can be aligned to a public reference set. It is possible to include all the variants present in the study data that are not in the reference set using the --keep option. After alignment the datasets can be investigated using a meta-analysis or can be merged into a single dataset. An example command is:

```
GenotypeHarmonizer.sh --input dataset1
--ref
dataset2 --output dataset1Aligned
--update-id --keep
```

Performance

GH requires 6:35 minutes to align a GWAS dataset consisting of 168,408 SNPs and 25,169 samples in binary PLINK format to another GWAS dataset with 528,969 SNPs and 11,950 samples, using a Linux system, a single core and 4 GB of RAM. Aligning the SHAPEIT2 results (25,169 and 19,321 variants on chromosome 1) to the Genome of The Netherlands imputation reference (499 samples, 1,536,126 SNPs on chromosome 1) [9] took 36 seconds using a single core and <1 GB of RAM.

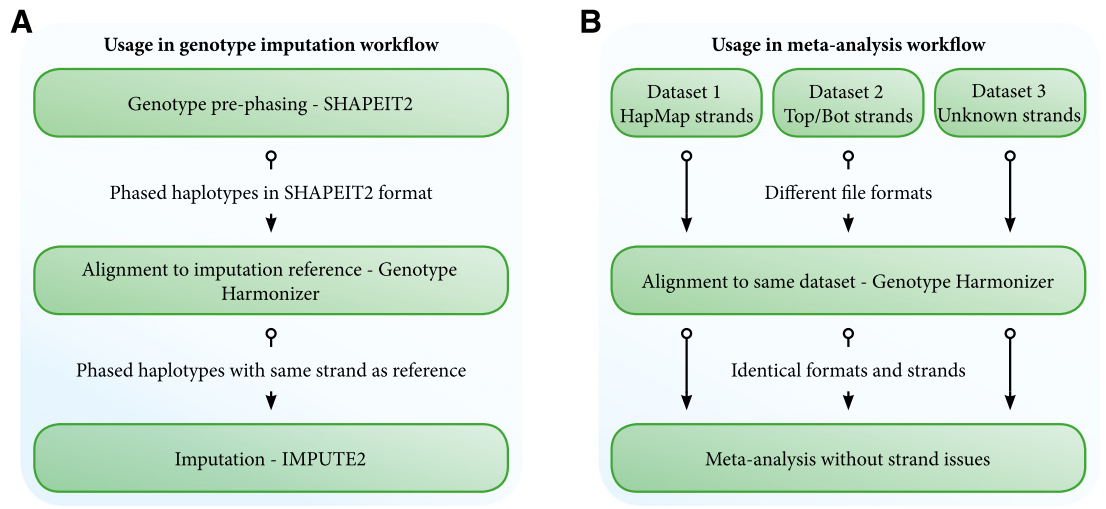

Figure 1 Usage of Genotype Harmonizer. A) GH can be applied after the pre-phasing of the genotypes, preventing the need to redo the phasing for each new version of a haplotype reference set. **B)** GH can be used to align and reformat genotype datasets allowing easy merging or meta-analysing of data. By aligning all datasets to a public reference, the genotype data can be kept private by consortia members.

Comparison using PLINK alignment

We compared the alignment of ambiguous variants using GH to the alignment using the flip-scan option in PLINK. We performed this analysis by using the latest HapMap3 data. We randomly assigned the samples into two equally sized sets, henceforth denoted as set1 and set2. In set1 we randomly changed the strand of roughly 50% of the A/T and G/C variants.

Set1 was aligned using GH by using set2 as the reference using the default settings. We successful aligned 40,617 out of the 55,517 swapped variants, 14 (0.03%) variants were aligned to the incorrect strand. In total 29,801 A/T and G/C variants (27% of the total ambiguous variants) were excluded since there were not enough variants in LD for accurate alignment. There were no variants swapped by GH that were not flipped in our test set.

For the analysis using PLINK we denoted the samples in set1 as cases and set2 as controls; we merged both sets and used the flip-scan option using the default settings. PLINK does not actually report which variants should be swapped but instead provides a log with information on which the decision to swap a variant can be based. Since the PLINK manual does not provide a recommendation on how to select the variants to swap based on this file, we used the same criteria as those used by the GH, i.e. there need to be at least 3 variants in LD, and then we assessed if there were more positive than negative correlations. This resulted in the successful alignment of 37,402 SNPs and the incorrect alignment of 54 SNPs (0.14%); 36,390 (33% of the total ambiguous variants) variants were excluded because of lack of variants in LD. We thus find that the number of incorrectly aligned SNPs increased by 40 SNPs and the number of excluded SNPs increased by 22% from 29,801 to 36,390 when using PLINK instead of GH.

Moreover, in one command GH covers many separate steps which require considerable manual work or scripting when using PLINK: manual alignment of non-ambiguous variants (which PLINK cannot do automatically), conversion of reference haplotypes to a PLINK supported format, merging the reference and study datasets, recoding using a fake phenotype file, running PLINK flip-scan to find swapped SNPs, and the selection and swapping of the SNPs on the wrong strand.

Conclusions

We have shown that using Genotype Harmonizer we can provide near perfect alignment of ambiguous SNPs without any prior knowledge of the strands. Compared to PLINK we have improved the strand alignment and limited the number of manual steps without sacrificing run-time performance. Another advantage of GH over PLINK is our support of file formats storing haplotype phase or genotype probability information, which also makes our software useful to employ within an imputation workflow or on data that has already been imputed.

GH uses an advanced LD-based method to perform the alignment of ambiguous SNPs and supports many genotype file formats. The underlying Genotype IO API is part of the MOLGENIS open source suite [10], which is also used by several other genetic analysis tools, and we expect the number of supported formats to grow in the future. These enhancements will be made available in later releases of GH. We have used GH to harmonize over 15 imputations and GWAS datasets [11-14]. GH is

now a standard part of our imputations and has been applied to over 25,000 samples (publications in preparation). We expect GH to be a major time saver for many research groups and to become a standard part of many analysis pipelines, as it alleviates manual steps when imputing data or when working with multiple GWAS datasets.

Availability and requirements

Project name: Genotype Harmonizer
Project home page: www.molgenis.org/systemsgenetics
Operating system(s): Platform independent
Programming language: Java
Other requirements: Java 1.6 or higher
License: LGPLv3
Any restrictions to use by non-academics: Free to use

Abbreviations

GH: Genotype harmonizer; GWAS: Genome-wide association study; SNP: Single nucleotide polymorphism; LD: Linkage disequilibrium.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PD, MJB, LF, HJW, MS designed the software. PD, MJB, KJV, EW, DH implemented the software. PD, MJB, MS wrote the manuscript. All authors read and approved the manuscript.

Acknowledgements

We thank Kate Mc Intyre and Jackie Senior for carefully reading and editing the manuscript and Alexandros Kanterakis for testing the software.

Funding

The research leading to these results received funding from BBMRI-NL, a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO project 184.021.007), to PD, MJB; the European Union Seventh Framework Programme (FP7/2007-2013) under grant 261433 (BioSHaRE-EU) to KJV; LifeLines/Target to EW; and TI Food and Nutrition (TIFN GH001) to MS.

Received: 7 November 2014 Accepted: 3 December 2014
Published: 11 December 2014

References

1. Evangelou E, Ioannidis JPA: **Meta-analysis methods for genome-wide association studies and beyond.** *Nat Rev Genet* 2013, **14:**379–389.
2. Marchini J, Howie B: **Genotype imputation for genome-wide association studies.** *Nat Rev Genet* 2010, **11:**499–511.
3. **"TOP / BOT" Strand and "A / B" Allele.** [http://res.illumina.com/documents/products/technotes/technote_topbot.pdf].
4. Roshyara N, Kirsten H, Horn K, Ahnert P, Scholz M: **Impact of pre-imputation SNP-filtering on genotype imputation results.** *BMC Genet* 2014, **15:**88.
5. Howie B, Donnelly P, Marchini J: **A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.** *PLoS Genet* 2009, **5:**e1000529.
6. Willer CJ, Li Y, Abecasis GR: **METAL: fast and efficient meta-analysis of genomewide association scans.** *Bioinformatics* 2010, **26:**2190–2191.
7. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81:**559–575.
8. Delaneau O, Zagury J-F, Marchini J: **Improved whole-chromosome phasing for disease and population genetic studies.** *Nat Genet* 2013, **10:**5–6.
9. The Genome of the Netherlands Consortium: **Whole-genome sequence variation, population structure and demographic history of the Dutch population.** *Nat Genet* 2014, **46:**818–825.
10. Swertz MA, Dijkstra M, Adamusiak T, van der Velde JK, Kanterakis A, Roos ET, Lops J, Thorisson GA, Arends D, Byelas G, Muilu J, Brookes AJ, de Brock EO, Jansen RC, Parkinson H: **The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button.** *BMC Bioinformatics* 2010, **11:**S12.
11. Deelen P, Menelaou A, van Leeuwen EM, Kanterakis A, van Dijk F, Medina-Gomez C, Francioli LC, Hottenga JJ, Karssen LC, Estrada K, Kreiner-Moller E, Rivadeneira F, van Setten J, Gutierrez-Achury J, Westra H-J, Franke L, van Enckevort D, Dijkstra M, Byelas H, van Duijn CM, Consortium G of the N, de Bakker PIW, Wijmenga C, Swertz MA: **Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'.** *Eur J Hum Genet* 2014, **22:**1321–1326.
12. Almeida R, Ricaño-Ponce I, Kumar V, Deelen P, Szperl A, Trynka G, Gutierrez-Achury J, Kanterakis A, Westra H-J, Franke L, Swertz MA, Platteel M, Bilbao JR, Barisani D, Greco L, Mearin L, Wolters VM, Mulder C, Mazzilli MC, Sood A, Cukrowska B, Núñez C, Pratesi R, Withoff S, Wijmenga C: **Fine mapping of the celiac disease-associated LPP locus reveals a potential functional variant.** *Hum Mol Genet* 2014, **23:**2481–2489.
13. Bonder MJ, Kasela S, Kals M, Tamm R, Lokk K, Barragan I, Buurman WA, Deelen P, Greve J-W, Ivanov M, Rensen SS, van Vliet-Ostaptchouk JV, Wolfs MG, Fu J, Hofker MH, Wijmenga C, Zhernakova A, Ingelman-Sundberg M, Franke L, Milani L: **Genetic and epigenetic regulation of gene expression in fetal and adult human livers.** *BMC Genomics* 2014, **15:**860.
14. Tigchelaar EF, Zhernakova A, Dekens JAM, Hermes G, Baranska A, Mujagic Z, Swertz MA, Muñoz AM, Deelen P, Cénit MC, Franke L, Scholtens S, Stolk RP, Wijmenga C, Feskens EJM: **An introduction to LifeLines DEEP: study design and baseline characteristics.** *bioRxiv* 2014, [http://dx.doi.org/10.1101/009217].

doi:10.1186/1756-0500-7-901
Cite this article as: Deelen *et al.*: Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Research Notes* 2014 **7**:901.