IMPUTE2 Final Project

Introduction to IMPUTE2

IMPUTE2 is a publicly available software to phase and impute genotypic data developed by Bryan Howie and Jonathan Marchini.

Phasing: Statistical estimation of haplotypes from genotype data. Colloquially, it is the process to estimate and assign haplotypes of the samples that were genotyped.

Imputation: Statistical inference of unobserved genotypes (untyped genetic variants) from observed genotypes (typed genetic variants) using a genomic reference panel. Imputation is achieved by using known haplotypes from a reference panel (i.e. HapMap or 1000 Genomes Projects) to predict the untyped genotypes in genomic data.

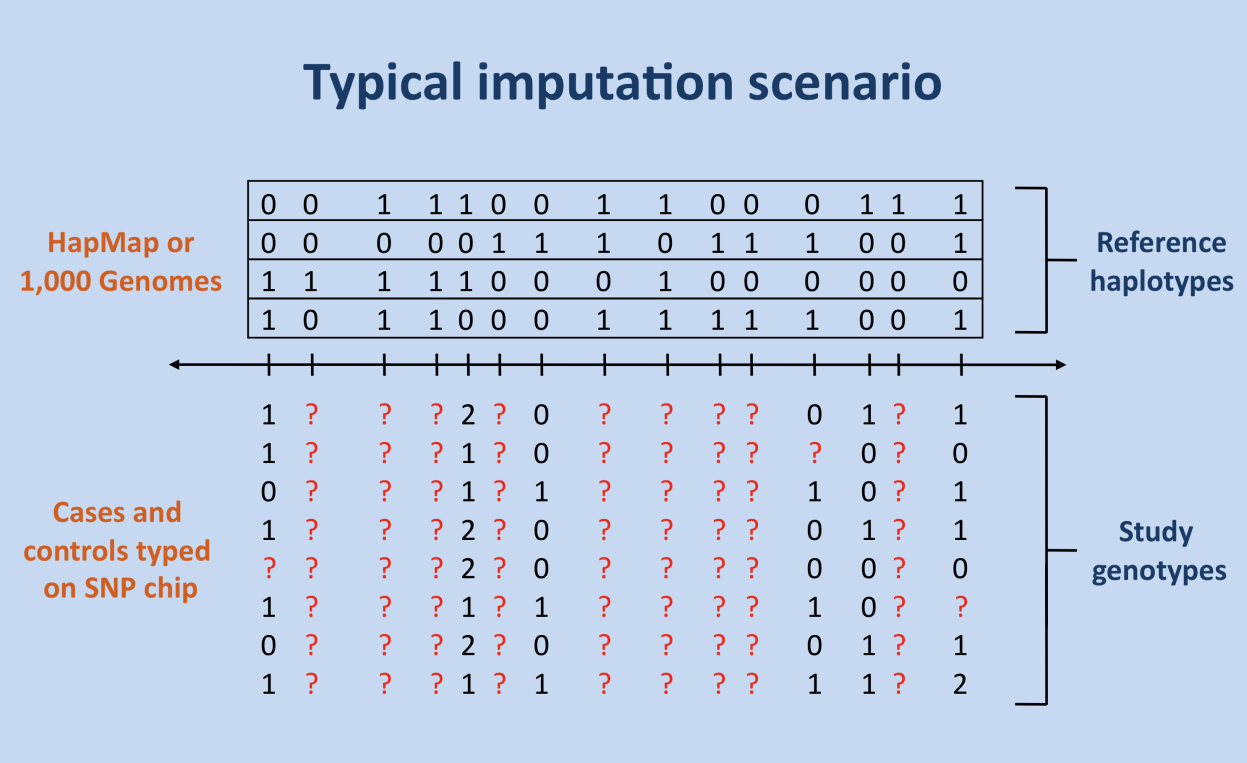

Figure 1: alt text

Why imputation is useful?

Imputation is mostly used in the context of genome-wide association studies (GWAS). Genotyping arrays used for GWAS are designed based on the idea of tag SNPs. A tag SNP is a representative SNP in a region of the genome with high linkage disequilibrium that represents a haplotype. Genotyping platforms only genotype these selected tag SNPs (typically 750K to 1M SNPs) and do not genotype every SNP in the genome. By using the genotype data obtained from tag SNPs and a genomic reference panel, an investigator can impute the rest of the genome, increasing the overall coverage genome-wide. Depending on the refrence panel used,

imputation can also predict insertions, deletions and rare variants based on the genotype data obtained from typed SNPs.

After quality control, some individuals may be missing the genotype information for some typed SNPs or a typed SNP may fail the quality control altogether for all individuals. Imputing can estimate and fill in the missing genotype information in the typed data.

Imputation also makes it possible to compare and/or integrate genotype data that was produced from different genotyping platforms (i.e. Illumina vs Affymetrix SNP arrays) as different platforms may use different sets of tag SNPs. By imputing genomic data from two different platforms, an investigator can yield a much larger overlap between these data sets to conduct meta analyses.

How IMPUTE2 works?

SNPs are first divided into two sets: a set T that is typed in both the study sample and reference panel, and a set U that is untyped in the study sample but typed in the reference panel. The algorithm involves estimating haplotypes at SNPs in T (using Hidden Markov Models as implemented in IMPUTE v1) and then imputing alleles at SNPs in U conditional on the current estimated haplotypes. By iterating these steps using a Markov chain Monte Carlo (MCMC) IMPUTE2 provides accurate results in a reasonable amount of time. (Marchini et. al 2012)

IMPUTE2 Tutorial with NARAC Data

Installing IMPUTE2

Imputation procedures heavily depend on the genomic locations to match typed data to a reference panel. Since gene locations are based on the human genome reference sequence, which is updated and changed by new releases, it is very important to select the correct reference panel that is mapped to the same genome build as the study data. Also in order to match the haplotypes, it is essential that both the study data and the reference panel have their allele codings aligned to a fixed reference (usually the human genome reference sequence).

Determine SNP locations based on given SNP IDs

HapMap 3 haplotypes – NCBI build 36 (hg18) coordinates , therefore SNPs should be mapped to a 36 build.

SNP locations for Homo sapiens (dbSNP Build 130)

Bioconductor version: Release (3.4)

SNP locations and alleles for Homo sapiens extracted from NCBI dbSNP Build 130. The source data files used for this package were created by NCBI on 5-6 May, 2009, and contain SNPs mapped to reference genome NCBI Build 36.1, which is identical to the hg18 genome from UCSC. Therefore, the SNPs in this package can be “injected” in BSgenome.Hsapiens.UCSC.hg18 and they will land at the correct location.

Citation (from within R, enter citation(“SNPlocs.Hsapiens.dbSNP.20090506”)): Pages H (2015). SNPlocs.Hsapiens.dbSNP.20090506: SNP locations for Homo sapiens (dbSNP Build 130). R package version 0.99.9. Installation

To install this package, start R and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("SNPlocs.Hsapiens.dbSNP.20090506")
```

conver ped and map to GEN

```
gtool -P --ped narac_hapmap3.ped --map narac_hapmap3.map --family --binary_phenotype --og narac_hapmap3
```

impute output

```
======================
 IMPUTE version 2.3.2
======================

Copyright 2008 Bryan Howie, Peter Donnelly, and Jonathan Marchini
Please see the LICENCE file included with this program for conditions of use.

The seed for the random number generator is 926121791.

Command-line input: impute2 -m narac_hapmap3_IMP2.map -h ../hapmap3/hapmap3_r2_b36_chr6.haps -l ../hapma

---------------------------------
 Nomenclature and data structure
---------------------------------

     Panel 0: phased reference haplotypes
     Panel 2: unphased study genotypes

For optimal results, each successive panel (0,1,2) should contain a subset of the SNPs in the previous p

-------------
 Input files
-------------

         Panel 0 haplotypes: ../hapmap3/hapmap3_r2_b36_chr6.haps
         Panel 0 hap legend: ../hapmap3/hapmap3_r2_b36_chr6.legend
          Panel 2 genotypes: narac_hapmap3.gen
                genetic map: narac_hapmap3_IMP2.map

--------------
 Output files
--------------

                main output: narac.chr6.impute2
                SNP QC info: narac.chr6.impute2_info
             sample QC info: narac.chr6.impute2_info_by_sample
                run summary: narac.chr6.impute2_summary
                warning log: narac.chr6.impute2_warnings

-----------------
 Data processing
-----------------

-reading genetic map from -m file
 --filename=[narac_hapmap3_IMP2.map]
 --read 196 SNPs in the analysis interval+buffer region
```

```
-reading Panel 2 genotypes from -g file
 --filename=[narac_hapmap3.gen]
 --detected 2062 individuals
 --read 196 SNPs in the analysis interval+buffer region

-reading Panel 0 haplotypes from -h and -l files
 --filename=[../hapmap3/hapmap3_r2_b36_chr6.haps]
 --filename=[../hapmap3/hapmap3_r2_b36_chr6.legend]
 --detected 2022 haplotypes
 --read 1671 SNPs in the analysis interval+buffer region

-removing SNPs that violate the hierarchical data requirements
 --no SNPs removed

-removing reference-only SNPs from buffer region
 --removed 714 SNPs

-checking strand alignment between Panel 2 and Panel 0 by allele labels
 --flipped strand due to allele mismatch at 93 out of 196 SNPs in Panel 2

-checking strand alignment between Panel 2 and Panel 0 by MAF at A/T and C/G SNPs lacking explicit stra
 --flipped strand due to allele frequency discordance at 0 out of 196 SNPs in Panel 2

-aligning allele labels between panels

-removing non-aligned genotyped SNPs
 --removed 0 out of 167 SNPs with data in multiple panels

--------------
 Data summary
--------------

[type 0 = SNP in Panel 0 only]
[type 1 = SNP in Panel 1]
[type 2 = SNP in Panel 2 and all ref panels]
[type 3 = SNP in Panel 2 only]

-Upstream buffer region
 --0 type 0 SNPs
 --0 type 1 SNPs
 --0 type 2 SNPs
 --0 type 3 SNPs
 --0 total SNPs

-Downstream buffer region
 --0 type 0 SNPs
 --0 type 1 SNPs
 --0 type 2 SNPs
 --0 type 3 SNPs
 --0 total SNPs

-Analysis region (as defined by -int argument)
 --790 type 0 SNPs
```

```
 --0 type 1 SNPs
 --167 type 2 SNPs
 --29 type 3 SNPs
 --986 total SNPs

-Output file
 --790 type 0 SNPs
 --0 type 1 SNPs
 --167 type 2 SNPs
 --29 type 3 SNPs

-In total, 986 SNPs will be used in the analysis, including 167 Panel 2 SNPs

-making initial haplotype guesses for Panel 2 by phasing hets at random and imputing missing genotypes r

-setting storage space
-setting mutation matrices
-setting switch rates

----------------
 Run parameters
----------------

        reference haplotypes: 2022 [Panel 0]
           study individuals: 2062 [Panel 2]
           sequence interval: [29700000,30300000]
                      buffer: 250 kb
                          Ne: 20000
           input call thresh: 0.900
     burn-in MCMC iterations: 10
       total MCMC iterations: 30 (20 used for inference)
      HMM states for phasing: 80 [Panel 2]
   HMM states for imputation: 500 [Panel 0->2]
                active flags: <-align_by_maf_g>

---------
 Run log
---------

MCMC iteration [1/30]

MCMC iteration [2/30]

MCMC iteration [3/30]

RESETTING PARAMETERS FOR "SURROGATE FAMILY" MODELING
-setting mutation matrices
-setting switch rates

MCMC iteration [4/30]

MCMC iteration [5/30]

MCMC iteration [6/30]
```

```
MCMC iteration [7/30]

MCMC iteration [8/30]

MCMC iteration [9/30]

MCMC iteration [10/30]

MCMC iteration [11/30]

MCMC iteration [12/30]

MCMC iteration [13/30]

MCMC iteration [14/30]

MCMC iteration [15/30]

MCMC iteration [16/30]

MCMC iteration [17/30]

MCMC iteration [18/30]

MCMC iteration [19/30]

MCMC iteration [20/30]

MCMC iteration [21/30]

MCMC iteration [22/30]

MCMC iteration [23/30]

MCMC iteration [24/30]

MCMC iteration [25/30]

MCMC iteration [26/30]

MCMC iteration [27/30]

MCMC iteration [28/30]

MCMC iteration [29/30]

MCMC iteration [30/30]

diploid sampling success rate: 0.998

haploid sampling success rate: (no haploid sampling performed)
```

```
---------------------------------
 Imputation accuracy assessment
---------------------------------

The table below is based on an internal cross-validation that is performed during each IMPUTE2 run. For

In the current analysis, IMPUTE2 masked, imputed, and evaluated 339959 genotypes that were called with

When the masked study genotypes were imputed with reference data from Panel 0, the concordance between
```

Interval	#Genotypes	%Concordance	Interval	%Called	%Concordance
[0.0-0.1]	0	0.0	[>= 0.0]	100.0	99.3
[0.1-0.2]	0	0.0	[>= 0.1]	100.0	99.3
[0.2-0.3]	0	0.0	[>= 0.2]	100.0	99.3
[0.3-0.4]	13	46.2	[>= 0.3]	100.0	99.3
[0.4-0.5]	119	63.0	[>= 0.4]	100.0	99.3
[0.5-0.6]	1485	61.5	[>= 0.5]	100.0	99.3
[0.6-0.7]	2321	83.3	[>= 0.6]	99.5	99.5
[0.7-0.8]	3901	91.1	[>= 0.7]	98.8	99.6
[0.8-0.9]	7522	96.1	[>= 0.8]	97.7	99.7
[0.9-1.0]	324598	99.8	[>= 0.9]	95.5	99.8