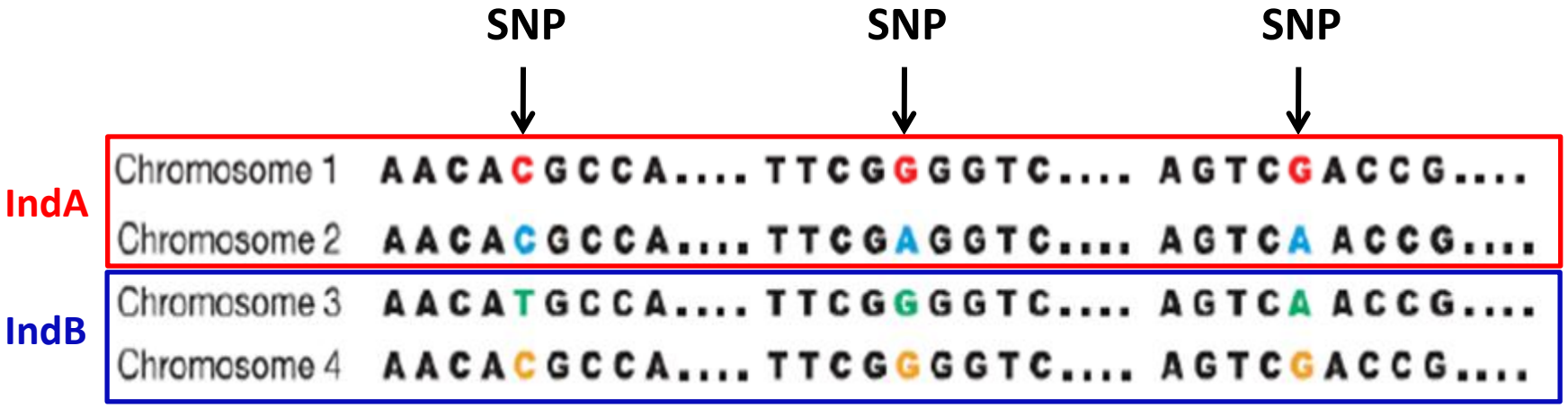
# Grupo de Imputação e Determinação Haplotípica

Wagner C S Magalhães
Eduardo Martín Tarazona Santos

# SNPs

- SNPs are the most common form of germline variation.
- Occur when a single nucleotide differs between paired chromosomes in an individual, or between individuals in a population.
- **~25 million SNPs in the human genome** with a minor allele frequency of >1%.
- Only a (very small) subset have pathogenic potential.



# Haplotype inference

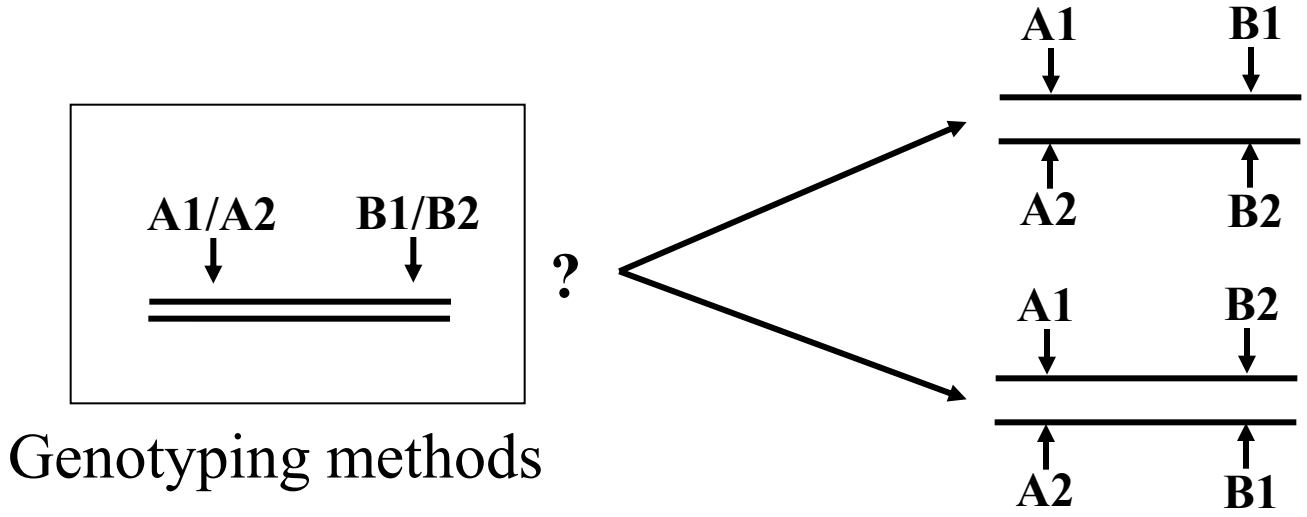
- Problem:
  - Most of experimental methods of sequencing or genotyping, only provide genotype information at each locus: the unordered  pair of alleles at the locus.

- A haplotype is a sequence of alleles that are on the same physical chromosome (i.e. that are inherited from the same parent).

# Haplotype inference

- Understanding the interplay of genetic variation and disease,
- Imputation  of untyped genetic variation,
- Detecting genotype error,
- Inferring human demographic history,
- Inferring points of recombination,
- Detecting recurrent mutation,
- Signatures of natural selection

**The phase problem with double (multiple) heterozygous sites:**
**which alleles are on the same chromosome?**




Genotyping methods

Solutions

-Labor intensive molecular / cellular techniques

-Long range PCR, Cell fusion

**-Statistical methods (work well, example: Phase algorithm or Expectation-Maximization methods embedded in gen-epi software)**

**-Segregation analysis by typing the parents**

# Haplotype inference

Son ATCA[G/C][C/G][A/C][T/A]CA[G/C]A.....

True haplotypes .....ATCAGCATCAGA.....

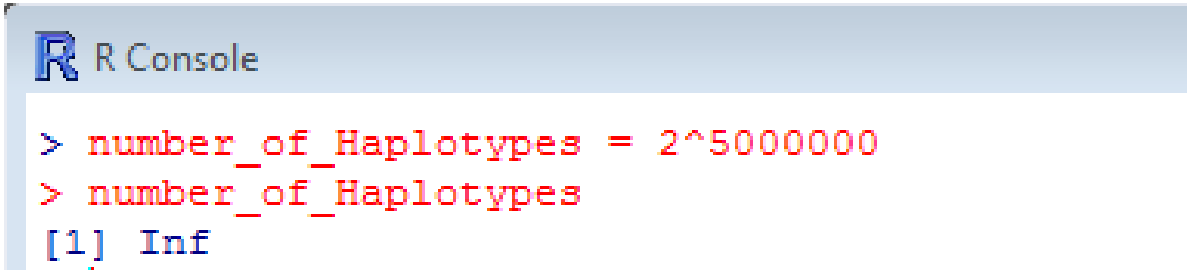
.....ATCACGCACACA.....

# Number of Haplotypes

- Possible haplotypes = 32

m = loci unphased $2^m$ haplotypes

Now imagine -------> m = 2.5 M or even 5 M



# Haplotype Phasing

- Unrelated individuals can be phased by considering sets of common haplotypes that can explain the observed genotype data. The number of unrelated individuals present in a sample is a crucial factor in determining how well the phase can be estimated: the more individuals, the better the estimation.

- Related individuals, by contrast, can be phased by considering haplotypes that are shared identical-by-descent between individuals within families. This within-family information on identity-by-descent (IBD)

One way to improve the power of GWAS is to infer haplotype phase and use a haplotype-based method for association testing, in addition to applying single-marker association testing methods

(Browning and Browning 2007).

| Group | Feature | BEAGLE | IMPUTE | MACH |
|---|---|---|---|---|
| **Accessibility** | Operating system | Java (platform independent) | Linux, Windows, MacOS X, Solaris | Linux, MacOS X |
| | Licence | Free | Free for academic use | Not clear |
| | Source code | Not available | Not available | Availability announced |
| | Documentation | Commendable | Clearly structured | Incomplete |
| | Authors' response | Quick and detailed | Quick | Quick |
| | GUI | No | No | No |
| **Input** | Genotype format | Discrete; custom format | Probabilities; custom format | QTDT (Linkage) |
| | Reference format | Custom format | Custom format; prepared HapMap reference available | HapMap format (custom format) |
| | Conversion utilities | Yes | Yes | No |
| **Processing** | Target of imputation | Chromosomes | Chromosomes or segments | Chromosomes |
| | Memory-saving mode | Yes | No | Yes |
| | Known checking errors | None | Missing probability and input check | Problematic handling of missing reference |
| | Runtime [chr. 6] | 350 minutes | 433 minutes | 2781 minutes |
| | Maximum memory allocation | 2 GB | 14 GB [<1 GB with ~10 MB segments] | 7 GB |
| | Memory-saving mode | Yes | No | Yes |
| | Strand orientation | Check | Check + autoflip | Check + autoflip |

# Data Imputation

- **Imputation** (Statistics)
  - Is the substitution of some value for missing data. Once all missing values have been imputed, the dataset can then be analyzed using standard techniques for complete data.
- **Imputation** (Genetics)
  - Is the estimation of missing genotype values by using the genotypes at nearby SNPs and the haplotype frequencies seen in other individuals.

**Genotype imputation is now an essential tool in the analysis of genome-wide association scans. This technique allows geneticists to accurately evaluate the evidence for association at genetic markers that are not directly genotyped.**

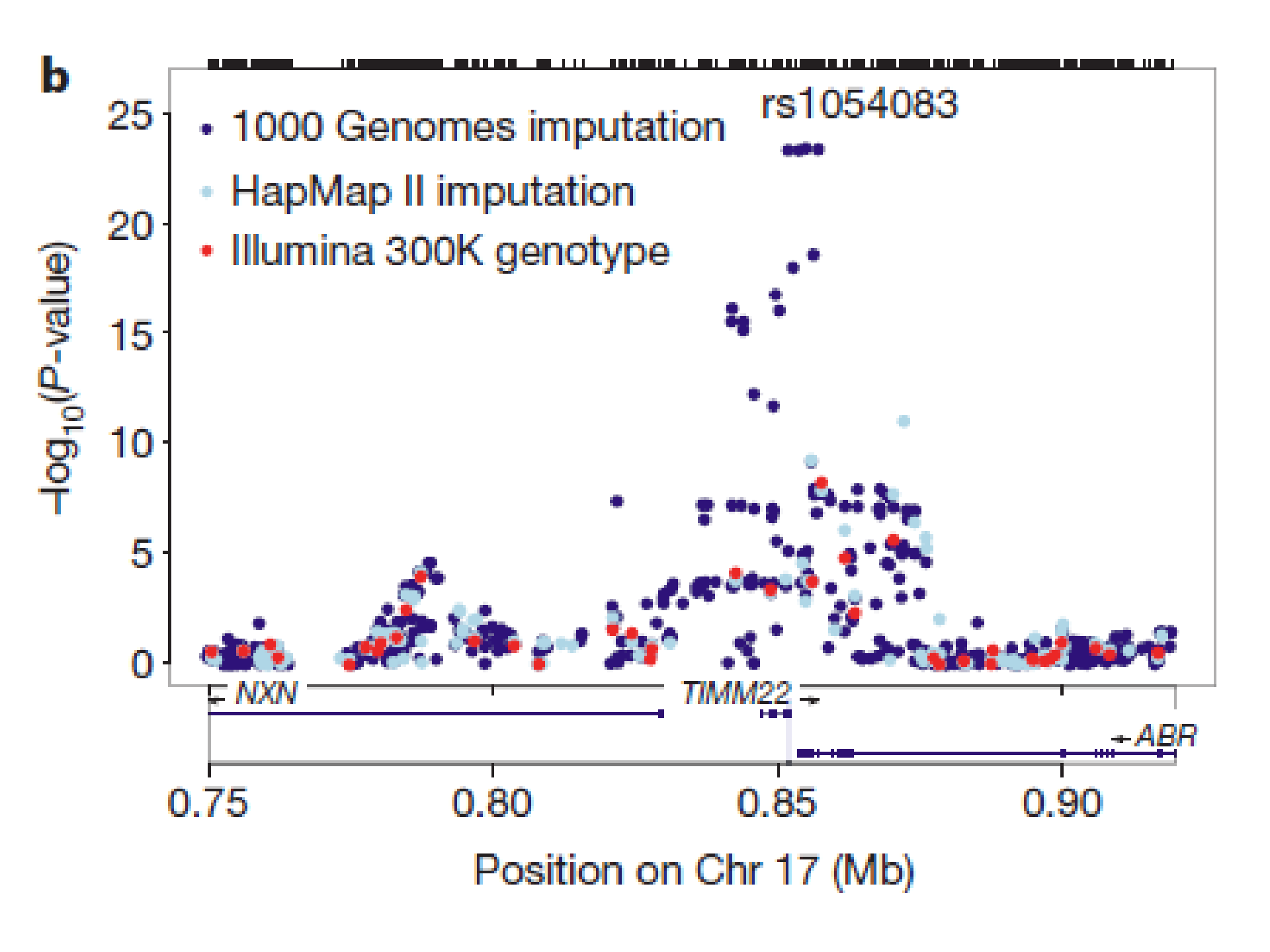
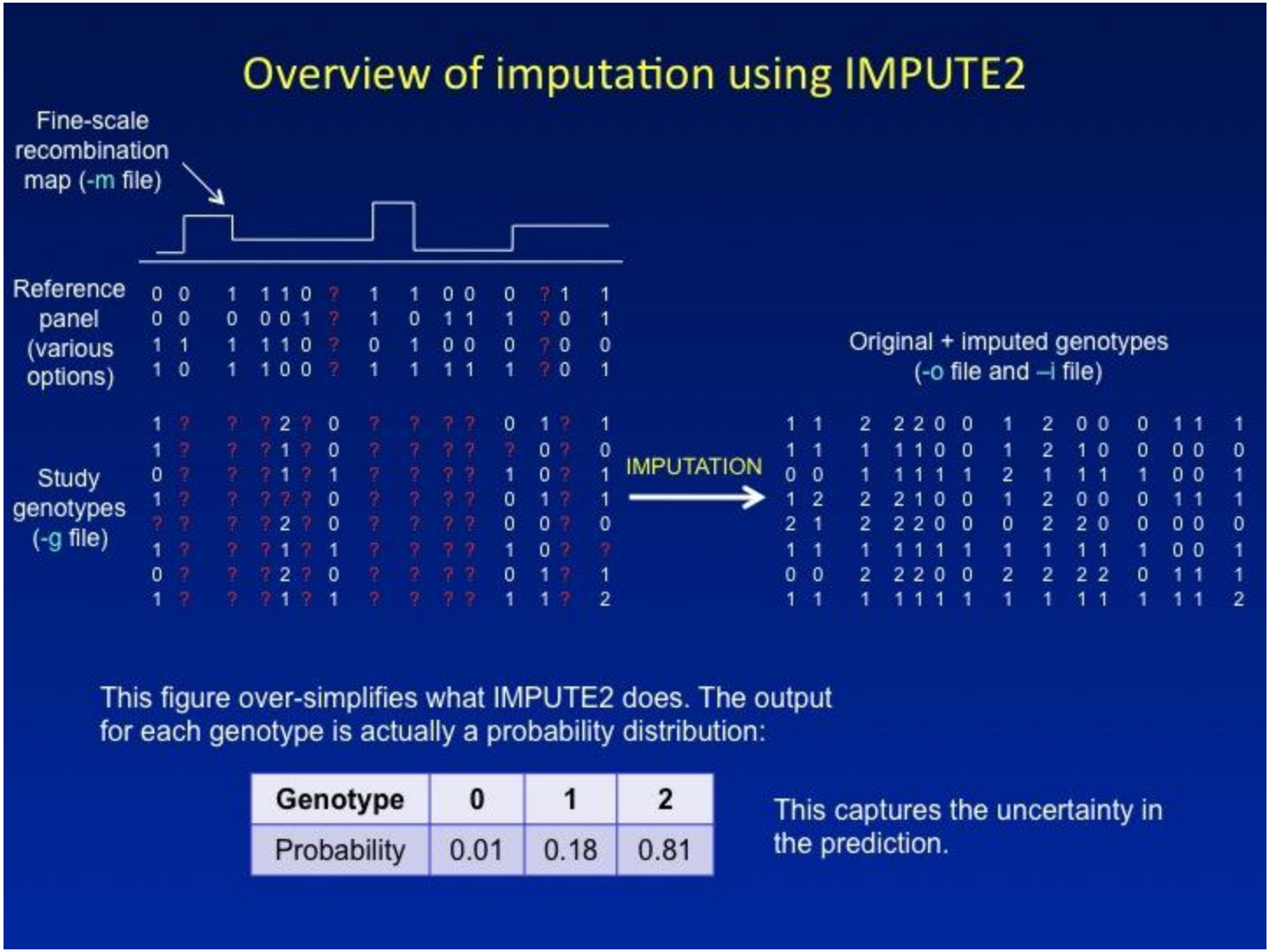
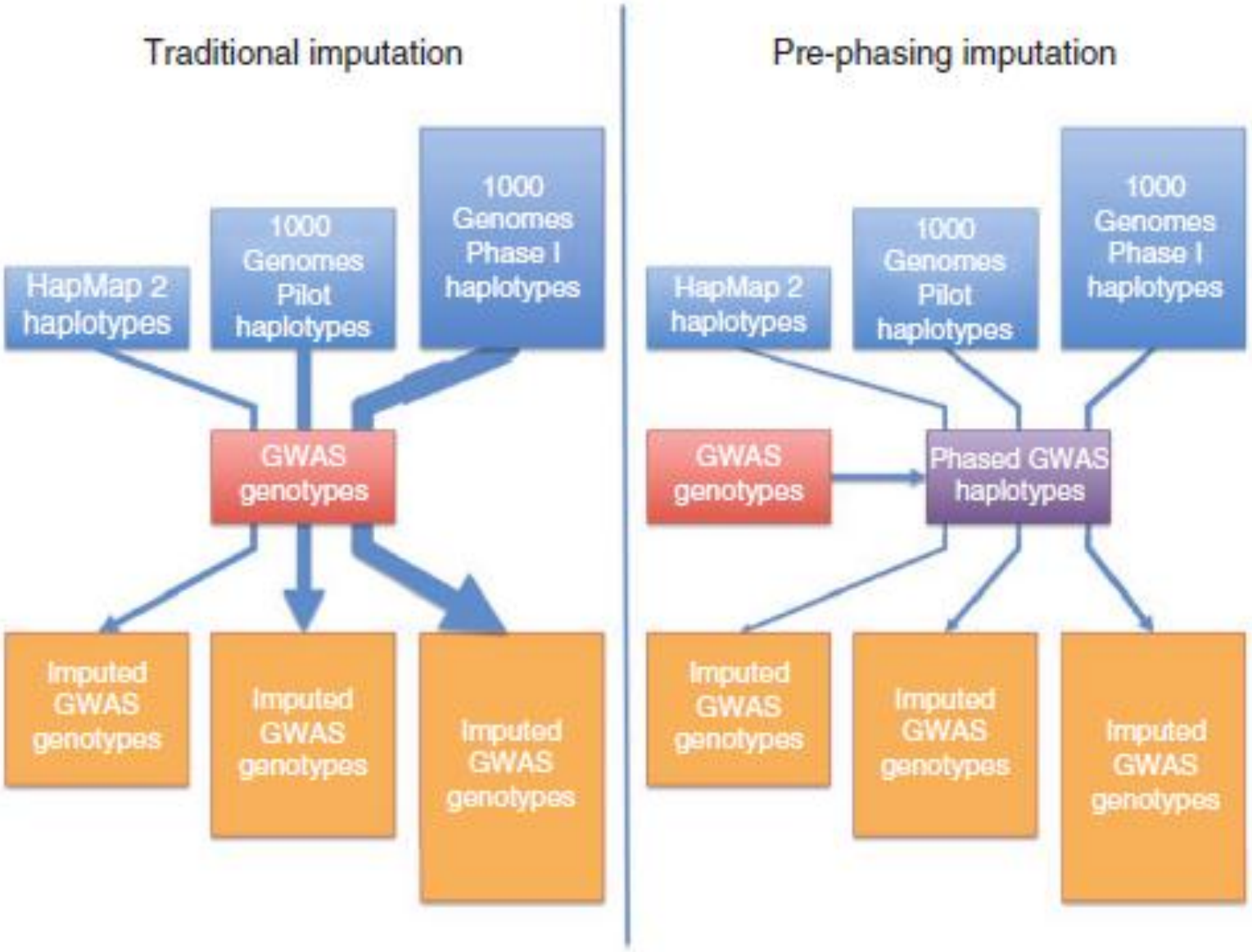

b
rs1054083
1000 Genomes imputation
HapMap II imputation
Illumina 300K genotype
$-\log_{10}$(P-value)
25
20
15
10
5
0
NXN
TIMM22
ABR
0.75
0.80
0.85
0.90
Position on Chr 17 (Mb)

**Figure 1. Schematic drawing of imputation Scenario A.**



# PRE-IMPUTATION FILTERING OF STUDY GENOTYPES

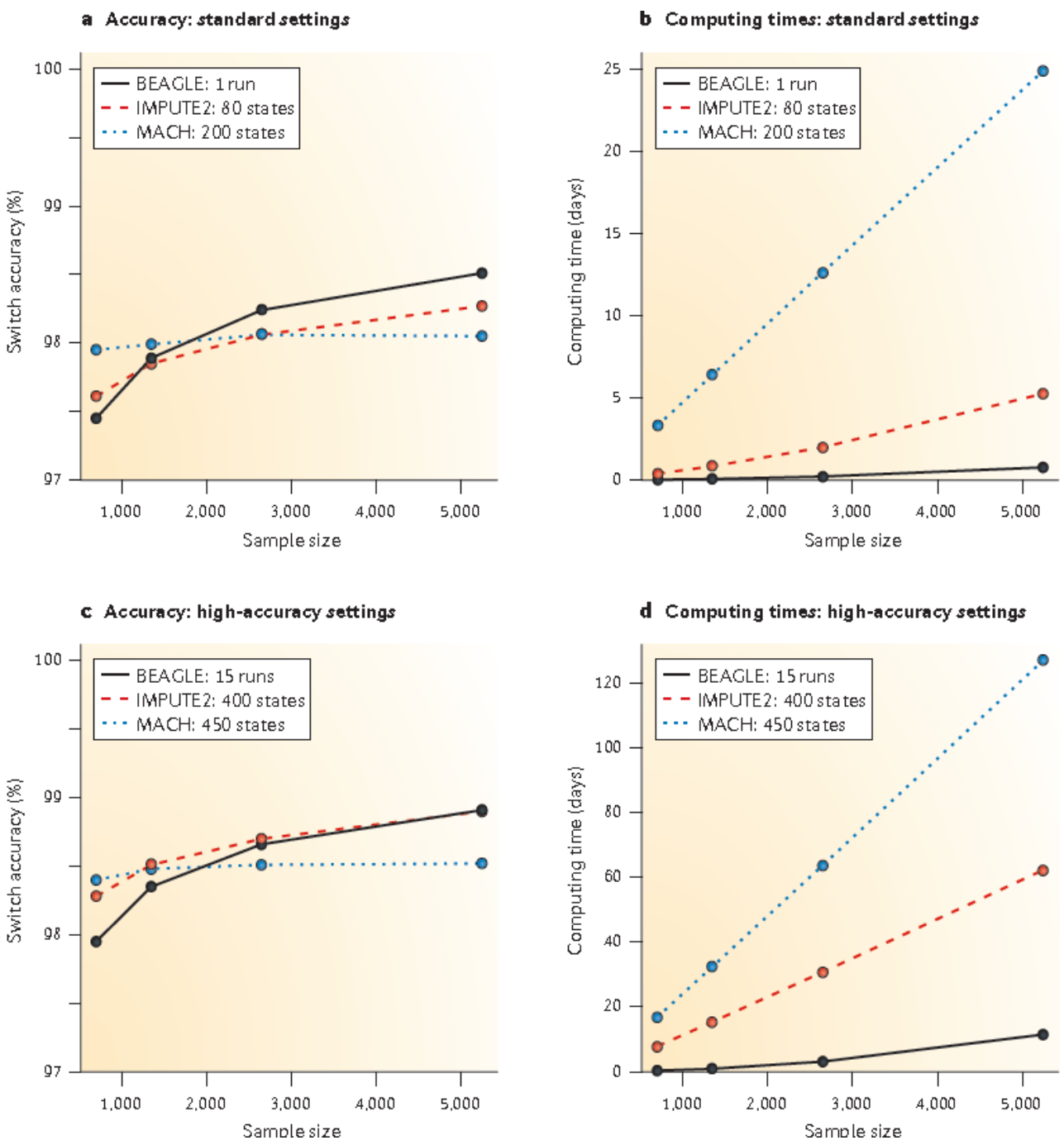
- Remove low-quality variants and individuals
- Take care of their positions on the chromosome
- Strand alignment between study and reference data
- Choose reference panels that match the ancestry of their study samples


Traditional imputation
HapMap 2 haplotypes
1000 Genomes Pilot haplotypes
1000 Genomes Phase I haplotypes
GWAS genotypes
Imputed GWAS genotypes
Imputed GWAS genotypes
Imputed GWAS genotypes
Pre-phasing imputation
HapMap 2 haplotypes
1000 Genomes Pilot haplotypes
1000 Genomes Phase I haplotypes
GWAS genotypes
Phased GWAS haplotypes
Imputed GWAS genotypes
Imputed GWAS genotypes
Imputed GWAS genotypes

| Link to download page | NCBI build | Haplotype release date | Release status |
|---|---|---|---|
| 1000 Genomes Phase I integrated variant set | b37 | Mar 2012 | Includes chrX; updated 19 Apr 2012 |
| 1000 Genomes Phase I (interim) | b37 | Jun 2011 | Includes chrX; updated 19 Apr 2012 |
| 1000 Genomes (2010 interim) | b37 | Dec 2010 | |
| 1000 Genomes Pilot + HapMap 3 | b36 | Jun 2010 / Feb 2009 | |
| 1000 Genomes Pilot | b36 | Jun 2010 | |
| HapMap 3 (release #2) | b36 | Feb 2009 | Includes chrX |
| HapMap 2 (release #24) | b36 | Oct 2008 | |
| HapMap 2 (release #22) | b36 | Jan 2008 | |
| HapMap 2 (release #21) | b35 | Jul 2006 | |

**Table 2 Recommended choices of HapMap reference panel haplotypes for imputing genotypes in Human Genome Diversity Panel different samples**

| These reference panel haplotypes… | …are best for imputing genotypes in these Human Genome Diversity Panel samples |
|---|---|
| CEU | **Europe**: Orcadian, Basque, French, Italian, Sardinian |
| | **Middle East**: Druze |
| CHB + JPT | **East Asia**: Han, Han-Nchina, Dai, Lahu, Miao, Oroqen, She, Tujia, Tu, Xibo, Yi, Mongola,[a] Naxi, Japanese |
| YRI | **Africa**: Bantu, Yoruba, San, Mandenka, MbutiPygmy, BiakaPygmy |
| Combined (CEU, CHB, JPT, YRI) | **Europe**: Adygei, Russian, Tuscan |
| | **Middle East**: Mozabite, Bedouin, Palestinian |
| | **Asian**: Balochi, Brahui, Makrani, Sindhi, Pathan, Burusho, Hazara, Uygur, Kalash |
| | **East Asia**: Daur, Hezhen, Mongola,[*] Cambodian, Yakut |
| | **Oceania**: Melanesian, Papuan |
| | **Americas**: Colombian, Karitiana, Surui, Maya, Pima |


a Accuracy: standard settings
BEAGLE: 1 run
IMPUTE2: 80 states
MACH: 200 states
Switch accuracy (%)
100
99
98
97
1,000
2,000
3,000
4,000
5,000
Sample size
b Computing times: standard settings
BEAGLE: 1 run
IMPUTE2: 80 states
MACH: 200 states
Computing time (days)
25
20
15
10
5
0
1,000
2,000
3,000
4,000
5,000
Sample size
c Accuracy: high-accuracy settings
BEAGLE: 15 runs
IMPUTE2: 400 states
MACH: 450 states
Switch accuracy (%)
100
99
98
97
1,000
2,000
3,000
4,000
5,000
Sample size
d Computing times: high-accuracy settings
BEAGLE: 15 runs
IMPUTE2: 400 states
MACH: 450 states
Computing time (days)
120
100
80
60
40
20
0
1,000
2,000
3,000
4,000
5,000
Sample size

## Table 1. Running times and memory requirements for various algorithms in Scenario A.

| Method | Avg. running time (min) | Avg. required RAM (MB) |
|---|---|---|
| BEAGLE | 56 | 3100 |
| BEAGLE (50iter) | 392 | 3200 |
| fastPHASE (K = 20) | 397 | 8 |
| fastPHASE (K = 30) | 855 | 16 |
| IMPUTE v1 | 43 | 1000 |
| IMPUTE v2 (k = 40) | 270 | 155 |
| IMPUTE v2 (k = 80) | 505 | 180 |
| MACH | 105 | 80 |

Running times are in minutes (min) and RAM requirements are in megabytes (MB). Each entry in the table is an average across four runs on different 7.5 Mb regions of chromosome 10. Each analysis included a reference panel of 120 chromosomes (CEU HapMap) and a study sample of 1,377 individuals genotyped on the Affymetrix 500 K SNP chip.
doi:10.1371/journal.pgen.1000529.t001

Howie BN, Donnelly P, Marchini J (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet 5(6): e1000529. doi:10.1371/journal.pgen.1000529
http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.1000529

AMIA Annual Symposium Proceedings Archive

AMIA Annu Symp Proc. 2011; 2011: 1013–1018.
Published online 2011 October 22.
PMCID: PMC3243280

# The Effect of Reference Panels and Software Tools on Genotype Imputation

Kwangsik Nho, PhD,[1,2] Li Shen, PhD,[2,3] Sungeun Kim, PhD,[2,3] Shanker Swaminathan, BTech,[2,4] Shannon L. Risacher, BS,[2] Andrew J. Saykin, PsyD,[2,3,4] and the Alzheimer's Disease Neuroimaging Initiative (ADNI)

Author information ► Copyright and License information ►

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 27 June 2012
doi: 10.3389/fgene.2012.00117

# Genotype imputation for Latinos using the HapMap and 1000 Genomes Project reference panels

*Xiaoyi Gao[1,2]*, Talin Haritunians[3], Paul Marjoram[2], Roberta Mckean-Cowdin[2], Mina Torres[1], Kent D. Taylor[3], Jerome I. Rotter[3], William J. Gauderman[2] and Rohit Varma[1,2]*

[1] *Department of Ophthalmology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA*
[2] *Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA*
[3] *Medical Genetics Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA*

***Edited by:***
*Rongling Wu, Pennsylvania State University, USA*

***Reviewed by:***
*Ashok Ragavendran, Purdue University, USA*
*Wei Hou, University of Florida, USA*

****Correspondence:***
*Xiaoyi Gao, Departments of Ophthalmology and Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA.*
*e-mail: xiaoyiga@usc.edu*

Genotype imputation is a vital tool in genome-wide association studies (GWAS) and meta-analyses of multiple GWAS results. Imputation enables researchers to increase genomic coverage and to pool data generated using different genotyping platforms. HapMap samples are often employed as the reference panel. More recently, the 1000 Genomes Project resource is becoming the primary source for reference panels. Multiple GWAS and meta-analyses are targeting Latinos, the most populous, and fastest growing minority group in the US. However, genotype imputation resources for Latinos are rather limited compared to individuals of European ancestry at present, largely because of the lack of good reference data. One choice of reference panel for Latinos is one derived from the population of Mexican individuals in Los Angeles contained in the HapMap Phase 3 project and the 1000 Genomes Project. However, a detailed evaluation of the quality of the imputed genotypes derived from the public reference panels has not yet been reported. Using simulation studies, the Illumina OmniExpress GWAS data from the Los Angles Latino Eye Study and the MACH software package, we evaluated the accuracy of genotype imputation in Latinos. Our results show that the 1000 Genomes Project AMR + CEU + YRI reference panel provides the highest imputation accuracy for Latinos, and that also including Asian samples in the panel can reduce imputation accuracy. We also provide the imputation accuracy for each autosomal chromosome using the 1000 Genomes Project panel for Latinos. Our results serve as a guide to future imputation based analysis in Latinos.

**Keywords: genotype imputation, Latino, HapMap Project, 1000 Genomes Project**

**Table 1 | Phased haplotypes downloaded from the MACH website.**

| Population | Code | HapMap phase 3 | 1000 Genomes project | |
|---|---|---|---|---|
| | | Number of haplotypes | Number of haplotypes | Group code |
| Mexican ancestry in Los Angeles, California | MEX | 104 | 132 | AMR |
| Colombian in Medellin, Colombia | CLM | | 120 | |
| Puerto Rican in Puerto Rico | PUR | | 110 | |
| CEPH in Utah residents | CEU | 234 | 174 | EUR |
| Tuscans in Italy | TSI | 176 | 196 | |
| Finnish individuals from Finland | FIN | | 186 | |
| British individuals from England and Scotland | GBR | | 178 | |
| Iberian populations in Spain | IBS | | 28 | |
| Yoruba in Ibadan, Nigeria | YRI | 230 | 176 | AFR |
| African ancestry individuals from Southwest, US | ASW | | 122 | |
| Luhya in Webuye, Kenya | LWK | | 194 | |
| Han Chinese in Beijing, China | CHB | 168 | 194 | ASN |
| Japanese in Tokyo, Japan | JPT | 172 | 178 | |
| Han Chinese South, China | CHS | | 200 | |
| Total | | 1084 | 2188 | |

*The population labels were obtained from the HapMap and the 1000 Genomes Project websites.*

Published in final edited form as:
*Nat Genet*. ; 44(1): 6–7. doi:10.1038/ng.1044.

# Improved Imputation of Common and Uncommon Single Nucleotide Polymorphisms (SNPs) with a New Reference Set

**Zhaoming Wang**[1,2], **Kevin B. Jacobs**[1,2], **Meredith Yeager**[1,2], **Amy Hutchinson**[1,2], **Joshua Sampson**[2], **Nilanjan Chatterjee**[2], **Demetrius Albanes**[2], **Sonja I. Berndt**[2], **Charles C. Chung**[2], **W. Ryan Diver**[3], **Susan M. Gapstur**[3], **Lauren R. Teras**[3], **Christopher A. Haiman**[4], **Brian E. Henderson**[4], **Daniel Stram**[4], **Xiang Deng**[1,2], **Ann W. Hsing**[2], **Jarmo Virtamo**[5], **Michael A. Eberle**[6], **Jennifer L. Stone**[6], **Mark P. Purdue**[2], **Phil Taylor**[2], **Margaret Tucker**[2], and **Stephen J. Chanock**[2]

| | Populations | | | | | Illumina Array | | |
|---|---|---|---|---|---|---|---|---|
| **Group** | **European American** | **African American** | **African** | **Asian** | **Hap660** | **Hap1** | **Omni1** | **Omni2.5** |
| **ATBC** | 246 | | | | | ✓ | ✓ | ✓ |
| **CPSII** | 227 | | | | | ✓ | ✓ | ✓ |
| **PLCO** | 255 | | | | | ✓ | ✓ | ✓ |
| **PLCO** | | 98 | | | | ✓ | | ✓ |
| **SHNX** | | | | 74 | ✓ | | | ✓ |
| **HapMap** | | | | | | | | |
| CEU | 116 | | | | | | | ✓ |
| CHB | | | | 44 | | | | ✓ |
| JPT | | | | 44 | | | | ✓ |
| TSI | 86 | | | | | | | ✓ |
| YRI | | | 59 | | | | | ✓ |
| **Total** | 930 | 98 | 59 | 162 | | | | |

# Literatura


- **J. Marchini**, B. Howie, S. Myers, G. McVean and P. Donnelly (2007) *A new multipoint method for genome-wide association studies via imputation of genotypes*. Nature Genetics 39 : 906-913

- B Servin and **M Stephens**. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet*, 3(7), Jul 2007.

- **Browning, S. R. and B. L. Browning**, 2007. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. American Journal of Human Genetics, 81:1084-1097

- **Browning, S. R. and B. L. Browning, 2011.** Haplotype phasing: existing methods and new developments. Nature Reviews Genetics, 12: 703-714.