

Towards Assessment of SNP Imputation Methods

Itsik Pe'er
on behalf of the

GAIN Imputation Working Group

SNP Imputation

- C - - G - - - T T - A -

- T - - A - - - G C - T -

A C C T G T G A T T A A G

C T T G A G A G G C T T A

- Guess unobserved genotypes from
 - ◆ Sparse SNPs in the analyzed individuals
 - ◆ Dense training in other samples

SNP Imputation

A C C T G T G A T T A A G

C T T G A G A G G C T T A

A C C T G T G A T T A A G

C T T G A G A G G C T T A

- Probabilities on unobserved genotypes

Motivation/Concerns

- Coverage
 - ◆ Accurately impute much of human variation
- Merge studies w/ different platforms
 - ◆ Differential ability to impute particular SNPs
 - ◆ Allele flipping
- Issues:
 - ◆ Robustness to training population?
 - ◆ Quality of imputation for classes of alleles?

Multiple Methods

- IMPUTE (Marchini)
- MACH (Abecasis)
- TUNA (Nicolae)
- COCAPHASE (Dudbridge)
- SNPMSTAT (Lin)
- PRETSSEL (Pe'er, Koller)

- Need to evaluate different strengths

Evaluation Criteria

- Accuracy
- Completeness
- Computing requirements

Evaluation Criteria

- Accuracy of probabilistic calls
 - ◆ in estimating allele frequency
 - ◆ of probabilities assigned
- Completeness
 - ◆ per accuracy level
 - ◆ per confidence level
- Computing requirements
 - ◆ Time/memory of different processing stages

Evaluation Experiment Design

- Observed:
 - ◆ Affy 500k x 1500 samples 58BC (WTCCC)
- Hidden:
 - ◆ 15,000 nsSNP + Illumina 550k
- Imputation:
 - ◆ Prob(Genotype) ; allele frequency estimates
- Independent evaluating group
(Neale, Daly)

Acknowledgements:

GAIN Imputation Workgroup

- Teri Manolio
- Peter Donnelly
- Nancy Cox
- WTCCC:
Paul Deloukas
Audrey Duncanson
- Methods
 - ◆ Jonathan Marchini
 - ◆ Goncalo Abecasis
 - ◆ Dan Nicolae
 - ◆ Frank Dudbridge
 - ◆ Danyu Lin
- Assessment:
 - ◆ Ben Neale
 - ◆ Mark Daly