To combine results from multiple studies...

- We need to have some methods to combine results from the testing of the same markers = **meta-analysis methods**

To combine results from multiple studies...

- We need to have some methods to combine results from the testing of the same markers = **meta-analysis methods**
- We need the same set of markers studied. This is not always the case (different chips used by different studies). Therefore we need methods to infer/impute the markers not typed in present study = **imputation methods**

Imputations

(and bits on analysis of imputed data)

Yurii Aulchenko

yurii [dot] aulchenko [at] gmail [dot] com

To combine results from multiple studies...

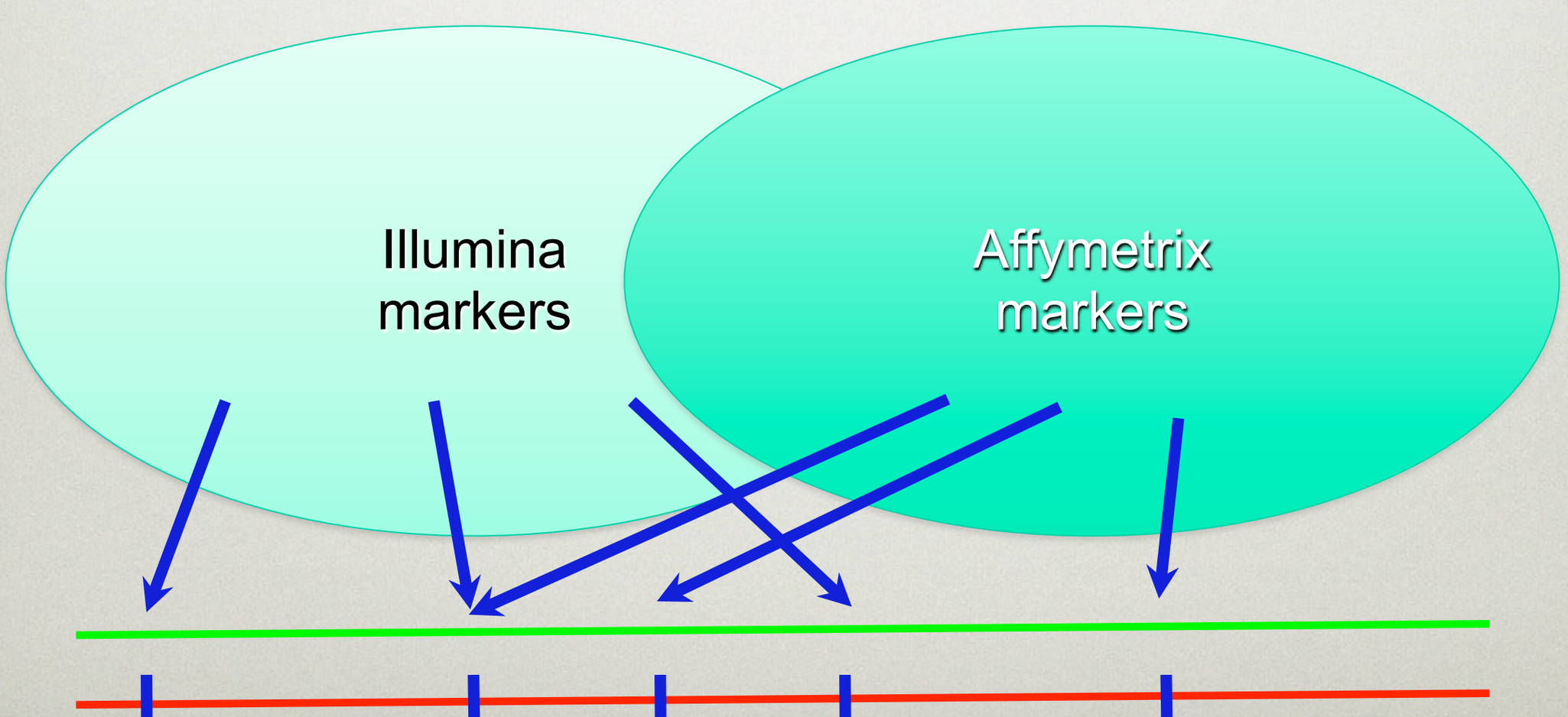
- We need to have some methods to combine results from the testing of the same markers = **meta-analysis methods**

To combine results from multiple studies...

- We need to have some methods to combine results from the testing of the same markers = **meta-analysis methods**
- We need the same set of markers studied. This is not always the case (different chips used by different studies). Therefore we need methods to infer/impute the markers not typed in present study = **imputation methods**

Harmonization through imputations

- Imputation to combine data across platforms
- Statistical models combining evidence across studies (meta-analysis)



Courtesy of Samuli Ripatti

Imputations in general

- For markers within genotyped set, based on flanking marker information
  - Predicts missing genotypes
  - Points out possible genotyping errors
- Extension: use a set of more densely typed samples (e.g. HapMap) to predict markers NOT typed in your set
- "Prediction" from imputations:
  - Not exact, but probabilistic prediction
  - E. g. some person at some SNP is predicted to have TT (90%), TG (7%), or GG (3%).

What is imputation?

Reference haplotypes:

Own data haplotype:

Courtesy of Samuli Ripatti

What is imputation?

Reference haplotypes:

Own data haplotype:

Courtesy of Samuli Ripatti

What is imputation?

Reference haplotypes:

Own data haplotype:

Courtesy of Samuli Ripatti

What is imputation?

Reference haplotypes:

Own data haplotype:

Courtesy of Samuli Ripatti

What is imputation?

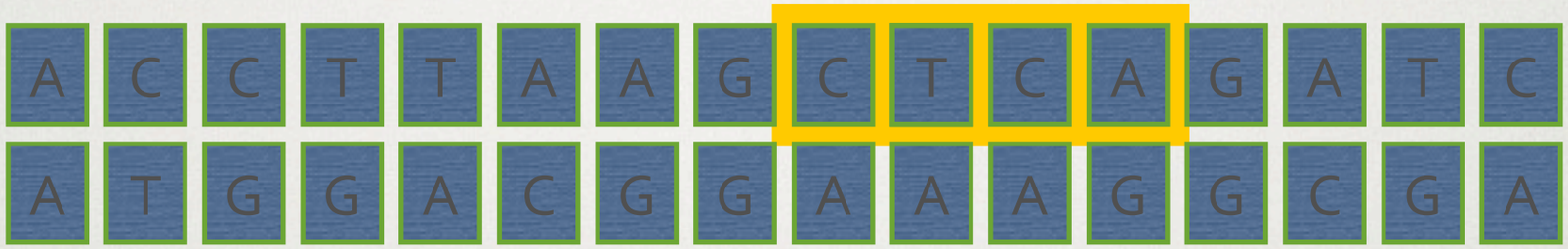
Reference haplotypes:

Own data haplotype:

Courtesy of Samuli Ripatti

What is imputation?

Reference haplotypes:

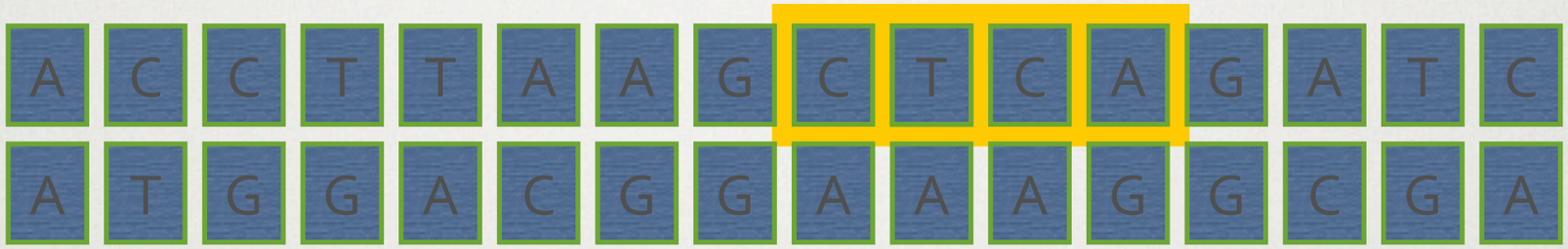


Own data haplotype:

Courtesy of Samuli Ripatti

What is imputation?

Reference haplotypes:



Own data haplotype:

Courtesy of Samuli Ripatti

What is imputation?

Reference haplotypes:

Own data haplotype:

Courtesy of Samuli Ripatti

What is imputation?

Reference haplotypes:

Own data haplotype:

Courtesy of Samuli Ripatti

Imputing: guess the "?"

Reference Haplotypes

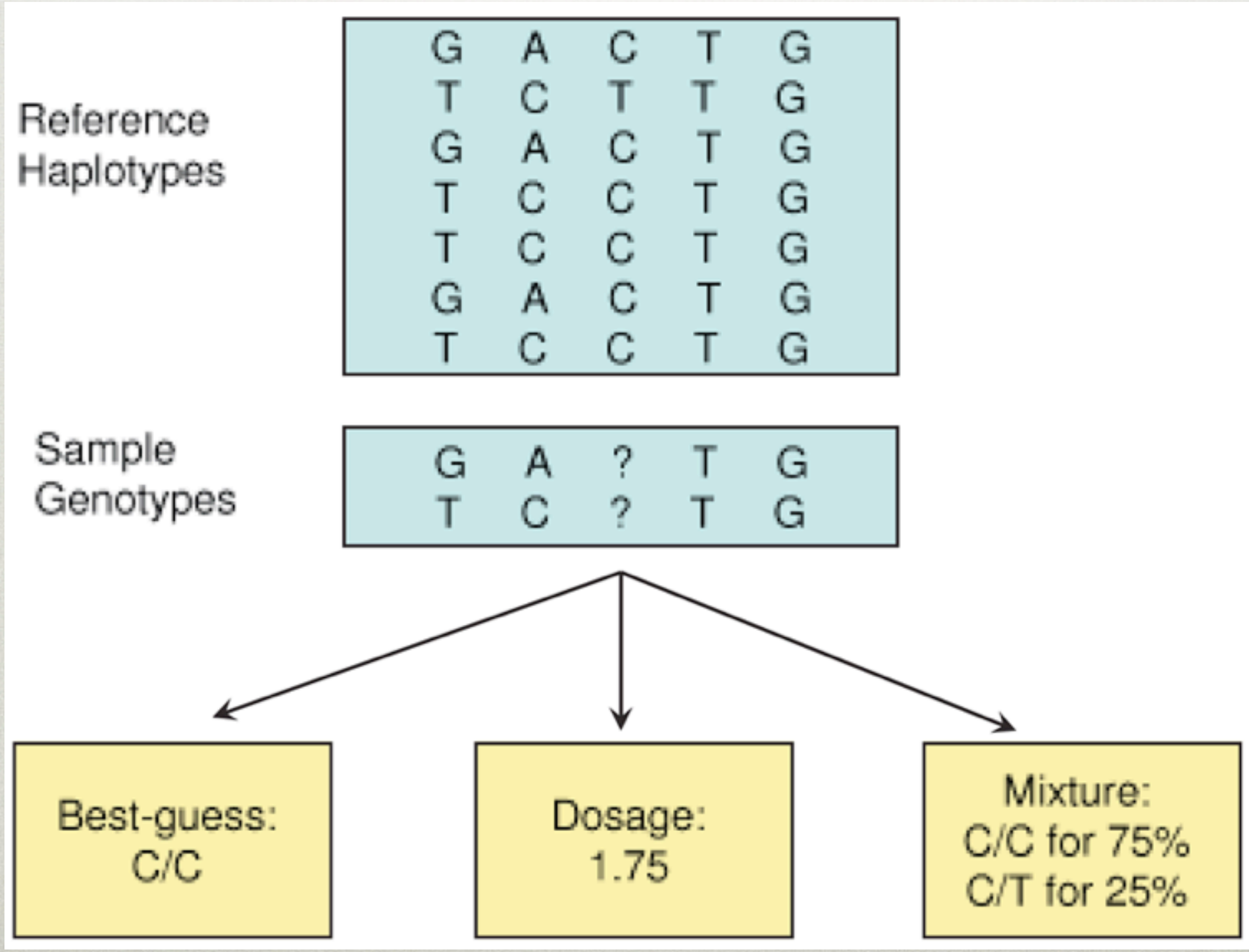
G	A	C	T	G
T	C	T	T	G
G	A	C	T	G
T	C	C	T	G
T	C	C	T	G
G	A	C	T	G
T	C	C	T	G

Sample Genotypes

G	A	?	T	G
T	C	?	T	G

Zheng et al., 2011

Imputing: guess the "?"



Zheng et al., 2011

Imputing: guess the "?"

In "real life", imputations are done using rather complex mathematical models, e.g. Hidden Markov Models, not simple guessing

Zheng et al., 2011

Imputed data

We can not tell the exact genotype, but can estimate posterior probability distribution: $P(g)=\{p_{AA},p_{AB},p_{BB}\}$

Directly typed SNPs: either AA, AB or BB. The probability distribution is *degenerate* (e.g. {*0,1,0*} that is to say AB)

For imputed SNPs, the distribution is not degenerate

How can we analyze imputed data?

- Instead of genotypes, we have probabilities that certain person at certain locus has certain genotype
- How do we relate these probabilities to an outcome (do GWAS)?

Best guess

- Best guess: take the genotype with maximal posterior probability and treat it as if it was true, directly typed
- Drawback: biased estimates, reduced power

Regression onto probabilities

Use the model

$$E[Y_i] = m + b_1 P(g_i=1) + b_2 P(g_i=2)$$

Note this is very similar to model for directly typed SNPs, with probabilities used instead of indicator variables. Different genetic models can be formulated in the same way.

Maximum likelihood based on probabilities

Define individual likelihood as

$$L_i = SUM_{gi=\{0,1,2\}}\ P(g_i)\ P(Y_i \mid g_i)$$

Where

$$P(Y_i \mid g_i) = Normal(E[Y_i \mid g_i], s2)$$

and

$$E[Y_i \mid g_i] = m + b_1\ I(g_i=1) + b_2\ I(g_i=2)$$

Maximum likelihood based on probabilities

Define joint likelihood as the product of individual likelihoods

Maximize the likelihood over the parameters involved

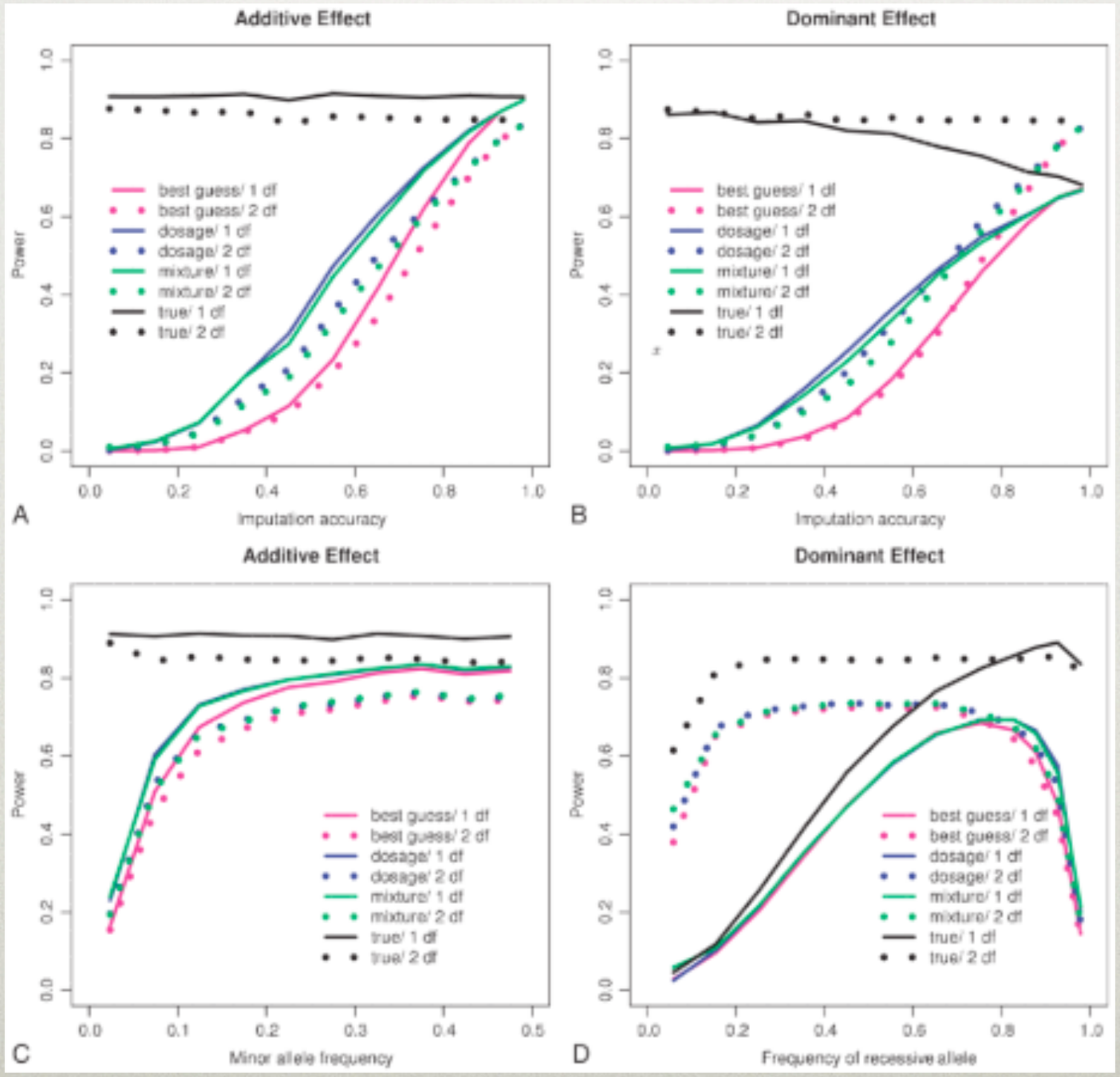
Maximum Likelihood Ratio test can be used to test nested models and draw statistical inferences

Power in large samples (small effects)

vs.

accuracy

vs. MAF

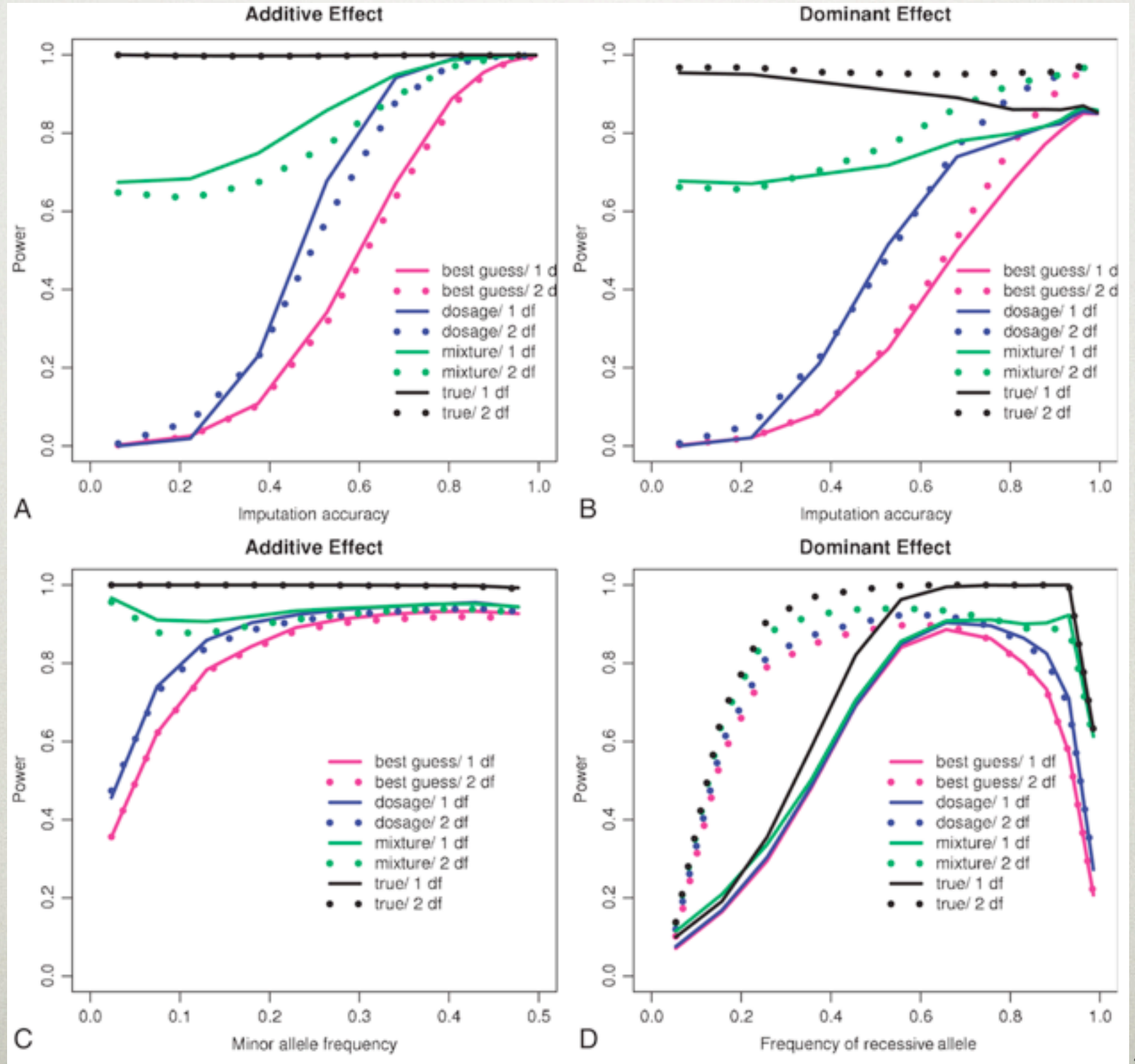


Zheng et al., 2011

Power in small samples (large effects)

vs. accuracy

vs. MAF



Zheng et al., 2011

Software for imputations

- MACH / mimimac
- BEAGLE
- IMPUTE2
- BIMBAM
- ...

Conclusions - Analysis of imputed data

- Using regression onto genotype probabilities is a valid and powerful method for standard scenarios
- Use of ML/mixture method can give extra power in case of small samples and large effects