Finding ancestry informative markers

BIOINFORMATICS INSTITUTE

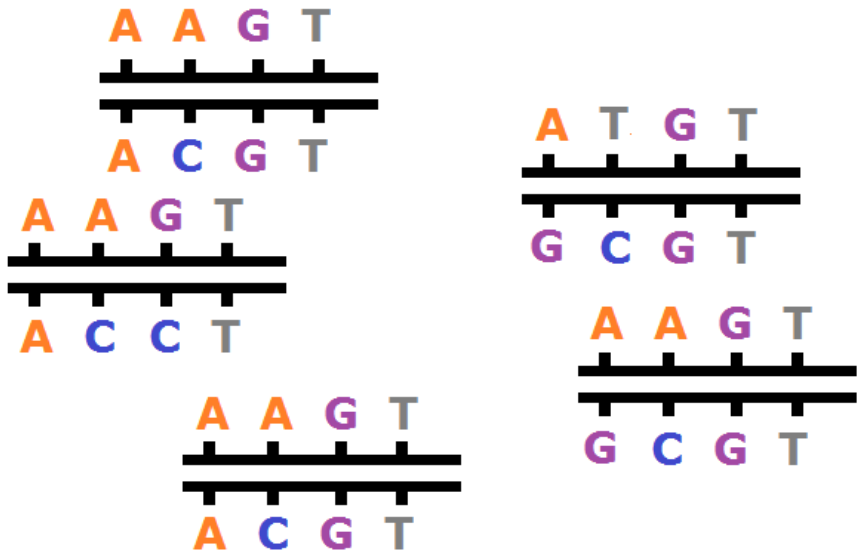
Yulia Kondratenko

Scientific advisor: Tatiana Tatarinova,
University of Southern California

Task

We need to assign individuals to populations based on their genotypes
We need to assess the reliability of such assignment

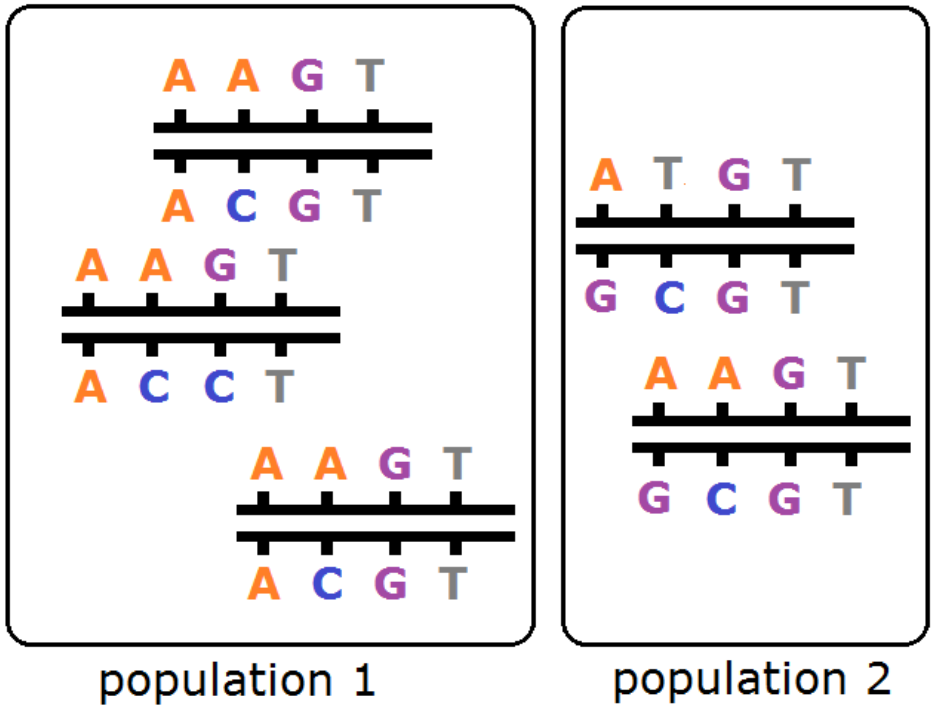
- Agriculture
- Personalized medicine
- Better reproducibility
- History



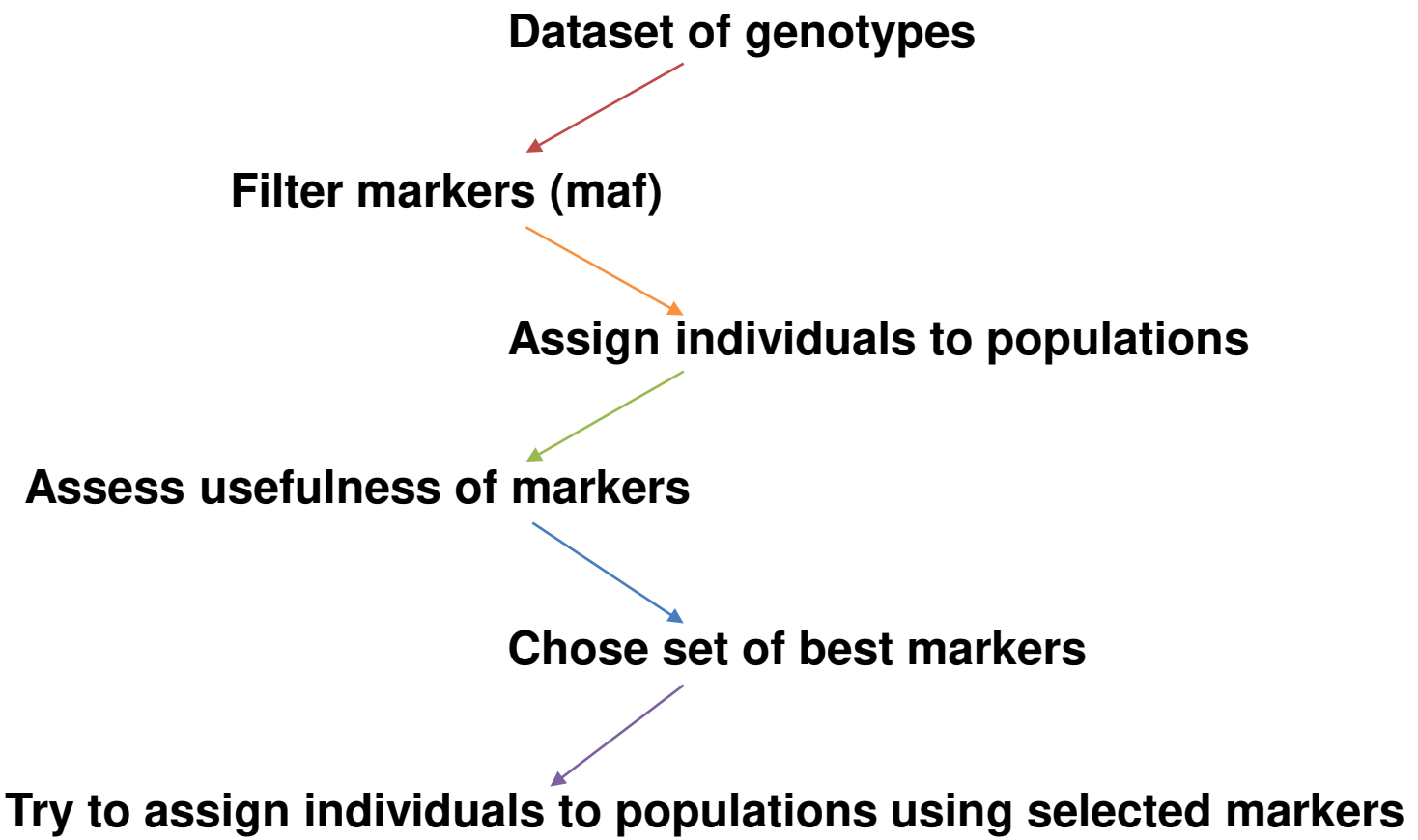
Task

We need to assign individuals to populations based on their genotypes
We need to assess the reliability of such assignment

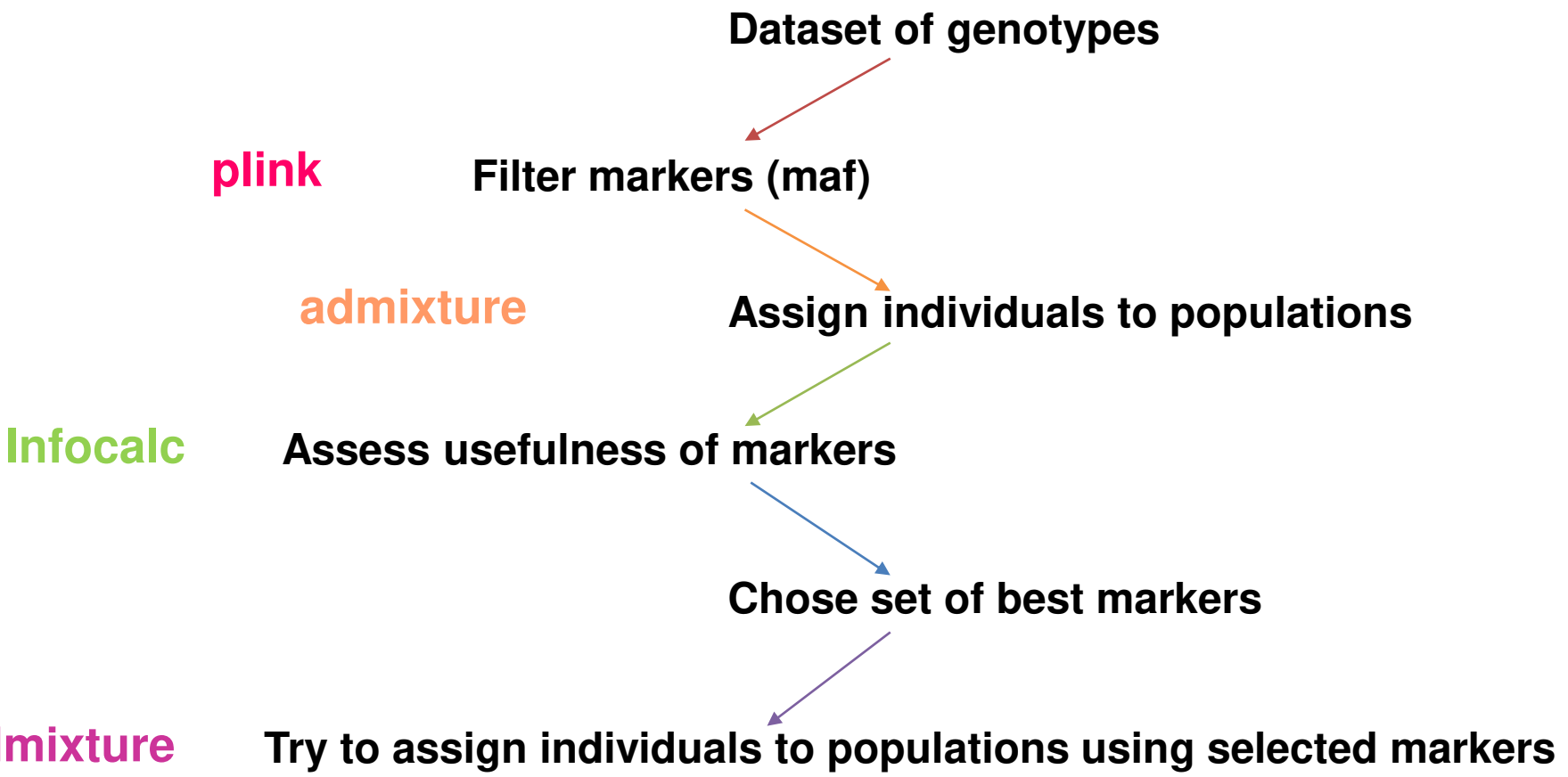
- Agriculture
- Personalized medicine
- Better reproducibility
- History



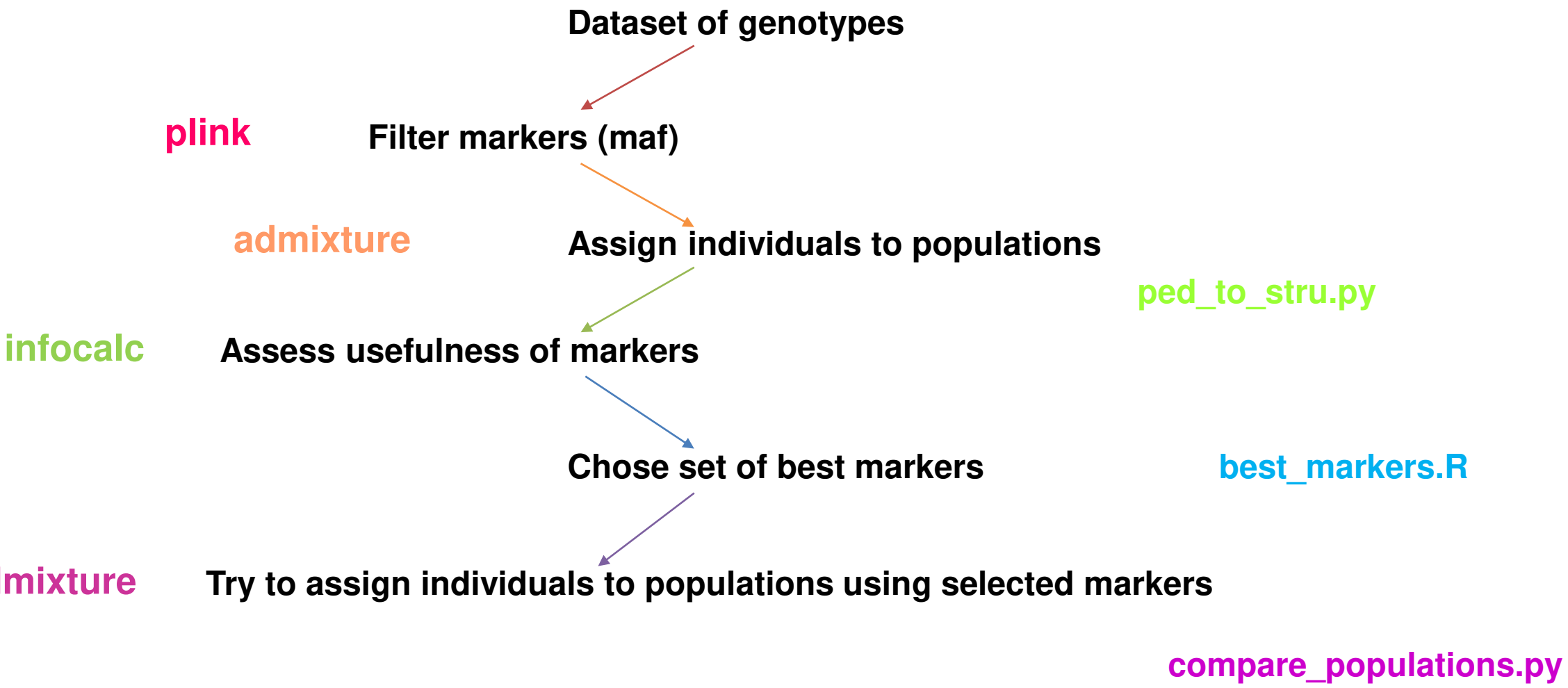
Identifying minimal panel of ancestry markers



Identifying minimal panel of ancestry markers



Identifying minimal panel of ancestry markers

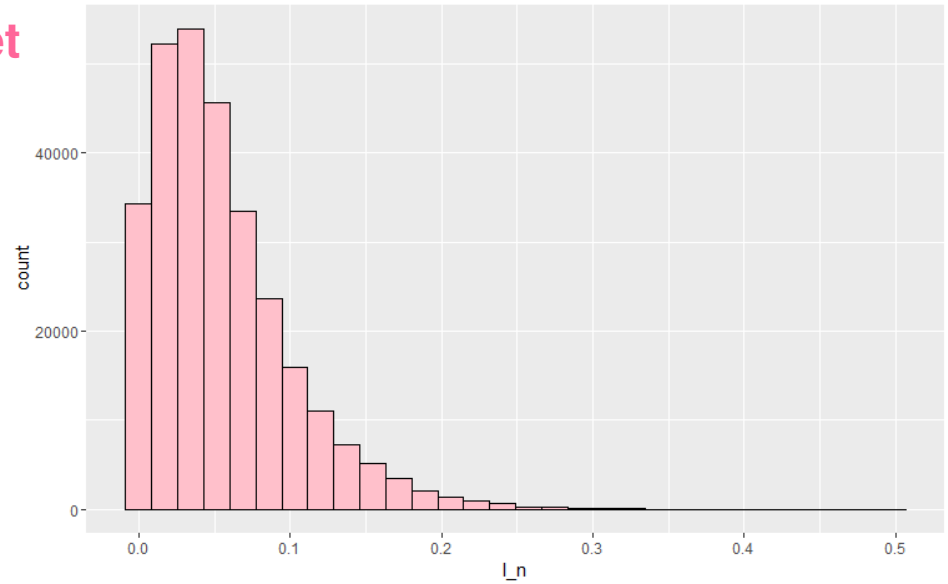


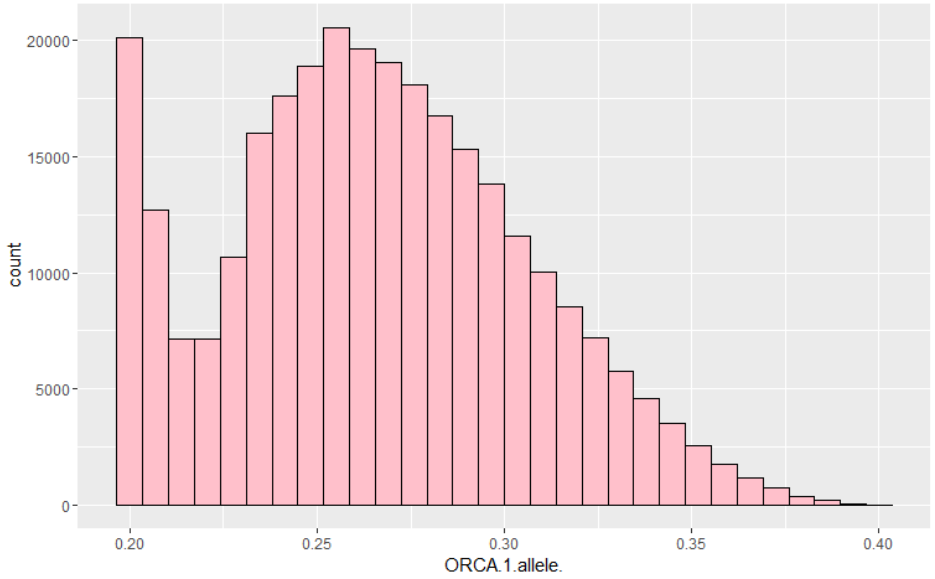
Challenges

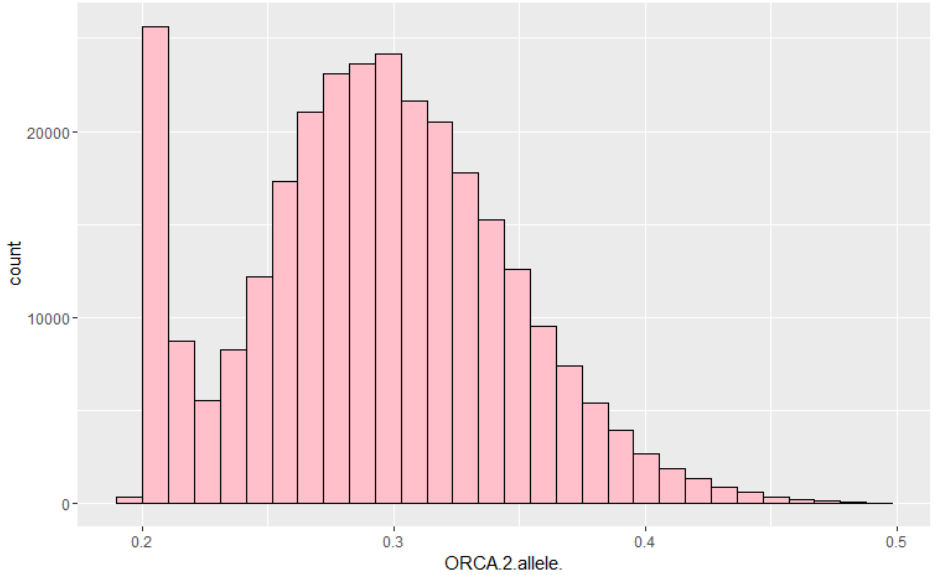
- Continuous distribution of markers usefulness – it is not clear, where to set a cut-off between good and bad markers
- Taking *n* individually best markers does not always yield the best panel of size *n*
- The more localized (specialized) populations are, the harder to describe their differences

Popular characteristics of marker usefulness

Human dataset







Distributions of these characteristics are not helpful for choosing appropriate number of markers

Datasets

	Human 1	Human 2	Arabidopsis	Rice
Individuals	702	3202	1135	755
Markers	291503	122780	778631	99993

Datasets

	Human 1	Human 2	Arabidopsis	Rice
Individuals	702	3202	1135	755
Markers	291503	122780	778631	99993

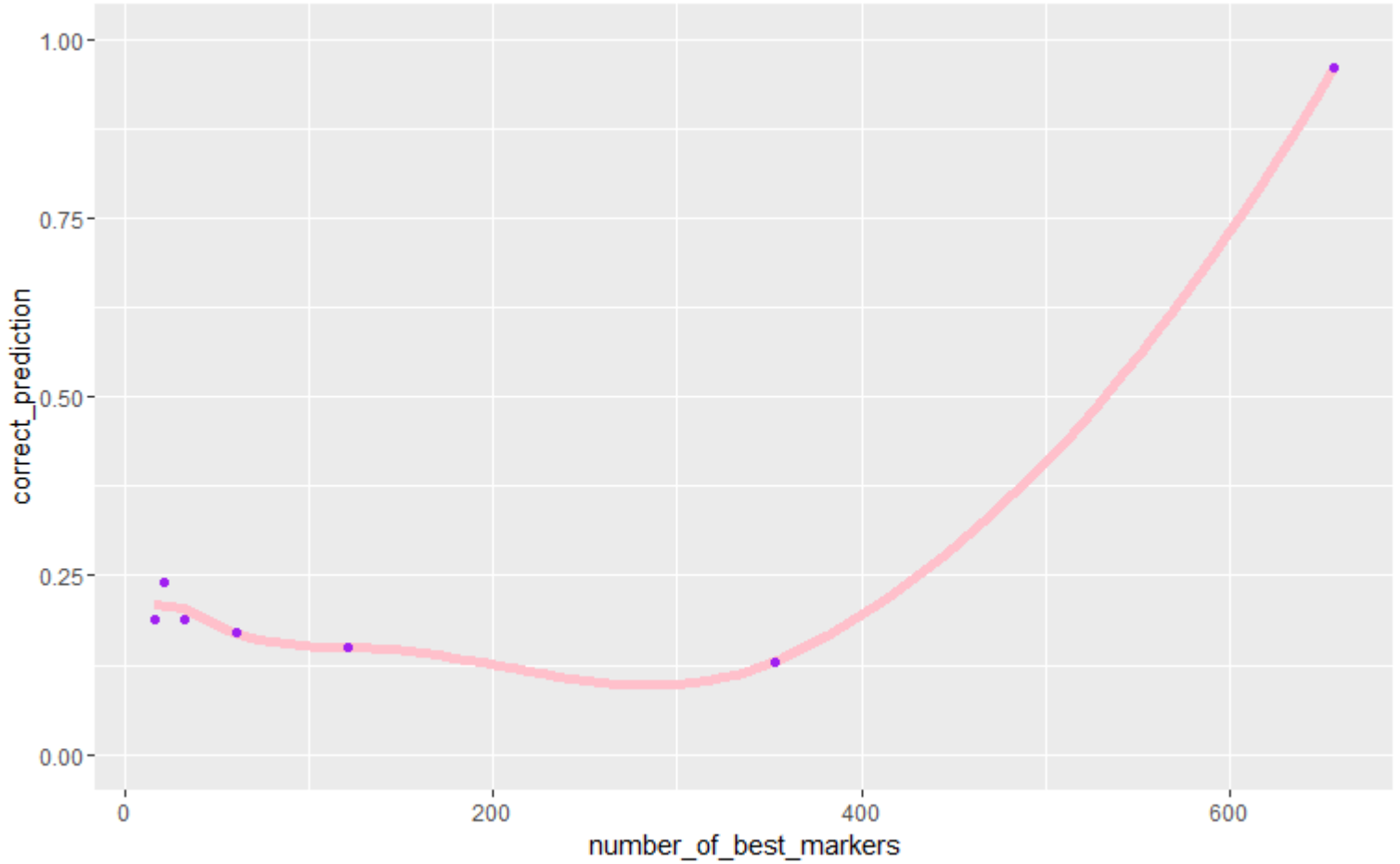
Representatives of more distant groups – Europe, Asia, Australia, Polynesia

Not that genetically distant individuals – Europe and Russia, but larger sample

Number of best markers & correct assignment

Human 1 dataset

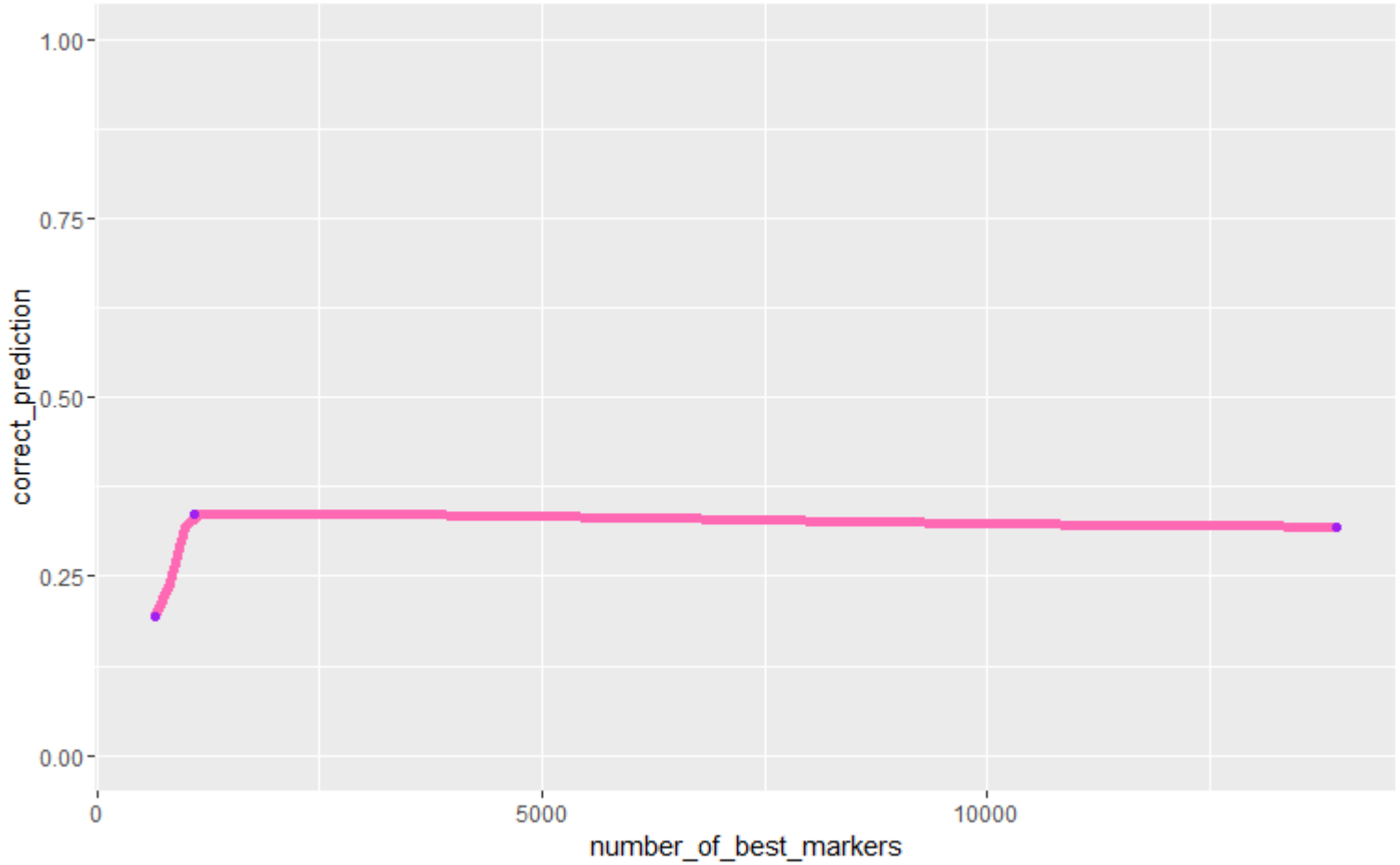
657 markers (from 291,503)
– 96% correct assignment



Number of best markers & correct assignment

Human 2 dataset

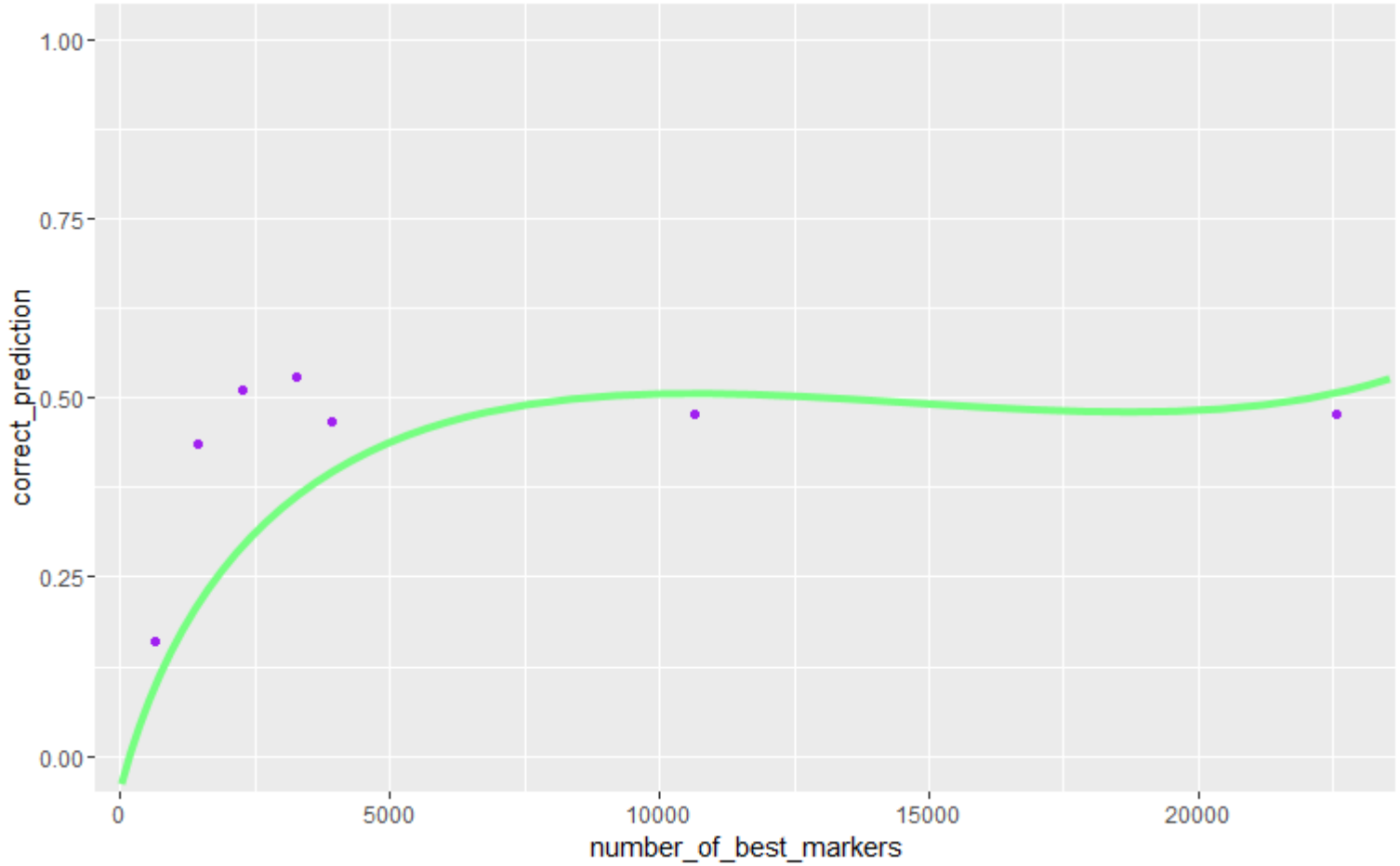
6,189 markers (from 122,780) – 33% correct assignment



Number of best markers & correct assignment

Arabidopsis dataset

3275 markers (from 778631) – 53% correct assignment

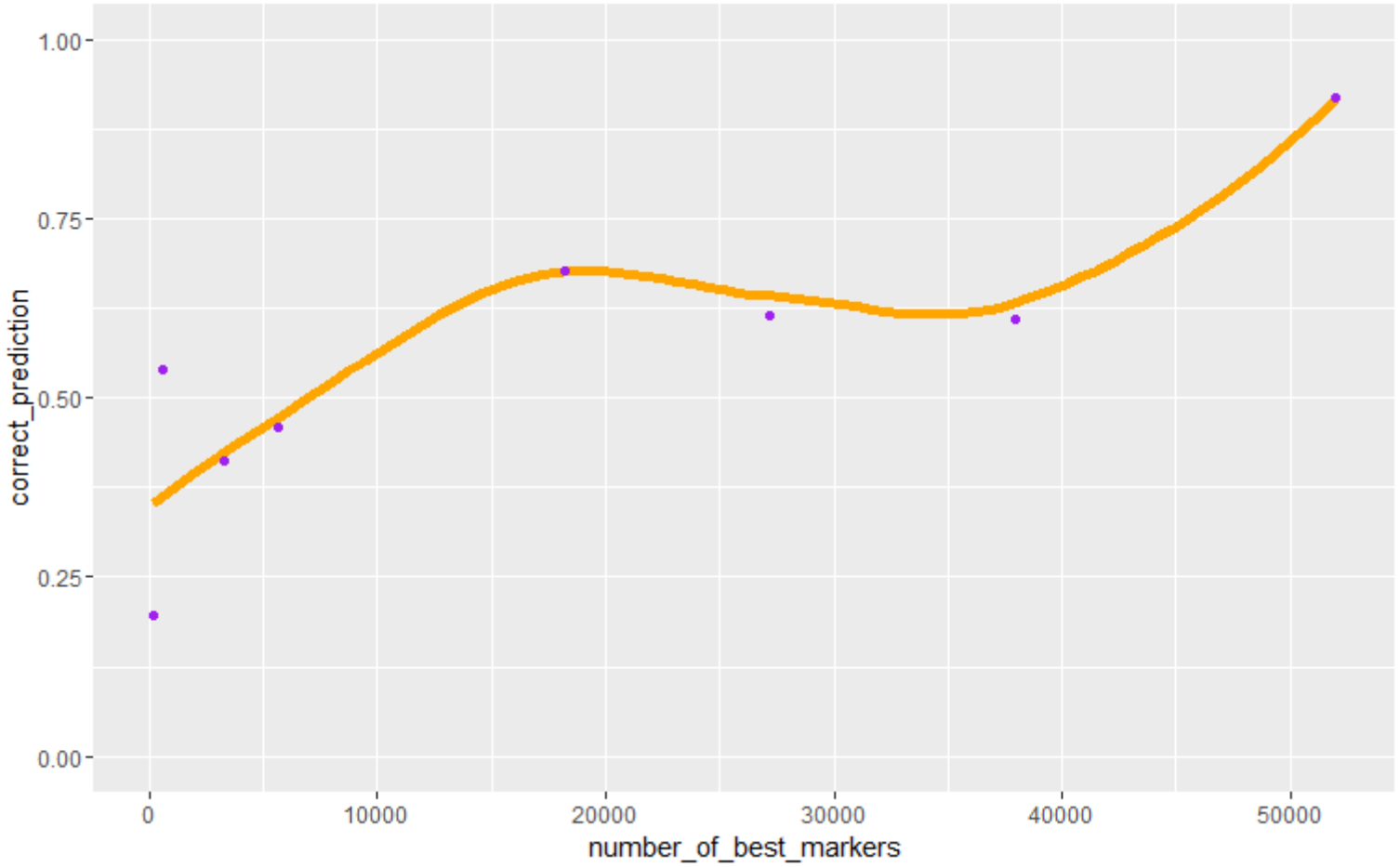


Number of best markers & correct assignment

Rice dataset

52007 markers (from 99993 – 92% correct assignment)

Local max – 607 markers – 54% correct assignment



Conclusions

- Sometimes it's impossible to describe genetic differences between populations using small panel of markers. Such complicated cases include search of ancestry markers in local human populations and cultivated plants.

- If it's necessary to build really small but maximally helpful panel – there are some local maximums on plots.

- Anyway, it's nice to reduce sets of markers to few tens of thousands, especially when raw datasets consist of hundreds of thousands or even millions of markers.

Thanks for your attention!