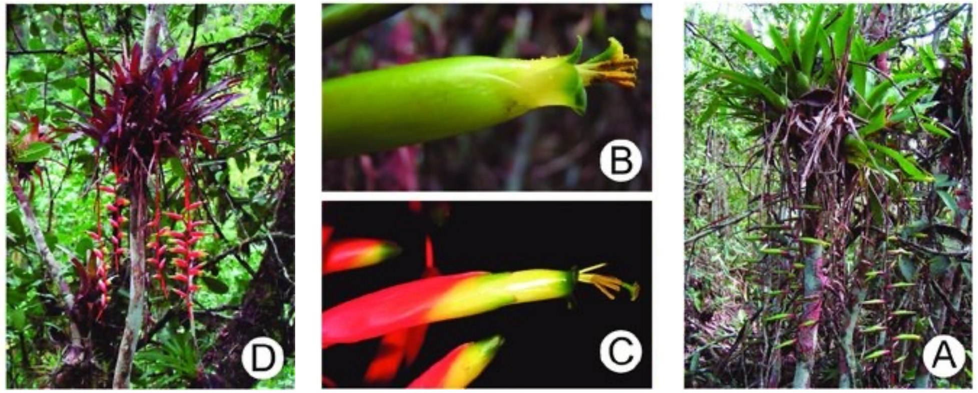
Population structure inference with ADMIXTURE

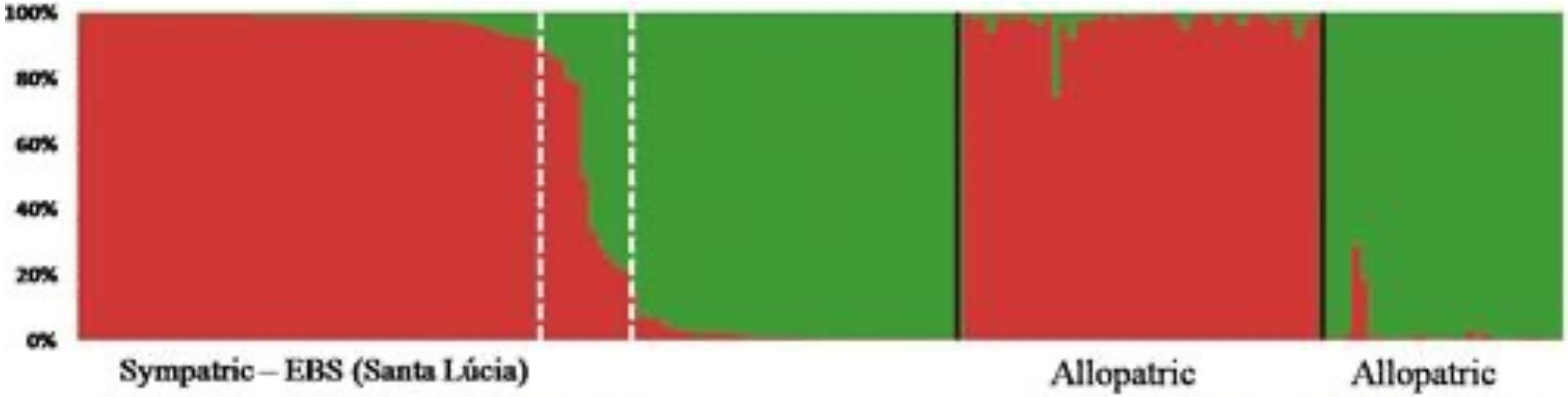
Joana Meier

Example of clustering methods to identify hybrids and infer ancestry proportions: Bromeliads



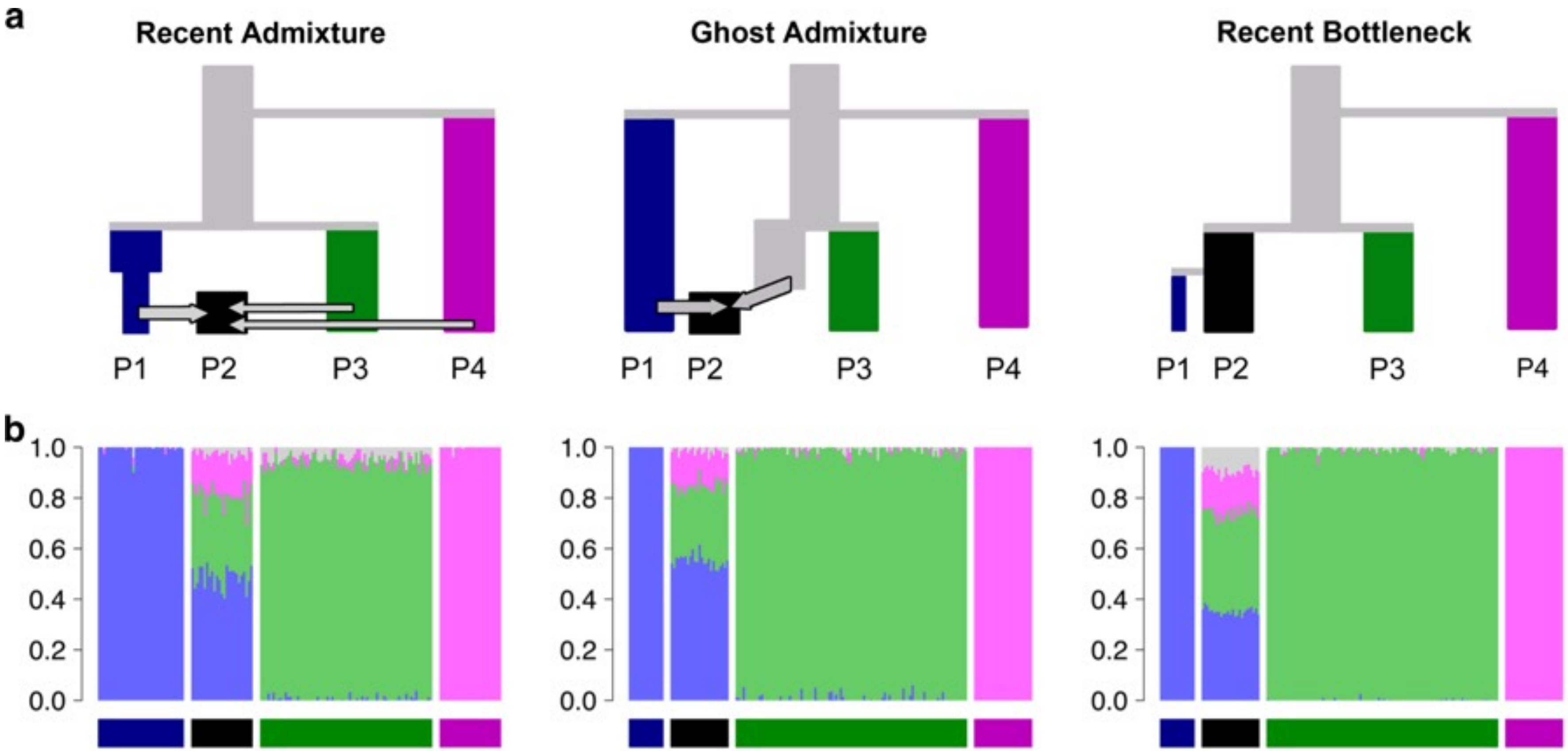
Vriesia simplex

Vriesia scalaris



Neri, et al., AoB Plants

Very different evolutionary histories can give the same ADMIXTURE plot -> badMixture



Very important: Use STRUCTURE/ADMIXTURE plots only to get a first overview of genetic structure. Putative cases of hybridization need to be followed up with other analyses such as D statistics and/or demographic modeling.

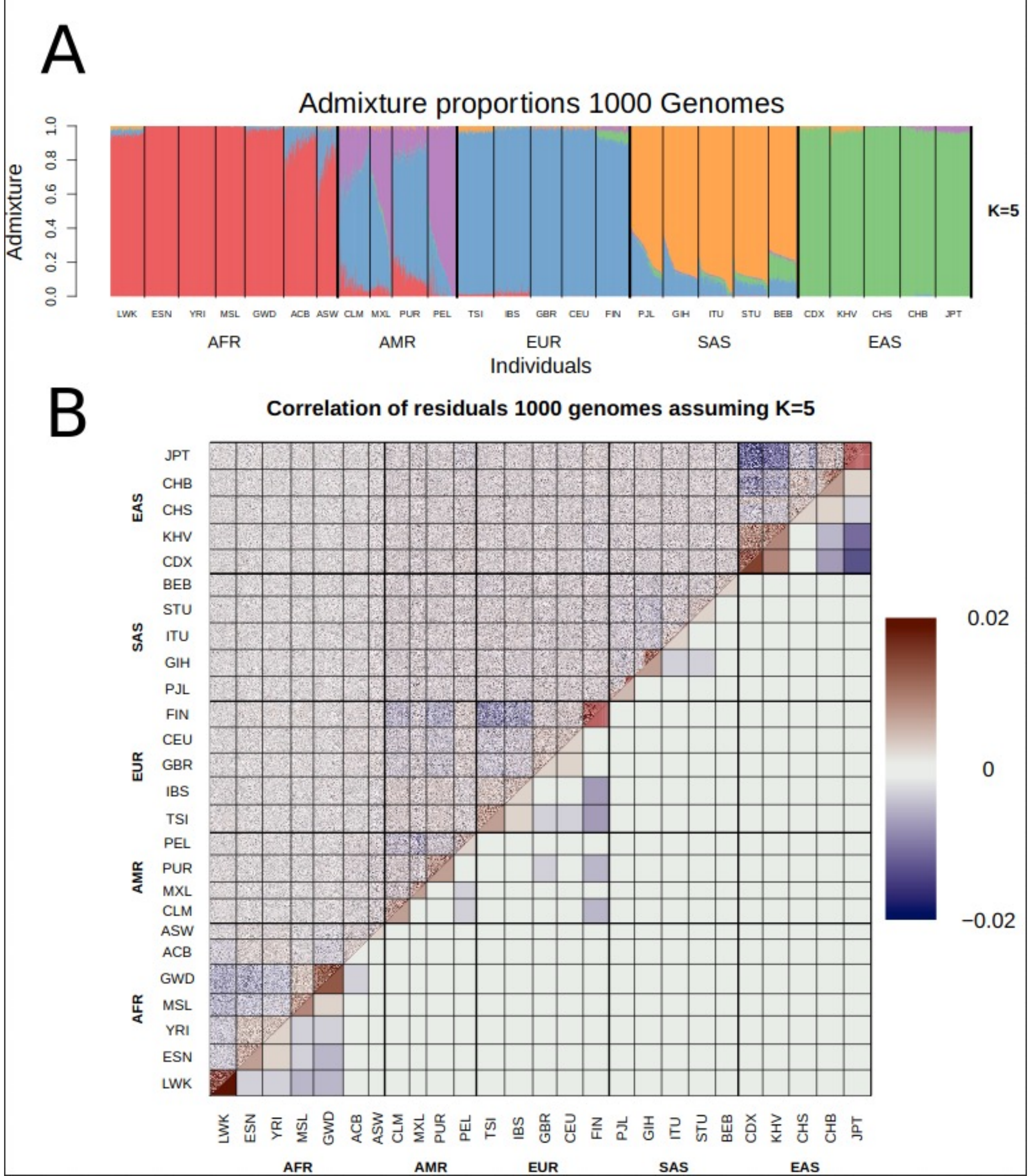
Lawson et al., 2018, NatComm

Is the ADMIXTURE plot a good fit for your data?

- **badmixture**
  (Dan Lawson)

- **Evaladmix**
  (Garcia-Erill & Albrechtsen)

  Positive correlation of residuals (red) indicates that the individuals in this group are more closely related than ADMIXTURE suggests, blue means that they are less closely related than modelled by ADMIXTURE.



Model assumptions

- SNPs are unlinked -> LD-pruning is very important
- Individuals are unrelated -> Do not include siblings
- Populations are represented by similar number of individuals -> strongly underrepresented populations may not be detected
- Ideally use at least 10,000 SNPs
- Sites are bi-allelic and singletons are removed

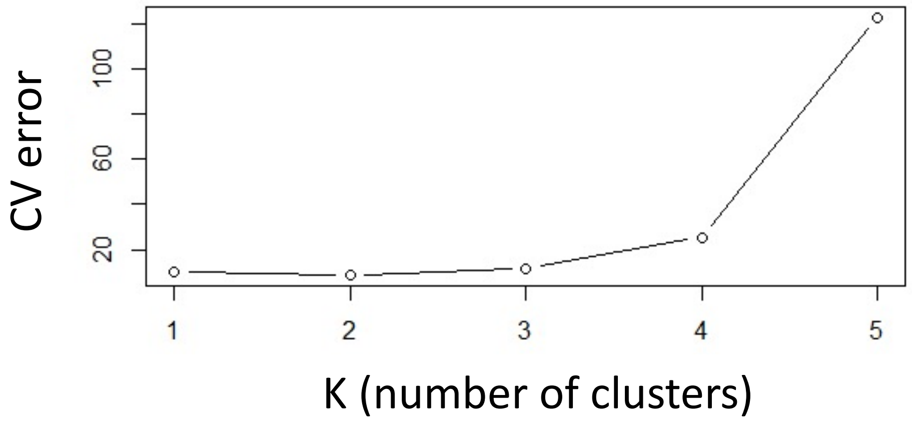
Limitations of ADMIXTURE and similar methods

ADMIXTURE and similar methods are great at inferring how many groups of individuals the dataset contains and which ones belong to which group. However, they struggle with the following:

- Detection of clusters that are strongly undersampled (e.g. just a single very divergent individual may not be picked up)
- Detecting ancestral admixture that is shared by all individuals of a cluster
- Handling individuals that are highly inbred or with loads of missing data

How to identify the best number of clusters?

ADMIXTURE uses a cross-validation (CV) procedure:
e.g. 5-fold CV: the genotypes are randomly split into five partitions. Each partition in turn is masked (set to missing) and then predicted with the individual cluster assignment and allele frequencies inferred for each cluster. The predicted genotypes are compared to the real genotypes to infer the prediction error. The prediction errors together are used to infer the cross validation error.
The lowest CV error indicates the best number of K but don't just use the CV error.



Make sure to always look at the plots with different numbers of clusters and also use your knowledge on the biology and collection places to identify which number of clusters is most informative.

Now, let’s try ADMIXTURE

STRUCTURE vs ADMIXTURE

- STRUCTURE is a model-based clustering approach which utilizes genotype data to infer the presence of distinct populations and assign individuals to populations. The number of populations (k) has to be predefined by the user. Usually, one tests multiple values and checks at which k the model posterior probability start to level out. Using an MCMC approach, STRUCTURE identifies k clusters that are each in HWE and assigns individuals to these clusters or a mix of them.
- ADMIXTURE works similarly but uses a maximum likelihood estimation to identify clusters and infer individual ancestries. The best value of k is determined with cross-validation.
- ADMIXTURE is much faster than STRUCTURE and thus ideal for large NGS datasets. STRUCTURE may be better suited for identification of very weak population structure also because it can use prior probabilities for the cluster assignments. However, ADMIXTURE allows to specification of unadmixed reference individuals with known ancestry.