

# Consistency study of a reconstructed genotype probability distribution via clustered bootstrapping in NORB pooling blocks

Camille Clouard<sup>1</sup> and Carl Nettelblad<sup>1</sup>

<sup>1</sup>Division of Scientific Computing, Department of Information Technology, Uppsala University

## Abstract

For applications with biallelic genetic markers, group testing techniques, synonymous to pooling techniques, are usually applied for decreasing the cost of large-scale testing as e.g. when detecting carriers of rare genetic variants. In some configurations, the results of the grouped tests cannot be decoded and the pooled items are missing. Inference of these missing items can be performed with specific statistical methods that are for example related to the Expectation-Maximization algorithm. Pooling has also been applied for determining the genotype of markers in large populations. The particularity of full genotype data for diploid organisms in the context of group testing are the ternary outcomes (two homozygous genotypes and one heterozygous), as well as the distribution of these three outcomes in a population, which is often ruled by the Hardy-Weinberg Equilibrium and depends on the allele frequency in such situation. When using a nonoverlapping repeated block pooling design, the missing items are only observed in particular arrangements. Overall, a data set of pooled genotypes can be described as an inference problem in Missing Not At Random data with nonmonotone missingness patterns. This study presents a preliminary investigation of the consistency of various iterative methods estimating the most likely genotype probabilities of the missing items in pooled data. We use the Kullback-Leibler divergence and the L2 distance between the genotype distribution computed from our estimates and a simulated empirical distribution as a measure of the distributional consistency.

## Background

### Purposes of group testing

Pooling is a group testing technique addressing how to confidently identify a category of items, called 'defectives', in a population, with as few tests as possible. Group testing has found numerous applications with DNA data for e.g. the purpose of large-scale sequencing or genotyping at reduced cost.

### A pooling algorithm for genetic data

In an other study [1], we have explored the usage of a Non Overlapping Repeated Block (NORB) design for simulation pooling on genotype data, similar to the design suggested by Erlich et al. [2]. Figure 1(a) presents the principle of such a pooling experiment.

The NORB procedure divides the population into  $B$  equally sized blocks of  $n_B$  individuals. In the encoding step of pooling, every block systematically defines how many pools are formed from the  $n_B$  items and the mapping of individual items to different pools. The genotype

of a pool is determined by the alleles that are detected among the pool members at the testing step. In the decoding step, the algorithm attempts to retrieve the genotype of any item based on the genotypes of the intersecting pools. In some cases, the decoding fails to confidently identify the genotype of an item and returns it as missing.

Erlich et al. [2] originally presented a NORB algorithm for decoding the genotypes into a binary response, that is, whether any genotype is a carrier of a rare variant or not. In our research, we extend the proposed decoding to a more general case of a ternary outcome, determining if the genotype is homozygous for the reference allele, heterozygous, or homozygous for the alternate allele. We suggest for this purpose an algorithm based on the Expectation-Maximization (EM) method that models all possible pooling configurations and computes the most likely genotype of every item involved in each configuration. The items that cannot not be confidently identified are assigned to a genotype which we call 'adaptive'. Such a genotype is a local consistent block-wise estimate of the genotype probabilities (GP) for these specific items. Figure 1(b) shows one example of our block-adaptive decoding algorithm, where a block configuration is identified by

its pooling pattern  $\psi$ .

In this study, we investigate the consistency of our adaptive estimates compared to the pre-pooling genotype values.

## Probabilistic formulation of the NORB pooling problem

### Data sets

#### Representation of the genotype data

We model the genotype data at any marker for a sample  $i$  as a probability simplex  $[p_{0i}, p_{1i}, p_{2i}]^\top$ , which stand, in this order, for the probability of the genotype being a homozygote for the reference allele, a heterozygote, and a homozygote for the alternate allele.

#### True genotype data

The pre-pooling data set, or true data set, consists of  $n$  genotypes at each genetic position. Each data point is a genotype  $x$  which is fully known, that is to say  $[p_{0i}, p_{1i}, p_{2i}]^\top$  is one of the three simplex in

$$\mathcal{X} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\},$$

At any position, the  $n$  individuals in the population are i.i.d. data points which sampled at frequencies  $\theta = [\theta_0, \theta_1, \theta_2]^\top$ . They form an empirical distribution  $\pi_n$

$$\mathbf{x} \sim \pi_n(x) \quad (1)$$

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \quad \forall i \in [1, n] \quad x_i = \begin{bmatrix} p_{0i} \\ p_{1i} \\ p_{2i} \end{bmatrix} \quad (2)$$

Under the assumption the HWE holds at each marker, the population-wide frequencies of the three genotypes at any marker are directly related to the alternate allele frequency (AAF) that we denote  $f$ . Therefore, the model of equation 1 can be reduced to a distribution which only depends on the variable  $f$

$$\mathbf{x} \sim \pi_n(x; f) \quad (3)$$

We note  $\hat{\theta}(f) = \mathbb{E}_n[\mathbf{x}; f]$  the empirical risk minimizer ERM estimating the mean of  $\pi_n(x; f)$ . Assuming HWE let the ERM be expressed as a single-variable parameter, that is

$$\theta(f) = \begin{bmatrix} (1-f)^2 \\ 2f(1-f) \\ f^2 \end{bmatrix} \quad (4)$$

As  $f$  is a continuous quantity, we discretize it for convenience in the simulation as the delimiting values of 21 equally-sized bins in the range  $[0.0, 1.0]$ . For each value of  $f$ , we simulate  $n = 160$  genotypes (10 pooling blocks of 16 samples) for 10 genetic positions, that is a number  $m = 200$  simulated markers.

The true genotypes  $\mathbf{x}$  are assigned to  $B = 100$  independent pooling blocks of  $n_B = 16$  samples

$$\mathbf{x}_B = (x_1, x_2, \dots, x_{16}), \quad (5)$$

and these blocks are used for simulating NORB pooling and decoding as the examples shown on Figure 1.

#### Pooled decoded genotype data

Let us describe pooling as a transformation  $t$  that maps the complete data  $\mathbf{x}$  to the incomplete data  $\mathbf{z}$  as follows

$$t: \mathcal{X} \longrightarrow \mathcal{Z} \quad (6)$$

$$\mathbf{x} \longmapsto \mathbf{z} \quad (7)$$

The vector  $\mathbf{z}$  consists of  $n$  genotypes resulting from simulating pooling and decoding on the true data  $\mathbf{x}$

$$\mathbf{z} = (z_1, z_2, \dots, z_n) \quad \forall i \in [1, n] \quad z_i = \begin{bmatrix} \tilde{p}_{0i} \\ \tilde{p}_{1i} \\ \tilde{p}_{2i} \end{bmatrix} \quad (8)$$

and, correspondingly to equation 5, the pooled data within a block  $b$  are denoted

$$\mathbf{z}_B = (z_1, z_2, \dots, z_{16}), \quad (9)$$

Depending on the pooling block configuration, the decoding is successful (or unambiguous) if the genotype  $z_i$  is a simplex as the ones in  $\mathcal{X}$  (white items on Figure 1). In this case,  $z_i$  is said to be determined. If the decoding is ambiguous, the genotype is said to be indeterminate and it is considered as missing (orange items on Figure 1).

We introduce  $V$  the vector of indices in  $\mathbf{z}$  for which the data is fully observed, and correspondingly  $\mathbf{y} = \{y_k\}$ ,  $k \in V$  the vector of observed genotypes i.e, determined after decoding. Conversely we use  $\bar{V}$  to denote the vector of indices in  $\mathbf{z}$  for which the data is unobserved, and  $\mathbf{u} = \{u_k\}$ ,  $k \in \bar{V}$  the vector of indeterminately decoded genotypes.

We are interested in studying the mappings  $t$  for any value of  $f$ . However, the pooling decoding process generating  $\mathbf{z}$  cannot be formulated in a closed-form expression. Therefore, we model  $\mathbf{z}$  as a sample from an unknown distribution

$$\mathbf{z} \sim \tilde{\pi}_n(z; f) \quad (10)$$

We consider that the distribution  $\tilde{\pi}_n(f)$  has an empirical mean  $\phi_n(f)$ .

### Characteristics of the missing data for the undecoded items in pooling blocks

The missing data  $\mathbf{u}$  can be categorized as Missing Not At Random (MNAR) data [3], since it inherently depends on the other genotypes observed in each pooling block, as well as on the unobserved AAF at the given genetic position. Because of the NORB setting used, the missingness patterns in the pooled decoded data are by design nonmonotone.

### Piece-wise estimates of the genotype probabilities in MNAR data with nonmonotone missing patterns

In another study (unpublished research), we propose a method for computing the most likely probability of each of the unobserved items  $\mathbf{u}$  by inverse transform sampling.

The finite set of possible nonmonotone missingness patterns can be categorized into subsets of block patterns  $\psi$ . All patterns with the same block pattern are just permutations of that pattern as illustrated on Figure 2.

The proposed method exhaustively enumerates all block patterns. For each pooling block having the pattern  $\psi$ , the probability of any genotype in  $\mathbf{z}_B$  is conditioned on  $\psi$ . The variable  $f$  is marginalized and depending on the algorithm version implemented, any missing item in  $\mathbf{u}$  is substituted with a fixed prior probability that can be initialized to any simplex. The missing data estimation problem over all patterns is solved piece-wise as a series of either Maximum Marginal Likelihood Estimation (MMLE) or EM [4].

Our method produces self-consistent estimates of the most likely genotype probabilities for any item in  $\mathbf{u}$ .

Using the computed estimates in place of any missing genotype in  $\mathbf{u}$ , we reconstruct a fully observable vector  $\underline{\mathbf{z}}$  as if the pooled genotypes would be sampled from a distribution

$$\underline{\mathbf{z}} \sim \hat{\pi}_n(z) \quad (11)$$

The different versions we have implemented and tested correspond to variations of an EM inference method:

0. The reconstructed distribution  $\underline{\mathbf{z}}$  corresponds to a naive uninformed completion of the data, where any item in  $\mathbf{u}$  is set to  $(1/3, 1/3, 1/3)$ . That is, all genotypes are equally likely, as they would be in the case of a "neutral" HWE and  $f = 0.5$  at any marker.
1. The GP are sampled based on the expected allele frequency  $f$  in the entire block, that is from a binomial distribution with parameters  $f$  and  $32 = 16 \times 2$  as each genotype is a pair of alleles. The expected allele frequency is initialized to  $f = 0.5$  and then deduced at each iteration from the priors for the genotypes e.g.  $f = 0.5 Pr(G = 1) + Pr(G = 2)$ . The posterior estimates for the GP are calculated with an iterative adjustment of the fixed priors. At each iteration, the posterior GP are divided by the prior and normalized in order to ensure the self-consistency of the algorithm. Moreover, since the heterozygotes estimates are degenerated, the posterior GP are rescaled by reweighing each genotype probability in the simplex and normalized in order to compensate for the degeneracy.
2. Similar to 1., but each of the 33 possible allele count outcomes in the block has an individual iteratively fitted probability. The lowest count of alleles is the case of a pooling block where all items have a genotype  $G = 0$ . Conversely, if all items have a genotype  $G = 2$ , the allele count sums up to 32. Therefore, there are in total  $33 = 32 + 1$  possible allele count outcomes in a pooling block. At every iteration, the alleles for every individual in the block are sampled from the allelic binomial distribution (reference or alternate allele), and the GP posterior is deduced from the allelic frequencies before rescaling it with the GP prior.
3. Similar to above, but the allelic proportions are used as such and not as binomial parameters.
4. Similar to 2., but the alleles are sampled geometrically. The posterior genotype frequencies are directly used as priors at next iteration, without rescaling them with the former prior. On the whole, this process is very close to an EM algorithm.

We approximate the mean of the reconstructed distribution with the empirical risk minimizer  $\hat{\phi}_n = \mathbb{E}_n[\underline{\mathbf{z}}]$ .

In this study, we evaluate the quality of the reconstructed empirical distributions  $\hat{\pi}_n$  from the various approaches presented above, with respect to the simulated

empirical distribution  $\pi_n$  from which they were generated. We conduct a preliminary study of the quality of the reconstructed data based on two consistency criteria.

## Clustered bootstrap sampling for pooled genotype data with a NORB design

### Statistics for studying the consistency of empirical distributions

We use the following statistics to do a preliminary study of the consistency of the reconstructed empirical distribution  $\hat{\pi}$ :

- The L2 norm  $\hat{\delta}_n = \|\hat{\pi}_n(z) - \pi_n(x)\|_2$  which has been suggested for testing goodness-of-fit for densities in e.g. [5].
- The Kullback-Leibler divergence  $\hat{\nu}_n = D_{KL}(\pi_n, \hat{\pi}_n)$  as suggested in e.g. [6], defined for the genotype data at one marker as

$$D_{KL} = \frac{1}{n} \sum_{i=1}^n \sum_{g=0}^2 -p_{g,i} \log \left( \frac{p_{g,i}}{\hat{p}_{g,i}} \right) \quad (12)$$

If the pooled reconstructed data  $\underline{\mathbf{z}}$  are consistent with  $\mathbf{x}$ , we expect  $\hat{\delta}_n \approx 0$  and  $\hat{\nu}_n \approx 0$ .

Statistics computed on every marker having frequency  $f$  in each  $f$ -bin.

### Pooled genotype data reconstruction in the case of infinite sample size

The simulated genotype data and the reconstructed data from point-wise estimates have finite sample size  $n$ . We assume the distribution  $\hat{\pi}_n$  is consistent with the distribution  $\pi_n$ . In the case of infinite sample size when  $n \rightarrow \infty$ , we expect the behavior

$$\hat{\pi}_n(z) \longrightarrow \pi^*(z) \quad (13)$$

For addressing the variability issue for the estimated statistics with a finite sample size  $n$ , we use a bootstrap resampling method to compute confidence intervals (CI) for both statistics  $\hat{\nu}_n$  and  $\hat{\delta}_n$ .

### Motivations for using clustered bootstrap sampling

Because of the NORB design chosen, the dependencies between the samples in the pooled genotype data vectors  $\mathbf{z}$  and  $\underline{\mathbf{z}}$  are particular. Every block is independent

from the  $B - 1$  other ones but within a pooling block, the samples are no longer i.i.d.

$$\forall k \in [1, B] \quad \forall j \in [1, n_B] \quad z_j^k \not\perp \{z_{-j}^k\} \quad (14)$$

where  $z_{-j}^k$  is any sample but the  $j$ -th one in the  $k$ -th block.

### Construction of the clustered bootstrap samples

Assimilating a pooling block to a cluster of data, we implement a specific bootstrap method for clustered data, largely based on the two-stage bootstrap described in [7]. However, if our block data are exchangeable (the order of the blocks does not matter) as in the two-stage bootstrap, the data within a block are not.

Let us form  $K$  bootstrap samples from the data  $\underline{\mathbf{z}}$  by randomly choosing with replacement  $C$  clusters in the  $B$  blocks

$$\forall k \in [1, K] \quad Z_k^* = \{Z_{k,1}^*, Z_{k,2}^*, \dots, Z_{k,C}^*\} \quad (15)$$

In each bootstrap sample, we randomly sample a single data point per block such that the equation (15) becomes

$$\forall k \in [1, K] \quad Z_k^* = \{z_{k,1}^*, z_{k,2}^*, \dots, z_{k,C}^*\} \quad (16)$$

For each bootstrap sample from  $\underline{\mathbf{z}}$ , we pick the same block and sample indices in the pre-pooling data  $\mathbf{x}$

$$\forall k \in [1, K] \quad X_k^* = \{x_{k,1}^*, x_{k,2}^*, \dots, x_{k,C}^*\} \quad (17)$$

We note the mean of the  $k$ -th bootstrap sample as

$$\bar{Z}_k^* = C^{-1} \sum_{c=1}^C z_{k,c}^* \quad (18)$$

similarly for  $\bar{X}_k^*$ , such that  $K \times C = N$ . The bootstrap estimators of the Kullback-Leibler divergence and the L2-norm are formed as

$$\forall k \in [1, K] \quad \hat{\nu}_{N,k} = D(\bar{X}_k^*, \bar{Z}_k^*) \quad (19)$$

$$\hat{\delta}_{N,k} = \|\bar{X}_k^* - \bar{Z}_k^*\|_2 \quad (20)$$

$\hat{\nu}_N$  has a bootstrap estimated variance of

$$\hat{\hat{\nu}}_N = K^{-1} \sum_{k=1}^K \left( \hat{\nu}_{N,k} - K^{-1} \sum_{k=1}^K \hat{\nu}_{N,k} \right)^2 \quad (21)$$

In practice, we choose  $K = \lfloor 0.8B \rfloor$ . The  $1 - \alpha$  CI for the bootstrap statistics is hence defined as

$$\bar{T}_\alpha^N = \left\{ \nu : |\nu - \hat{\nu}_N| \leq \sqrt{\hat{V}_N q_\alpha} \right\}, \quad (22)$$

similarly for  $\hat{\delta}_N$ .

## Results

We do not claim to do any hypothesis test about the consistency, we are rather interested in preliminary results that let us visualize the dissimilarity between the simulated and reconstructed empirical distributions. The results presented are to be considered in the perspective of genotype imputation. Most methods achieving genotype imputation essentially consist in a HMM-based inference of the missing genotype data in a population and for a given set of markers. They generally assume that the genotypes at a marker in the population are at HWE. Genotype imputation produces posterior genotype probability estimates for each individual at each marker. The imputation algorithms internally double the prior genotype probability for the heterozygotes, so we need to rescale the *simpool* estimates by doubling the heterozygotes and normalizing every probability simplex in order to render how the reconstructed distribution would be used in the imputation method. Therefore, the results presented compare statistics calculated from the rescaled reconstructed distribution.

We use a ternary plot for representing a genotype probability simplex. Ternary plots, synonymously de Finetti diagrams, are a standard representation of 3-dimensional data [8, 9, 10]. This representation provides a first intuitive visualization of the distributions as well as the distances between the different genotype estimates. The ternary plots presented are produced with a specific Python package [11].

Figure 3 shows an example of annotated ternary plot for the estimates computed in a pooling block of pattern  $((2, 2, 0), (2, 2, 0))$ . Table 1 gives the coordinates of every data point projected on the ternary axes on Figure 3 in order to facilitate the interpretation of the ternary plot.

Figure 4 shows the empirical means of the rescaled data in each AAF-bin on a ternary plot. The 'true' line represents the distribution from which the data  $\mathbf{x}$  is sampled. The heterozygotes are under-represented in all reconstructed distributions, which indicates that the *simpool* algorithm tends to favor the inference of homozygous genotypes. For example, in place of two missing items, two opposite homozygotes are more likely than two heterozygotes. The closest reconstruction of the pooled distribution is achieved with the version 4 of *simpool*.

The L2 distance measures shown on Figure 5 present the same characteristics as the  $D_{KL}$  measures. Since the

reference and alternate alleles for biallelic markers have symmetrical properties, the allelic frequency is commonly presented as Minor Allele Frequency (MAF) rather than AAF. The minor allele is either the alternate or the reference one depending on its frequency. The L2 distance is a commonly used metrics that reveals how far the data points forming the empirical 'true' and the empirical reconstructed distribution are. If the metrics is equal to 0, the data points have identical coordinates. This is for example the case if  $MAF = 0$  on Figure 5, that is to say the population studied is purely homozygous for the major allele at the marker considered. Given the NORB pooling design used in this experiment, all genotypes are decoded as homozygous for a pure homozygous population. Therefore the decoded data are identical to the true one and the L2 norm is null. However, we are more interested in studying the distributional consistency between the 'true' and the reconstructed distributions than the distance between single points. Indeed, since the GP estimates are to be used as prior probabilities for genotype imputation, we need to consider the dissimilarity between the distribution from this perspective. In genotype imputation in a population and especially at the phasing step, the posterior genotype probabilities computed for the indeterminate markers depend on the Linkage Disequilibrium (LD) between the markers. The LD renders the probability that the genotypes of a sequence of markers in parent individuals are inherited together by the offspring. This metrics is correlated to the physical distance between the markers in the DNA but the relationship is not linear, such that the L2 distance is not the most well-suited metrics for apprehending how the prior genotype probabilities might impact the imputation. The concept of  $D_{KL}$  is more relevant for studying the distributional consistency and quantifying the information loss between probabilities, therefore we prefer to focus on describing the divergence results of the bootstrap resampling.

The divergence  $D_{KL}$  between  $\pi_n$  and  $\hat{\pi}_n$  across range of the MAF values is shown on Figure 6. It presents the same characteristics as the confidence intervals for the L2 distance.  $D_{KL}$  quantifies as a single measure the dissimilarity between the reconstructed distribution and the 'true' distribution it was pooled from. Overall, all CI-envelopes reveal a correlation between  $D_{KL}$  and the MAF. As for the L2 distance, the minimum is observed if the data is purely homozygous ( $MAF = 0$ ) since all items are fully decoded to homozygotes. The least divergence is also achieved if both alleles are in equal proportions ( $MAF = 0.5$ ). Around  $MAF = 0.5$ , all pooled genotypes are very likely to be missing and this results in nearly uniform estimates ( $\hat{\pi} \sim (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ ) regardless of the version of *simpool* that is used. After rescaling, the

---

estimates are almost equal to (0.25, 0.5, 0.25) which are the HWE proportions used for generating the 'true' data set when  $MAF = 0.5$ . The situation is very similar to the case of fully missing data with equally likely genotypes for each of the unassayed markers, which is on the whole the assumption made by most imputation methods. The divergence of the reconstructed distributions reaches a maximum around 0.05 for  $MAF = 0.2$ , except from the reconstruction with the version 4 of *simpool* for which the maximum is shifted to  $MAF = 0.3$ . When  $MAF = 0.2$ , the homozygotes for the major allele are dominating in the true data, whereas  $MAF = 0.3$  coincides with the frequency at which the heterozygotes are the most frequent in a population at HWE.

When designing *simpool*, we expected our estimates to be closer to the true distribution than the default case of uniform data completion (version 0). Figure 6 however shows that the estimates from the versions 1, 2, 3 are almost identical to the naive version 0. In other words, the computed estimates do not add much information about the most likely genotype. As it is already suggested by Figure 4, the reconstructed distribution is the most consistent with the 'true' data when using estimates computed with the version 4 of *simpool*. This reconstruction is also the most accurate, as the narrow curve envelope indicates. The version 4 of *simpool* was implemented while conducting this study as we noticed that the earlier versions 1, 2, 3 were not satisfying. The version 4 intends to improve the consistency of the reconstructed pooled distribution. While  $D_{KL}$  still correlates to the MAF, it is significantly lower (at most 0.012) than with the previous versions (up to 0.055) and is almost null for  $MAF = 0.5$ . The divergence measures reveals that we have succeeded in capturing better the 'true' distribution when reconstructing the pooled data.

## Conclusions

Many studies have proposed powerful algorithms for decoding binary outcomes from pooled data and NORB is one pooling design example that has been investigated. When the test outcomes are ternary ( $G = 0, 1, 2$ ) as for genotyping biallelic markers, the DNA Sudoku method described in [2] is not robust enough for decoding the pooled genotypes. We implement in [1] various EM-based estimation methods specifically tailored for reconstructing the incomplete genotype data from a NORB pooling design.

The findings of the present study should be put in the context of the genotype imputation that we are interested in with our research [1]. Indeed, it is essential that the GP estimates forming the reconstructed distri-

bution favor the downstream imputation of the correct genotype. In this perspective, the consistency between the reconstructed distribution and the 'true' one is more relevant than the physical closeness of the point-wise GP estimates. Therefore, the quality metrics should not only focus on the divergence from the true data but also reward the information gain they bring to the pooled data.

In this paper, we made a preliminary analysis of the consistency of these various GP reconstruction methods in with a divergence and a distance measure. In order to account for the limitations of the numerical representation of the genotypes in *simpool*, we have introduced the concept of heterozygotes degeneracy. However, the first versions of *simpool* did not appear to be satisfying and we therefore explored variations with the explicit intent to minimize the values of the KL divergence. The later versions of *simpool* were implemented as we started investigating the consistency of the reconstructed distribution. Thanks to a geometrical sampling of the alleles at each iteration, the improved *simpool* algorithm manages to capture better the allelic distribution at the level of the pooling block, as well as the derived genotype frequencies. The version 4 of *simpool* uses the same initial prior probabilities for each missing item regardless of the pooling pattern in a block. Another strategy, possibly improving the consistency of the reconstructed pooled distribution, could choose the initial allelic priors depending on the pooling pattern observed.

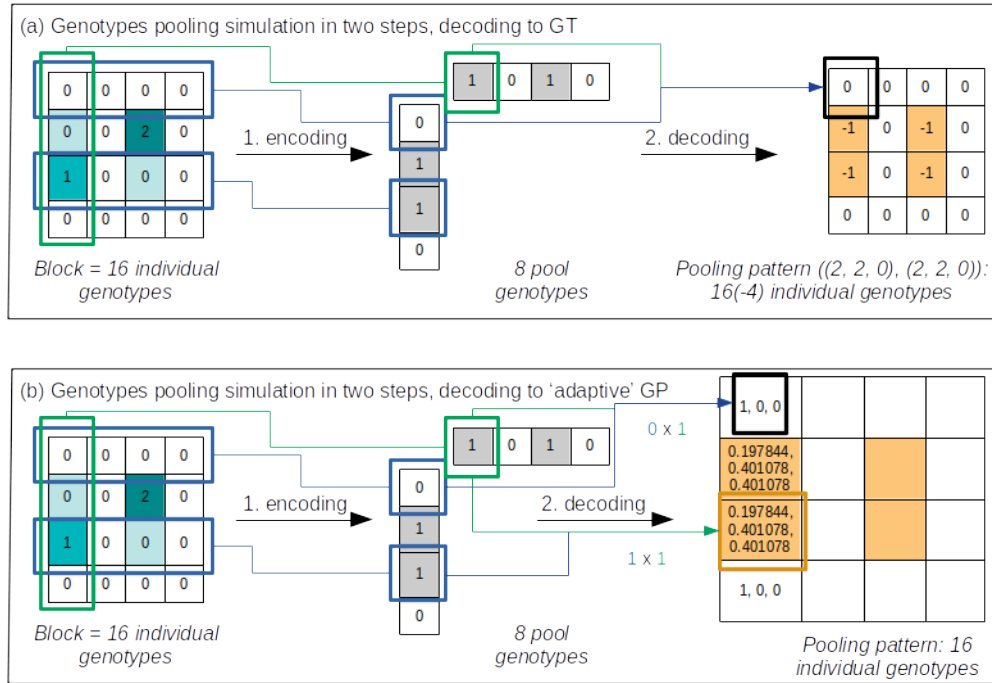
It is difficult to assess from  $D_{KL}$  only which of the heterozygous or homozygous genotypes contribute the most to the divergence. A further analysis of the divergence in relation to the heterozygosity rate might enlight new improvements that could be made in the GP estimation method. Moreover, a broader investigation of the information gain brought by our adaptive GP estimates to imputation would be suitable, especially compared to a naive uninformed completion. We suggest for this purpose to study the imputation results we obtained in an earlier paper [1] with the results we would obtain for the same pipeline but replacing the reconstructed estimates with values of later *simpool* versions e.g. version 4 as they have the highest consistency.

## References

- [1] C. Clouard, K. Ausmees, and C. Nettelblad. *A Joint Use of Pooling And Imputation For Genotyping SNPs*. 2021. DOI: 10.21203/rs.3.rs-1131930/v1. URL: <https://doi.org/10.21203/rs.3.rs-1131930/v1>.

- 
- [2] A. Gordon et al. Y. Erlich K. Chang. “DNA Sudoku—harnessing high-throughput sequencing for multiplexed specimen analysis”. In: *Genome Research* 19 (2009), pp. 1243–1253.
- [3] D. B. Rubin. “Inference and missing data”. In: *Biometrika* 63 (1976), 581–592.
- [4] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum Likelihood from Incomplete Data Via the EM Algorithm”. In: *Journal of the Royal Statistical Society* 39.1 (1977), pp. 1–22.
- [5] M. H. Neumann and E. Paparoditis. “On bootstrapping L2-type statistics in density testing”. In: *Statistics & Probability Letters* 50.2 (2000), pp. 137–147. ISSN: 0167-7152. DOI: [https://doi.org/10.1016/S0167-7152\(00\)00091-2](https://doi.org/10.1016/S0167-7152(00)00091-2). URL: <https://www.sciencedirect.com/science/article/pii/S0167715200000912>.
- [6] A. Lindholm et al. “Data Consistency Approach to Model Validation”. In: *IEEE Access* 7 (2019), 59788–59796. DOI: 10.1109/ACCESS.2019.2915109.
- [7] C. A. Field and A. H. Welsh. “Maximum Likelihood from Incomplete Data Via the EM Algorithm”. In: *Journal of the Royal Statistical Society* 69.Part 3 (2007), pp. 369–390.
- [8] C. Cannings and A. W. Edwards. “Natural selection and the de Finetti diagram”. In: *Ann Hum Gen* 31 (1968), 421–428. DOI: <https://doi.org/10.1111/j.1469-1809.1968.tb00575.x>.
- [9] Richard J. Howarth. “Sources for a history of the ternary diagram”. In: *The British Journal for the History of Science* 29.3 (1996), 337–356. DOI: 10.1017/S000708740003449X.
- [10] A. W. Edwards. *Foundations of Mathematical Genetics 2nd Edition*. Cambridge University Press, 2000. ISBN: 978-0-521-77544-1.
- [11] Marc Harper et al. “python-ternary: Ternary Plots in Python”. In: *Zenodo* 10.5281/zenodo.594435 (2015). DOI: 10.5281/zenodo.594435. URL: <https://github.com/marcharper/python-ternary>.

# Artwork



## Example of pooling simulation with a NORB algorithm.

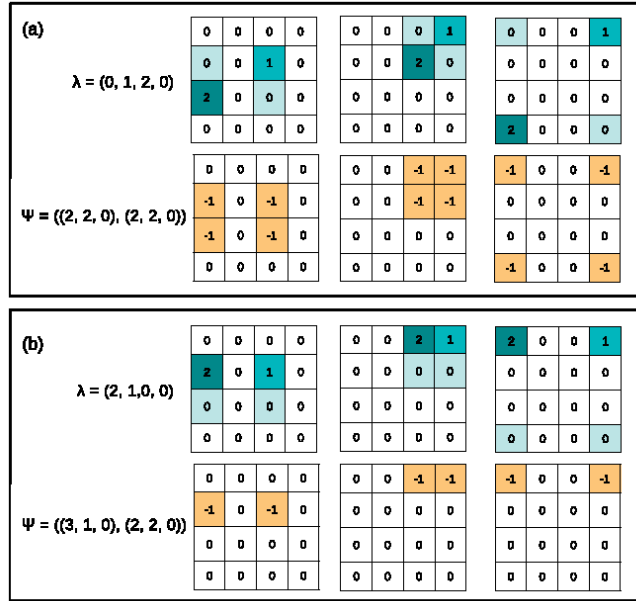
A pooling block of  $n_B = 16$  samples is modelled as a square matrix, the rows and the columns form 8 intersecting pools of 4 samples each. The encoding step assigns the 4 individual genotypes to a pool and pooling is done as follows: the genotype of a pool is 0 (resp. 2) iff all samples are 0 (resp. 2), as for example the top row-pool. In all others cases (allelic-heterogeneous pools), the genotype of the pool is 1, as for example the leftmost column-pool (green frame). The decoding step reconstructs the genotype of every sample based on the intersecting pools. Decoding is successful if at least 1 homogeneous pool (genotype 0 or 2) is involved. Otherwise, the genotype of the sample is indeterminate and considered as missing.

A block is described by its pooling pattern  $\psi = (n_{G_{rows}}, n_{G_{columns}})$  where  $n_{G_{rows}}$  (respectively  $n_{G_{columns}}$ ) gives the number of row-pools (resp. column-pools) in the block having the genotype 0, 1, and 2.

Figure 1:

**Subfigure (a):** the pooled genotypes are decoded into integer genotypes (GT format) in  $\{0, 1, 2, -1\}$  representing, respectively, a homozygote for the reference allele, a heterozygote, a homozygote for the alternate allele, or a missing item. In this example, there are 4 indeterminate samples. The pooling pattern  $\psi$  is  $((2, 2, 0), (2, 2, 0))$ , and the sample highlighted by a black square is intersected by pools having genotype 0 and 1.

**Subfigure (b):** the pooled genotypes are decoded to adaptive genotype probabilities (GP format) that are computed with a Maximum Marginal Likelihood estimation method. We qualify the genotype probabilities as 'adaptive', as we estimate them relatively to the pattern of the pooling block that the samples are part of. Four samples have an ambiguous genotype, for which none the genotype probabilities is 1.  $\psi$  is the same as on the subfigure (a).



### Permutations of block patterns.

Examples of two pooling patterns obtained from two distinct permutations from the same set of genotypes.  $\lambda$  denotes the subvector of blue-colored genotypes that are possible completions of  $\mathbf{z}$ .

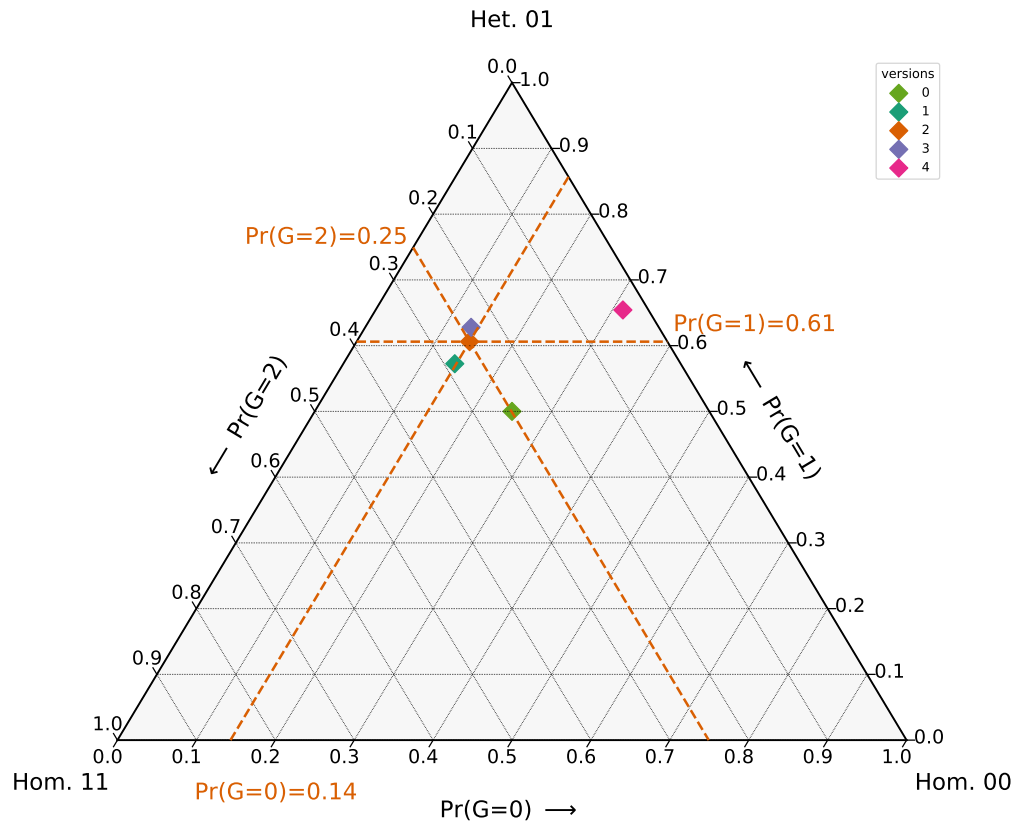
Figure 2:

**Subfigure (a):** The carriers of the alternate allele e.g. having the genotype 1 or 2 are located on different rows and different columns, such that they never show up in the same pool. In the three pooling blocks shown, the pooling pattern  $\psi = ((2, 2, 0), (2, 2, 0))$  is the same while they result from different permutations of the completed data  $\mathbf{z}$ .

**Subfigure (b):** The carriers of the alternate allele are located on different columns but the same rows, such they are genotyped together in the row pool. The three pooling blocks shown have the same pooling pattern  $\psi = ((3, 1, 0), (2, 2, 0))$ .

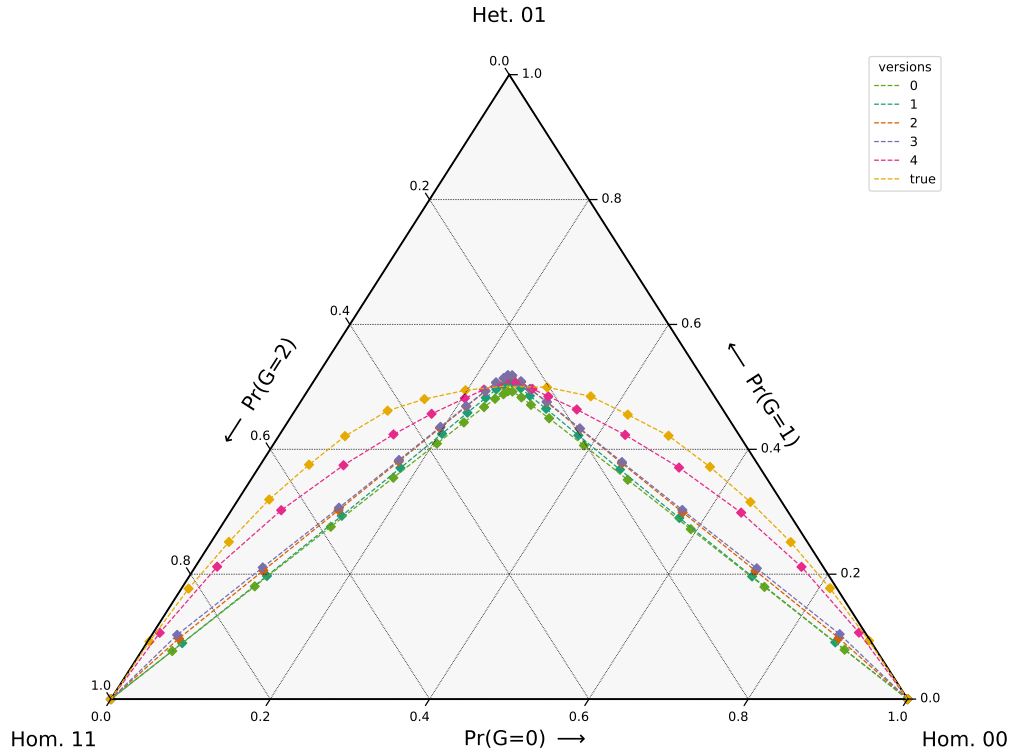
version	Pr(G=0)	Pr(G=1)	Pr(G=2)
0	0.250000	0.500000	0.250000
1	0.141208	0.572528	0.286264
2	0.143231	0.606085	0.250684
3	0.134242	0.627877	0.237881
4	0.313383	0.654295	0.032322

Table 1: Rescaled most likely genotype probabilities computed by different versions of the *simpool* algorithm for undecoded items in a pooling with pattern  $\psi = ((2, 2, 0), (2, 2, 0))$



**Example of de Finetti diagram: Genotype probabilities estimates for the missing data in a pooling block with pattern  $\psi = ((2, 2, 0), (2, 2, 0))$ .**

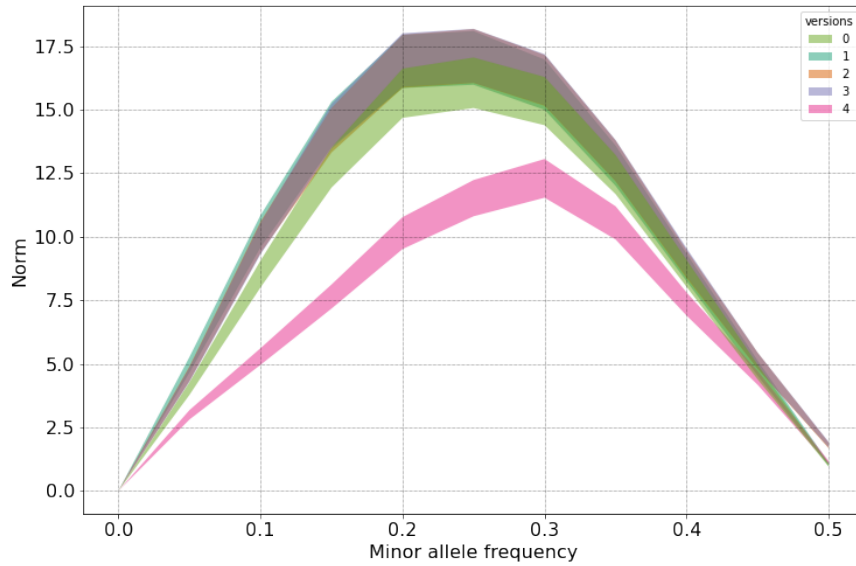
Figure 3: The annotations on the three axes indicate the coordinates of the orange point that is the GP estimate computed with the version 2 of *simpl*. The orange lines represent the projection of the data point on the axes. The values for all the data points displayed are given in Table 1. Each of the tops of the triangle is the position for a fully known genotype, either homozygous or heterozygous.



**De Finetti diagram of the averaged genotype probabilities in true and reconstructed pooled data for each allele frequency bin.**

The series of points represent the mean genotype probabilities computed from genetic markers with increasing allele frequency  $f$ : the smallest frequencies (from  $f = 0.05$ ) are located at the bottom right corner Hom. 00 ( $Pr(G = 0)$  is almost 1) and the largest frequencies (up to  $f = 0.95$ ) are located at the bottom left corner Hom. 11 ( $Pr(G = 2)$  is almost 1). The 'true' points and line in yellow show the bin-averaged genotype probabilities from the true data set used to simulate pooling. The other points show the bin-averaged genotype probabilities from rescaled pooled data that was completed with different versions of *simpool*.

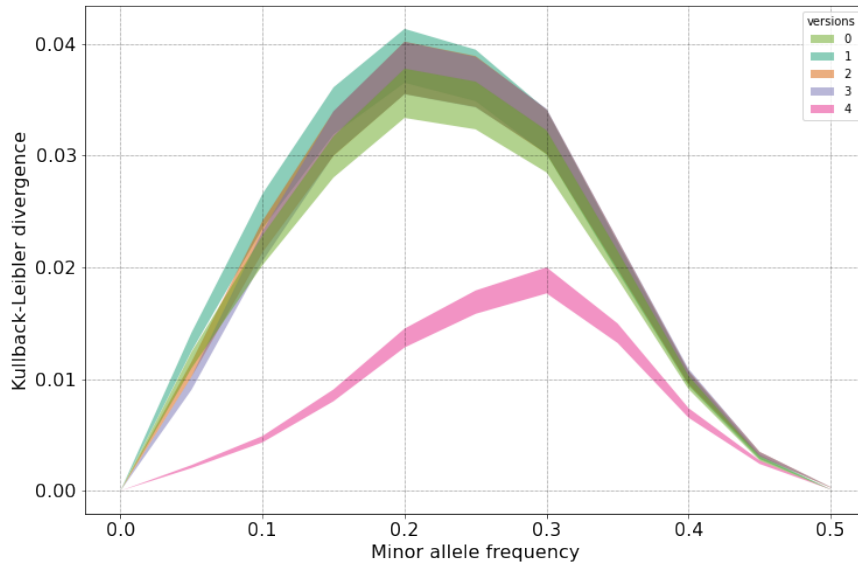
Figure 4:



**95% bootstrap confidence intervals for the L2 distance.**

The distributional L2 distance is computed between a 'true' empirical distribution and reconstructed empirical distributions. The true data consists of genotypes sampled under the HWE assumption, and used for simulating genotype pooling experiments. The reconstructed distributions consist of decoded pooled data and different estimates of the missing data that are computed with various versions of the *simpool* algorithm. Each data point in the reconstructed distribution is rescaled before averaging the genotype probabilities in each MAF-bin. The rescaling takes into account the heterozygotes degeneracy. The allele frequency is presented as MAF since the reference and the alternate alleles have symmetrical properties when the genotype data are pooled. A null value for the L2 distance indicates that the reconstructed distribution is perfectly consistent with the true one. The L2 distance computed from a reconstructed distribution based on the *simpool* version 4 has a different shape from all other versions and is the most consistent one. This is the only version of *simpool* that uses a geometrical resampling of the genotypes at each iteration of the algorithm.

Figure 5:



**95% bootstrap confidence intervals for the Kullback-Leibler divergence.**

The distributional divergence is computed between a 'true' empirical distribution and reconstructed empirical distributions. The true data consists of genotypes sampled under the HWE assumption, and used for simulating genotype pooling experiments. The reconstructed distributions consist of decoded pooled data and different estimates of the missing data that are computed with various versions of the *simpool* algorithm. Each data point in the reconstructed distribution is rescaled before averaging the genotype probabilities in each MAF-bin. The rescaling takes into account the heterozygotes degeneracy. The allele frequency is presented as MAF since the reference and the alternate alleles have symmetrical properties when the genotype data are pooled. A null value for the divergence indicates that the reconstructed distribution is perfectly consistent with the true one. Notably, the  $D_{KL}$  computed from a reconstructed distribution based on the *simpool* version 4 has a different shape from all other versions and is the most consistent one. This is the only version of *simpool* that uses a geometrical resampling of the genotypes at each iteration of the algorithm.

Figure 6: