Genotipificación del maíz usando *Genotyping by Sequencing* (GbS) y Big Data

Yareli Morán[1] J. Andrés Christen[2]
Sarah Hearne[1]

[1]Centro Internacional para el Mejoramiento del Maíz y el Trigo, CIMMYT, Texcoco, Mexico

[2]Centro de Investigación en Matemáticas , CIMAT, Guanajuato, Mexico

XIII Escuela de Probabilidad y Estadística,
CIMAT, 18 de Marzo de 2015.

Introduction

The International Maize and Wheat Improvement Centre (CIMMYT, `www.cimmyt.org`) is an international research for development organization headquartered in Mexico. CIMMYT houses the worlds largest and most diverse collection of maize **landrace** germoplasm, with more than 50% of the collection originating from Mexico, the centre of domestication and primary centre of diversity for maize.

> *A landrace is a domesticated, regional ecotype, a locally adapted, traditional variety of a domesticated species of animal or plant that has developed over time, through adaptation to its natural and* **cultural environment** *of agriculture[...] Specimens of a landrace tend to be relatively genetically uniform, but are more diverse than members of a standardized or formal breed.*
> *`http://en.wikipedia.org/wiki/Landrace`*

Introduction

Figure: Maize "Accessions", stored at CIMMYT at 4^{o}C.

Introduction

Currently selection of potentially valuable germplasm from the collection is constrained by limitation in information of relevance to breeders e.g. stress resistance.

As part of a multi year initiative called Seeds of discovery (SeeD) funded by the Mexican government the entire genebank collection is being characterized genotypically to provide a uniform framework of information across the varying accessions in the collection and to enable the application of genomics tools in breeder selection.

Introduction

Currently selection of potentially valuable germplasm from the collection is constrained by limitation in information of relevance to breeders e.g. stress resistance.

As part of a multi year initiative called Seeds of discovery (SeeD) funded by the Mexican government the entire genebank collection is being characterized genotypically to provide a uniform framework of information across the varying accessions in the collection and to enable the application of genomics tools in breeder selection.

Introduction

Genotyping by sequencing (GbS) is a relative new **fingerprinting** technique employing next generation sequencing. GbS permits the identification of single nucleotide polymorphisms (SNP's) in a scaleable (10^3 to 10^6 genomes), cost effective manner.

There are 28,000 accessions within the CIMMYT genebank and in order to effectively analyze all samples a composite method for fingerprinting was implemented that requires new statistical analysis methods. We have worked on GbS data from 10,000 accessions covering more than 250,000 SNP's.

Introduction

Genotyping by sequencing (GbS) is a relative new **fingerprinting** technique employing next generation sequencing. GbS permits the identification of single nucleotide polymorphisms (SNP's) in a scaleable (10^3 to 10^6 genomes), cost effective manner.

There are 28,000 accessions within the CIMMYT genebank and in order to effectively analyze all samples a composite method for fingerprinting was implemented that requires new statistical analysis methods. We have worked on GbS data from 10,000 accessions covering more than 250,000 SNP's.

Introduction: Single Nucleotide Polymorphisms (SNP's)

Maize is a diploid organism, with chromosomes with two DNA strings.
Accessions i:

$i = \mathbf{2321}$ ... ACGGGTACCTAATCG ...
$i = \mathbf{2321}$ ... ACGGGTACCTAATCG ...

$i = \mathbf{8742}$ ... ACGGGTACCTACTCG ...
$i = \mathbf{8742}$ ... ACGGGTACCTACTCG ...

$i = \mathbf{4637}$ ... ACGGGTACCTAATCG ...
$i = \mathbf{4637}$ ... ACGGGTACCTACTCG ...

There is a Single Nucleotide Polymorphisms (SNP) at locus l (eg. = 3758473). Reference allele is A and alternative allele is C. Accession 2321 is homocygotic with the reference allele, Accession 8742 is homocygotic with the alternative allele and Accession 4637 is heterocygotic.

Introduction: Single Nucleotide Polymorphisms (SNP's)

Maize is a diploid organism, with chromosomes with two DNA strings. Accessions i:

$i = \mathbf{2321}$ ... ACGGGTACCTAATCG ...
$i = \mathbf{2321}$ ... ACGGGTACCTAATCG ...

$i = \mathbf{8742}$ ... ACGGGTACCTACTCG ...
$i = \mathbf{8742}$ ... ACGGGTACCTACTCG ...

$i = \mathbf{4637}$ ... ACGGGTACCTAATCG ...
$i = \mathbf{4637}$ ... ACGGGTACCTACTCG ...

There is a Single Nucleotide Polymorphisms (SNP) at locus l (eg. = 3758473). Reference allele is A and alternative allele is C. Accession 2321 is homocygotic with the reference allele, Accession 8742 is homocygotic with the alternative allele and Accession 4637 is heterocygotic.

Introduction: GbS data

The DNA strings are separated and cut into millions of small strings, which are duplicated and from this GbS counts the number of reference allele and alternative allele found.

	A	B	C	D	E	F	G	H	
1			SEEDDIV1	SE	SE	SE	SE	SE	S
2			SEEDDIV MAIZE SH4	SE	SE	SE	SE	SE	S
3			A1	B1	C1	D1	E1	F1	G
4			853027	###	###	###	###	###	#
5	Allele_ID	Allele_Sequence	517891	###	###	###	###	###	#
6	100009226|F|0--33:A>G	TGCAGCCTTGAGTCA	0	0	1	0	0	0	
7	100009226|F|0--33:A>G	TGCAGCCTTGAGTCA	0	0	0	0	0	0	
8	100009252|F|0--1000092	TGCAGCTGAAACTTA	2	0	0	2	2	1	
9	100009252|F|0--1000092	TGCAGCCGAAACTTA	1	4	1	1	0	1	
10	100009262|F|0--58:A>C	TGCAGCTCTGTTCAT	14	3	0	0	0	8	
11	100009262|F|0--58:A>C	TGCAGCTCTGTTCAT	0	0	0	0	0	0	

Figure: GbS data.

CIMMYT could not analyzed individual plants, but mixed material from $m = 30$ plants arising from one accession.

Introduction: GbS data

The DNA strings are separated and cut into millions of small strings, which are duplicated and from this GbS counts the number of reference allele and alternative allele found.

	A	B	C	D	E	F	G	H	
1			SEEDDIV1	SE	SE	SE	SE	SE	S
2			SEEDDIV MAIZE SH4	SE	SE	SE	SE	SE	S
3			A1	B1	C1	D1	E1	F1	G
4			853027	###	###	###	###	###	#
5	Allele_ID	Allele_Sequence	517891	###	###	###	###	###	#
6	100009226|F|0--33:A>G	TGCAGCCTTGAGTCA	0	0	1	0	0	0	
7	100009226|F|0--33:A>G	TGCAGCCTTGAGTCA	0	0	0	0	0	0	
8	100009252|F|0--100009	TGCAGCTGAAACTTA	2	0	0	2	2	1	
9	100009252|F|0--100009	TGCAGCCGAAACTTA	1	4	1	1	0	1	
10	100009262|F|0--58:A>C	TGCAGCTCTGTTCAT	14	3	0	0	0	8	
11	100009262|F|0--58:A>C	TGCAGCTCTGTTCAT	0	0	0	0	0	0	

Figure: GbS data.

CIMMYT could not analyzed individual plants, but mixed material from $m = 30$ plants arising from one accession.

Basic model

After all required preprocessing the data may be described as follows. Let $i = 1, 2, \ldots, C$ denote the accessions and $l = 1, 2, \ldots, M$ denote all loci.

For the i-th accession and locus l, the GBS method obtains counts n_{li}^1 and n_{li}^0 which correspond to the number of counted reference and alternative allele, respectively.

Let then $N_{li} = n_{li}^1 + n_{li}^0$ be the total number of counts observed for locus l for accession i. Let also $S_i = \sum_{l=1}^{M} N_{li}$ be the sum of all counts for accession i, $T_l = \sum_{i=1}^{C} N_{li}$ the sum of all counts (for all accessions) for locus l and $Z = \sum_{i=1}^{C} S_i$ the total sum of all counts for the data base.

Basic model

Note that ideally, that is with no counting error involved, only certain counts n_{li}^1 and n_{li}^0 are possible. Namely, if only one individual is analyzed then we should only have, for example, for

$$N_{li} = 30, (n_{li}^1, n_{li}^0) = (30, 0), (0, 30) \quad \text{or} \quad (15, 15).$$

That is, homocygotic with the reference allele, homocygotic with the alternative allele and heterocygotic.

However, we often see counts like (25,5), (28,2), (13,18) etc. The total N_{li} varies, depending on the locus and the accession.

Alele proportion estimates

Since we are analyzing 30 individuals (bulked), and these represent a population of homo and hetero cygotic genomes, what we aim at is to estimate the proportion of the the reference allele in the accession, for each locus. Let for the locus l the reference allele be denoted by X and the alternative by x. Let $A^j_{li} = 1$ if the j-th DNA string containing the locus, if such locus is X and $A^j_{li} = 0$ if the locus is x. Then

$$A^j_{li} \sim \text{Ber}(p_{li});$$

that is Bernoulli with parameter p_{li}, where p_{li} is the proportion of the reference allele, for locus l, within accession i.

Alele proportion estimates

However, we do not observe A_{li}^{j} without error. Counting errors are modeled as follows Let B_{li}^{j} be such that

$$B_{li}^{j} = \begin{cases} A_{li}^{j} & \text{with probability} \quad 1-e \\ 1 - A_{li}^{j} & \text{with probability} \quad e \end{cases} \tag{1}$$

that is, with some error e, the locus polymorphism is changed from X to x and viceversa. We do not observe A_{li}^{j} but rather B_{li}^{j}. This is our error model.

Alele proportion estimates

Using more notation

$$n_{li}^1 = \sum_{j=1}^{N_{li}} B_{li}^j \qquad \text{and} \qquad n_{li}^0 = \sum_{j=1}^{N_{li}} 1 - B_{li}^j,$$

($N_{li} = n_{li}^1 + n_{li}^0$). Therefore

$$n_{li}^1 | N_{li}, p_{li}, e \sim Bi(N_{li}, p_{li}(1-e) + (1-p_{li})e), \tag{2}$$

that is, the observed count for a reference allele has a binomial distribution including the error rate e. Note that N_{li} is known while p_{li} and e are unknown.

e is the same for all loci and all accessions.

Alele proportion estimates

Considering that we are working with bulked samples of size m the allele *theoretical* proportions, ie. the p_{li}'s, are

$$p_{li} = k\frac{1}{2m}, \quad k = 0, 1, 2, \ldots, 2m. \tag{3}$$

As explained in Section 1 the CIMMyT maize germoplams GBS effort has used bulked samples of size $m = 30$.

However, other smaller data sets are available using other individual agreggation numbers and for *double haploids* $m = 1$ and $p_{li} = 1, 0$ only, where estimation of the error rate e is specially suited.

Alele proportion estimates

For the moment we consider e fixed and proceed to estimate p_{li}.

Assuming $N_{li} > 0$ and establishing a prior distribution for p_{li}, the corresponding posterior distribution is informative and straightforward to obtain, namely

$$\begin{aligned} \mathrm{P}(p_{li} = p | n^1_{li}, n^0_{li}, e = e_0) &= \mathrm{P}(p_{li} = p | n^1_{li}, N_{li}, e = e_0) \\ &= \frac{h(p) C^{N_{li}}_{n^1_{li}} \{p(1-e_0) + (1-p)e_0\}^{n^1_{li}} \{(1-p)(1-e_0) + pe_0\}^{n^0_{li}}}{\sum_{a \in V} h(a) C^{N_{li}}_{n^1_{li}} \{a(1-e_0) + (1-a)e_0\}^{n^1_{li}} \{(1-a)(1-e_0) + ae_0\}^{n^0_{li}}} \end{aligned} \tag{4}$$

where the set $V(s) = \{s\frac{1}{2m} : s = 0, 1, \ldots, 2m\}$, are the possible values of p_{li}. $h(p)$ is the prior distribution of p_{li} and will be fixed uniform on $V(s)$ and e is fixed to e_0.

Alele proportion estimates

For the moment we consider e fixed and proceed to estimate p_{li}.

Assuming $N_{li} > 0$ and establishing a prior distribution for p_{li}, the corresponding posterior distribution is informative and straightforward to obtain, namely

$$\begin{aligned} \mathrm{P}(p_{li} = p | n_{li}^1, n_{li}^0, e = e_0) &= \mathrm{P}(p_{li} = p | n_{li}^1, N_{li}, e = e_0) \\ &= \frac{h(p) C_{n_{li}^1}^{N_{li}} \{p(1-e_0) + (1-p)e_0\}^{n_{li}^1} \{(1-p)(1-e_0) + pe_0\}^{n_{li}^0}}{\sum_{a \in V} h(a) C_{n_{li}^1}^{N_{li}} \{a(1-e_0) + (1-a)e_0\}^{n_{li}^1} \{(1-a)(1-e_0) + ae_0\}^{n_{li}^0}} \end{aligned} \tag{4}$$

where the set $V(s) = \{s\frac{1}{2m} : s = 0, 1, \ldots, 2m\}$, are the possible values of p_{li}. $h(p)$ is the prior distribution of p_{li} and will be fixed uniform on $V(s)$ and e is fixed to e_0.

Alele proportion estimates

Note that this posterior distribution may be calculated independently of other accessions and loci, is simple to implement since m is small, and is discrete on the support $V(s)$.

However, saving this complete distribution for all accessions and loci is totally infeasible given the size of the current data base.

An estimator, that is, one single number, is necessary to produce as an outline of this distribution. We decided to work with the posterior mean, that is

$$\hat{p}_{li} = \sum_{p \in V(s)} p \mathrm{P}(p_{li} = p | n_{li}^1, n_{li}^0, e = e_0).$$

Alele proportion estimates

Note that this posterior distribution may be calculated independently of other accessions and loci, is simple to implement since m is small, and is discrete on the support $V(s)$.

However, saving this complete distribution for all accessions and loci is totally infeasible given the size of the current data base.

An estimator, that is, one single number, is necessary to produce as an outline of this distribution. We decided to work with the posterior mean, that is

$$\hat{p}_{li} = \sum_{p \in V(s)} p \mathrm{P}(p_{li} = p | n_{li}^1, n_{li}^0, e = e_0).$$

Alele proportion estimates

Note that this posterior distribution may be calculated independently of other accessions and loci, is simple to implement since m is small, and is discrete on the support $V(s)$.

However, saving this complete distribution for all accessions and loci is totally infeasible given the size of the current data base.

An estimator, that is, one single number, is necessary to produce as an outline of this distribution. We decided to work with the posterior mean, that is

$$\hat{p}_{li} = \sum_{p \in V(s)} p\mathrm{P}(p_{li} = p | n_{li}^1, n_{li}^0, e = e_0).$$

Alele proportion estimates

Any chosen estimator will have issues unless, for example, the posterior distribution is peaked and concentrated around one single value. The posterior mean does not necessarily belong to the support $V(s)$, and care should be taken in the interpretation of this estimator, as with any other. Perhaps a dispersion estimate, or some hieghtest posterior probability range could be used instead, but would duplicate the size of a data base already extremely large.

Alele proportion estimates

To illustrate our methodology we present some examples of the corresponding posterior distribution for various observed counts and a fixed error rate $e = 0.06$ in Figure 3. The posteriors shown correspond to samples with $m = 30$ individuals, therefore there are 60 possible values for these discrete distributions. As the total count N_{li} increases the corresponding posterior becomes more peaked around the true value.

Alele proportion estimates

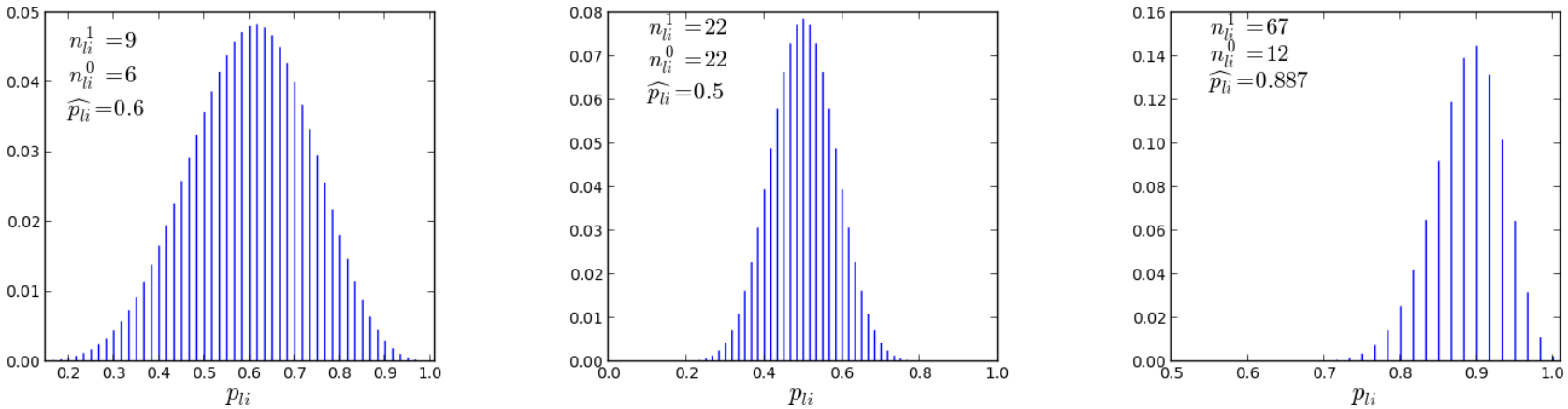


Figure: Three examples of the posterior distribution and the corresponding posterior mean estimator for p_{li} with $e = 0.06$. Since $m = 30$ the support $V(s)$ has 60 points and is a discrete distribution on $[0, 1]$. The posterior distribution is more concentrated as the total count ($N_{li} = n_{li}^1 + n_{li}^0$) increases, and the posterior mean becomes a better outline of the posterior.

Error estimation using double haploids

In principle we have a joint posterior distribution of all p_{li}'s and the error rate e, for all accessions and all loci. Note that error rates are only possible to estimate when differences are seen from possible values of p_{li}. For example, if $m = 2$ and a count ratio of $n^1_{i,l}/N_{i,l} = 20/30$ is observed, a counting error most be present since the proportion $2/3$ is not possible (only either $1/2$ or $3/4$).

The likelihood in this case will have information on e. However, if we increase the bulk size to $m = 30$, and without knowing the true theoeretical proportion of the reference allele, the likelihood in this case will have only marginal information on e given that the possible values for the p_{li}'s are 60 in the $[0, 1]$ interval, since any deviance from these possible values are already relatively small.

Instead we will use a special data subset of the original set of accessions called **double haploids**.

Error estimation using double haploids

In principle we have a joint posterior distribution of all p_{li}'s and the error rate e, for all accessions and all loci. Note that error rates are only possible to estimate when differences are seen from possible values of p_{li}. For example, if $m = 2$ and a count ratio of $n^1_{i,l}/N_{i,l} = 20/30$ is observed, a counting error most be present since the proportion $2/3$ is not possible (only either $1/2$ or $3/4$).

The likelihood in this case will have information on e. However, if we increase the bulk size to $m = 30$, and without knowing the true theoeretical proportion of the reference allele, the likelihood in this case will have only marginal information on e given that the possible values for the p_{li}'s are 60 in the $[0, 1]$ interval, since any deviance from these possible values are already relatively small.

Instead we will use a special data subset of the original set of accessions called **double haploids**.

Error estimation using double haploids

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...
$i = \mathbf{0956}b$ ... ACGGGTACCTATGCTCG ...

Separate chains and take one:
$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...

Duplicate:
$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...

Join to form a new DNA genome (!!!):
$i = \mathbf{new}a$ ... ACGGGCACCTATGGTCG ...
$i = \mathbf{new}b$ ... ACGGGCACCTATGGTCG ...
¡perfectly homocygotic!

Error estimation using double haploids

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...
$i = \mathbf{0956}b$ ... ACGGGTACCTATGCTCG ...

Separate chains and take one:

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...

Duplicate:

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...

Join to form a new DNA genome (!!!):

$i = \mathbf{new}a$ ... ACGGGCACCTATGGTCG ...
$i = \mathbf{new}b$ ... ACGGGCACCTATGGTCG ...

¡perfectly homocygotic!

Error estimation using double haploids

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...
$i = \mathbf{0956}b$ ... ACGGGTACCTATGCTCG ...

Separate chains and take one:

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...

Duplicate:

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...

Join to form a new DNA genome (!!!):

$i = \mathbf{new}a$ ... ACGGGCACCTATGGTCG ...
$i = \mathbf{new}b$ ... ACGGGCACCTATGGTCG ...

¡perfectly homocygotic!

Error estimation using double haploids

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...
$i = \mathbf{0956}b$ ... ACGGGTACCTATGCTCG ...

Separate chains and take one:

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...

Duplicate:

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...

Join to form a new DNA genome (!!!):

$i = \mathbf{new}a$ ... ACGGGCACCTATGGTCG ...
$i = \mathbf{new}b$ ... ACGGGCACCTATGGTCG ...

¡perfectly homocygotic!

Error estimation using double haploids

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...
$i = \mathbf{0956}b$ ... ACGGGTACCTATGCTCG ...

Separate chains and take one:

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...

Duplicate:

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...

$i = \mathbf{0956}a$ ... ACGGGCACCTATGGTCG ...

Join to form a new DNA genome (!!!):

$i = \mathbf{new}a$ ... ACGGGCACCTATGGTCG ...
$i = \mathbf{new}b$ ... ACGGGCACCTATGGTCG ...

¡perfectly homocygotic!

Error estimation using double haploids

We have more than 1,100 double haploid accessions. To simplify notation we will consider in this section that M is the total number of double haploids. The joint posterior distribution (assuming an uniform prior for all p_{li}'s)

$$f(p_{li}\text{'s}, e \mid n^1_{i,l}\text{'s and } n^0_{i,l}\text{'s}) \propto \pi(e) \prod_{l=1}^{M} \prod_{i=0}^{C} Bi(n^1_{i,l} | N_{i,l}, p_{li}),$$

where $Bi(\cdot|N, p)$ is the Binomial pmf from (2) and π is the prior distribution of e. Ideally we would like to work with the marginal distribution of p_{li}, summing over the rest of the p_{li}'s and e. However, this will render all calculations too complex and computationally unfeasible (an alternative MCMC method will be needed to calculate all marginals). Instead we proceed as follows.

Error estimation using double haploids

Summing over all p_{li}'s the marginal posterior distribution of e is

$$f(e \mid n^1_{i,l}\text{'s and } n^0_{i,l}\text{'s}) = K\pi(q)\prod_{l=1}^{M}\prod_{i=0}^{C} K_{i,l}(q),$$

where K is a normalizing constant and

$$K_{i,l}(e) = \sum_{p=0,1} C^{N_{i,l}}_{n^1_{i,l}}\{p(1-q)+(1-p)q\}^{n^1_{i,l}}\{(1-p)(1-q)+pq\}^{n^0_{i,l}}$$

(the normalizing constant in (4)). For a series of values of e and a repeated evaluation over various grids, this marginal distribution can be envisaged for e within a grid of $[0, 1]$. However, we propose to perform this calculation only once.

Error estimation using double haploids

If this marginal distribution for e is sufficiently peaked over its maximum e_0 then indeed

$$\int f(p_{li}\text{'s}, e \mid n^1_{i,l}\text{'s and } n^0_{i,l}\text{'s})de \approx f(p_{li}\text{'s} \mid e = e_0, n^1_{i,l}\text{'s and } n^0_{i,l}\text{'s}), \tag{5}$$

and we may proceed as in the previous Section, by conditioning over $e = e_0$. We then propose to make an estimate of e and proceed to fix it in subsequent calculations of $\hat{p}_{li}$.

We are in process of estimating e; we have seen it is within 0.01 to 0.06.

Estimating PAV's

Presence Absence Variation is an important feature in all genomes that constitute a crucial factor phenotypic separation [REF?]. PAV's refer to missing parts of the genome in some types or subtypes of a species. In our case it bounds to missing loci in accessions.

N_{li} not zero for a missing locus l in accession i and $N_{li} = 0$ even of the former is indeed present in the genome accession.

Estimating PAV's

Presence Absence Variation is an important feature in all genomes that constitute a crucial factor phenotypic separation [REF?]. PAV's refer to missing parts of the genome in some types or subtypes of a species. In our case it bounds to missing loci in accessions.

N_{li} not zero for a missing locus l in accession i and $N_{li} = 0$ even of the former is indeed present in the genome accession.

Estimating PAV's

$i = \mathbf{012321}a$ ... ACGGGTACCTCACGTAAGTTCG ...
$i = \mathbf{012321}b$ ... ACGGGTACCTGACGTAAGTTCG ...

$i = \mathbf{126734}a$ ... ACGGGTA...missing...TAAGTTCG ...
$i = \mathbf{126734}b$ ... ACGGGTA...missing...TAAGTTCG ...

Here we present an approach to estimate the probability that any locus is present given all observed counts.

Estimating PAV's

Let $\mathbf{N}_i = (N_{1i}, N_{2i}, \cdots, N_{Mi})$. We may think as the total count for accession i, $S_i = \sum_{l=1}^{M} N_{li}$, should be spread over the M categories (loci) with unknown probabilities $\mathbf{q}_i = (q_{1,i}, \ldots, q_{l,i}, \ldots, q_{M,i})$. That is

$$\mathbf{N}_i \mid S_i, \mathbf{q}_i \sim MN_M(S_i, \mathbf{q}_i); \tag{6}$$

that is a Multinomial distribution with M classes, total trials S_i and success probabilities $\mathbf{q}_i$. This this the basis of our modeling approach.

Estimating PAV's

To estimate PAV's we will assume that if a locus is missing in an accession, its corresponding cell probability will be lower. Accordingly, let now

$$x_{li} = \begin{cases} 1 & \text{if accession } i \text{ has locus } l \text{ missing} \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

We assume each locus l has a factor or strength r_l to be observed, and this is a feature of the DNA string associated with this locus, and independent of the particular accession being observed.

Estimating PAV's

If for accession i locus l is missing, then the associated cell in the multinomial model gets 'masked' reducing the probability of success of such cell; ie. simply, it is less likely to record an observation for such cell (locus). That is

$$q_{li} = c\lambda^{x_{li}} r_l$$

where c is a normalization constant to make $\sum_{l=1}^{M} q_{li} = 1$ and $\lambda \in (0, 1)$. λ represents the mask or reducing factor on the original measuring strength of the locus r_l, present only if the locus is missing for the accession, ie. $x_{li} = 1$.

Estimating PAV's

In this case, the vector of success probabilities for our Multinomial model is

$$\mathbf{q}_i = c\ (\lambda^{x_{1,i}} r_1, \ldots, \lambda^{x_{li}} r_l, \ldots, \lambda^{x_{M,i}} r_M), \tag{8}$$

where $c = 1/\sum_{l=1}^{M} \lambda^{x_{li}} r_l$.
Note that we are assuming one single factor λ for all accessions and all loci, but assuming a different λ for each accession will be a simple modification in our model.

Estimating PAV's

Using these assumptions, model (6) is completly specified and we may proceed to do our inference using all accession data S_i's and $\mathbf{N}_i$'s. However, this full model has $CM + M + 1$ parameters that in the current data base amounts to more than 10^8 parameters. Dealing with this full model is prohibitive and we need to follow a series of simplifications to be explain next.

Estimating PAV's

Indeed, the main interest are the parameters $\mathbf{x}_i = (x_{1,i}, \ldots, x_{li}, \ldots, x_{M,i})$ since λ and the r_l's may be considered as nuisance. Our simplification relies on fixing the r_l's to a simple estimate and assume λ as known. A sensitivity analysis will be performed on the former to verify our results. r_l will be estimated with the proportion T_l/Z. Since $T_l \sum_{i=1}^{C} N_{li}$ and Z os the overall total, then the ratio T_l/Z represents the overall strength or relative propensity to observe locus l accros all accessions. As a first approximation we fix $r_l = T_l/Z$.

Estimating PAV's

Following these further assumptions our posterior distribution is

$$
\begin{aligned}
P(\mathbf{x}_i|\mathbf{N}_i, S_i, \lambda, \mathbf{r}) &\propto \pi(\mathbf{x}_i) P(\mathbf{N}_i|S_i, \lambda, \mathbf{r}, \mathbf{x}_i) \\
&= \pi(\mathbf{x}_i) \frac{S_i!}{N_{1,i}! \cdots N_{M,i}!} (c\lambda^{x_{1,i}} r_1)^{N_{1,i}} \cdots (c\lambda^{x_{M,i}} r_M)^{N_{M,i}} \\
&= \pi(\mathbf{x}_i) \frac{S_i!}{N_{1,i}! \cdots N_{M,i}!} c^{\sum_{l=1}^{M} N_{li}} \lambda^{\sum_{l=1}^{M} x_{li} N_{li}} r_1^{N_{1,i}} \cdots r_M^{N_{M,i}} \\
&\propto \pi(\mathbf{x}_i) \frac{\lambda^{\sum_{l=1}^{M} x_{li} N_{li}}}{(\sum_{l=1}^{M} \lambda^{x_{li}} r_l)^{S_i}}
\end{aligned}
$$

where $\pi(\cdot)$ is the prior distribution of vector $\mathbf{x}_i$ and $\mathbf{r} = (r_1, \ldots, r_M)$ and $\lambda = 0.1$.

Estimating PAV's

Note that this posterior distribution may be calculated independently of other accessions, once λ and $\mathbf{r}$ are considered fixed. As a default prior for $\mathbf{x}_i$ we consider that, for each accession i, $h = P(x_{li} = 0)$ for all $l \in \{1, \ldots, M\}$. The posterior distribution reduces to

$$\begin{aligned} P(\mathbf{x}_i|\mathbf{N}_i, S_i, \lambda, \mathbf{r}) &\propto \prod_{l=1}^{M} h^{1-x_{li}}(1-h)^{x_{li}} \frac{\lambda^{\sum_{l=1}^{M} x_{li} N_{li}}}{(\sum_{l=1}^{M} \lambda^{x_{li}} r_l)^{S_i}} \\ &= h^M \left(\frac{1-h}{h}\right)^{\sum_{l=1}^{M} x_{li}} \frac{\lambda^{\sum_{l=1}^{M} x_{li} N_{li}}}{(\sum_{l=1}^{M} \lambda^{x_{li}} r_l)^{S_i}}. \end{aligned} \tag{9}$$

Certainly this expression does not belong to any known family and we will resort to a MCMC method to simulate from this posterior distribution. This is explained in the next section. Further calculations are based on a fixed $\lambda = 0.1$.

MCMC

Since each accession i will be analyzed independently we denote a particular value of $\mathbf{x}_i = \mathbf{x} = (x_1, \ldots, x_M)$ $(x_l = 0, 1)$ dropping the dependency in i to ease notation. We define a Metropolis-Hastings algorithm on the posterior

$$P(\mathbf{x}_i = \mathbf{x}|\mathbf{N}_i, S_i, \lambda, \mathbf{r}) \propto h^M \left(\frac{1-h}{h}\right)^{\sum_{l=1}^{M} x_l} \frac{\lambda^{\sum_{l=1}^{M} x_l N_{li}}}{(\sum_{l=1}^{M} \lambda^{x_l} r_l)^{S_i}}.$$

Let $E(\mathbf{x}) = K - \log P(\mathbf{x}_i = \mathbf{x}|\mathbf{N}_i, S_i, \lambda, \mathbf{r})$ for some arbitrary constant K, relevant only for numerical estabillity. $E(\mathbf{x})$ is the so called posterior 'energy'.

MCMC

Our MCMC algorithm is as follows

1. Fix a total number of iterations R.
2. Randomly set $\mathbf{x}^{(0)}$ as initial value for the Markov chain by drawing from the prior $\pi(\cdot)$.
3. Calculate the initial Energy:

$$E_0 = E(\mathbf{x}^{(0)}) = -A\log(\lambda) + S_i\log(B) - M\log(h) - s\log\left(\frac{1-h}{h}\right),$$

where $A = \sum_{l=1}^{M} x_l N_{li}$, $B = \sum_{l=1}^{M} \lambda^{x_l} r_l$ and $s = \sum_{l=1}^{M} x_l$.

4. Iterate $j = 1, \ldots, R$:
   a) Given the current point $\mathbf{x}^{(j-1)} = (x_1, \ldots, x_M)$, propose a new value $\mathbf{x}^*$ by randomly selecting a locus k, swapping the current value x_k from 0 to 1 and viceversa to create x_k^*, and set $\mathbf{x}^* = (x_1, \ldots, x_k^*, \ldots, x_M)$. Note that this simple proposal is symetric.
   b) Calculate the new energy

$$E^* = E(\mathbf{x}^*) = E_{j-1} + \text{Simple update involving only } x_k^* \text{ and } x_k.$$

   c) Accept the proposal $\mathbf{x}^{(j)}$ with probability

$$\rho(\mathbf{x}^{(j-1)}, \mathbf{x}^*) = \min\left\{1, e^{E_{j-1} - E_j}\right\}$$

   and set $\mathbf{x}^{(j)} = \mathbf{x}^*$ and the new energy to $E_j = E^*$, otherwise reject the proposal and set $\mathbf{x}^{(j)} = \mathbf{x}^{(j-1)}$ and $E_j = E_{j-1}$.

MCMC

This is a simple MCMC. We decided to explain the details since the algorithm is feasible only by correctly updating the energy in step 4(b). Although the expression is simple, calculating the full energy in step 3 is a very time consuming process given that $M \approx 25 \times 10^4$. Our MCMC updates one x_l at each iterations and the calculations involved are very simple indeed.

This permits doing several million iterations in a matter of several hours. We cannot save all iterations $\mathbf{x}^{(j)}$, instead, by correct book keeping we only save the number of 1's in the chain for each variable, to obtain an approximation of the marginal posterior probability $P(x_{li} = 1|\mathbf{N}_i, S_i, \lambda, \mathbf{r})$. By comparing several runs in few accessions, we decided to set R to 400 million, with a burn.in of 10 million.

MCMC

This is a simple MCMC. We decided to explain the details since the algorithm is feasible only by correctly updating the energy in step 4(b). Although the expression is simple, calculating the full energy in step 3 is a very time consuming process given that $M \approx 25 \times 10^4$. Our MCMC updates one x_l at each iterations and the calculations involved are very simple indeed.

This permits doing several million iterations in a matter of several hours. We cannot save all iterations $\mathbf{x}^{(j)}$, instead, by correct book keeping we only save the number of 1's in the chain for each variable, to obtain an approximation of the marginal posterior probability $P(x_{li} = 1|\mathbf{N}_i, S_i, \lambda, \mathbf{r})$. By comparing several runs in few accessions, we decided to set R to 400 million, with a burn.in of 10 million.

MCMC

This is our based model, and although simplified, the MCMC takes at least 40 hours to run in one single accession. There are more that 10,000 accessions and in the next section we present a simplified calculation procedure that will be calibrated using the MCMC output of some 35 accessions.

Binomial approximation

From the Multinomial model in (6), $\mathbf{N}_i \mid S_i, \mathbf{q}_i \sim MN_M(S_i, \mathbf{q}_i)$, a well knwon fact is that the marginal distribution of each N_{li} is a Binomial distribution, specifically

$$N_{li}|S_i, q_{li} \sim Bi(S_i, q_{li}).$$

Let $b^j_{li}|x_{li} \sim Be(q_{li}(x_{li}))$ such that, given x_{li}, $N_{li} = \sum_{j=1}^{S_i} b^j_{li}$. That is, there are S_i b^j_{li}'s and these account for the individual DNA strings, with $b^j_{li} = 1$ if the j^{th} string has the locus l.

Binomial approximation

By total probability we do have

$$P(b_{li}^j = 1) = P(b_{li}^j = 1|x_{li} = 0)P(x_{li} = 0) + P(b_{li}^j = 1|x_{li} = 1)P(x_{li} = 1), \tag{10}$$

where $q_{li}(x) = P(b_{li}^j = 1|x_{li} = x)$. It is not unreasonable to see this marginal probability as approximately the strength of observing DNA strings associated with the locus l; therefore

$$P(b_{li} = 1) \approx r_l \approx \frac{T_l}{Z}.$$

Equating $P(b_{li} = 1) = T_l/Z$ may indeed be debatable but we are only trying to generate an approximate procedure, that will later be calibrated with the better justified Multinomial approach, presented in the previous Section.

Binomial approximation

Since $h = P(x_{li} = 1)$ (the prior for x_{li}), then

$$\frac{T_l}{Z} = P(b^j_{li} = 1 | x_{li} = 1)\{h + \alpha_{il}(1 - h)\}$$

where $\alpha_{il} P(b^j_{li} = 1 | x_{li} = 1) = P(b^j_{li} = 1 | x_{li} = 0)$. Certainly we expect $P(b^j_{li} = 1 | x_{li} = 1) > P(b^j_{li} = 1 | x_{li} = 0)$ and therefore $\alpha_{il} \in [0, 1]$. Our next approximation strategy is to consider $\alpha_{il} = \alpha_i$, that is, all α_{il}'s fixed to one single value, dependent only on the accession but independent of all loci.

Binomial approximation

This α_i value will be our main calibrating factor to obtain an acceptable approximation procedure, as will be explain in Section 4. Note that now we have expressions for $q_{li}(0) = P(b^j_{li} = 1|x_{li} = 0) = \frac{T_l/Z}{h+\alpha_i(1-h)}$ and $q_{li}(1) = P(b^j_{li} = 1|x_{li} = 1) = \frac{\alpha_i T_l/Z}{h+\alpha_i(1-h)}$. From this (indeed, as a function of α_l) we may calculate the posterior probability

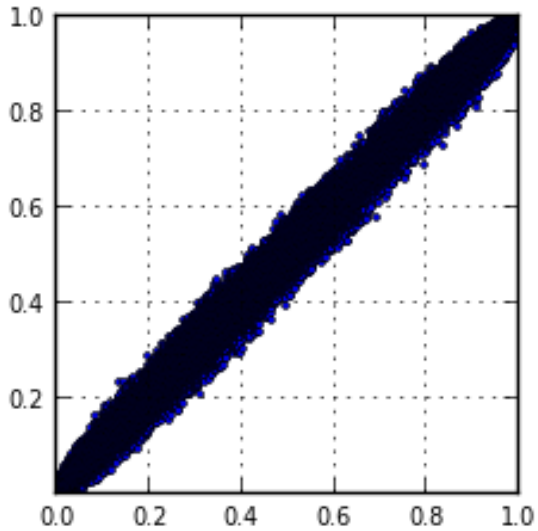
$$P(x_{li} = 0|N_{li}) = \frac{P(N_{li}|S_i, x_{li} = 0)h}{P(N_{li}|S_i, x_{li} = 0)h + P(N_{li}|S_i, x_{li} = 1)(1-h)} \tag{11}$$

where $N_{li}|S_i, x_{li} \sim Bi(S_i, q_{li}(x_{li}))$.

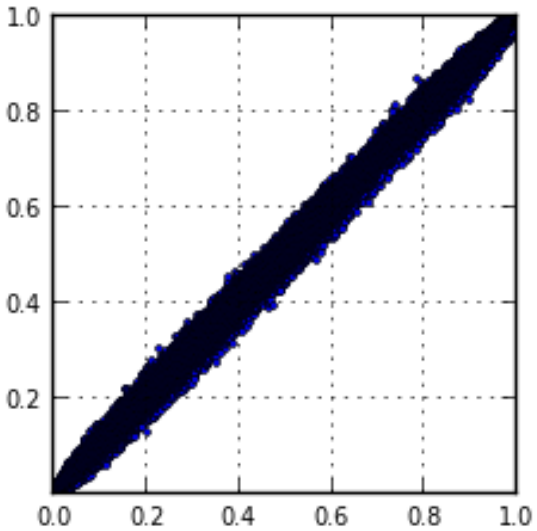
Binomial approximation

Calculating this posterior probability is now straightforward and is done independently of other loci and accessions, does not involve a MCMC method and analyzing all 25×10^4 loci in one accession takes approximately 10min. In the next section we present our results, with strategy to calibrate α_i.

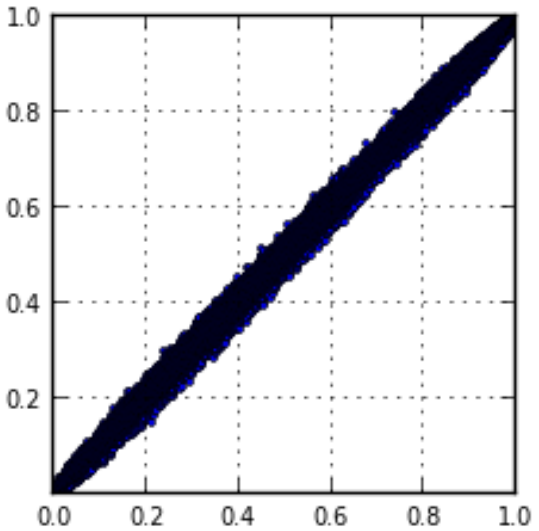
Results



(a)



(b)



(c)

Figure: Approximated $P(x_{il} = 0|Data)$ probabilities with two MCMC runs of (a) 200×10^6, (b) 400×10^6 and (c) 600×10^6 iterations, for accessions $i = 3634$. Little further improvement in precision is seen from (b) to (c) and therefore we decided to fix the number of runs to 400 million.

Results

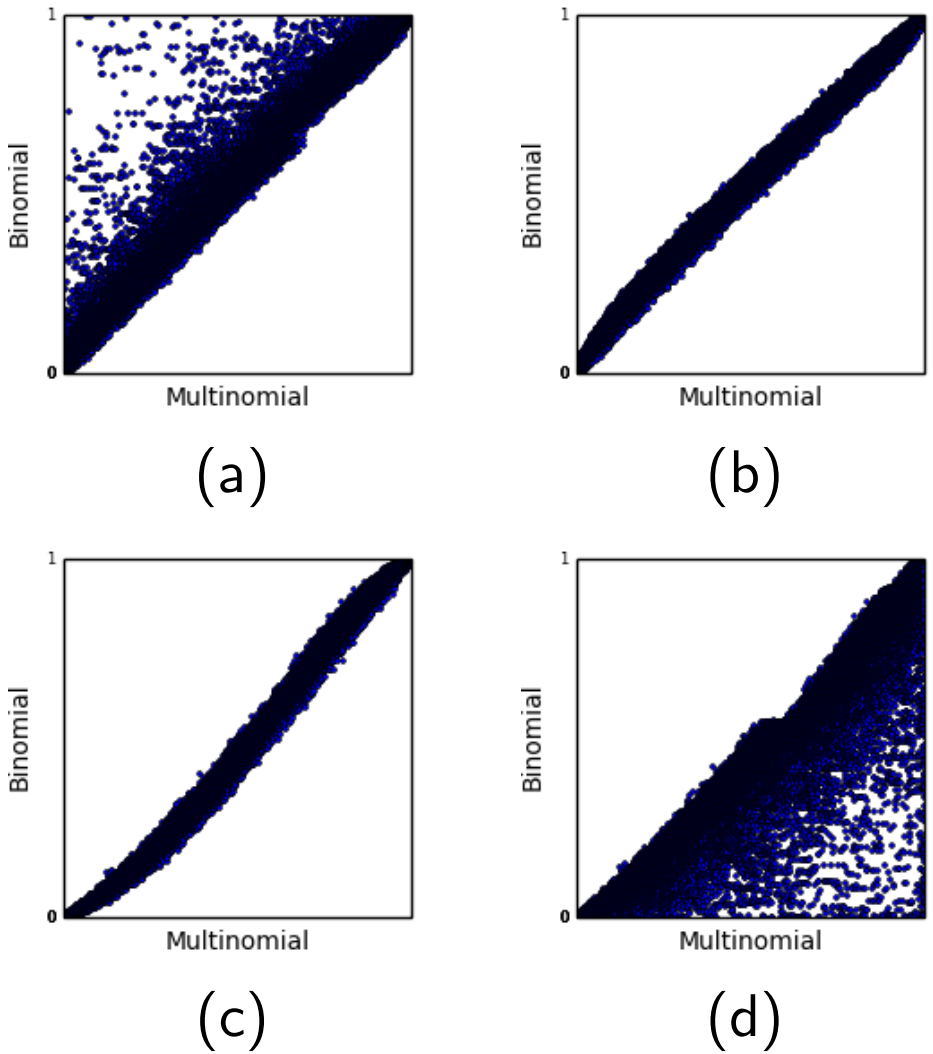


Figure: There is small range of α's that lead to very good approaximations.

Results

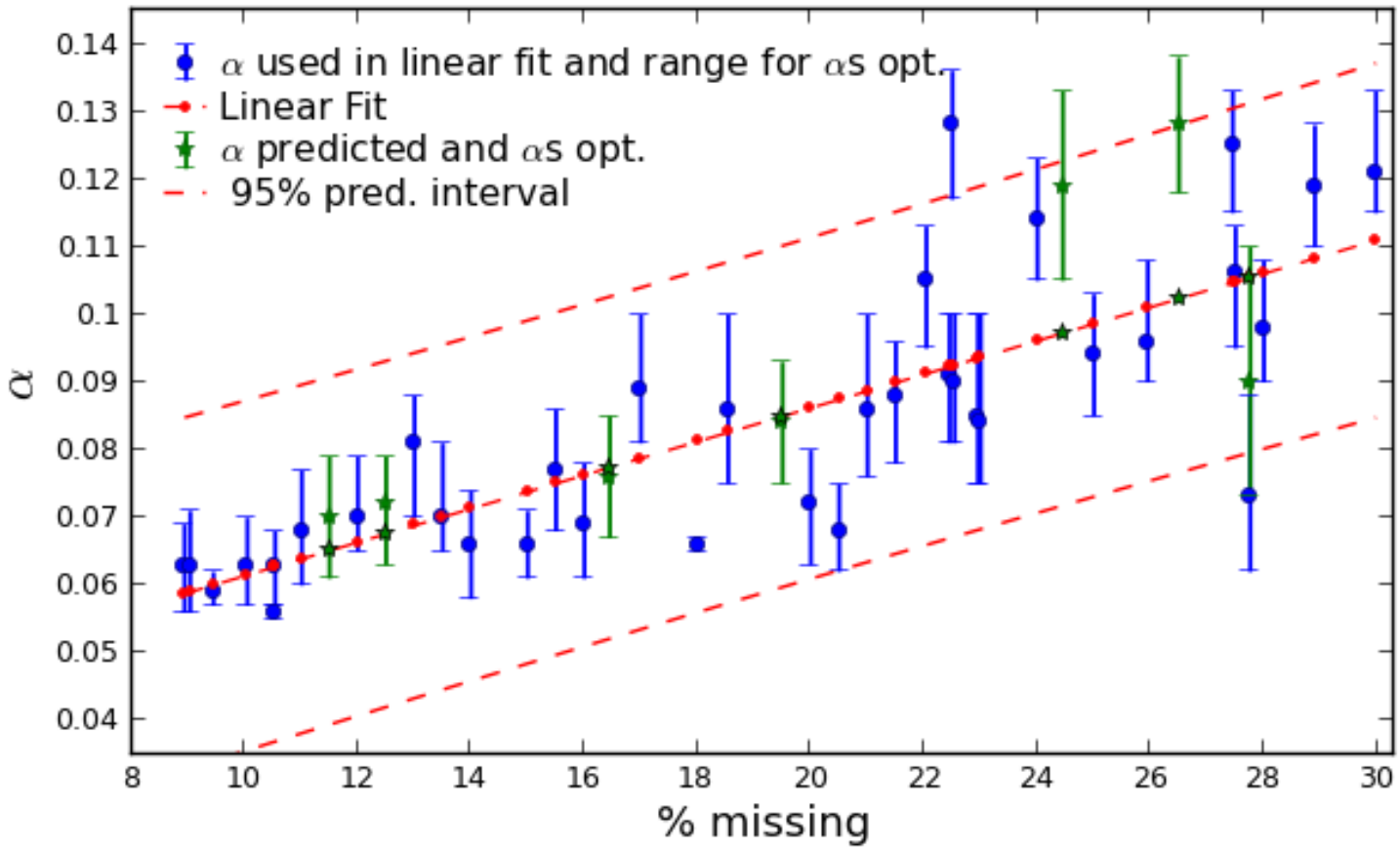


Figure: .

Discussion

- The data set now comprises 1,000,000 loci and 20,000 accessions!
- Analyzing for allele proportions and PAV's is possible using this technique.
- However, even the simplest manipulation is prohibitive!!
- The great benefit will be when this genotyping is mapped with phenotypes (grow in dry areas, tropical maize, temperate maize, etc.), to classify all maize landraces at CIMMYT.
- Moreover, once this basic mapping is at hand, *in silico* hybridization couls be a possibility.

Discussion

GRACIAS!