Watterson estimators for Next Generation Sequencing: from trios to autopolyploids

Luca Ferretti\textsuperscript{a,b,*}, Sebástian E. Ramos-Onsins\textsuperscript{c}

\textsuperscript{a}Systématique, Adaptation et Evolution (UMR 7138), UPMC Univ Paris 06, CNRS, MNHN, IRD, Paris, France
\textsuperscript{b}CIRB, Collège de France, Paris, France
\textsuperscript{c}Centre for Research in Agricultural Genomics (CRAG), 08193 Bellaterra, Spain

Abstract

Several variation of the Watterson estimator of variability for Next Generation Sequencing (NGS) data have been proposed in the literature. We present a unified framework for generalized Watterson estimators based on Maximum Composite Likelihood, which encompasses most of the existing estimators. We propose this class of unbiased estimators as generalized Watterson estimators for a large class of NGS data, including pools and trios. We also discuss the relation with the estimators that have been proposed in the literature and show that they admit two equivalent but seemingly different forms, deriving a set of combinatorial identities as a byproduct. Finally, we give a detailed treatment of Watterson estimators for single or multiple autopolyploid individuals.

Keywords: Coalescent theory, Site frequency spectrum, Population genetics, Summary statistics, Maximum likelihood

1. Introduction

The rescaled mutation rate per base $\theta$ is an important quantity for population genetics models. Its definition is $\theta = 2pN_e\mu$ where $\mu$ is the mutation rate per base, $N_e$ is the effective population size and $p$ is the ploidy of the species. In the context of the SNM at low mutation rate, the most used
estimator is the Watterson estimator \[1\] based on the number of segregating sites \(S\) in a sample from the population. It is a good estimator since it is unbiased and it corresponds to the Maximum Composite Likelihood estimator for \(\theta\), moreover it is a sufficient statistics for \(\theta\) for large sequences \[2\]. As shown there, for unequal mutation rates between different alleles, Watterson estimator actually measures the net rate of mutation towards different alleles, rescaled by population size.

Today, genomic scale variability analysis are affordable thanks to NGS technologies. NGS technologies can sequence a single complete genome at relatively high redundancy but the sequencing of many samples multiplies substantially the cost of the experiment. Several strategies have been used to obtain sequence data from many individuals (e.g., from restriction reduced libraries to/and pooled samples) which reduce substantially the coverage across the genome (i.e. how many regions are sequenced) or/and the sequence redundancy per nucleotide base (i.e. how many sequences for each nucleotide are sequenced). In that sense, the study of variability on NGS data can be seen as a missing data problem, where information about several samples is missing at covered positions. The Watterson \(\theta\) estimator has been generalized to missing data by Ferretti et al. \[3\]. However this covers just a limited number of cases.

Estimators of variability based on NGS data should take into account the relative high probability of sequence error in order to not overestimate the number of variants detected. The sequence redundancy at a given position is fundamental for eliminating sequence errors but also for detecting homozygote or heterozygote positions in diploid individuals. NGS estimators for estimating variability were first proposed by Lynch \[4\] for a single diploid individual using the sequence error rate and the number of reads observed at each variant. Also Hellmann et al. \[5\] and Jiang et al. \[6\] proposed Watterson estimators for multiple individuals considering the combinatorics of having reads from different individuals and homologous chromosomes. Later, Futschik and Schlötterer \[7\] and Ferretti et al. \[8\] developed variability estimators for pooled sample data using a Method of Moments (MM) and a Maximum Composite Likelihood (MCL) methods, respectively. We observe that for an analysis of a single diploid individual, they all reduce to the same estimator except for \[7\]. In this work, the relations between these estimators are studied and new Watterson estimators for other kinds of NGS data are developed, as for example the case of trios or autopolyploids.

In our framework, we treat all these cases within an unified approach.
Data are represented by NGS or Sanger sequences coming from several units. Units can be haploid, diploid or polyploid individuals or pools, each one containing a different number of lineages. We derive Maximum Composite Likelihood estimators for a large class of data. These MCL estimators are not unbiased, but we show that they can be well approximated by unbiased estimators similar to the Watterson estimator. We propose this class of unbiased estimators as generalized Watterson estimators for a large class of NGS data, such as multiple haploid/diploid/polyploid individuals, pools, trios, inbred lines, and combinations of these. We discuss their relation with other estimators that have been proposed in the literature, showing that many of our estimators admit two equivalent but seemingly different forms. As a byproduct, this equivalence implies a set of combinatorial identities. Finally, we treat in details the case of single or multiple autopolyploid individuals and we provide the Watterson estimator for autopolyploids.

2. General Watterson estimators

2.1. Maximum Composite Likelihood estimators

Composite Likelihood Estimation of parameters have been extensively used in population and quantitative genetics for estimating the linkage disequilibrium among positions [9, 10] and for estimating evolutionary parameters as the level of variation [5], the recombination parameter [11], [12], the effect of positive selection [13, 14] or demographic parameters [15]. Maximum Composite Likelihood is an appropriate method to estimate the nucleotide variability across large regions because it turns out to be based on the same statistics as the Watterson estimator (the total number of segregating sites S) and it actually reduces to the Watterson estimator if the data are represented by complete sequences. Furthermore, it has minimum variance for large regions with recombination, since in this case the Composite Likelihood is a good approximation for the exact likelihood.

We summarize the SNP features, like allele frequencies, with the index \( \xi \in \Xi \) and the features of each site that do not depend on the allelic content, like the read depth, with the index \( \varphi \in \Phi \). Both indices indicate non-overlapping, mutually exclusive features.

We denote by \( p_{\varphi, \xi}(\theta) \) the probability that a site with features \( \varphi \) contains an observed SNP with features \( \xi \) for the sample studied. For small values of \( \theta \) (that is, infinite site model), these probabilities are proportional to \( \theta \)
multiplied by a quantity that depends on the population model and the sequencing setup:

\[ p_{\varphi,\xi}(\theta) = \theta Z_{\varphi,\xi} \]  \hspace{1cm} (1)

We denote by \( S_{\varphi,\xi} \) the number of segregating sites with features \( \xi \) in positions with features \( \varphi \), and by \( L_{\varphi} \) the number of sites with features \( \varphi \). We also define the quantities \( S_{\varphi} = \sum_{\xi \in \Xi} S_{\varphi,\xi} \) and \( Z_{\varphi} = \sum_{\xi \in \Xi} Z_{\varphi,\xi} \), the total number of segregating sites \( S = \sum_{\varphi \in \Phi} S_{\varphi} \) and total length \( L = \sum_{\varphi \in \Phi} L_{\varphi} \).

Under the composite approximation, all sites are independent. The Composite Likelihood is therefore

\[
CL(\theta) = \left[ \prod_{\varphi \in \Phi} \prod_{\xi \in \Xi} p_{\varphi,\xi}(\theta)^{S_{\varphi,\xi}} \right] \cdot \left[ \prod_{\varphi \in \Phi} \left( 1 - \sum_{\xi \in \Xi} p_{\varphi,\xi}(\theta) \right)^{L_{\varphi} - S_{\varphi}} \right] \]  \hspace{1cm} (2)

Substituting \( p_{\varphi,\xi}(\theta) \) and taking the log, we obtain

\[
\log(CL(\theta)) = S \log(\theta) + \sum_{\varphi \in \Phi} \sum_{\xi \in \Xi} S_{\varphi,\xi} \log(Z_{\varphi,\xi}) - \sum_{\varphi \in \Phi} (L_{\varphi} - S_{\varphi}) \log(1 - \theta Z_{\varphi}) \]  \hspace{1cm} (3)

For \( S > 0 \), the loglikelihood is always negative and tends to \(-\infty\) both for \( \theta \to 0 \) and \( \theta \to 1/\max_{\varphi \in \Phi}(Z_{\varphi}) \), so the maximum can be obtained from the zeros of the first derivative of the above equation. After rearrangements, we obtain the equation that defines the MCLE:

\[
L = \sum_{\varphi \in \Phi} \frac{L_{\varphi} - S_{\varphi}}{1 - \theta_{\text{MCLE}} Z_{\varphi}} \]  \hspace{1cm} (4)

which is valid when the second derivative in \( \theta = \hat{\theta}_{\text{MCLE}} \) is negative, that is, in the range \( \sum_{\varphi \in \Phi}(L_{\varphi} - S_{\varphi})Z_{\varphi}(1 - 2\theta_{\text{MCLE}} Z_{\varphi})/(1 - \theta_{\text{MCLE}} Z_{\varphi})^2 > 0 \). This is always true for \( \theta \) small enough, that is, \( \theta < (2 \max_{\varphi \in \Phi}(Z_{\varphi}))^{-1} \).

For the simplest case of the original Watterson estimator \( \hat{\theta}_W \), all sites are equivalent and there is no site feature \( \varphi \), so the above equation can be easily rewritten as \( \hat{\theta}_{\text{MCLE}} = \hat{\theta}_W = S/LZ \). The SNP features \( \Xi \) correspond simply to the derived allele counts \( i = 1 \ldots n - 1 \) and the probability of a SNP of frequency \( i/n \) is related to the frequency spectrum \( \xi_i \) by \( p_i(\theta) = E(\xi_i)/L \). For the standard neutral model, \( p_i(\theta) = \theta/i \), therefore \( Z = \sum_{i=1}^{n-1} Z_i = \sum_{i=1}^{n-1} p_i(\theta)/\theta \) is given by the harmonic number \( a_n = \sum_{i=1}^{n-1} 1/i \). Then the MCLE in this case corresponds precisely to the unbiased Watterson estimator.
The estimator (4) is defined implicitly, so it is not particularly convenient. An explicit, approximate MCL estimator can be derived in two equivalent ways: either by expanding equation (4) at first order in the small parameters $\theta$ and $S_\phi/L_\phi$ with constant ratio $S_\phi/\theta L_\phi$, or by taking the small $\theta$, large $L$ limit of the likelihood (2) with $\theta L$ and $L_\phi/L$ constant. In this limit, the $S_{\phi,\xi}$ are Poisson distributed random variables with mean $L_\phi \theta Z_{\phi,\xi}$.

\[
CL(\theta) \approx \prod_{\phi \in \Phi} \prod_{\xi \in \Xi} \frac{(L_\phi \theta Z_{\phi,\xi})_{S_{\phi,\xi}}}{S_{\phi,\xi}!} e^{-L_\phi \theta Z_{\phi,\xi}} = \theta^S e^{-\theta \sum_{\phi \in \Phi} L_\phi Z_{\phi}} \left[ \prod_{\phi \in \Phi} \prod_{\xi \in \Xi} \frac{(L_\phi Z_{\phi,\xi})_{S_{\phi,\xi}}}{S_{\phi,\xi}!} \right] \]

and by the factorization theorem $S$ is a sufficient statistics for $\theta$, as already observed in [2]. We define the resulting approximate MCL estimator as the generalized Watterson estimator:

\[
\hat{\theta}_W = \frac{S}{\sum_{\phi \in \Phi} L_\phi Z_{\phi}} \tag{6}
\]

This estimator depends only on the total number of segregating sites, like the original Watterson estimator, since $S$ is a sufficient statistics for small $\theta$. Furthermore, it is unbiased. In fact, $E(S) = \sum_{\phi \in \Phi} \sum_{\xi \in \Xi} E(S_{\phi,\xi}) = \sum_{\phi \in \Phi} \sum_{\xi \in \Xi} L_\phi p_{\phi,\xi}(\theta) = \sum_{\phi \in \Phi} \theta L_\phi Z_{\phi}$.

Both the implicit estimator in equation (4) and the formula (6), which is unbiased, can be used. The error of the Watterson estimator (6) with respect to the MCLE (4) is very small, in practice of order $\theta^2$ multiplied by a weighted covariance between $Z$ and the relative fluctuations of $\hat{\theta}_W$ for different $\phi$:

\[
\hat{\theta}_{MCL} - \hat{\theta}_W \approx -\hat{\theta}_W^2 \sum_{\phi \in \Phi} \frac{L_\phi}{L} Z_{\phi} \left( Z_{\phi} - \frac{\sum_{\phi' \in \Phi} L_{\phi'} Z_{\phi'}}{L} \right) \left( \frac{S_{\phi}/L_\phi Z_{\phi} - \hat{\theta}_W}{\hat{\theta}_W} \right) \tag{7}
\]

which is usually negligible.

The only information needed to compute (4) or (6) are the factors $Z_{\phi} = \sum_{\xi \in \Xi} p_{\phi,\xi}/\theta$, which depends both on the model and on the sequencing setup. In the rest of the paper, we will specialize these factors for several combinations of NGS data.

2.2. Estimators for Next Generation Sequencing

In this section we deal with combinations of both complete sequences and NGS data from different sources in a unified way. Our data are represented
by reads or sequences, aligned to a reference genome. Each read/sequence originates from a single unit: units can be individuals of different ploidy, or pools of individuals. Complete sequences are considered as sequences coming from an haploid unit, so the two complete sequences of the two homologous chromosomes from a diploid organism are equivalent to two different haploid units.

We denote by $U$ the number of units. The features associated to the units are the number of copies of homologous chromosomes present in each unit and their evolutionary relationships. We denote the number of copies of homologous chromosomes in the $i$th unit by $c_i$, $i = 1\ldots U$. Diploid individuals will have $c_i = 2$, polyploid individuals will have $c_i$ equal to their ploidy and pools will have $c_i$ equal to the number of individuals in the pool multiplied by their ploidy. The evolutionary relationships are well described by the probabilities that some of the sequenced chromosome derived by the same lineage, either because they are actually from the same individual or because are identical by descent. We denote this information as $\pi_U$.

We assume that the number of reads covering each position depends only on the sequencing process and not on the allelic composition of the sequence. In this case, we associate to each position $x$ the read depth of the $i$th unit $r_i(x)$, $i = 1\ldots U$, i.e. the number of reads or sequences from the $i$th unit that cover position $x$.

We can derive the Watterson estimator for this case by using the definition (6) with $\varphi = \{r_i\}_{i=1\ldots U}$. $Z$ can be computed by conditioning on the number of independent homologous chromosomes -or lineages- that are actually sequences, denoted by $j$, and then averaging over $j$:

$$\theta Z_{\{r_i\}_{i=1\ldots U}} = \sum_{j=2}^{\infty} P(\text{SNP} | \{r_i\}_{i=1\ldots U}, j) \cdot P_c(j | \pi_U, \{c\}, \{r\})$$

where $P(\text{SNP}|\ldots)$ is the probability of observing a SNP and $P_c(j|\ldots)$ is the distribution of $j$. These probabilities may depend also on $\{c_i\}_{i=1\ldots U}$ and $\pi_U$. The probability of observing a SNP among $j$ independent lineages depends just on $j$ and on the expected site frequency spectrum $E(\xi_k|j)$ of the model, being equal to $P(\text{SNP}|j) = \sum_{k=1}^{j-1} E(\xi_k|j) / L$. In the case of the standard neutral Wright-Fisher model we have $E(\xi_k|j) = \theta L / k$ and therefore $P(\text{SNP}|j) = \theta \sum_{k=1}^{j-1} 1/k = \theta a_j$.

We use the short form $\{r\} = \{r_i\}_{i=1\ldots U}$ and $\{c\} = \{c_i\}_{i=1\ldots U}$ for the information about the site features. The general Watterson estimator for the
Standard Neutral Model (SNM) is then

$$\hat{\theta}_W = \frac{S}{\sum_{\{r\}} L_{\{r\}} \sum_{j=2}^{\infty} P_c(j|\pi_U, \{c\}, \{r\}) \cdot a_j}$$

(9)

This form was found by [5] and [16, 8] in specific cases, but it holds for a very large class of data as shown. In the next section we will find the expression of $P_c(j|\ldots)$ for the most common sequencing setups.

### 3. Distributions of the number of sequenced lineages

As discussed in the previous sections, the distribution of the number of sequenced lineages $P_c(j|\ldots)$ is actually enough to define the Watterson estimator. Before deriving its expression for a number of cases, we introduce some notation.

We denote the Stirling numbers of second kind for $j$ sets from $r$ objects by $S(r, j)$. We define the probability distribution

$$P^*(j|c, r) = \frac{c! S(r, j)}{(c - j)! \cdot c^r}$$

(10)

that corresponds to the probability of extracting exactly $j$ different objects with $r$ extractions (with repetitions) from a set of $c$ objects [8]. It will often appear in the formulae for the distribution of the number of lineages.

We denote by $I(x)$ the indicator function that takes the value 1 if $x$ is true and 0 otherwise. We also denote by $\delta_{i,j}$ the Kronecker delta, that is, the identity matrix $\delta_{i,j} = I(i = j)$.

#### 3.1. General case: independent lineages

Assume that all lineages in these units are independent. This corresponds to sequencing many unrelated individuals in a population without inbreeding. If there are $U$ units, $c_i$ is the number of lineages/homologous chromosomes in the $i$th unit, and $r_i$ is the number of reads/sequences coming from the $i$th unit, the probability $P_c(j|\{c\}_{i=1..U}, \{r\}_{i=1..U})$ in the Watterson estimator is

$$P_c(j|\{c\}, \{r\}) = \sum_{i_1=0}^{c_1} \cdots \sum_{i_U=0}^{c_U} I\left(j = \sum_{p=1}^{U} i_p\right) \prod_{q=1}^{U} P^*(i_q|c_q, r_q)$$

(11)

In Section 4 we will present an alternative form for these Watterson estimators. In the rest of this section we specialize the expression (11) to the most common scenarios.
### 3.1.1. Multiple haploid individuals

In this case all individuals have $c_i = 1$, therefore $P^*(i_q|1, r_q) = I(r_q > 0)$ and the estimator reduces to the one proposed for missing data in [3], which is equivalent to

$$P_c(j|c = n, \{r_i\}_{i=1..n}) = I\left(j = \sum_{i=1}^{n} I(r_i > 0)\right)$$

(12)

This choice was implicitly suggested also in [2].

### 3.1.2. Multiple diploid individuals

In this case all individuals have $c_i = 2$ and the MCLE was already derived by Hellmann et al. [5]:

$$P_c(j|c = 2n, \{r_i\}_{i=1..n}) = \sum_{i_1=0}^{2} \cdots \sum_{i_n=0}^{2} I\left(j = \sum_{p=1}^{n} i_p\right) \prod_{q=1}^{n} P^*(i_q|2, r_q)$$

(13)

### 3.1.3. Pools

In this case, there is a single unit of $c$ chromosomes. The probability was derived by Ferretti et al. [8]

$$P_c(j|c, r) = P^*(j|c, r)$$

(14)

but see also Section 4 for a simpler formula.

For multiple pools, the probability follows closely equation (11).

### 3.2. Related lineages

Sequencing of many unrelated individuals, either pooled together or sequenced separately, is the most common experimental setup for variability studies as described in the previous section, but it is not the only one. There are several cases where lineages in different units are related by identity (for example, the same individual sequenced both alone and in a pool with other individuals) or identity-by-descent, like for trios or inbred lines from a population. We develop estimators for these cases.
3.2.1. Trios

A trio is a (diploid) family of mother, father and son that are sequenced separately. We restrict our analysis to autosomes, where the alleles of the son are the copies of a paternal and a maternal allele. For complete sequences, the probability is just $P(j) = \delta_{j,4}$ since there are four independent lineages.

For NGS data, we denote by $r_M$, $r_F$ and $r_S$ the read depths of mother, father and son respectively. We obtain $P_c(j|r_M, r_F, r_S)$ by conditioning on the number of lineages $j'$ sequenced in the parents. We can rewrite it in terms of the probability $P_c(j'\mid c = 4, r_M, r_F)$ of the parents alone (eq.13):

$$P_c(j| r_M, r_F, r_S) = \sum_{j' = 0}^{4} P_t(j\mid j') P_c(j'\mid c = 4, r_M, r_F) \quad (15)$$

where $P_t(j\mid j')$ is obtained case by case depending on the probability that the sequences of the son could contain new alleles with respect to the parental sequences:

$$P_t(j\mid 0) = P^*(j\mid 2, r_S) \quad (16)$$
$$P_t(j\mid 1) = \frac{1}{2} P^*(j - 1\mid 2, r_S) + \frac{1}{2} (\delta_{j,1} 2^{-r_S} + \delta_{j,2} (1 - 2^{-r_S}))$$
$$P_t(j\mid 2) = \frac{1}{2} (1 + I(r_M r_F = 0)) (\delta_{j,2} 2^{-r_S} + \delta_{j,3} (1 - 2^{-r_S})) + \frac{1}{4} I(r_M r_F > 0) (\delta_{j,2} + P_c(j - 2\mid 2, r_S))$$
$$P_t(j\mid 3) = \frac{1}{2} \delta_{j,3} + \frac{1}{2} (\delta_{j,3} 2^{-r_S} + \delta_{j,4} (1 - 2^{-r_S}))$$
$$P_t(j\mid 4) = \delta_{j,4}$$

Multiple trios can be dealt with by replacing the probability $P^*(i_q\ldots)$ in equation (11) with the probability (15) and replacing $c_q$ with 4.

3.2.2. Pooled trios

A pooled trio is a family of mother, father and son that are sequenced in a pool. We consider $n$ unrelated families, each one sequenced separately from the others, and denote by $r_i$ the total read depths. $P_c(j\mid \{r_i\}_{i=1\ldots n})$ is given by

$$P_c(j\mid \{r_i\}_{i=1\ldots n}) = \sum_{i_1 = 0}^{4} \cdots \sum_{i_n = 0}^{4} \delta \left( j = \sum_{p=1}^{n} i_p \right) \prod_{q=1}^{n} P_{pt}(i_q \mid r_q) \quad (17)$$
where $P_{pt}(i|r)$ is the probability for a single pool. It can be derived case-by-case conditioning on the number of sequenced chromosomes (identical or not) for a pool, obtaining

$$P_{pt}(i|r) = \left( 4 \binom{2}{i-2} + 4 \binom{2}{i-1} + \binom{2}{i} \right) \frac{P^*(i|6,r)}{\binom{6}{i}} +$$

$$+ \left( 4 \binom{2}{i-2} + 2 \binom{2}{i-1} \right) \frac{P^*(i+1|6,r)}{\binom{6}{i+1}} + \left( 2 \binom{2}{i-1} \right) \frac{P^*(i+2|6,r)}{\binom{6}{i+2}}$$

(18)

### 3.2.3. Pools and complete sequences with overlapping individuals

A potentially useful setup is the combination of complete sequences of few individuals and a pool of several individuals from the same population. In some cases, there could be individuals in the pool for which the complete sequence is also available.

Here we consider the haploid case, but can be easily adapted to the diploid case by considering a diploid individual as a pair of haploids. Denote by $m$ the number of individuals completely sequenced, by $n$ the number of individuals pooled, and by $o$ the overlapping between the two groups of individuals, i.e. the individuals in the pool that have also been sequenced separately. $r$ is the read depth of the pool. The distribution of $j$ can be obtained by conditioning on the number $l$ of individuals that are actually sequenced among the $n-o$ exclusive to the pool:

$$P_c(j|r) = \sum_{l=0}^r \binom{r}{l} \left( 1 - \frac{o}{n} \right)^l \left( \frac{o}{n} \right)^{r-l} P^*(j-m|n-o,l)$$

(19)

See also Section 4 for a simpler formula.

### 3.2.4. Inbred lines

Consider a population from which $n$ inbred lines are derived. For each line, a diploid individual is sequenced. The degree of inbreeding is measured by the inbreeding coefficient $F = (H - H_{inbred})/H$, that is the relative decrease in heterozygosity due to inbreeding, and is assumed to be known. $F$ is also equal to the probability of identity by descent for the inbred line. Our aim is to estimate the heterozygosity of the initial population from the sequences of the inbred lines.

If complete sequences are available,

$$P_c(j|c = 2n) = \binom{n}{2n-j} F^{2n-j} (1 - F)^j (2n-j)$$

(20)
If instead we have NGS reads, the distribution of sequenced chromosomes is

\[ P_c(j|c = 2n, \{r_i\}_{i=1}^{n}) = \sum_{i_1=0}^{2} \cdots \sum_{i_n=0}^{2} I \left( j = \sum_{p=1}^{n} i_p \right) \cdot \prod_{q=1}^{n} [I(r_q = 0)\delta_{i_q,0} + I(r_q > 0) \left( F + (1 - F)2^{-r_q+1} + (i_q - 1) \left( 2(1 - F)(1 - 2^{-r_q+1}) - 1 \right) \right)] \]

Note that in this case, the formula (6) can be inverted to give the expected variability \( E(S) \) for inbred lines with a given inbreeding coefficient \( F \).

4. An equivalent form for Watterson estimators

4.1. Equivalence between the estimator of Jiang et al. and the Watterson estimator for pools

An unbiased estimator of \( \theta \) based on \( S \) was proposed in [6] for NGS data of multiple diploid individuals, even if this is not appropriate, as we will see immediately. The estimator is

\[ \hat{\theta}_J = \frac{S}{\sum_{r=2}^{\infty} L_r \sum_{k=1}^{c-1} \frac{1}{k} \left( 1 - \left( \frac{k}{c} \right)^r - \left( 1 - \frac{k}{c} \right)^r \right)} \]  (22)

where \( c \) is twice the sample size and \( r \) is the total read depth. This estimator is unbiased since the mean of \( S \) is given by the probability \( \theta/k \) of a SNP of frequency \( k \) in the sample multiplied by the probability of detecting it in a random extraction of \( r \) alleles, that is \( 1 - \left( \frac{k}{c} \right)^r - \left( 1 - \frac{k}{c} \right)^r \).

A first observation is that this estimator is not actually unbiased for reads coming from multiple individuals sequenced separately. In fact, it takes into account only the total number of reads, while an unbiased estimator would depend on how they are distributed among individuals. However, it is an unbiased estimator of \( \theta \) for pooled sequences, since in that case information about the origin of the reads is lost.

Furthermore, there is only a single unbiased estimator proportional to \( S \), since the proportionality constant is fixed by the bias of \( S \). This means that the estimator \( \hat{\theta}_J \) is actually the Watterson estimator \( \hat{\theta}_W \) for pools proposed in [8]. The two different forms derive from different intermediate conditioning for \( Z \): on the allele frequency \( k \) in the sample in the first case, on the number of lineages actually sequenced \( j \) in the second.
Note that in the light of this equivalence, the conclusions of [6] about the differences between their estimator and Hellmann’s one when applied to individual data are at least doubtful. They found both estimators to be biased and the variance of Hellmann’s one to be significantly larger, but the numerical simulations performed in [8] showed almost no bias, no sensible difference in variance and a very good correlation between them.

From the mathematical point of view, the equality between \( \hat{\theta}_J \) and \( \hat{\theta}_W \) for pools and the related equalities that we will present in the next section depend on a family of combinatorial identities. We discuss them in Appendix A.

4.2. General alternative form for the Watterson estimators

The above form of [6] for the Watterson estimator for pools can be generalized to the whole family of estimators for units of independent lineages, described by equations (9) and (11). We follow the same notation as before, but we denote the total number of lineages by \( c = \sum_{i=1}^{U} c_i \). The general form for these estimators is

\[
\hat{\theta}_W = \frac{S}{\sum_{\{r\}} L_{\{r\}} \sum_{k=1}^{c-1} \frac{1}{k} \Pi_k(\{c\}, \{r\})}
\]

\[
\Pi_k(\{c\}, \{r\}) = \sum_{k_1=0}^{c_1} \ldots \sum_{k_U=0}^{c_U} I \left( k = \sum_{i=1}^{U} k_i \right) \prod_{i=1}^{U} \left( \binom{c_i}{k_i} \right) \left[ 1 - \prod_{i=1}^{U} \left( \left( \frac{k_i}{c_i} \right)^{r_i} + \left( 1 - \frac{k_i}{c_i} \right)^{r_i} \right) \right]
\]

where the multi-hypergeometric distribution \( \prod_{i=1}^{U} (\binom{c_i}{k_i}) / (\binom{c}{k}) \) describes how the alleles are assigned to the different units and the term \( \prod_{i=1}^{U} \left( \left( \frac{k_i}{c_i} \right)^{r_i} + \left( 1 - \frac{k_i}{c_i} \right)^{r_i} \right) \) is the probability of extracting just one of the two alleles. All the estimators of section 3.1 can be rewritten in this form. This form is often more convenient computationally than the combinatorics in equations (9), (11).

We can also find an estimator similar to (22) for a combination of a pool of \( n \) (haploid) individuals and \( m \) complete sequences, \( o \) of which are overlapping. In this case simple combinatorial reasoning on the probability of detecting a SNP of frequency \( k \) among the \( n + m - o \) individuals leads to

\[
\hat{\theta}_W = \frac{S}{\sum_{r=2}^{\infty} L_r \sum_{k=1}^{n+m-o-1} \left( 1 - \frac{\binom{n-o}{k}}{\binom{n+m-o}{k}} \right) \left( 1 - \frac{k}{n} \right)^r - \frac{\binom{n-o}{k}}{\binom{n+m-o}{k}} \left( 1 - \frac{k-m-o}{n} \right)^r \frac{1}{k}}
\]

that is equivalent to the case (19) of estimator (9).
5. Watterson estimators for autopolyploids

A particularly interesting and challenging set of data is represented by polyploid genomes. Species with ploidy greater than 2 are highly interesting from an evolutionary point of view, as well as economically in agrobiotech and breeding since it involves many commercial species of plants (e.g. potato, sugar cane) and fishes (e.g. Salmonidae).

Polyploid species are difficult both to sequence and to analyze, due to the complex homology/paralogy relation between the constituent genomes. However, some polyploids can be treated by the methods developed here. In particular, autotetraploid populations follow the standard coalescent as shown by [17], and this can be extended to autopolyploids that have similar transition probability matrices. Here we present estimators of variability for populations of autopolyploid species.

Polyploids can be considered as pools with number of lineages equal to their ploidy. Multiple polyploids can then be considered as combinations of pools, but they can be pooled themselves. We consider a species with ploidy \( p \). The estimator for autopolyploids is given by

\[
\hat{\theta}_W = \frac{S}{\sum_{r=2}^{\infty} L_r \sum_{k=1}^{p-1} \frac{1}{k} \left( 1 - \left( \frac{k}{p} \right)^r - \left( 1 - \frac{k}{p} \right)^r \right)}
\]

(25)

for a single polyploid individual, where \( r \) is the read depth, and by the formula

\[
\hat{\theta}_W = \frac{S}{\sum_{\{r\}} L_{\{r\}} \sum_{k=1}^{np-1} \frac{1}{k} \Pi_k(n, p, \{r\})}
\]

(26)

\[
\Pi_k(n, p, \{r\}) = \sum_{k_1=1}^{p} \ldots \sum_{k_n=1}^{p} I \left( k = \sum_{i=1}^{n} k_i \right) \frac{\Pi_{i=1}^{n} \left( \frac{k_i}{np} \right)}{\Pi_{i=1}^{n} \left( \frac{p}{p} \right)} \left[ 1 - \prod_{i=1}^{n} \left( \left( \frac{k_i}{p} \right)^{r_i} + \left( 1 - \frac{k_i}{p} \right)^{r_i} \right) \right]
\]

for sequences from \( n \) polyploid individuals, where \( \{r\} = \{r_i\}_{i=1..n} \) are the read depths per individual. This is also equivalent to the formula (9) for \( \hat{\theta}_W \) with

\[
P_c(j|\{r_i\}_{i=1..n}) = \sum_{i_1=0}^{p} \ldots \sum_{i_n=0}^{p} I \left( j = \sum_{l=1}^{n} i_l \right) \prod_{q=1}^{n} P^r(i_q|p, r_q)
\]

(27)

The estimator for a pool of polyploid individuals is the same as in the general case for pools with \( c = np \), where \( n \) is the number of individuals in
the pooled sample:

\[
\hat{\theta}_W = \frac{S}{\sum_{r=2}^{\infty} L_r \sum_{k=1}^{n_p-1} \frac{1}{k} \left(1 - \left(\frac{k}{np}\right)^r - \left(1 - \frac{k}{np}\right)^r\right)}
\]  

(28)

6. Discussion

In this paper we presented a large family of generalized Watterson estimators that are suited for different types of NGS data, from haploids to polyploids, pools and trios, or a mix of NGS/Sanger data. These estimators are built on the Maximum Composite Likelihood approach; furthermore they are unbiased and depend linearly on \( S \), which is a sufficient statistics for small \( \theta \). The general theory presented here includes all these estimators and many others. Existing estimators are assigned to the proper place in this unified framework.

We pay special attention to estimators for single and multiple autopolyploid individuals. Sequencing of these species has proved to be hard, but more and more projects will soon be devoted to some of the more interesting polyploid species from a commercial point of view, especially among domesticated plants [18, 19, 20]. Autopolyploids without a strong inbreeding follow the dynamics of the usual coalescent, so our theory is applicable to these species. On the other hand, allopolyploid cannot be studied by the same technique since the differences between homologous chromosomes from different constituent species are much stronger and the divergence time between them is often of order of the divergence between species. Specific methods have to be developed for the analysis of variability in allopolyploids [21, 18]. A simple approach could be the study of the variability of each constituent genome and, independently, the genetic differentiation between them.

We did not discuss an important issue with NGS data, that is, base errors. Sequencing errors and misalignments occur with an high rate in NGS data. The bases with lower quality can be removed from the reads or the sequences, however sequencing errors or similar effects can often generate false SNPs at low frequency and it could be difficult to distinguish them from true low frequency alleles. In this case, filtering or SNP calling is usually applied to the data, resulting in an unknown \( Z_{\varphi,\xi} \) for these alleles. Denote by \( (\Phi, \Xi)_\epsilon \) the set of features \( \{\varphi, \xi\} \) strongly affected by sequencing errors. If it is not possible to estimate the contribution of the errors, a good practice is
to discard the corresponding $S_{\varphi, \xi}$, \{\varphi, \xi\} $\in (\Phi, \Xi)_\epsilon$ and to work with the approximate MCL estimator for this case, that is
\[
\hat{\theta}_W = \frac{\sum_{\{\varphi, \xi\} \notin (\Phi, \Xi),} S_{\varphi, \xi}}{\sum_{\{\varphi, \xi\} \notin (\Phi, \Xi),} L_{\varphi} Z_{\varphi, \xi}}
\] as proposed in [22] for sequence data and [7], [8] for pooled reads. The only alternative is to estimate $Z_{\varphi, \xi}$ by heuristic methods.

The estimator proposed here assume the standard Wright-Fisher neutral model for the allele frequency spectrum. However, an arbitrary expected frequency spectrum $E(\xi_k|n) = \theta L \bar{\xi}_{k,n}$ could be used in the place of the neutral spectrum $\theta L/k$. It is sufficient to replace $a_j$ by $\sum_{i=1}^{j-1} \xi_{i,j}$ in the denominator of equation (9) or to replace $1/k$ by $\bar{\xi}_{k,c}$ in the denominator of equation (23). This extends previous adaptations of the original Watterson estimator to null scenarios with demography or varying population size (e.g. [23], [24]).

In this paper we used the composite likelihood approximation to derive the estimators of variability. However, the variance of these Watterson estimators depends on recombination. The usual formulae for ML work only for unlinked sites. In this case, in the limit $\theta \to 0$ and $\theta L$ constant, the variance of $S_{\varphi, \xi}$ is Poisson, i.e. $\text{Var}(S_{\varphi, \xi}) = E(S_{\varphi, \xi})$ and therefore $\text{Var}(\hat{\theta}_W) = \theta/\sum_{\varphi \in \Phi} L_{\varphi} Z_{\varphi}$. In the same limit, the variance for linked sites contain a term $\theta^2 L^2$ coming from the covariances between sites [25, 3, 8]. An exact formula for this term of the variance is available only for a few cases: complete sequences, sequences with missing data [3] and pooled NGS reads [8]. In this case, that is, completely linked sites and known variance, these estimators could be improved [26], also by shrinkage methods [27].

Finally, the theoretical framework developed in this paper allowed to obtain an interesting set of combinatorial identities. This is another example of the way research on theoretical population genetics is highly connected to some fields of mathematics, e.g. combinatorial optimization [28] and could lead to further mathematical insights.

Acknowledgments

We thank A. Fonseca Amaral, M. Pérez-Enciso, W. Burgos and B. Nevado for useful discussions and M. Mamino for providing an explicit proof of the main combinatorial identity. LF acknowledges support from ANR-12-JSV7-0007 (ANR, France). The project was funded by Grants CGL2009-09346
Appendix A. Combinatorial identities

The identity of the two estimators \( \hat{\theta}_J \) and \( \hat{\theta}_W \) for pools implies identity of their denominators. The reasoning in section 4 is equivalent to a double counting proof of the combinatorial identity

\[
\sum_{j=1}^{\min(c,r)} \frac{c!}{(c-j)!} S(r, j) a_j = \sum_{k=1}^{c-1} \frac{c^r - k^r - (c-k)^r}{k}
\]  

(A.1)

valid for all pairs of integers \((c,r)\) such that \(c \geq 1\) and \(r \geq 1\). The identity involves Stirling numbers \( S(r,j) \) and harmonic numbers \( a_j \) in a nontrivial way. Note that both sides of the identity are integers. This identity can also be proved directly.

Appendix A.1. More combinatorial identities

We can extend the previous identity (A.1) to a set of identities derived from the same equivalence of estimators but for an arbitrary frequency spectrum. The fundamental identities are obtained by double counting technique, considering a single frequency in the population, computing the Taylor expansion of both sides and equating the coefficients of the \( l \)th power:

\[
\sum_{j=2}^{\min(c,r)} \frac{c!}{(c-j)!} S(r, j) \sum_{k=1}^{j-1} (-1)^k \binom{j}{k, l-k, j-l} = \sum_{k=1}^{c-1} (-1)^k \binom{c}{k, l-k, c-l} (c^r - k^r - (c-k)^r) \quad (A.2)
\]

for integers \((r,c,l)\) with \(r \geq 1\) and \(1 \leq l \leq c\). They involve Stirling numbers and multinomials. Any other identity in this family can be obtained as a linear combinations of these ones. Note that since the l.h.s. is 0 for \(l > r\), these identities reduce to

\[
\sum_{k=1}^{c} (-1)^k \binom{c}{k, l-k, c-l} (c^r - k^r - (c-k)^r) = 0 \quad (A.3)
\]

for \(r < l \leq c\).
References

[1] G. Watterson, On the number of segregating sites in genetical models without recombination., Theoretical Population Biology 7 (2) (1975) 256.

[2] A. RoyChoudhury, J. Wakeley, Sufficiency of the number of segregating sites in the limit under finite-sites mutation, Theoretical Population Biology 78 (2) (2010) 118–122.

[3] L. Ferretti, E. Raineri, S. Ramos-Onsins, Neutrality tests for sequences with missing data, Genetics 191 (4) (2012) 1397–1401.

[4] M. Lynch, Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects, Molecular Biology and Evolution 25 (11) (2008) 2409.

[5] I. Hellmann, Y. Mang, Z. Gu, P. Li, M. Francisco, A. Clark, R. Nielsen, Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals, Genome Research 18 (7) (2008) 1020–1029.

[6] R. Jiang, S. Tavaré, P. Marjoram, Population genetic inference from resequencing data, Genetics 181 (1) (2009) 187.

[7] A. Futschik, C. Schlotterer, The next generation of molecular markers from massively parallel sequencing of pooled dna samples., Genetics 186 (1) (2010) 207–218. doi:10.1534/genetics.110.114397

[8] L. Ferretti, S. Ramos-Onsins, M. Perez-Enciso, Population genomics from pool sequencing, Molecular Ecology, accepted.

[9] B. Devlin, N. Risch, K. Roeder, Disequilibrium mapping: composite likelihood for pairwise disequilibrium., Genomics 36 (1) (1996) 1–16. doi:10.1006/geno.1996.0419

[10] B. S. Weir, Inferences about linkage disequilibrium., Biometrics 35 (1) (1979) 235–254.

[11] R. Hudson, Two-locus sampling distributions and their application, Genetics 159 (4) (2001) 1805.
[12] G. McVean, P. Awadalla, P. Fearnhead, A coalescent-based method for detecting and estimating recombination from gene sequences., Genetics 160 (3) (2002) 1231–1241.

[13] Y. Kim, W. Stephan, Detecting a local signature of genetic hitchhiking along a recombining chromosome., Genetics 160 (2) (2002) 765–777.

[14] L. Zhu, C. D. Bustamante, A composite-likelihood approach for detecting directional selection from DNA sequence data., Genetics 170 (3) (2005) 1411–1421. doi:10.1534/genetics.104.035097.

[15] D. Garrigan, Composite likelihood estimation of demographic parameters., BMC Genet 10 (2009) 72. doi:10.1186/1471-2156-10-72.

[16] M. Pérez-Enciso, L. Ferretti, Massive parallel sequencing in animal genetics: wherefroms and wheretos, Animal Genetics 41 (6) (2010) 561–569.

[17] B. Arnold, K. Bomblies, J. Wakeley, Extending coalescent theory to autotetraploids., Genetics 192 (1) (2012) 195–204. doi:10.1534/genetics.112.140582.

[18] R. Brenchley, M. Spannagl, M. Pfeifer, G. L. A. Barker, R. D’Amore, A. M. Allen, N. McKenzie, M. Kramer, A. Kerhornou, D. Bolser, S. Kay, D. Waite, M. Trick, I. Bancroft, Y. Gu, N. Huo, M.-C. Luo, S. Sehgal, B. Gill, S. Kianian, O. Anderson, P. Kersey, J. Dvorak, W. R. McCombie, A. Hall, K. F. X. Mayer, K. J. Edwards, M. W. Bevan, N. Hall, Analysis of the bread wheat genome using whole-genome shotgun sequencing., Nature 491 (7426) (2012) 705–710. doi:10.1038/nature11650.

[19] Y. Han, Y. Kang, I. Torres-Jerez, F. Cheung, C. D. Town, P. X. Zhao, M. K. Udvardi, M. J. Moneros, Genome-wide SNP discovery in tetraploid alfalfa using 454 sequencing and high resolution melting analysis., BMC Genomics 12 (2011) 1–11. doi:10.1186/1471-2164-12-350.

[20] X. Wang, H. Wang, J. Wang, R. Sun, J. Wu, S. Liu, Y. Bai, J.-H. Mum, I. Bancroft, F. Cheng, S. Huang, X. Li, W. Hua, J. Wang, X. Wang, M. Freeling, J. C. Pires, A. H. Paterson, B. Chalhoub, B. Wang, A. Hayward, A. G. Sharpe, B.-S. Park, B. Weisshaar, B. Liu, B. Li, B. Liu, C. Tong, C. Song, C. Duran, C. Peng, C. Geng, C. Koh, C. Lin, D. Edwards, D. Mu, D. Shen, E. Soumpourou, F. Li, F. Fraser, G. Conant,
[21] J. T. Page, M. D. Huynh, Z. S. Liechty, K. Grupp, D. M. Stelly, A. M. Hulse, H. Ashrafi, A. Van Deynze, J. F. Wendel, J. A. Udall, Insights into the evolution of cotton diploids and polyploids from whole-genome re-sequencing., G3 (Bethesda) doi:10.1534/g3.113.007229.

[22] G. Achaz, Testing for neutrality in samples with sequencing errors, Genetics 179 (3) (2008) 1409.

[23] M. Rafajlović, A. Klassmann, A. Eriksson, T. Wiehe, B. Mehlig, Demography-adjusted tests of neutrality based on genome-wide SNP data, ArXiv e-prints arXiv:1307.0337.

[24] L. Ferretti, M. Perez-Enciso, S. Ramos-Onsins, Optimal neutrality tests based on the frequency spectrum, Genetics 186 (1) (2010) 353.

[25] Y.-X. Fu, Statistical properties of segregating sites, Theoretical Population Biology 48 (2) (1995) 172–197.

[26] Y.-X. Fu, Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of dna sequences., Genetics 138 (4) (1994) 1375–1386.

[27] A. Futschik, F. Gach, On the inadmissibility of Watterson’s estimator, Theoretical Population Biology 73 (2) (2008) 212–221.

[28] R. Arratia, A. D. Barbour, S. Tavaré, Logarithmic Combinatorial Structures: A Probabilistic Approach, European Mathematical Society, 2003.