A generalized Watterson estimator for next-generation sequencing: From trios to autopolyploids

Luca Ferretti, Sebastian E. Ramos-Onsins

A generalized Watterson estimator for next-generation sequencing: from trios to autopolyploids

Luca Ferretti[a,b,*], Sebástian E. Ramos-Onsins[c]

[a] *Systématique, Adaptation et Evolution (UMR 7138), UPMC Univ Paris 06, CNRS, MNHN, IRD, Paris, France*
[b] *CIRB, Collège de France, Paris, France*
[c] *Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Edifici CRAG, Campus Universitat Autònoma. Bellaterra 08193, Spain*

Abstract

Several variations 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: Site frequency spectrum, Population genetics, Summary statistics, Maximum likelihood, Composite likelihood

1. Introduction

The rescaled mutation rate per base θ plays an important role in population genetics models. Its definition is $\theta = 2pN_e\mu$ where μ denotes the mutation rate per base, N_e the effective population size and p the ploidy of the species. In the context of the Standard Neutral Model (SNM) at low

*Corresponding author
Email address: luca.ferretti@gmail.com (Luca Ferretti)

mutation rate, commonly used estimator include the Watterson estimator [1], based on the number of segregating sites S in a sample from the population, and the pairwise nucleotide diversity Π [2], defined as the average number of differences per base between two individual sequences. Despite its simple interpretation and its popularity, Π is an inconsistent estimator of θ, while the Watterson estimator is a good estimator since it is unbiased and it corresponds to the Maximum Composite Likelihood estimator for θ. Furthermore, as shown in [3], it is a sufficient statistics for θ for large sequences. For unequal mutation rates between different alleles, the Watterson estimator actually measures the net rate of mutation towards different alleles, rescaled by population size.

Today, variability analyses on a genome-wide scale can be easily done by Next Generation Sequencing (NGS) technologies. NGS technologies can sequence a single complete genome at relatively high redundancy, but the sequencing of many samples increases substantially the cost of the experiment. Several strategies have been used to obtain sequence data from many individuals, from restriction reduced libraries to pooled samples. These strategies may reduce substantially the coverage across the genome (i.e. how many regions are sequenced) and the sequence redundancy per nucleotide base (i.e. how many sequences cover each position). From this point of view, 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 θ estimator has been generalized to missing data by Ferretti et al. [4], but it covers just a limited number of cases.

Estimators of variability based on NGS data should take into account the relatively high probability of sequencing errors in order not to overestimate the number of variants. Sequence redundancy at a given position is fundamental for detecting and removing errors but also for detecting correctly homozygote or heterozygote positions in diploid individuals. NGS estimators of variability were first proposed by Lynch [5] for a single diploid individual using the sequence error rate and the number of reads observed at each variant. Hellmann et al. [6] and Jiang et al. [7] proposed Watterson estimators for multiple individuals taking into account the combinatorics of reads from different individuals and homologous chromosomes. Later, Futschik and Schlötterer [8] and Ferretti et al. [9] developed variability estimators for pooled sample data using Method of Moments (MM) and Maximum Composite Likelihood (MCL) methods, respectively. We observe that for the analysis of a single diploid individual, they all reduce to the same estima-

tor except for [8] (the differences between this estimator and the MCL have been detailed in [9]). In this work, we study the relations between these estimators and present a common framework for estimators of variability. We develop new Watterson estimators for other kinds of NGS data, like trios or autopolyploids.

In our framework, we deal with all these cases as special instances of a general, 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. In section 2, we derive Maximum Composite Likelihood estimators for a generic class of data. These MCL estimators are not unbiased, but we show that they can be well approximated by unbiased estimators that share the same functional form as the Watterson estimator. In section 3, we propose these 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 them. We present explicit formulae for these estimators in section 4. We discuss their relation with other estimators that have been proposed in the literature, showing in section 5 that many of our estimators admit two equivalent but seemingly different forms. As a byproduct, this equivalence implies a set of combinatorial identities. Finally, in section 6 we treat in details the case of single and 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 has been extensively used in population and quantitative genetics for estimating the linkage disequilibrium among positions [10, 11] and for estimating evolutionary parameters as the level of variation [6], the recombination rate [12], [13], the strength of positive selection [14, 15] or demographic parameters [16]. Maximum Composite Likelihood is an appropriate method to estimate the nucleotide variability across large regions because it has minimum mean squared error for large recombining regions, since in this case the Composite Likelihood is a good approximation for the exact likelihood and therefore the estimator is approximately asymptotically efficient. Furthermore, 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.

In this paper, we consider allelic variants represented by segregating sites, or Single Nucleotide Polymorphisms (SNPs), but the methods can be applied to generic variants with low mutation rate.

For each site, there are features that depend on an eventual SNP (for example, allele frequencies) and features that do not depend on the allelic content (for example the read depth, i.e. the number of sequences that contain data for a given site). We summarize the SNP features with the index $\xi \in \Xi$ and the features of each site that do not depend on the allelic content with the index $\varphi \in \Phi$. Both indices indicate mutually exclusive, collectively exhaustive features.

We denote by $p_{\varphi,\xi}(\theta)$ the probability that a site with features φ contains an observed SNP with features ξ for the sample studied. For small values of θ (that is, in the infinite site model), we can expand it in Taylor series and using the fact that there are no SNPs without mutations, i.e. $p_{\varphi,\xi}(0) = 0$, we find that these probabilities are proportional to θ multiplied by a quantity that depends on the population model and the sequencing setup:

$$p_{\varphi,\xi}(\theta) \simeq \theta Z_{\varphi,\xi} \ . \tag{1}$$

We denote by $S_{\varphi,\xi}$ the number of segregating sites with features ξ in positions with features φ, and by L_φ the number of sites with features φ. 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 the simple product of probabilities:

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

Substituting eq. (1) for $p_{\varphi,\xi}(\theta)$ and taking the log, we obtain

$$\log(CL(\theta)) = S\log(\theta) + \sum_{\varphi\in\Phi}\sum_{\xi\in\Xi} S_{\varphi,\xi}\log(Z_{\varphi,\xi}) + \sum_{\varphi\in\Phi}(L_\varphi - S_\varphi)\log(1-\theta Z_\varphi) \ . \tag{3}$$

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 $\log(CL)$ in equation (3). After rearrangements, we obtain the equation that defines the Maximum Composite Likelihood Estimator (MCLE):

$$L = \sum_{\varphi \in \Phi} \frac{L_\varphi - S_\varphi}{1 - \hat{\theta}_{MCLE} Z_\varphi} \ , \tag{4}$$

which is valid for all values of $\theta < 1/\max_{\varphi\in\Phi}(Z_\varphi)$ since the second derivative in $\theta = \hat{\theta}_{MCLE}$ is negative if $\sum_{\varphi\in\Phi}(L_\varphi - S_\varphi)Z_\varphi/(1-\hat{\theta}_{MCLE}Z_\varphi)^2 > 0$, which is always verified.

For the simplest case of the original Watterson estimator [1], all sites are equivalent and there is no site feature φ, so the above MCLE equation (4) can be easily rewritten as $L = (L-S)/(1-\hat{\theta}_{MCLE}Z)$, i.e. $\hat{\theta}_{MCLE} = \hat{\theta}_W = S/(LZ)$. The SNP features Ξ correspond simply to the derived allele counts $i = 1 \dots n-1$ and the probability of a SNP of frequency i/n is related to the expected frequency spectrum ξ_i - defined as the count of SNPs of frequency i/n in the sample - by $p_i(\theta) = \mathrm{E}(\xi_i)/L$. 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 $\hat{\theta}_W = S/(L\sum_{i=1}^{n-1} 1/i) = S/(La_n)$.

The estimator (4) is defined implicitly, so it is not easy to use. An explicit, approximate MCL estimator can be derived in two equivalent ways: either (i) by expanding equation (4) at first order in the small parameters θ and S_φ/L_φ with constant ratio $S_\varphi/(\theta L_\varphi)$, or (ii) by taking the small θ, large L limit of the likelihood (2) with θL and L_φ/L constant; in this limit, the $S_{\varphi,\xi}$ are Poisson distributed random variables with mean $L_\varphi \theta Z_{\varphi,\xi}$

$$CL(\theta) \simeq \prod_{\varphi\in\Phi}\prod_{\xi\in\Xi} \frac{(L_\varphi\theta Z_{\varphi,\xi})^{S_{\varphi,\xi}}}{S_{\varphi,\xi}!} e^{-L_\varphi\theta Z_{\varphi,\xi}} = \theta^S e^{-\theta\sum_{\varphi\in\Phi} L_\varphi Z_\varphi} \left[\prod_{\varphi\in\Phi}\prod_{\xi\in\Xi} \frac{(L_\varphi Z_{\varphi,\xi})^{S_{\varphi,\xi}}}{S_{\varphi,\xi}!}\right] \tag{5}$$

and since the dependence on θ lies in the first term which is a function of the statistics S only, S is a sufficient statistics for θ by the Fisher-Neyman factorization theorem, as already observed in [3].

Both ways lead to the same estimator. We define the resulting approximate MCL estimator as the generalized Watterson estimator:

$$\hat{\theta}_W = \frac{S}{\sum_{\varphi\in\Phi} L_\varphi Z_\varphi} \ . \tag{6}$$

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

Both the implicit estimator in equation (4) and the formula (6), which is unbiased, can be used. The relative error of the Watterson estimator (6) with respect to the MCLE (4) is very small. By expanding eq. (4) at first order in $\hat\theta_{MCLE} - \hat\theta_W$ and at first nonzero order in $\hat\theta_W$ and $S_\phi/(L_\phi Z_\phi)$, we obtain an error estimate of order θ multiplied by a weighted covariance between Z and the relative fluctuations of $\hat\theta_W$ for different φs:

$$\frac{\hat\theta_{MCLE} - \hat\theta_W}{\hat\theta_W} \simeq -\hat\theta_W \sum_{\varphi\in\Phi} \frac{L_\varphi Z_\varphi}{\sum_{\varphi'\in\Phi} L_{\varphi'} Z_{\varphi'}} \left(Z_\varphi - \frac{\sum_{\varphi'\in\Phi} L_{\varphi'} Z_{\varphi'}}{L} \right) \left(\frac{S_\varphi/(L_\varphi Z_\varphi) - \hat\theta_W}{\hat\theta_W} \right) \tag{7}$$

up to terms of order $(\hat\theta_W, S_\phi/(L_\phi Z_\phi))^2$. This error is usually negligible, since the r.h.s. is suppressed by a factor of $\theta \ll 1$; furthermore it has mean 0, since it correlates the fluctuations of the mutation process with the fluctuations of the sequencing process, but the two processes are independent.

The only information needed to compute (4) or (6) are the factors $Z_\varphi = \sum_{\xi\in\Xi} p_{\varphi,\xi}/\theta$, 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.

3. Application to NGS and sequence data

3.1. The data: sequences aligned to a reference genome

In this section we deal with combinations of complete sequences, genotypes and NGS data from different sources in a unified way. Our data are represented by reads, sequences or genotypes[1], aligned to a reference genome. Each read/sequence/genotype 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.

[1]Some genotyping methods (like DNA microarrays) preselect the possible SNPs or the positions containing a segregating site. These methods give biased estimates of variability and cannot be meaningfully combined with unbiased methods like Sanger or NGS.

We denote by U the number of units. The features associated to the units are (i) the number of copies of homologous chromosomes present in each unit and (ii) the evolutionary relationships between the units. We denote the number of copies of homologous chromosomes in the ith unit by c_i, $i = 1 \dots U$. 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. We denote the set of numbers $\{c_i\}_{i=1\dots U}$ by $\{c\}$.

The evolutionary relationships (denoted here by the generic symbol χ) include all the available information relevant for 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. For example, two different pools could contain two genetically identical individuals, or two individuals could be parent and offspring, or a single diploid individual could originate from a single inbred line with given inbreeding coefficient, and so on.

We assume that the number of NGS 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 ith unit $r_i(x)$, $i = 1 \dots U$, i.e. the number of reads or sequences from the ith unit that cover position x. For Sanger sequences and genotyping, we define an "effective read depth' $r_i(x) = 0$ for positions with missing data and $r_i(x) = +\infty$ otherwise. We denote the set of read depths $\{r_i\}_{i=1\dots U}$ by $\{r\}$.

An example of the data is given in Table 1.

3.2. Estimators for Next Generation Sequencing

We consider NGS data like the ones described above. We can derive the general Watterson estimator for this case by using the definition (6) with $\varphi = \{r\} = \{r_i\}_{i=1\dots U}$, the set of read depths of the different units at a given site. We use the short form $\{r\} = \{r_i\}_{i=1\dots U}$ and $\{c\} = \{c_i\}_{i=1\dots U}$ for the information about the site features. Z can be computed by conditioning on the number of unrelated homologous chromosomes or lineages which actually contribute to the data, denoted by j, and then averaging over j:

$$\sum_{\xi \in \Xi} p_{\{r\},\xi} = \theta Z_{\{r\}} = \sum_{j=2}^{\infty} P(\mathrm{SNP}|\chi, \{c\}, \{r\}, j) \cdot P_c(j|\chi, \{c\}, \{r\}) \qquad (8)$$

where $P(\mathrm{SNP}|\dots)$ is the probability of observing a SNP among j independent lineages and $P_c(j|\dots)$ is the distribution of j, which depends also on $\{c\}$

Reference		A	C	A	C	G	T	A	A	T	C	G	C
Sanger:	(unit 1)	A	C	A	C	**G**	T	T	A	**A**	C	G	**C**
	(unit 2)	A	C	A	C	**C**	T	-	A	**T**	C	G	**C**
Genotyping:	(unit 3)	$\frac{A}{A}$	$\frac{C}{C}$	-	$\frac{C}{C}$	$\frac{C}{G}$	$\frac{T}{T}$	$\frac{T}{T}$	$\frac{A}{A}$	$\frac{T}{T}$	$\frac{C}{C}$	$\frac{G}{G}$	$\frac{C}{G}$
NGS reads:	(unit 4)			A	C	**G**	T	T	A				
	(unit 4)	A	C	A	C	**C**							
	(unit 4)				C	**C**	T	T	A	**A**			
	(unit 5)							T	A	**A**	C	G	**C**
Read depths:													
unit 1: $c_1 = 1$	$r_1 =$	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
unit 2: $c_2 = 1$	$r_2 =$	∞	∞	∞	∞	∞	∞	0	∞	∞	∞	∞	∞
unit 3: $c_3 = 2$	$r_3 =$	∞	∞	0	∞	∞	∞	∞	∞	∞	∞	∞	∞
unit 4: $c_4 = 2$	$r_4 =$	1	1	2	3	3	2	2	2	1	0	0	0
unit 5: $c_5 = 2$	$r_5 =$	0	0	0	0	0	0	1	1	1	1	1	1

Table 1: Example (not real) of data from five sequencing units from different sources and technologies. The data come from four diploid individuals, of which two are parent and child. The data are aligned to the reference genome (the sequence at the top). There are two complete Sanger sequences (units 1 and 2) with a missing base in the 7th position, one sequence of genotypes (unit 3) with a missing genotype at the 3rd position, and four NGS read (three coming from unit 4 and one from unit 5). The allelic content of the positions with SNPs is shown in bold red. At the bottom, we report the "effective' read depths for all units. All units are diploid individuals (therefore $c = 2$). Units 1 and 2 are the two homologous sequences of the parent (hence $c_1 = c_2 = 1$) and unit 5 comes from the child ($c_5 = 2$), while units 3 and 4 come from unrelated individuals ($c_3 = c_4 = 2$).

and on their relationships χ. The probability of observing a SNP among j independent lineages depends just on j, not on the redundancy of each lineage in the sequences, and on the expected site frequency spectrum $\mathrm{E}(\xi_k|j)$ of the model, being equal to $P(\mathrm{SNP}|j) = \sum_{k=1}^{j-1} \mathrm{E}(\xi_k|j)/L$. In the case of the standard neutral Wright-Fisher model we have $\mathrm{E}(\xi_k|j) = \theta L/k$ and therefore $P(\mathrm{SNP}|j) = \theta \sum_{k=1}^{j-1} 1/k = \theta a_j$.

The general Watterson estimator for the Standard Neutral Model (SNM) is then

$$\hat{\theta}_W = \frac{S}{\sum_{\{r\}} L_{\{r\}} \sum_{j=2}^{\infty} P_c(j|\chi, \{c\}, \{r\}) \cdot a_j} \,. \tag{9}$$

This form was found by [6] and [17, 9] 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. Note that this estimator does not take into account sequencing errors. The treatment of sequencing error will be presented in the Discussion.

Complete sequences and genotyping data can be also analyzed by the above formulae, provided that an effective read depth $r = +\infty$ is considered for positions with data, and $r = 0$ for positions with missing data. In fact, if there are missing data, the information is equivalent to the absence of NGS data ($r = 0$), while presence of data means that the genotype is known with certainty, as it would be with large read depth for NGS data ($r = \infty$).

Finally, the general Watterson estimator for an arbitrary model with spectrum $\mathrm{E}(\xi_k|j) = \theta L \bar{\xi}_{k,j}$ is

$$\hat{\theta}_W = \frac{S}{\sum_{\{r\}} L_{\{r\}} \sum_{j=2}^{\infty} P_c(j|\chi, \{c\}, \{r\}) \sum_{i=1}^{j-1} \bar{\xi}_{i,j}} \,. \tag{10}$$

3.3. A simple example: data from a single diploid individual

In this section we present the case of a single diploid individual [Lynch]. The single individual represents a single unit $U = 1$, and being diploid (assuming unrelated parents) we have two homologous sequences, therefore $c_1 = 2$. Assuming that each read is extracted at random from one of the two homologous sequences, the estimator is specified by the probabilities

$$P_c(j = 1|c_1 = 2, r_1) = 2^{-r_1+1} \tag{11}$$

$$P_c(j = 2|c_1 = 2, r_1) = 1 - 2^{-r_1+1} \tag{12}$$

Reference	A	G	A	C	C	A	T
NGS reads:			**A**	C	C	**T**	T
		G	**C**	C	C	**A**	T
	A	G	**C**	C			
Read depths: $r_1 =$	1	2	3	3	2	2	2

Table 2: Example (not real) of data from a single diploid individual. The reads are aligned to the reference genome (the sequence at the top). Positions with SNPs are shown in bold red.

since the probability of extracting all r_1 reads from a given sequence is 2^{-r_1}. Therefore the estimator from equation (9), taking into account the relevant harmonic factors $a_1 = 0$ and $a_2 = 1$, is

$$\hat{\theta}_W = \frac{S}{\sum_{r_1=2}^{\infty} L_{r_1}(1 - 2^{-r_1+1})} \,. \tag{13}$$

As an example, consider the sequences in Table 2. In the sequence of length $L = 7$ there are $S = 2$ segregating sites. There are bases with read depth 1, 2 and 3 and their numbers are $L_1 = 1$, $L_2 = 4$ and $L_3 = 2$ respectively. The value of the Watterson estimator for this sequence is therefore $\hat{\theta}_W = 2/(1 \cdot 0 + 4 \cdot 1/2 + 2 \cdot 3/4) = 4/7 \simeq 0.57$.

4. 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

$$P^*(j|c, r) = \frac{c!\ S(r, j)}{(c - j)!\ c^r} \tag{14}$$

that corresponds to the probability of extracting exactly j different objects with r extractions (with repetitions) from a set of c objects [9]. In fact, the number of possible extractions from a set of c objects is c^r, while the number of extractions of precisely j objects is given by the product of the number

of ordered choices of j objects out of c, that is $c!/(c-j)!$, multiplied by the number of ways to distribute the j objects across r extractions $S(r,j)$. Since all extractions are equiprobable, the ratio gives the the probability (14). This equation 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 $\delta_{i,j} = I(i=j)$. Note that $P^*(j|c,0) = \delta_{j,0}$.

4.1. 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 ith unit, and r_i is the number of reads/sequences coming from the ith unit, the probability $P_c(j|\{c\},\{r\})$ in the Watterson estimator is

$$P_c(j|\{c\},\{r\}) = \sum_{i_1=0}^{c_1} \dots \sum_{i_U=0}^{c_U} I\left(j = \sum_{p=1}^{U} i_p\right) \prod_{q=1}^{U} P^*(i_q|c_q,r_q) \qquad (15)$$

which is a product of probabilities for each unit of sequencing $i_1 \dots i_U$ chromosomes respectively, summed over all combinations resulting in j independent chromosomes.

In Section 5 we will present an alternative form for these Watterson estimators. In the rest of this section we specialize the expression (15) to the most common scenarios.

4.1.1. Multiple haploid individuals

In this case all individuals have $c_i = 1$, therefore $P^*(i_q|c_q=1,r_q) = I(i_q = I(r_q>0))$ and the estimator reduces to the one proposed for missing data in [4], which is equivalent to

$$P_c(j|c=n,\{r\}) = I\left(j = \sum_{i=1}^{n} I(r_i>0)\right) . \qquad (16)$$

This choice was implicitly suggested also in [3].

4.1.2. Multiple diploid individuals

In this case all individuals have $c_i = 2$ and the MCLE was already derived by Hellmann et al. [6]:

$$P_c(j|c = 2n, \{r\}) = \sum_{i_1=0}^{2} \dots \sum_{i_n=0}^{2} I\left(j = \sum_{p=1}^{n} i_p\right) \prod_{q=1}^{n} P^*(i_q|c_q = 2, r_q) \ . \quad (17)$$

4.1.3. Pools

In this case, there is a single unit of c chromosomes. The probability was derived by Ferretti et al. [9]

$$P_c(j|c, r) = P^*(j|c, r) \quad (18)$$

but see also Section 5 for a simpler formula.

For multiple pools, the probability follows closely equation (15) where $c_1 \dots c_U$ are the numbers of (haploid) individuals inside each pool and $r_1 \dots r_U$ are the read depths of the pools at the position considered.

4.2. Related lineages

Sequencing 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. In this section we develop estimators for these cases.

4.2.1. Trios

A trio is a (diploid) family of mother, father and child that are sequenced separately. We assume that father and mother are two unrelated individuals from the same population. We restrict our analysis to autosomes, where the two alleles of the child are the copies of one paternal and one 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_C the read depths of mother, father and child respectively. We obtain $P_c(j|r_M, r_F, r_C)$ by conditioning on

the number of lineages j' sequenced in the parents. We can rewrite it in terms of the probability $P_c(j'|c=4, r_M, r_F)$ of the parents alone (eq.(17)):

$$P_c(j|r_M, r_F, r_S) = \sum_{j'=0}^{4} p_t(j|j', r_M, r_F, r_C) P_c(j'|c=4, r_M, r_F) \tag{19}$$

where $p_t(j|j', r_M, r_F, r_C)$ is the probability of sequencing j independent lineages in the trio given the number of independent lineages j' sequenced from the parents. $p_t(j|j', r_M, r_F, r_C)$ is obtained case by case depending on the probability that the sequences of the child could contain new alleles with respect to the parental sequences and the probability to detect them. For example, consider the case $j'=3$. Then there is only a single allele in the parents that has not been sequenced. This allele is absent in the child with 50% probability (in this case $j=3$ because no new alleles are sequenced in the child) or it could be present with 50% probability, but not sequenced (then $j=3$ with probability 2^{-r_c}) or could be sequenced (then $j=4$ with probability $1-2^{-r_c}$).

The complete probability is

$$\begin{aligned}
p_t(j|j'=0, r_M, r_F, r_C) =& P^*(j|2, r_S) \qquad (20)\\
p_t(j|j'=1, r_M, r_F, r_C) =& \frac{1}{2} P^*(j-1|2, r_S) + \frac{1}{2}(\delta_{j,1} 2^{-r_S} + \delta_{j,2}(1-2^{-r_S}))\\
p_t(j|j'=2, r_M, r_F, r_C) =& \frac{1}{2}(1 + I(r_M r_F = 0))(\delta_{j,2} 2^{-r_S} + \delta_{j,3}(1-2^{-r_S})) +\\
&+ \frac{1}{4} I(r_M r_F > 0)\,(\delta_{j,2} + P^*(j-2|2, r_S))\\
p_t(j|j'=3, r_M, r_F, r_C) =& \frac{1}{2}\delta_{j,3} + \frac{1}{2}(\delta_{j,3} 2^{-r_S} + \delta_{j,4}(1-2^{-r_S}))\\
p_t(j|j'=4, r_M, r_F, r_C) =& \delta_{j,4}\ .
\end{aligned}$$

Multiple unrelated trios can be dealt with by replacing the probability $P^*(i_q|\ldots)$ in equation (15) with the probability (19) and replacing c_q with 4.

4.2.2. Pooled trios

A pooled trio is a family of mother, father and child that are pooled together and sequenced. We consider n unrelated families, each family sequenced separately from the others, and denote by r_i the total read depths.

$P_c(j|\{r_i\}_{i=1\dots n})$ is given by

$$P_c(j|\{r_i\}_{i=1\dots n}) = \sum_{i_1=0}^{4} \dots \sum_{i_n=0}^{4} I\left(j = \sum_{p=1}^{n} i_p\right) \prod_{q=1}^{n} P_{pt}(i_q|r_q) \tag{21}$$

where $P_{pt}(i|r)$ is the probability of sequencing i homolog chromosomes for a single pooled trio. It can be derived case-by-case conditioning on the number i' of sequenced chromosomes (identical or not) for a pool, obtaining

$$P_{pt}(i|r) = \sum_{i'} p_{pt}(i|i') P^*(i'|c = 6, r) \ . \tag{22}$$

$p_{pt}(i|i')$ can be found by conditioning on the number of non-inherited chromosomes sequenced, obtaining

$$p_{pt}(i|i') = \delta_{i,i'} \frac{4\binom{2}{i'-2} + 4\binom{2}{i'-1} + \binom{2}{i'}}{\binom{6}{i'}} + \delta_{i,i'-1} \frac{4\binom{2}{i'-3} + 2\binom{2}{i'-2}}{\binom{6}{i'}} + \delta_{i,i'-2} \frac{\binom{2}{i'-4}}{\binom{6}{i'}} \tag{23}$$

and finally

$$\begin{aligned} P_{pt}(i|r) = &\left[4\binom{2}{i-2} + 4\binom{2}{i-1} + \binom{2}{i}\right] \frac{P^*(i|6,r)}{\binom{6}{i}} + \\ &+ \left[4\binom{2}{i-2} + 2\binom{2}{i-1}\right] \frac{P^*(i+1|6,r)}{\binom{6}{i+1}} + \binom{2}{i-2} \frac{P^*(i+2|6,r)}{\binom{6}{i+2}} \ . \end{aligned} \tag{24}$$

4.2.3. 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 this situation, there could be individuals in the pool for which the complete sequence is also available.

Here we deal with the haploid case, but the results 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. Denote by r the read depth of the pool. The distribution of j can be obtained by conditioning on the number l of pooled reads that come

actually from the $n-o$ individuals exclusive to the pool. The distribution of l is a binomial with probability o/n and r extractions, therefore

$$P_c(j|r) = \sum_{l=0}^{r} P^*(j-m|n-o,l)\binom{r}{l}\left(1-\frac{o}{n}\right)^l\left(\frac{o}{n}\right)^{r-l} = \tag{25}$$
$$= \frac{(n-o)!}{n^r(n-o-j+m)!}\sum_{l=0}^{r}\binom{r}{l}o^{r-l}S(l,j-m)\ .$$

See also Section 5 for a simpler formula.

4.2.4. Inbred lines

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

If complete sequences are available, since each individual has an independent probability F of being homozygote because of inbreeding, the distribution of the number of homozygotes is just a binomial. But the number of sequenced chromosomes is $2n$ minus the number of homozygotes, therefore

$$P_c(j|c=2n) = \binom{n}{2n-j}F^{2n-j}(1-F)^{j-n}\ . \tag{26}$$

If instead we have NGS reads, the distribution of sequenced chromosomes should account for the " effective homozygote probability" $F+(1-F)2^{-r_q+1}$ due to sampling:

$$P_c(j|c=2n,\{r\}) = \sum_{i_1=0}^{2}\dots\sum_{i_n=0}^{2} I\left(j=\sum_{p=1}^{n} i_p\right)\cdot\prod_{q=1}^{n}\left[I(r_q=0)\delta_{i_q,0}+\right. \tag{27}$$
$$\left.+I(r_q i_q>0)\left(F+(1-F)2^{-r_q+1}+(i_q-1)\left(2(1-F)(1-2^{-r_q+1})-1\right)\right)\right]\ .$$

Note that in this case, the formula (6) with (26), (27) can be inverted to give the expected variability $\mathrm{E}(S)$ for individuals from inbred lines with a given inbreeding coefficient F.

5. An equivalent form for Watterson estimators

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

An unbiased estimator of θ based on S was proposed in [7] for NGS data of multiple diploid individuals, even if this is not the most appropriate setup, as we will see immediately. The estimator is

$$\hat{\theta}_J = \frac{S}{\sum_{r=2}^{\infty} L_r \sum_{k=1}^{c-1} \frac{1}{k} \left(1 - \left(\frac{k}{c}\right)^r - \left(1 - \frac{k}{c}\right)^r\right)} \tag{28}$$

where c is twice the sample size (for diploids) and r is the total read depth. This estimator is unbiased since the mean of S is given by the probability θ/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 θ 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 [9]. 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 [7] 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 theory suggests that they are unbiased and the variance of Hellmann's one should be lower. In fact, numerical simulations performed in [9] 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. The identity of the two estimators $\hat{\theta}_J$ and $\hat{\theta}_W$ for pools implies identity of their denominators. The

reasoning in this section is equivalent to a double counting proof of the combinatorial identity

$$\sum_{j=1}^{\min(c,r)} \frac{c!}{(c-j)!} S(r,j) a_j = \sum_{k=1}^{c-1} \frac{c^r - k^r - (c-k)^r}{k} \tag{29}$$

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 [M.Mamino, persona communication]. This identity is a combination of a family of related identities for a general spectrum, presented in Appendix A.

5.2. General alternative form for the Watterson estimators

The above form of [7] 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 (15). 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

$$\hat{\theta}_W = \frac{S}{\sum_{\{r\}} L_{\{r\}} \sum_{k=1}^{c-1} \frac{1}{k} \Pi_k(\{c\},\{r\})} , \tag{30}$$

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

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 $\prod_{i=1}^{U} \left(\frac{k_i}{c_i}\right)^{r_i} + \prod_{i=1}^{U} \left(1 - \frac{k_i}{c_i}\right)^{r_i}$ is the probability of extracting just one of the two alleles. All the estimators of section 4.1 can be rewritten in this form. This form is often more convenient computationally than the combinatorics in equations (9), (15).

For a generic frequency spectrum $\mathrm{E}(\xi_k|n) = \theta L \bar{\xi}_{k,n}$, equation (30) should be replaced by

$$\hat{\theta}_W = \frac{S}{\sum_{\{r\}} L_{\{r\}} \sum_{k=1}^{c-1} \bar{\xi}_{k,n} \Pi_k(\{c\},\{r\})} . \tag{31}$$

We can also find an estimator similar to (28) 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 (the SNP is detected unless all complete sequences share the same allele) leads to the unbiased estimator

$$\hat{\theta}_W = \frac{S}{\sum_{r=2}^{\infty} L_r \sum_{k=1}^{n+m-o-1} \left(1 - \frac{\binom{n-o}{k}}{\binom{n+m-o}{k}} \left(1-\frac{k}{n}\right)^r - \frac{\binom{n-o}{k-m}}{\binom{n+m-o}{k}} \left(\frac{k-m+o}{n}\right)^r\right) \frac{1}{k}} \tag{32}$$

that is equivalent to the case (25) of estimator (9).

6. 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, autopolyploids are polyploid organisms with different homologous chromosomes from the same species. Autotetraploid populations follow the standard coalescent as shown by [18], 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 estimators for autopolyploids are given by

$$\hat{\theta}_W = \frac{S}{\sum_{r=2}^{\infty} L_r \sum_{k=1}^{p-1} \frac{1}{k} \left(1 - \left(\frac{k}{p}\right)^r - \left(1-\frac{k}{p}\right)^r\right)} \tag{33}$$

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

$$\hat{\theta}_W = \frac{S}{\sum_{\{r\}} L_{\{r\}} \sum_{k=1}^{np-1} \frac{1}{k} \Pi_k(n,p,\{r\})}\ , \tag{34}$$

$$\Pi_k(n,p,\{r\}) = \sum_{k_1=0}^{p} \dots \sum_{k_n=0}^{p} I\left(k = \sum_{i=1}^{n} k_i\right) \frac{\prod_{i=1}^{n} \binom{p}{k_i}}{\binom{np}{k}} \left[1 - \prod_{i=1}^{n} \left(\frac{k_i}{p}\right)^{r_i} - \prod_{i=1}^{n} \left(1-\frac{k_i}{p}\right)^{r_i}\right]$$

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

$$P_c(j|\{r\}) = \sum_{i_1=0}^{p} \ldots \sum_{i_n=0}^{p} I\left(j = \sum_{l=1}^{n} i_l\right) \prod_{q=1}^{n} P^*(i_q|p, r_q) \ . \tag{35}$$

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:

$$\hat{\theta}_W = \frac{S}{\sum_{r=2}^{\infty} L_r \sum_{k=1}^{np-1} \frac{1}{k} \left(1 - \left(\frac{k}{np}\right)^r - \left(1 - \frac{k}{np}\right)^r\right)} \ . \tag{36}$$

7. Discussion

In this paper we have 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 statistic for small θ. 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 [19, 20, 21]. Autopolyploids without a strong inbreeding follow the dynamics of the usual coalescent, so our theory is applicable to these species. On the other hand, allopolyploids (whose genome derives from different species) 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 [22, 19]. 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 at 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. In some cases it is possible to estimate $Z_{\varphi,\xi}$ or to correct $S_{\varphi,\xi}$ based on quality scores for called SNPs or known error rates. On the other hand, 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

$$\hat{\theta}_W = \frac{\sum_{\{\varphi,\xi\}\notin(\Phi,\Xi)_\epsilon} S_{\varphi,\xi}}{\sum_{\{\varphi,\xi\}\notin(\Phi,\Xi)_\epsilon} L_\varphi Z_{\varphi,\xi}} \tag{37}$$

as proposed in [23] for sequence data and [8],[9] for pooled reads. The only alternative is to estimate $Z_{\varphi,\xi}$ by heuristic methods.

Generalizing equation (37), it is also possible to extend the results of this paper to generic sums of the frequency spectrum, for example estimators of the form $\sum_{i=1}^{k}\xi_k/\sum_{i=1}^{k}1/k$ which consider only the lowest frequencies.

The estimator proposed here assume the standard Wright-Fisher neutral model for the allele frequency spectrum. However, an arbitrary expected frequency spectrum $\mathrm{E}(\xi_k|n) = \theta L\bar{\xi}_{k,n}$ could be used in the place of the neutral spectrum $\theta L/k$. It is sufficient to replace a_j by $\sum_{i=1}^{j-1}\bar{\xi}_{i,j}$ in the denominator of equation (9) or to replace $1/k$ by $\bar{\xi}_{k,c}$ in the denominator of equation (30). This extends previous adaptations of the original Watterson estimator to null scenarios with demography or varying population size (e.g. [24], [25]).

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 θL constant, $S_{\varphi,\xi}$ is Poisson distributed, i.e. $\mathrm{Var}(S_{\varphi,\xi}) = \mathrm{E}(S_{\varphi,\xi})$ and therefore $\mathrm{Var}(\hat{\theta}_W) = \theta/\sum_{\varphi\in\Phi} L_\varphi Z_\varphi$. In the same limit, the variance for linked sites contain a term $\theta^2 L^2$ coming from the covariances between sites [26, 4, 9]. An exact formula for this term of the variance is available only for a few cases: complete sequences, sequences with missing data [4] and pooled NGS reads [9]. In the case of completely

linked sites and known variance, these estimators could be improved [27], also by shrinkage methods [28]. Note that the variance of the MCLE could be estimated by the bootstrapping methods described in [29].

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. combinatorics [30] and could lead to further mathematical insights.

Acknowledgments

We thank A. Fonseca Amaral, M. Pérez-Enciso, W. Burgos, B. Nevado and G. Achaz for useful discussions, M. Mamino for providing an explicit proof of the main combinatorial identity and two anonymous referees for their constructive comments. LF acknowledges support from ANR-12-JSV7-0007 (ANR, France). The project was funded by Grants CGL2009-09346 (MICINN, Spain) and AGL2013-41834-R (MEC, Spain) to SERO and by a Consolider Grant from Spanish Ministry of Research, CSD2007-00036 "Centre for Research in Agrigenomics".

Appendix A. Combinatorial identities

We can extend the previous identity (29) 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.

We considering a frequency spectrum concentrated around a single frequency $\tilde{f}$ in the population (i.e. $\xi(f) = \delta\left(f - \tilde{f}\right)$, or $\xi_k = \binom{n}{k}\tilde{f}^k\left(1-\tilde{f}\right)^{n-k}$ for the sample spectrum). By double counting, the two Watterson estimators of the form (10) and (31) should be equal, and therefore we can equal their denominators. By computing the Taylor expansion in the variable $\tilde{f}$ of both sides and equating the coefficients of the lth power, we obtain:

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

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 (including (29)) 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

$$\sum_{k=1}^{l} (-1)^k \binom{c}{k, l-k, c-l} \left(c^r - k^r - (c-k)^r\right) = 0 \tag{A.2}$$

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] F. Tajima, Evolutionary relationship of dna sequences in finite populations, Genetics 105 (2) (1983) 437–460.

[3] 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.

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

[5] 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.

[6] 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.

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

[8] 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.

[9] L. Ferretti, S. E. Ramos-Onsins, M. Pérez-Enciso, Population genomics from pool sequencing, Molecular ecology 22 (22) (2013) 5561–5576.

[10] 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.

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

[12] R. Hudson, Two-locus sampling distributions and their application, Genetics 159 (4) (2001) 1805.

[13] G. McVean, P. Awadalla, P. Fearnhead, A coalescent-based method for detecting and estimating recombination from gene sequences, Genetics 160 (3) (2002) 1231–1241.

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

[15] 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.

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

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

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

[19] 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.

[20] Y. Han, Y. Kang, I. Torres-Jerez, F. Cheung, C. D. Town, P. X. Zhao, M. K. Udvardi, M. J. Monteros, 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.

[21] X. Wang, H. Wang, J. Wang, R. Sun, J. Wu, S. Liu, Y. Bai, J.-H. Mun, 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, G. Lassalle, G. J. King, G. Bonnema, H. Tang, H. Wang, H. Belcram, H. Zhou, H. Hirakawa, H. Abe, H. Guo, H. Wang, H. Jin, I. A. P. Parkin, J. Batley, J.-S. Kim, J. Just, J. Li, J. Xu, J. Deng, J. A. Kim, J. Li, J. Yu, J. Meng, J. Wang, J. Min, J. Poulain, J. Wang, K. Hatakeyama, K. Wu, L. Wang, L. Fang, M. Trick, M. G. Links, M. Zhao, M. Jin, N. Ramchiary, N. Drou, P. J. Berkman, Q. Cai, Q. Huang, R. Li, S. Tabata, S. Cheng, S. Zhang, S. Zhang, S. Huang, S. Sato, S. Sun, S.-J. Kwon, S.-R. Choi, T.-H. Lee, W. Fan, X. Zhao, X. Tan, X. Xu, Y. Wang, Y. Qiu, Y. Yin, Y. Li, Y. Du, Y. Liao, Y. Lim, Y. Narusaka, Y. Wang, Z. Wang, Z. Li, Z. Wang, Z. Xiong, Z. Zhang, The genome of the mesopolyploid crop species brassica rapa, Nature Genetics 43 (10) (2011) 1035–1039. doi:10.1038/ng.919.

[22] 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.

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

[24] D. Živković, T. Wiehe, Second-order moments of segregating sites under variable population size, Genetics 180 (1) (2008) 341–357.

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

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

[27] 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.

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

[29] A. RoyChoudhury, Composite likelihood-based inferences on genetic data from dependent loci, Journal of Mathematical Biology 62 (1) (2011) 65–80.

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