Abstract

Motivation: Unique molecular identifiers (UMIs) are added to DNA fragments before PCR amplification to discriminate between alleles arising from the same genomic locus and sequencing reads produced by PCR amplification. While computational methods have been developed to take into account UMI information in genome-wide and single-cell sequencing studies, they are not designed for modern amplicon based sequencing experiments, especially in cases of high allelic diversity. Importantly, no guidelines are provided for the design of optimal UMI length for amplicon-based sequencing experiments.

Results: Based on the total number of DNA fragments and the distribution of allele frequencies, we present a model for the determination of the minimum UMI length required to prevent UMI collisions and reduce allelic distortion. We also introduce a user-friendly software tool called AmpUMI to assist in the design and the analysis of UMI-based amplicon sequencing studies. AmpUMI provides quality control metrics on frequency and quality of UMIs, and trims and deduplicates amplicon sequences with user specified parameters for use in downstream analysis. AmpUMI is open-source and freely available at http://github.com/pinellolab/AmpUMI.

1 Introduction

Next-generation sequencing technologies have enabled the rapid and cost-effective translation of biological DNA or RNA sequences to short sequencing reads that can be used to analyze and understand the genome. In most library preparation protocols, genomic material must be amplified using PCR to ensure successful sequencing. Exponential amplification is sequence dependent and differences in amplification rate may arise from variation in sequence composition (Airdt et al., 2011). Duplicate sequencing reads resulting from PCR amplification may lead to biases in results or incorrect conclusions about the actual frequency of that read. Many approaches have been developed for the deduplication of next-generation reads from across the genome. Some tools such as FastUniq (Xu et al., 2012) or Fulcrum (Burriesci et al., 2012) filter raw sequencing output in FASTQ format and remove any reads with the same sequence. Commonly used tools such as Picard MarkDuplicates (Li et al., 2009), samtools rmdup (Li, 2011), and SEAL (Pireddu et al., 2011) align the reads to the genome, and identify reads aligning to the same genomic position as duplicates. However, if two DNA fragments from different cells produce reads aligning to the same location, they will be incorrectly called as PCR duplicates, even though they originated from two different molecules. As sequencing depth increases, so does the chance of finding apparently duplicate reads that align to the same genomic location but are from different DNA fragments. In addition, nonuniform genomic fragment
formation (e.g., short exons, restriction enzyme digestion, etc) can also increase the apparent rate of PCR duplicates if sequence identity and alignment coordinates are the only criteria for identifying PCR duplicates.

Although PCR deduplication may not affect experimental outcome in cases of calling reference variants from genome wide sequencing (Ebbert et al., 2016), library preparation strategies that rely on many PCR amplification cycles (such as single-cell experiments) are susceptible to errors introduced by amplification bias and should be corrected (Islam et al., 2014). In libraries with potentially low sequence complexity, PCR amplification bias may go undetected and lead to inaccurate interpretation of sequencing results. For example, accurate allele quantification may be distorted when identifying rare cancer mutations (Kinde et al., 2011; Kukita et al., 2015; Mansukhani et al., 2017) or when measuring the genomic outcomes of CRISPR genome editing experiments (Pinello et al., 2016) through amplicon sequencing of a specific target region.

In order to address the problem of duplicate reads from a biological perspective, library generation protocols have been developed that tag DNA fragments with a sequence of randomly selected nucleotides or partially degenerate nucleotides, called a unique molecular index (UMI). After sequencing, reads arising from PCR duplicates will all have the same UMI and all but one of these reads are marked as duplicates as shown in Figure 1a. In downstream analysis these reads are normally ignored, but they can also be used in error correction or in estimating sequencing error rates.

Although there are many benefits to using UMIs, there is a dearth of documented design rationale for selecting sufficient UMI length, and no existing tools to aid experimental designers who may not feel comfortable addressing the probabilities of UMI collision. In one case, a UMI length is chosen so that the number of unique UMIs ‘far exceeds’ the number of primer molecules used in the experiment (Kou et al., 2016). However, this may lead to overestimation of the required UMI length.

Several tools to analyze UMI datasets are available, but most are designed for genome-wide sequencing analysis, and require aligned sequence files in bam format as input. UMI-Reducer (Mangul et al., 2017), Je (Girardot et al., 2016), Debacrer (Stahlberg et al., 2016), and UMI-tools (Smith et al., 2017) are tools that take UMI information into account in the deduplication process, but are designed for genome-wide sequencing analyses. These methods first trim the UMI and associated sequencing adapters from the sequenced read, and append the UMI to the read name. Next, the trimmed sequencing reads are mapped to the genome. Finally, each read is scanned to identify other duplicate reads that align to the same genomic location and have the same UMI. However, these existing algorithms and methods are not well-suited to amplicon sequencing, where each basepair has an equal probability of sequencing error, longer UMI sequences are more likely to accumulate sequencing errors, which could result in ambiguity about whether sequences with similar UMIs result from sequencing errors or are unique molecular events.

To quantify the effects of UMI length on UMI collisions, we have derived an equation, Equation (4), that allows us to compute the probability of observing no UMI collisions as a function of UMI length, number of reads, and the number and proportions of alleles. The calculation of this probability for the worst case is implemented in our software and can be calculated using the AmpUMI Collision command. In some cases, a small number of UMI collisions may be tolerable as long as the final allelic proportions are within a certain error. In this case, Equation (13) can be used to determine the total allelic fraction distortion for a given UMI length and the true underlying allelic composition. The calculation of allelic distortion can be computed using the AmpUMI Distortion command. AmpUMI can also be used to determine the minimum UMI length required to have a probability of observing a UMI collision below a given threshold, or to produce an allelic distortion below a given percentage. These commands can be accessed using the AmpUMI Collision or AmpUMI Distortion command as before, except that the desired cutoff is given as a parameter instead of the UMI length.

We note that there is a difference between the number of reads and the number of DNA fragments being sampled. While it is relatively easy
to predict the number of reads that will be produced by a sequencing run based on the machine type and the abundance of the sample among those being sequenced, it is impossible to predict the allele frequency of the most frequent allele. With this in mind, we suggest that for experiments in which the number of input molecules and multiple rounds of PCR have been performed. This experimental design will yield multiple copies of each UMI-molecule pair, and sequencing errors will be apparent as rare mutations in the amplicon sequence. During deduplication, reads will be grouped by UMI, and the most frequent amplicon sequence will be accepted as the consensus sequence and the other sequences will be assumed to come from sequencing error and will be discarded. The second type of error correction is optional and can be used to correct errors in the UMI sequence by filtering rare UMI-amplicon sequence pairs that could arise from sequencing errors. First, FASTQ sequences with the same UMI and amplicon sequence are grouped. Only pairs of UMI-amplicon sequence that have been seen multiple times (a parameter specified by the user) are used for downstream analysis, while the remainder of rare UMI-amplicon sequence pairs are discarded. As this parameter is increased, the confidence in the UMI and amplicon sequence is increased, but the quantification accuracy of allele distribution is decreased, especially for rare alleles.

3 Methods

3.1 Problem formulation and assumptions

We are interested in determining the probability of no collisions of any two different alleles associated with the same UMI. In what follows, we make a few motivated assumptions to derive a closed form expression for our probability of interest.

Let \( X \) be a finite set of \( |X| \) DNA fragments partitioned into \( I \) types (i.e. alleles) and let \( Y \) be a finite set of \( |Y| \) UMI primers partitioned into \( J \) UMIs (i.e. barcodes), such that fragments or UMI primers within each class are indistinguishable from each other. Here, the number of UMIs \( J \) equals the number of the four nucleotides \( A, G, C, T \) raised to the power of the UMI’s sequence length \( K \). That is, \( J = 4^K \), where the value of \( K \) usually ranges from 8 to 12bp. Accordingly, the two sets of interest can be partitioned into \( X = \{X_1, \ldots, X_I\} \) and \( Y = \{Y_1, \ldots, Y_J\} \) where \( |X_i| = M_i \) for \( i = 1, \ldots, I \) and \( |Y_j| = U_j \) for \( j = 1, \ldots, J \).

Assumption 1 (Sampling with Replacement). Given that the total number of UMI primers \( U \) is much greater than the number of unique UMIs \( J \) and the total number of DNA fragments \( M \) is also much greater than the number of alleles \( I \), we can assume that both the UMI primers and DNA fragments are sampled with replacement. Note that the number of reads (samples) \( n \), which we introduce later, is also assumed to be much less than \( U \). For the UMIs, the assumption is motivated by the observation that in the PCR amplification step, there are typically \( 3 \times 10^3 \) UMI-labeled primers in 0.5 pmol of solution and so for UMIs of length \( K = 10 \), we would, on average, have approximately 286,100 copies for each UMI.

Assumption (1) implies that the probability of sampling either allele \( i \) or UMI \( j \) is given by the categorical distribution (a generalization of the
Bermoulli distribution to multiple categories). That is
\[ x \mid \text{rn} = (x_1, \ldots, x_J) \sim \text{Categorical}(I, \text{rn}) \]
\[ y \mid u = (y_1, \ldots, y_J) \sim \text{Categorical}(J, u) \]
where \( x \in \{1, 2, \ldots, J\} \), \( \text{rn} = (m_1, \ldots, m_J) \), \( y \in \{1, 2, \ldots, J\} \), and \( u = (u_1, \ldots, u_J) \). The proportion \( m_i = \frac{\text{rn}_i}{\sum_{i=1}^J \text{rn}_i} \) represents the probability of observing allele \( i \) such that \( \sum_{i=1}^J m_i = 1 \). Similarly, the proportion \( u_j = \sum_{i=1}^J u_j \) represents the probability of observing UMI \( j \) such that \( \sum_{j=1}^J u_j = 1 \).

Now, the total number of reads (samples) \( n \) equals the number of UMI-allele pairs—of which there are \( Q = |Q| = I \times J \) unique combinations. To determine the distribution over unique UMI-allele pairs, we first define a joint sample space \( \Omega = \{(1,1), (1,2), \ldots, (J,J)\} \) to enumerate all \( Q \) possible pairing outcomes and we let the random variable \( z \) be defined over \( \Omega \).

**Assumption 2 (Independence).** Since there is no indication that a given UMI has a preference for any allele, we can assume that the pairing process is unbiased such that sampling \( z \) is independent of sampling \( y \).

Assumption (2) implies that the probability mass function for \( z \) is given by a multinomial distribution with support \( z = \{z_1, \ldots, z_Q\} \in \mathbb{N}^Q \mid z_1 + \cdots + z_Q = n \) where \( n \) is the number of samples and \( \text{rn} = \text{sm} \cup z = [m_1 u_1, \ldots, m_J u_J] \) is the probability vector for the \( Q \) possible pairings, where \( \text{sm} \) is the Kronecker product. That is,
\[
z \sim \text{Multinomial}(n, \text{rn})
\]

\[ \Pr(z = i \text{ and } y = j) = \Pr(z = i) \Pr(y = j) = m_i u_j = p_{ij} \]
where \( q = 1, \ldots, Q \) and \( \sum_{q=1}^Q p_{q} = 1 \).

**Example:** For the case of two alleles \( J = 2 \), three UMI sequences \( I = 3 \), and three observations (i.e. reads) \( n = 3 \), the probability of observing two reads from Allele 1 paired with UMI sequence \( z_1 \) and one read from Allele 2 paired with UMI sequence \( z_2 \) is
\[
\Pr(z = \{2,0,0,0,1\}) = \binom{3}{1} = \frac{3!}{1!2!} = 3
\]
\[
\frac{\left(\left(m_1 u_1\right)^2 \left(m_2 u_2\right)\right)}{3 \left(1 - m_1^2 u_1^2 - m_2^2 u_2^2\right)}
\]
To derive the probability of obtaining no collisions, we note that the multinomial distribution for a sample size \( n \) can be viewed geometrically as the normalized \( n \)th component of a \( Q \)-dimensional Pascal’s simplex, \( \mathbb{S}_Q \), whose \( n \)th component, \( \mathbb{S}_Q^n \), consists of the coefficients of the multinomial expansion of the polynomial \( |p|_{Q} = \sum_{q=1}^Q p_q \) raised to the power \( n \). In particular, according to the multinomial theorem, we have
\[
|p|_Q^n = \sum_{n_{1}, \ldots, n_{Q} = 0}^{n} \binom{n}{n_{1}, \ldots, n_{Q}} p_{1}^{n_{1}} \cdots p_{Q}^{n_{Q}}
\]
where \( p \in [0,1]^Q \), \( z \in \mathbb{S}_Q^n \), \( n \in \mathbb{N}_0 \), \( Q \in \mathbb{N} \), and \( |p|_Q^{n} = 1 \).

Therefore, the number of multinomial coefficients is equal to the number of coefficients in the \( (Q-1) \)-simplex \( \mathbb{S}_Q^{Q-1} \) and is given by the \((n+1)\)th simplicial \((Q-1)\)-polytope number
\[
S = P_{Q-1}(n+1) = \left(\begin{pmatrix} n+1 \\ Q-1 \end{pmatrix}\right) = \frac{(n+1)(Q-1)}{Q-1}
\]
Each of the \( S \) components of the multinomial expansion represents the probability of a particular counts outcome event (i.e. \( z = z \)), which is given by the multinomial distribution
\[
p(z = z; n, p) = \binom{n}{z_1, \ldots, z_Q} p_{z_1}^{z_1} \cdots p_{z_Q}^{z_Q}
\]
We are interested in a subset of the \( S \) events that correspond to having no collisions. We denote the set of no collisions by
\[
\mathcal{T} := \{z \in \mathbb{S}_Q^n \mid \forall_{z_1, \ldots, z_Q \in [0,1]} \text{ and } \|z\|_1 = n\}
\]
Note that the cardinality of \( \mathcal{T} \) is equal to the number of ways a total of \( n \) \( 1 \)'s are selected from a group of \( n \) \( 1 \)'s and \((Q - n)\) 0's. More specifically,
\[
|\mathcal{T}| = \binom{Q}{n}
\]
Now since for all \( T \) events, the product of the terms in denominator of multinomial coefficient is equal to 1, we can simplify the expression for the probability of observing \( z^* \in \mathcal{T} \) as
\[
p(z = z^*; n, p) = \binom{Q}{n}(p^*)^n
\]
\[(1)\]

Note that when \( n > Q \), the probability of observing any of the \( T \) no-collision events is zero, otherwise when \( n \leq Q \) the probability of observing no collision can be expressed concisely as
\[
Pr(z \in \mathcal{T}) = \binom{Q}{n}(p^n)^n
\]
\[(2)\]

**Assumption 3 (Equal Proportions of UMIs).** Since the relative abundance of UMIs can be controlled during UMI synthesis, we can assume that the proportion of UMIs to be uniform such that \( u_j = \frac{1}{Q} \) for all \( j, j = 1, \ldots, J \).

Assumption (3) allows us to further simplify Equation (1) as such
\[
p(z; n, p) = n! \frac{1}{J^n} \prod_{q=1}^Q \frac{m_q^n}{n_q!}
\]
and so the probability of observing any of the \( T \) no-collision events, Equation (2), becomes
\[
Pr(z \in \mathcal{T}) = n! \frac{1}{J^n} \sum_{z \in \mathcal{T}} \left( \text{rn} \otimes 1_J \right)^z
\]
\[(3)\]

Computing the probability given by Equation (3) is intractable for large values of \( T \). Nonetheless, the expression can further be simplified if we assume the following.

**Assumption 4 (Upper Limit on the Number of Reads).** When the UMI sequence length \( K \) is 10, \( J \) already exceeds 1 million. For some experiments then, we can assume that the number of reads is less than or equal to the number of UMIs \( n \leq J \).

Using Assumption (4), we can partition the set \( T \) into \( E \) equivalent classes, essentially reducing the set into an \( I \)-dimensional discrete simplex
\[
E = \Delta_i^n := \{p \in [0,1]^Q \mid \forall_{\pi \in \{1, \ldots, I\}} \sum_{j \in \pi} \frac{Q}{n_{\pi}}\}
\]
with cardinality
\[
E = \left(\binom{I}{n}\right) = \left(\frac{n + 1}{n}\right)
\]
such that
\[
T = \frac{1 \times J}{n} \sum_{(r_1, \ldots, r_{I}) \in \mathcal{E}} \left(\prod_{i=1}^I \left(\frac{J}{n_i}\right)\right)
\]
so now the probability of observing any of the \( T \) no-collision events is given by this general formula.
\[
Pr(z \in \mathcal{T}) = n! \frac{1}{J^n} \sum_{(r_1, \ldots, r_{I}) \in \mathcal{E}} \left(\prod_{i=1}^I \left(\frac{J}{n_i}\right)\right)^n
\]
\[(4)\]
3.2 Special cases

Case 1 (One Allele (I = 1, Worst Case)). When I = 1, the discrete simplex $E$ has just a single component. Subsequently, Equation (4) reduces to

$$Pr(z \in T) = n! \left( \frac{1}{J} \right)^n \binom{J}{J}$$

$$= \frac{n!}{J^n (J-n)!} \binom{J}{J}$$

$$= \frac{(J)!}{J^n} \frac{1}{(1-\frac{1}{J})^n}$$

$$= \prod_{i=1}^{n-1} (1 - \frac{1}{J})^n$$

(5)

To find the minimum UMI length $k^*$ for a given sample size $n^*$ such the probability of no collisions is above a certain threshold $p^*$, we first need to compute the quantity given by Equation (5) for a given $n^*$ for each UMI length up to an upper bound $k_{max}$ (say 20). Let us denote the vector of computed probabilities by

$$\pi = (\pi_1, \pi_2, \pi_3, \ldots, \pi_{k_{max}})$$

Note the elements of $\pi$ corresponding to values of $K$ for which $n < 4^K$ are zero and do not require computing. The minimum UMI length $k^*$ can then be formally written as

$$k^* = \text{arg where}(\pi, p^*) = \text{min}\{i \in [1, k_{max}] | \pi_i \geq p^*\}$$

(6)

where $\text{arg where}(\pi, p^*)$ gives back the index of the first vector’s element whose value exceeds $p^*$. Note that since the value of $\pi_i$ increases as $i \to k_{max}$, we only need to return the index of the first element exceeding the threshold.

Case 2 (Equal Proportion of Alleles). Equations (3) and (4) have the simplest form when $m$ is a vector of equal probabilities. That is, when $m_i = \frac{1}{I}$ for all $i, 1 \leq i \leq I$. The probability of observing any of the $T$ no-collision events simplifies as such.

$$Pr(z \in T) = n! \sum_{z \in T} \frac{1}{J} \times \frac{1}{J} \times \frac{1}{J}$$

$$= n! \left( \frac{1}{J^3} \right)^n \binom{J}{J}$$

$$= \frac{(J)!}{J^n} \frac{1}{(1-\frac{1}{J})^n}$$

$$= \prod_{i=1}^{n-1} (1 - \frac{1}{J})^n$$

(7)

where $(Q)_n$ denotes the falling factorial. Note that the probability of no collision approaches 0 very rapidly as $n$ goes from 1 to $Q$. Also, note that when $n > Q$, the probability of collision is 1 for any $Q$.

Note that whereas the number of multinomial coefficients is given by $S$ for the set of all events and $T$ for the set of no-collision events, the sum of the coefficients is given by $A = Q^n$ for the set of all events and by

$$B = n!(Q)_n$$

(8)

for the set of no collisions. Alternatively, the probability can be derived as a ratio of $B$ over $A$

$$\frac{B}{A} = \frac{Q^n}{Q^n (Q-n)!} = \frac{(Q)_n}{Q^n}$$

Case 3 (Two Alleles (I = 2)). When I = 2, the discrete simplex $E$ has $n$ components. Subsequently, Equation (4) reduces to

$$Pr(z \in T) = \left( \frac{1}{J^2} \right) \sum_{(n_1,n_2) \in E} \frac{J}{n_1} \frac{J}{n_2} \frac{1}{n_1^{n_1}} \frac{1}{n_2^{n_2}}$$

(8)

where the values of $r = (n_1, n_2)$ vary over the anti-diagonal indices of an $n \times n$ matrix and the discrete simplex is given by

$$E = \Delta^2 = \{ r \in \mathbb{N}_0^2 | (0, n), (1, n - 1), \ldots, (n - 1, 1), (n, 0); n \leq J \}$$

3.3 Effects of Collisions on Allelic Diversity

Let $y = (y_{11}, y_{12}, \ldots, y_{1J}, \ldots, y_{IJ}, \ldots, y_{IJ})$ denote the $q$-dimensional vector of deduplicated reads for all UMI-allele pairing such that $y_{ij} \in \{0, 1\}$ refers to the read corresponding to the $i$th allele and $j$th UMI. The deduplication process can be formulated as a thresholded latent variable model, where an indicator function $I$ thresholds each element of the unobservable random vector $z$ into an observable binary variable $y_{ij}$. That is,

$$z \sim \text{Multinomial}(n, p)$$

$$y_{ij} = I(z_{ij} > 0)$$

where $z \in \mathbb{N}_0^Q$, $z_{ij}$ is the number of deduplicated reads for all UMI-allele pairing such that $y_{ij} \in \{0, 1\}$ refers to the read corresponding to the $i$th allele and $j$th UMI. The expected number of collisions is given as

$$\text{E}(y_{ij}) = \text{E}(\sum_{j=1}^{J} y_{ij}) = \sum_{j=1}^{J} \text{E}(y_{ij})$$

where $\text{E}(y_{ij})$ denotes the expected number of collisions for the $i$th allele and the $j$th UMI.

Expected Number of Deduplicated Reads

We are interested in the expected number of deduplicated reads (i.e., one-valued elements) in each of the $I$ allele partitions of $y$. We begin by letting $M$ denote the $I \times (I \times J)$ transformation matrix that gives the $J$-dimensional vector of deduplicated read counts for each allele. That is,

$$M = \begin{bmatrix} 1^T & 0^T & \ldots & 0^T \\ 0^T & 1^T & \ldots & 0^T \\ \vdots & \vdots & \ddots & \vdots \\ 0^T & 0^T & \ldots & 1^T \end{bmatrix}$$

Accordingly, the vector of expected deduplicated read counts $e$ can be obtained as such

$$e = \text{E}(My)$$

$$= \begin{bmatrix} \text{E}(\sum_{j=1}^{J} y_{1j}) \\ \text{E}(\sum_{j=1}^{J} y_{2j}) \\ \vdots \\ \text{E}(\sum_{j=1}^{J} y_{IJ}) \end{bmatrix} = \sum_{j=1}^{J} \begin{bmatrix} \text{E}(y_{1j}) \\ \text{E}(y_{2j}) \\ \vdots \\ \text{E}(y_{IJ}) \end{bmatrix}$$

$$= \begin{bmatrix} J - \sum_{j=1}^{J} \text{E}(y_{1j} = 0) \\ J - \sum_{j=1}^{J} \text{E}(y_{2j} = 0) \\ \vdots \\ J - \sum_{j=1}^{J} \text{E}(y_{IJ} = 0) \end{bmatrix}$$

$$= \begin{bmatrix} 1 - \left( 1 - \frac{1}{J} \right)^n \\ 1 - \left( 1 - \frac{2}{J} \right)^n \\ \vdots \\ 1 - \left( 1 - \frac{J}{J} \right)^n \end{bmatrix}$$

(9)

The penultimate step was obtained by observing that the expectation of an event defined by an indicator function is equal to the probability of that event.
4 Results

4.1 Allelic fraction distortion resulting from inadequate UMI length

One of the purposes of performing amplicon sequencing is to accurately estimate the frequency of each allele variant in the population. However, we show here that the observed allelic fraction of each variant can be distorted if the UMI length is too short and the UMI complexity is too low. In the case in which allele variants are present at roughly the same proportion in the population (\(m_1 \approx m_2 \approx m_3 \approx \ldots\)), this distortion will be small, as UMI-molecule collisions will affect each molecule type equally. However, as the proportion of allele variants becomes more imbalanced (e.g., \(m_1 \gg m_2 \gg m_3 \ldots\)), the observed frequency of rare alleles will increase, and the observed frequency of frequent alleles will be lower than the actual frequency in the population. Intuitively, this is because collisions are more frequent in the more frequent alleles, where a larger UMI diversity is required to avoid UMI-molecule collisions. Rare alleles with few reads are less likely to exhaust the available unique UMIs.

As an example, we simulated biological library preparations with \(n = 100,000\) reads from a population with an underlying allelic diversity defined as \(m = (0.1, 0.1, 0.3, 0.5)\). In each simulation, alleles were paired with a UMI. Samples were simulated using UMIs of length between 1 and 18bp long. After the simulated sample was assembled, reads with the same UMI and allele were deduplicated, leaving only one unique UMI-allele combination in the sample. The proportion of each allele was reported. As seen in Figure 2, if the UMI is too short and UMI complexity is too low then the proportion of each allele, including rare alleles will be approximately equal. This has the effect of overestimating the frequency of rare alleles in the population, while underestimating the frequency of abundant alleles as shown on the left side of Figure 2. With sufficient UMI diversity (right side of Figure 2), the true, underlying allele frequencies are recapitulated in the simulated, deduplicated populations. As expected, the allelic frequencies produced by our model are in perfect agreement with those obtained with computationally expensive simulations.

We then considered the number of collisions in our simulated samples (Figure 3). The number of collisions is defined as the number of simulated reads (UMI-molecule combination) that had been previously observed in the simulation. For example, We simulated 100,000 reads with a UMI of length 1 (4 possible UMIs), and measured the collision rate as \(p = 0.93\). We note that the accuracy of recapitulating the actual allelic fraction as shown in Figure 2 is mirrored in the collision rate shown in Figure 3.

4.2 Probability of observing no UMI collisions

Next, we address the problem of finding the probability of observing no UMI collisions. Equation (5) describes the probability of observing no collisions in the worst case scenario in which we are most likely to observe a UMI collision—when only one allele is present in our DNA fragment pool. Figure 4 shows the probability of having no UMI collisions for a variety of UMI lengths. When we construct 100,000 reads consisting of a random UMI and a DNA fragment, an 18bp UMI yields a sample with no collisions with probability \(p < 0.93\), while shorter UMIs are less likely to yield a sample with no collisions.

4.3 Rates of no UMI collisions in simulated samples

We wanted to validate our model with simulated data, so we simulated samples of size \(n = 100,000\) with a UMI of length between 60bp and 17bp long. For each simulation sample, we simulated a set of DNA fragments consisting of five different alleles. The fractional presence of each allele was randomly selected for each simulation sample to explore collision rates in cases without prior knowledge on allelic composition of a sample. DNA fragments were paired with UMIs and the number of UMI collisions...
Fig. 2. Association between distortion of allelic frequency and UMI length. Colored bars show simulated allelic fractions of four alleles after deduplication of reads with the same UMI and allele. Simulated samples consisted of \( n = 100,000 \) reads and were drawn from a population with an allelic diversity given by \( m = (0.1, 0.3, 0.3, 0.5) \). Reads were generated using UMI of length between 1 and 18bp long. For each UMI length, the average simulation proportion of each allele is shown after removing UMI-allele collisions. 100 samples were simulated for each UMI length. The right column marked ‘Truth’ shows the underlying allelic diversity from which the simulated samples were drawn. Dots connected by lines show the predicted allele frequency given our model (Equation (12)) and are in complete concordance with the simulation results. The gray histogram at the bottom of the plot shows the total allelic fraction distortion (TAFD) (Equation (13)) for each UMI length.

Fig. 3. Distribution of collisions in simulated populations. The number of collisions in each simulated sample used in Figure 2 were aggregated by UMI length. Boxplots show the median (thick line), interquartile range (box), and the range of the data (whiskers). The count of collisions is defined as the number of simulated reads (UMI-molecule combination) that had already been observed in the simulation.

was recorded. A total of 1000 simulations were performed for each UMI length. Figure 5 shows the average percent of simulated samples in which no UMI collisions were recorded. The red line in this figure shows the predicted probability for the worst case in which there is only one type of molecule. As expected, the model-predicted worst case (red line) is always below the average of all simulated samples (blue lines) for every UMI length.

We were surprised to observe that the length of UMI required to achieve no UMI collisions was so long—with a minimum of 18bp long for a sample of 100,000 DNA fragments. This indicates that for experiments performed
sequences allows researchers to adapt the software to their specific UMI
prevent UMI collisions. Next, flexible trimming of the UMI and adapter
of UMIs and to determine whether the UMI pool had adequate diversity to
sequencing using UMIs. AmpUMI first allows users to assess the quality
can be chosen to keep the allelic distortion below a specified threshold.
resulting from these collisions, and an appropriate UMI length
probability of observing a collision to an acceptable threshold as
probability of having no collisions for a given experiment, and can be
UMIs and for preparing data for downstream analysis.

Table 1. Number of reads, alleles, and the percent of reads matching the reference allele that are present before and after
AmpUMI analysis of a dataset of deep amplicon sequencing (Kou
et al., 2016).

|                  | Unprocessed | Deduplicated - Step 1 | Deduplicated - Step 2 |
|------------------|-------------|-----------------------|-----------------------|
| No. Reads        | 6,440,216   | 426,447               | 269,564               |
| No. Alleles      | 20,760      | 2,524                 | 176                   |
| Pct. Ref Allele  | 95.77       | 97.05                 | 99.19                 |

with this number of fragments and a shorter UMI (length 8bp or length
12bp), a small number of undetectable PCR duplicates are likely present
in these libraries.

We note that the probability of having no collisions may be somewhat
conservative for some experimentalists who may have a certain tolerance
for an allowable number of collisions in an amplicon sequencing
experiment. In this case, Equation (13) or the AmpUMI Distortion
command can be used to determine the required UMI length after selecting
a tolerable allelic distortion for their sequencing application.

4.4 AmpUMI validation on experimental data
To validate our deduplication software, we analyzed a publicly-available
dataset of deep amplicon sequencing (Kou et al., 2016). This dataset was
created to measure PCR amplification and sequencing errors, and utilized
a 22bp UMI. The amplicon sequence was verified by Sanger sequencing
to match the reference sequence. We tested to see whether AmpUMI could
correctly remove erroneous, non-reference reads. Table 1 shows data for
the unprocessed data and after the first and second steps of data processing
using AmpUMI. In Step 1, reads with the same UMI are deduplicated, and
only the most-frequent allele is kept for each UMI. In Step 2, UMI-
sequence pairs that appear less than 10 times are filtered out. Table 1 shows
the number of total reads, the number of unique alleles, and the percentage
of reads that represent the true, unmodified, reference allele. Although only
95.77% of reads in the unprocessed sample match the unmodified allele
sequence (the remainder of reads contain PCR and sequencing artifacts),
after both steps of processing by AmpUMI, 99.19% of reads match the
unmodified allele, showing a substantial reduction in contaminating reads
resulting from PCR or sequencing artifacts.

5 Discussion
In this study, we address the problem of UMI design and outline a statistical
framework for calculating the probability of collisions of different molecules having the same UMI. We also introduce AmpUMI, a flexible
software for analyzing next-generation sequencing from experiments using
UMIs and for preparing data for downstream analysis.

Our statistical framework allows researchers to calculate the probability of having no collisions for a given experiment, and can be
used to determine an acceptable number of UMIs in order to reduce the probability of observing a collision to an acceptable threshold as
determined by the user. In cases where a small number of UMI collisions are acceptable, our framework can also be used to estimate the allelic
distortion resulting from these collisions, and an appropriate UMI length can be chosen to keep the allelic distortion below a specified threshold.

We show that AmpUMI is a valuable tool for analyzing amplicon sequencing using UMIs. AmpUMI first allows users to assess the quality
of UMIs and to determine whether the UMI pool had adequate diversity to
prevent UMI collisions. Next, flexible trimming of the UMI and adapter
sequences allows researchers to adapt the software to their specific UMI
and sequencing design. AmpUMI can also perform sequencing error
correction to reduce noise and mitigate sequencing errors. Finally, PCR
duplicate reads are removed from the sample FASTQ for easy integration
into downstream analysis.

Although many analysis tools exist for deduplicating next-generation sequencing data, none are designed to use UMI information for deep
amplicon sequencing in cases where preservation of allelic frequency is important. AmpUMI will fill a gap in the available analysis tools that
are currently available to the community. We anticipate that this tool will be useful in a variety of applications. For example, in the genome editing
community, much effort is being placed on the quantification and characterization of CRISPR activity in the genome and the identification of
off-targets (Tsai et al., 2015; Kim et al., 2015; Tsai et al., 2017). The gold
standard for analyzing mutations at targets is by deep amplicon sequencing
of the target which allows for quantification of editing frequency and
analyzing genetic editing events (Pinello et al., 2016). After performing
genomic editing, the rate of modified reads is compared to the rate of
unmodified reads to determine editing efficiency at that location. Because
a large percent of the resulting amplicon sequences are the unmodified
sequence, simply removing duplicate sequences would greatly distort the
measure of editing frequency, and it is impossible to tell which sequence are
 arising from PCR duplication or unique molecules. Sequencing strategies
that use UMIs could greatly improve the accuracy and confidence in this
measure.

When performing amplicon sequencing on a sample with unknown
allele frequency, we recommend using the collision rate calculated for a
sample with only one allele, and acknowledge that this will yield a very
conservative estimate of minimum UMI length. In a laboratory setting, this
conservative UMI length could be used in a pilot study sequenced at low
sequencing depth to better understand the underlying allele distribution,
after which a more practical UMI length can be established for full-scale
sequencing analysis.

In conclusion, we provide AmpUMI as an open source tool that we
hope will benefit the scientific community, particularly in the design of
appropriate UMIs and in the processing of sequencing reads produced by
amplicon sequencing using UMIs. AmpUMI is open-source and freely
available at http://github.com/pinellolab/AmpUMI.

Funding
L.P. is supported by the National Science Foundation National
Institutes of Health R01HG008399 and the Defense Advanced Research
Projects Agency HR0011-17-2-0042. D.E.B. supported by NIDDK,
R01DK109232, NHLBI (DP2OD022716, P01HL32262), Burroughs
Wellcome Fund, Doris Duke Charitable Foundation, and ASH Scholar
Award.

Author Contributions
R.F. and K.C. conceived the methodology; K.C. developed the software;
R.F. developed the statistical formalism and worked out the mathematical
derivations; R.F. and K.C. created the visualization plots; K.C., R.F., D.E.B.,
and L.P. wrote, reviewed, and contributed to the final version of
the manuscript; K.C., R.F., D.E.B., and L.P. conceived the ideas in the
paper. L.P. supervised the project.

References
Aird, D. et al. (2011). Analyzing and minimizing PCR amplification bias
in Illumina sequencing libraries. Genome Biol., 12(2), R18.
Burtruci, M. S. et al. (2012). Fulcrum: Condensing redundant reads from high-throughput sequencing studies. Bioinformatics, 28(10), 1324–1327.

Ebbert, M. T. et al. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics, 17 Suppl 7, 239.

Girardot, C. et al. (2016). Jx, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. BMC Bioinformatics, 17(1), 419.

Islam, S. et al. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods, 11(2), 163–166.

Kim, D. et al. (2015). Digencode-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. Nat. Methods, 12(3), 237–243.

Kinde, I. et al. (2011). Detection and quantification of rare mutations with massively parallel sequencing. Proc. Natl. Acad. Sci. U.S.A., 108(23), 9530–9535.

Kou, R. et al. (2016). Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations. PLoS ONE, 11(1), e0146638.

Kukita, Y. et al. (2015). High-fidelity target sequencing of individual molecules identified using barcode sequences: de novo detection and absolute quantitation of mutations in plasma cell-free DNA from cancer patients. DNA Res., 22(4), 269–277.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27(21), 2987–2993.

Li, H. et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2078.

Mangul, S. et al. (2015). Umi-reducer: Collapsing duplicate sequencing reads via unique molecular identifiers. bioRxiv.

Mansukhani, S. et al. (2017). Ultra-sensitive mutation detection and genome-wide dna copy number reconstruction by error corrected circulating tumour dna sequencing. bioRxiv.

Pinello, L. et al. (2016). Analyzing CRISPR genome-editing experiments with CRISPResso. Nat. Biotechnol., 34(7), 695–697.
Pireddu, L. et al. (2011). SEAL: a distributed short read mapping and duplicate removal tool. Bioinformatics, 27(15), 2159–2160.

Slugay, M. et al. (2017). MAGERI: Computational pipeline for molecular-barcoded targeted resequencing. PLoS Comput. Biol., 13(5), e1005480.

Smith, T. et al. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res., 27(3), 491–499.

Stahlberg, A. et al. (2016). Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. Nucleic Acids Res., 44(11), e105.

Stahlberg, A. et al. (2017). Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing. Nat Protoc, 12(4), 664–682.

Tsai, S. Q. et al. (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas9 nucleases. Nat. Biotechnol., 33(2), 187–197.

Tsai, S. Q. et al. (2017). CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. Nat. Methods, 14(6), 607–614.

Vander Heiden, J. A. et al. (2014). pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics, 30(13), 1930–1932.

Xu, C. et al. (2017). Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller: BMC Genomics, 18(1), 5.

Xu, H. et al. (2012). FastUniq: a fast de novo duplicates removal tool for paired short reads. PLoS ONE, 7(12), e52249.