PCR duplicate proportion estimation and consequences for DNA copy number calculations

Andy G. Lynch[1,2], Mike L. Smith[3] Matthew D. Eldridge[4], and Simon Tavaré[5],
on behalf of the OCCAMS consortium

[1] School of Mathematics and Statistics, University of St Andrews, Mathematical Institute, North Haugh, St Andrews, KY16 9SS. andy.lynch@st-andrews.ac.uk
[2] School of Medicine University of St Andrews, North Haugh, St Andrews, KY16 9TF
[3] EMBL Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany.
mike.smith@embl.de
[4] CRUK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, England. matthew.eldridge@cruk.cam.ac.uk
[5] Irving Institute for Cancer Dynamics, Columbia University, Schermerhorn Hall Suite 601, 1190 Amsterdam Avenue, New York, NY 10027, USA.
st3193@columbia.edu

Abstract. The volume of DNA in a sequencing experiment is often amplified by PCR, leading to the possibility that the same original DNA fragment will be sequenced twice - a 'PCR duplicate'. Sometimes indistinguishable from these are multiple sequences arising from identical but independent molecules, which can lead to an over-estimation of the PCR duplicate proportion. The PCR duplicate proportion, and other measures derived from it, are important statistics for quality assurance, experimental design, and interpretation of sequencing experiments. Here we provide a full likelihood basis for a combinatorial approach using heterozygous SNPs as implemented in our R package, and demonstrate the efficacy of the approach. We also discuss the association with DNA copy number, and demonstrate the impact on a question of inferring mitochondrial DNA copy number that has recently been a feature of several high-profile cancer studies. This is explored through a simulation study.

Keywords: Whole-genome sequencing, DNA copy number, Likelihood, Quality Control, Mitochondria, Cancer

1 Duplicate Sequencing Reads

A simple DNA whole-genome sequencing (WGS) experiment might consist of sampling DNA from several cells, breaking the DNA up into fragments, increasing the number of fragments by creating copies, using a sequencer to identify the sequences of a random sample of those fragments, and mapping them to a reference genome. Once this is done, we can assume that the number of reads mapping to a genomic region is proportional to the average DNA copy number for that region, and that the degree of evidence for a particular feature of the genome is measured in the number of sequenced fragments (reads) that support the feature.

It is desirable for the accuracy of these quantitative methods that no original small molecule ends up being counted more than once in the analysis, for which reason 'duplicate reads' are typically removed from analyses.

Typically, duplicate reads are defined by the locations to which they map in the genome. Broadly, there are three ways in which such reads can arise. The first is an error in the imaging or image processing (hereafter referred to as "optical duplicates", although our definition may be broader than that usually associated with this term). The second is that the same original DNA fragment can give rise to multiple clusters on the sequencing flow cell - most likely because the fragment was duplicated in a Polymerase Chain Reaction (PCR) amplification step - and so we refer to these as PCR duplicates. The third is that two independent DNA molecules happen to fragment in the same positions and both give rise to clusters on the flow cell (hereafter referred to as "fragmentation duplicates").

Since fragmentation duplicates represent independent molecules, we wish to retain them in analyses. By contrast, ideally we would wish to retain only one of a set of PCR duplicates, and so all others are typically removed. As fragmentation duplicates are generally indistinguishable from PCR duplicates but fewer in number, fragmentation duplicates are typically removed along with PCR duplicates. Optical duplicates by their nature are typically identifiable and can be removed as a separate process, but could also reasonably be combined with the PCR duplicates in an "undesirable" duplicate category. We will ignore optical duplicates for the rest of the manuscript.

1.1 An Example Data Set

We illustrate this article with WGS data sets previously published by the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) consortium [14]. In particular we focus on 22 'control' WGS data sets that we can assume to be broadly diploid (i.e. they have two copies of each of the autosomal chromosomes). These were generated from blood ($n = 12$) and non-cancerous oesophageal tissue ($n = 10$) from 22 Oesophageal Adenocarcinoma (OAC) patients. DNA from oesophageal tissue was extracted using the DNeasy kit (Qiagen) and from blood using the NucleonTM Genomic Extraction kit (Gen-Probe) (according to the manufacturers' instructions). A single library was created for each sample and was sequenced to a nominal depth of 50x, using paired-end reads of length 100. Read-pairs were aligned to the human reference genome GRCh37 using the Burrows-Wheeler Aligner (BWA) [9].

The median duplicate percentage (counting both PCR and Fragmentation) across these samples, as derived using Picard [3], is 6.5% (ranging from 3.73% to 14.49%). As might be anticipated, these values are skewed by areas of the genome that are not diploid (e.g. mitochondria), or which (due to problems with aligning reads and discrepancies between the reference genome and the true genome) do not behave as diploid (e.g. telomeres). Such regions need to be removed before calculating and correcting the values.

2 Approaches to Separating Out the Duplicate Types

When classifying specific duplicate reads as being fragmentation duplicates is not possible, i.e. when random tags have not been appended to the original molecules in the mix, it is sufficient for some analyses merely that we can estimate the numbers in each class.

A probabilistic approach for estimating the numbers of fragmentation duplicates, through considering the distribution of insert sizes, read lengths and coverage has been presented in the context of high-coverage targeted sequencing experiments [16]. This makes such reasonable assumptions about the independence of reads, the independence of insert size and depth of coverage, and the lack of external limiting factors (e.g. constraints imposed by starting with few molecules). While it is possible to apply an approach comparable to Zhou et al's [16] to WGS data, the nature of these data allows for an empirical estimate of the proportion of fragmentation duplicates.

Specifically, we take advantage of knowing that many loci in a WGS experiment will be heterozygous, with known allele fraction (often 0.5), and that the definition of duplicate reads makes use of their genomic locations rather than their sequences. PCR duplicates covering a heterozygous site should show the same allele, while fragmentation duplicates are neither constrained to show the same allele nor compelled not so to do. This was a characteristic that we exploited in our 2016 software and the update accompanying this manuscript [11], and which has been similarly exploited by others [1]. This latter application is notable for suggestions of application also to RNA-seq data.

3 A Likelihood Approach Based on Allele Patterns at Heterozygous Loci

Here we set out a likelihood methodology for estimating the proportion of duplicate reads that are PCR duplicates (or equivalently the proportion that are fragmentation duplicates). This is the same approach as implemented in our software [11].

3.1 A Simple Approach Using only Pairs of Duplicates

We first simplify the problem by imagining that where duplicate reads exist, there are precisely two reads mapping to the locus and no more.

We will consider such pairs of reads that overlay the sites of heterozygous Single Nucleotide Polymorphisms (SNPs). In each case, one of the pair will have been marked as a duplicate. In practice we will consider only a pre-defined set of potential sites of heterozygous SNPs in order to simplify computations. Our aim is to identify the proportion, P_D, of duplicate reads that do not represent an observation arising from a novel molecule and to separate this from the proportion that do.

If we assume that we are dealing only with a pair of fragments, then with probability P_D (the quantity we wish to estimate) they are observations of the same original molecule, while with probability $F_D = 1 - P_D$ they are observations representing different starting molecules. We exploit the fact that at these locations, if we are restricting ourselves to parts of the genome that are in allelic balance (i.e. the same number of copies of paternal and maternal sequence) then we can make the following statements:

If the two reads are observations of the same original molecule then they should report the same nucleotide at the locus of the heterozygous SNP (excepting for sequencing errors, the inclusion of which we assume we can control by filtering on the base-calling quality score). This scenario we denote AA regardless of the allele being reported.

If the reads arise from different starting molecules, then they will report the same nucleotide (AA) half of the time and different nucleotides (denoted AB) half of the time (assuming that the number of cells contributing to the sequencing library is such that removing one molecule from one cell does not noticeably affect the balance of available alleles).

If we observe counts of N_{AA} pairs of reads where the duplicate is reporting the same nucleotide, and N_{AB} where it is reporting different nucleotides, giving a total number $N = N_{AA} + N_{AB}$, then equating the observed and expected proportions of AA and AB patterns gives

$$\begin{aligned} N_{AA}/N &= P_D \times 1 + F_D \times 0.5 \\ N_{AB}/N &= F_D \times 0.5 \end{aligned}$$

which we can rearrange to gain an estimate of P_D:

$$P_D = 1 - 2 \times N_{AB}/N. \tag{1}$$

3.2 A Likelihood Approach for Pairs of Duplicates

We can explicitly frame this in terms of a likelihood model. There are $Q = 2$ distinct observable allele patterns ($AP_1 = AA$ and $AP_2 = AB$). We wish to calculate the probabilities of observing each of the Q allele patterns given a value of P_D, denoted $\Pr(AP_k|P_D)$ for allele pattern k of Q. Coupled with the observed counts of each allele pattern, $N(AP_k)$, these allow us to define the log-likelihood of P_D to within an additive constant:

$$l(P_D) = \sum_{k=1}^{Q} N(AP_k) \log \Pr(AP_k \mid P_D), \tag{2}$$

We can write down $\Pr(AP_k \mid P_D)$ in a straightforward manner. When $Q = 2$,

$$\begin{aligned} \Pr(AA \mid P_D) &= \frac{1}{2}(1 + P_D) \\ \Pr(AB \mid P_D) &= \frac{1}{2}(1 - P_D). \end{aligned}$$

The log-likelihood is then

$$l(P_D) = N_{AA} \log((1 + P_D)/2) + N_{AB} \log((1 - P_D)/2), \tag{3}$$

and if we seek the maximum likelihood estimate by equating the first derivative to zero, we obtain $0 = N_{AA}(1 - \hat{P}_D) + N_{AB}(1 + \hat{P}_D)$, whence

$$\hat{P}_D = (N_{AA} + N_{AB})/N = 1 - 2 \times N_{AB}/N$$

as required to match the estimate in (1).

3.3 The Full Model

If we have more than two fragments in our duplicate set, then we can extend the log-likelihood approach in a natural manner. For each size M of duplicate set, we still sum over all $Q = Q_M$ potential allele patterns the number of times that allele pattern was seen multiplied by the log of the probability of seeing that allele pattern. We simply have to extend this by summing also over all values of M. The challenge is to calculate the probabilities of the allele patterns, $\Pr(AP_k \mid P_D)$.

This calculation can be facilitated by conditioning on the underlying partition of the M reads into the, at most M, original molecules contributing to the set. We consider every possible partitioning of M fragments into m non-identifiable bins representing m original molecules, to obtain

$$\Pr(AP_k \mid P_D) = \sum_i \Pr(AP_k \mid \mathrm{PART}_i) \Pr(\mathrm{PART}_i \mid P_D), \tag{4}$$

allowing us to calculate the log-likelihood and to find the maximum likelihood estimate of P_D.

Determining the Number of Partitions Details of the sequence of partition numbers can be found at http://oeis.org/A000041/. Since the task need only be performed once, and the largest value of M observed will typically not be very large, the numbers can be determined by recursively deriving all possible partitions.

Determining the Probability of an Allele Pattern Given a Partition Given a partition, we know the number of molecules present, and the number of read-pairs each molecule contributes (this is in essence our definition of a partition). Every pattern of assignment of alleles (“A” or “B”) to the molecules is given equal probability. Without loss of generality, we can initially assign “A” to the first molecule, so the number of allele assignments to be considered is only 2^{m-1} where m is the number of molecules in the partition. Similarly, for reasons of identifiability, if necessary we relabel the alleles within a pattern so that the number of “A” alleles is at least as great as the number of “B” alleles. See Fig.

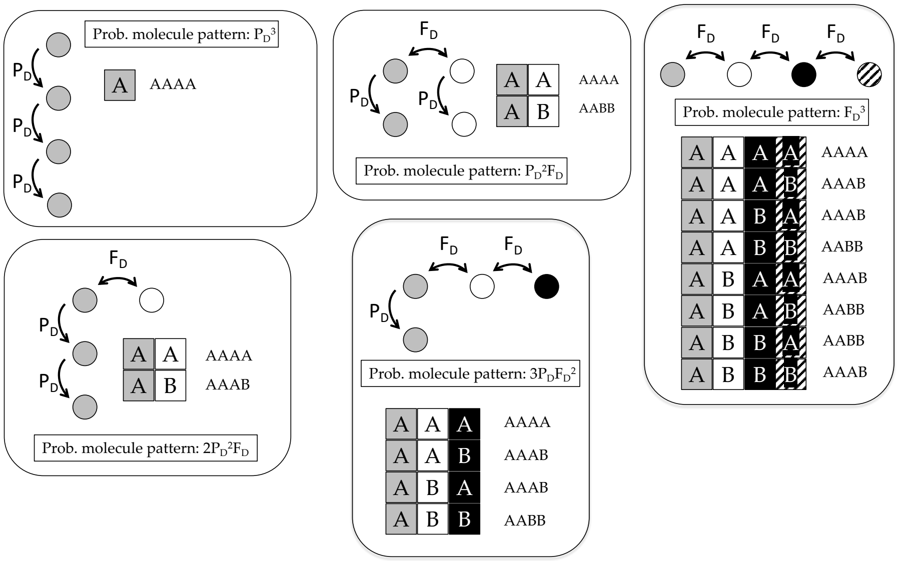

Fig. 1. Details for the case when $M = 4$. For example if the 4 read-pairs arose from two original molecules then they must partition into a 3 and 1, or into 2 lots of 2. In the latter case the only two patterns that can be seen are "AAAA" and "AABB" depending on whether the two original molecules exhibited the same or different alleles respectively. Each has equal probability when arising from this partitioning, but the same observed patterns can also result from other partitionings.

1 and supplementary materials for example calculations. The probability of an allele pattern is the sum of probabilities of assignments that give rise to that pattern.

When M read-pairs are partitioned amongst m molecules, it is straightforward to see that the form of $\Pr(\text{PART}_i \mid P_D)$ must be:

$$\Pr(\text{PART}_i \mid P_D) = KF_D^{(m-1)} P_D^{(M-m)} \tag{5}$$

since m molecules implies that we have $(m-1)$ fragmentation duplicates (the "-1" since one read-pair is regarded as an original and not a duplicate of anything) and the remaining $(M - m)$ read-pairs in the set must be PCR duplicates.

The value of K is

$$K = \left(\sum_j \nu_j \right)! \bigg/ \prod_j \nu_j! \tag{6}$$

where ν_j is the number of molecules from which exactly j read-pairs have originated. This may be intuited via combinatorial arguments, or a proof is given in the appendix.

As a concrete example, consider the case when there are four fragments ($M = 4$) partitioned amongst three molecules ($m = 3$) such that two read-pairs arise from one molecule and one read-pair arises from each of the other two (as in case (D) of Fig. 1). Then $\nu_1 = 2$ and $\nu_2 = 1$ ($\nu_j = 0 \; \forall j > 2$). The value of K is then $3!/(2!1!) = 3$.

3.4 Application to our Example data

Our method relies on identifying heterozygous SNPs in regions of constant copy number. In this case, we seek regions that are well-behaved and diploid. We also require an observation of the proportion of duplicates that we can correct. The basic observation of percentage of duplicates for the samples (the total number of duplicates seen, minus optical duplicates, divided by the total number of read-pairs examined) varies from 3.7% to 14.5% with a median of 6.5%. However, as highlighted above, the proportion of duplicates seen is affected by the inclusion of regions that are not representative of the regions in which our SNPs are located, and we may then choose to replace our basic observation with a “masked” observation.

With the removal of masked regions, the median duplicate percentage seen in our 22 samples is 5.0% (ranging from 2.3% to 13.2%). The reduction in the percentage is quite uniform across samples (Fig. 2) but, additionally, we have a third observation. To apply the methods described above, we investigate 2,500 common SNPs in anticipation of identifying 1,000 heterozygous sites for each sample: In fact, the numbers seen per sample vary from 942 to 1,093. From these approximately $1,000$ heterozygous SNPS, we can calculate an observed proportion of duplicates that directly relates to the correction we will make. On average there are 70,000 read-pairs considered per sample, which is sufficient to estimate the duplicate proportion well.

As seen in Fig. 2, masking the genome brings the observed proportion much more in line with that observed from the SNP loci (median 4.5%, range 1.9% to 12.1%), but the observations remain over-estimates of the duplicate proportions at the SNPs. Therefore, our calculations that follow will take the direct observation at the SNP loci as the combined PCR and fragmentation duplicate rate.

In total, $78,173$ sets of duplicates are observed across our 22 samples, with the largest set containing six duplicate read pairs. The percentage of observed duplicates attributed to fragmentation varied from 1.7% to 19.8% and was higher in blood samples (which have a lower observed duplicate proportion than the tissue samples). As seen in Fig. 2, the “corrected” estimate for PCR duplicate proportion is overestimated by more than a tenth in several samples if the fragmentation contribution is not removed, but more noticeably it is overestimated by a factor of up to 2.5 if, additionally, over-influential regions of the genome are not masked, or some other approach to establishing a representative observation of duplicate proportion is not used.

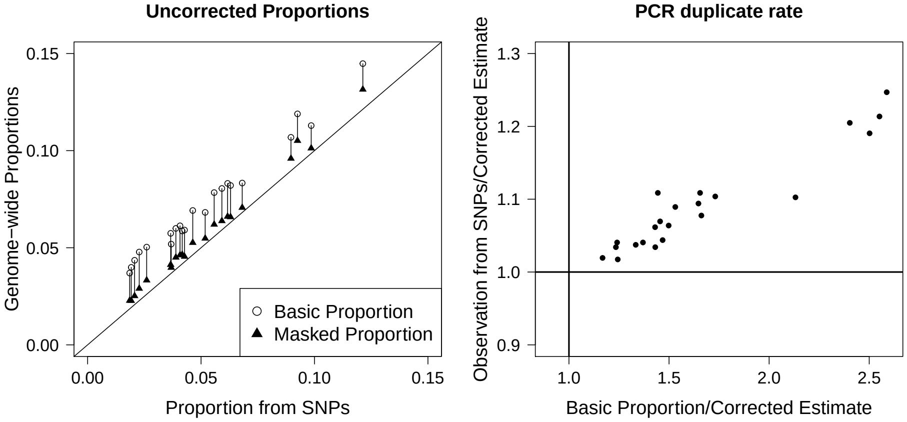

Fig. 2. Left: Showing the basic and masked observations of duplicate proportions to be unsatisfactory estimates of the proportions at our SNP loci. Right: Showing the degree to which the basic observation and observation from our SNP loci over-estimate our best, corrected, estimate.

Single vs Paired End Sequencing. We have, to this point, been considering read-pairs, those reads arising from DNA fragments where both ends of the fragment are sequenced. It is possible also to conduct sequencing such that only one end of the fragment is read ("Single-end sequencing"). In the case of single-end sequencing the definition of a duplicate is based on only the coordinate of one end of the fragment, and not the length of the fragment. With this laxer criterion reads will be classed as fragmentation duplicates more readily.

We can simulate a single-end read data set by discarding the second end from each read in our example data. Crucially, in doing this, we are simulating a single end data set with the same PCR duplicate proportion as the paired-end data set, because these are the same DNA fragments represented in both. One property of our correction method then is that it should return the same value when applied to each data set in turn.

In Fig. 3 we see that the observed duplicate proportion is indeed substantially higher in the single end data, and not even highly correlated with the observation in the paired-end data. After applying the correction methods presented here, the estimates show remarkable agreement (also Fig. 3).

A Cancer Sample We have until now been considering 'normal' diploid samples, but much interest lies in the study of cancer samples which are not diploid. We will illustrate now the application of this methodology to an OAC sample. Specifically, we consider sequencing library SS6003314 from the same study [14] as the blood and benign tissue samples that have been the examples so far.

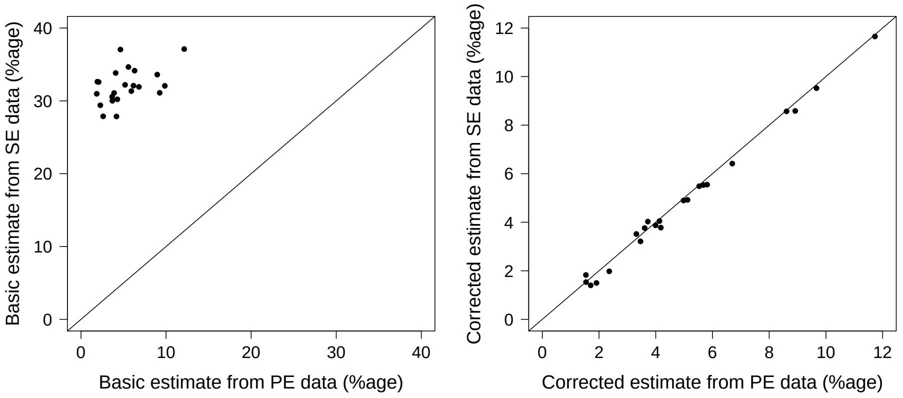

Fig. 3. Left: Comparing the basic observations of duplicate proportions from our paired-end example data and simulated single-end sequencing data. Right: Showing the agreement between corrected estimates of PCR duplicate proportion from the two data sets.

SS6003314 appears to be generated from a broadly tetraploid tumour with approximately 74% tumour purity (i.e. 26% of cells in the sample are contaminating normal tissue). While a copy number stage of "AABB" (i.e. two copies of both the paternal and maternal genome) is most frequently observed (Fig. 4, there are noticeable regions with copy number ranging from one to six and ranging from balanced to exhibiting loss of heterozygosity (i.e. for a given region only one of the maternal or paternal genome is present). Importantly for this analysis, there are regions present with inferred copy number state "AB".

So long as the SNP loci selected are from regions with the same copy number, and are balanced, then we can apply the methods outlined in this paper. For this example, we can apply them independently to the "AB" and "AABB" copy number states. For the "AABB" regions we have identified 8, 396 heterozygous SNPs to use in the analysis, while for the "AB" regions we have identified 759.

The results from the analysis in Table 1 show that the inferred proportion of duplicates due to fragmentation is still, for this data set, low at a copy number of four. Therefore the duplicate proportions are similar before and after correction and also when comparing "AABB" and "AB" regions. While still small, the proportion of duplicates due to fragmentation did increase going from "AB" to "AABB".

Note that the likelihood models presented here could be extended to any copy number state where both alleles are present, but a) this is more complicated and b) most cases have a region that can be identified as balanced. Note also, that even if the assignment of copy number states was wrong (perhaps we have actually been considering regions that were "AABB" and "AAABBB") we are

	Observed duplicates	Proportion due to fragmentation	PCR duplicates	Fragmentation duplicates
AABB	3.76%	0.047	3.58%	0.18%
AB	3.68%	0.017	3.62%	0.06%

Table 1. Reporting the percentage of observed duplicates and the estimated PCR and Fragmentation duplicate percentages both for regions of the genome that have copy number pattern "AABB" and those with copy number pattern "AB".

not affected, because all we have made use of is the knowledge that the regions were balanced.

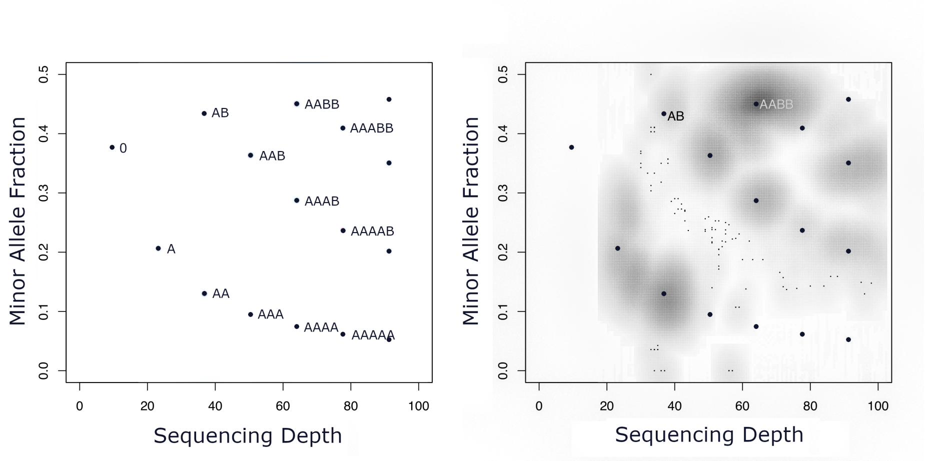

Fig. 4. Left: Illustrating the expected patterns to be seen when plotting minor allele frequency against sequencing depth for a sample that is 74% tumour. Genomic regions that, in the tumour, share the same copy number state in all cells are expected to appear at the points indicated, and the copy number state associated with each point is annotated. For example "AAABB" indicates a region of copy number five with three paternal and two maternal copies or vice versa, while "0" indicates regions that are entirely missing in the tumour genome. Note that since the minor allele fraction is bounded at 0.5, the expected value for balanced regions has to be less than this, and the bias is more extreme the lower the copy number count. Right: A scatter plot illustrating the observed relationship between smoothed minor allele fraction and smoothed sequencing depth for sequencing library SS6003314; an oesophageal adenocarcinoma sample with inferred tumour purity of 74%. Darker regions indicate that more of the genome lies at this position. The grid from the left hand plot is superimposed. Clouds of points lying off the grid may indicate regions of the genome that do not have a common copy number state in all cancer cells, artefacts from the smoothing, or that the tumour purity has been misidentified.

4 Effects on the estimation of DNA Copy Number

While the effects of copy number may sometimes be minimal when comparing diploid and tetraploid genomes, there are circumstances in considering duplicate proportions when it is important to distinguish DNA copy number and depth of sequencing coverage despite the linear relationship we anticipate (as in Fig. 4).

Local to a region of constant copy number, the PCR duplicate proportion will be trivially linked to depth of coverage since duplicate sequences count towards that coverage, and if fewer than two reads are present then there cannot be a duplicate. There may also be factors such as GC content that have the potential to influence both properties locally within a genome. Beyond this, there is no reason to expect the PCR duplicate proportion to vary with copy number. The variation of our observed duplicate proportion with copy number we then attribute to the fragmentation duplicates.

At high values, it is clear that depth of sequencing will drive the proportion of fragmentation duplicates we see. There are only a finite number of positions in which sequencing read pairs can be positioned, and that there must be a depth of coverage, beyond which, all additional reads will be classed as duplicates. That is, we achieve saturation, and there is a depth beyond which our sequencing will reveal only additional duplicates. Consequently, if we remove duplicates from an analysis, we place an artificial threshold on the copy number we can call. Moreover, as one approaches that threshold, reads will be classed as duplicates more frequently. Depth of sequencing for a region of the genome will depend on the DNA copy number locally (as shown in Fig. 4) and the overall number of sequences generated for the sample.

For a given DNA copy number, there are three key aspects of the sequencing that determine the numbers of reads that are lost after being classed as fragmentation duplicates. These are a) the depth of sequencing associated with a region of copy number one: More depth is generally a good thing in sequencing experiments, but leads us to problems with fragmentation duplicates sooner. b) the standard deviation (or more generally the distribution) of the fragment sizes: Less variable fragment lengths are conceptually useful for identifying some types of structural variant, but increase the numbers of fragment duplicates. c) The lengths of the sequencing reads: For most purposes it is beneficial that these are long, and minimizing fragmentation duplicates is no exception.

If we simulate some not-unrealistic data, where the sequencing coverage associated with one copy of DNA is 30 reads (achieved using paired-end 100 base pair (bp) reads and a DNA fragment length standard deviation of 50, and with no PCR duplicates) then we see that the removal of duplicates (all due to fragmentation) has an increasingly large effect as the copy number increases. At moderate copy numbers, taking a copy number of 2 as the baseline, there is little effect - regions of copy number 8 for example will be estimated to have a copy number of 7.98 in these conditions. A true copy number of 50 would be estimated to be 49, 100 would be estimated to be 96, 500 would be estimated to be 410, and a copy number of 1000 would be estimated to be 686. This ‘compression’ of observed copy numbers in high-copy-number-states will naturally result in increased uncertainty in the inference of true copy number, even before saturation is reached.

In Fig. 5 we show the percentages of reads that are lost through being fragmentation duplicates for three simulated examples. One example matches the scenario given above where the parameters provide an approximation to our real data. The second example is a more extreme case with shorter reads and a tighter distribution of insert sizes, while the third shows the effect for a case with longer reads, more variable fragment length, but lower coverage per DNA copy (perhaps because the sample being studied is tetraploid not diploid). From this figure we can see that the effects can vary from extreme to possibly ignorable.

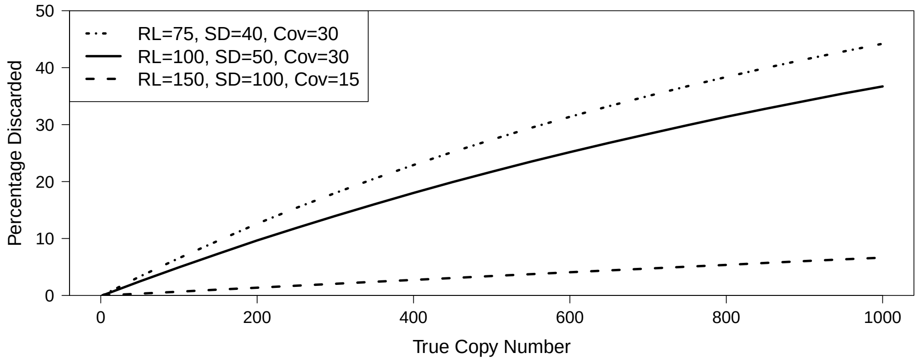

Fig. 5. Showing the percentages of reads incorrectly discarded due to being identified as fragmentation duplicates under three simulation schemes. The parameters varied are read length (RL), standard deviation of the fragment lengths (SD), and depth of coverage for a region of copy number one (Cov).

Although most of the genome is of a copy number where these effects are small, it is worth noting that a) the PCR duplicate proportion is also small, and the effect on this (and downstream characteristics such as inferred sequencing library size) can be considerable, b) that only a small region of genome at very high copy number can greatly increase the number of fragmentation duplicates present in a samples, and c) the inference of copy number states can sometimes be finely balanced between multiple credible solutions, and any discrepancy between the assumptions of the model and the true nature of the data, could affect the proffered solution.

5 The Estimation of Mitochondrial DNA Copy Number

Mitochondria are organelles within a cell that contain their own small (∼17 kB) genome (mtDNA). There will be many mitochondria within a cell and each can have several copies of the mtDNA. Thus the mtDNA is expected to be present in a cell at high copy number.

That different cell types have different mtDNA copy numbers has been known for nearly half a century [2] and 'next-generation' sequencing data have been used to investigate this since the early days of the technology using targeted sequencing [13] or WGS [5]. A review of the changes in mtDNA copy number in cancers of various tissues highlights the potential importance of this quantity and also highlights the range of copy numbers that are possible (0-100,000) [12]. A recent pan-cancer analysis of over 2000 tumour samples estimated values from 8 in a pancreatic cancer to > 1750 in a cancer originating in the central nervous system [15].

Clearly then, we can be of an order of copy number where the saturation effects described above could have an effect, in underestimating the mtDNA copy number.

5.1 PCAWG Copy Number

Assuming that appropriate measures of coverage for the nuclear genome and mitochondrial genome can be identified, the recent Pan Cancer Analysis of Whole Genomes (PCAWG) survey of mitochondrial changes in cancer [15] used the following approach for the estimation of mtDNA copy number (mtDNA-CN):

$$\text{mtDNA-CN}_{\text{tumour}} = \frac{\text{mtDNA coverage}}{\text{nuclear coverage}} \times \text{mean nuclear copy number} \tag{7}$$

where the mean nuclear copy number is defined as

$$f \times \text{`Tumour mean nuclear copy number'} + (1 - f) \times 2, \tag{8}$$

f being the proportion of tumour within the sample.

Tumour purity We note that the term

$$\frac{\text{nuclear coverage}}{\text{mean nuclear copy number}} \tag{9}$$

is simply a measure of the coverage per copy number, and while it has been calculated taking into account the tumour purity, when we rearrange Equation 7 it becomes clear that there is no further correction for tumour purity in the mitochondrial copy number:

$$\text{mtDNA-CN}_{\text{tumour}} = \frac{\text{mtDNA coverage}}{\text{coverage per copy number}} \tag{10}$$

Thus the estimated mtDNA copy number is that averaged over the tumour and contaminating benign tissue. This is natural if, as is often the case, the mtDNA copy number for benign tissue is not known (unlike for the nuclear genome where the copy number for benign tissue can be assumed to be 2), but since it has long been known that mtDNA copy number can differ between malignant and benign tissue [8], trends between estimated mtDNA copy number and tumour purity would seem inevitable.

In particular the tumour mtDNA-CN will be shrunk towards the benign value. A process that accentuates any bias due to duplicate removal if the mtDNA-CN is higher in the tumour than surrounding benign tissue, but may apparently compensate for it if the mtDNA-CN is lower. Nevertheless, in that scenario, the two competing biases cannot be relied upon to 'cancel out' (Fig. 6). These two effects clearly have the potential to mask or reduce the differences in mtDNA-CN observed between groups.

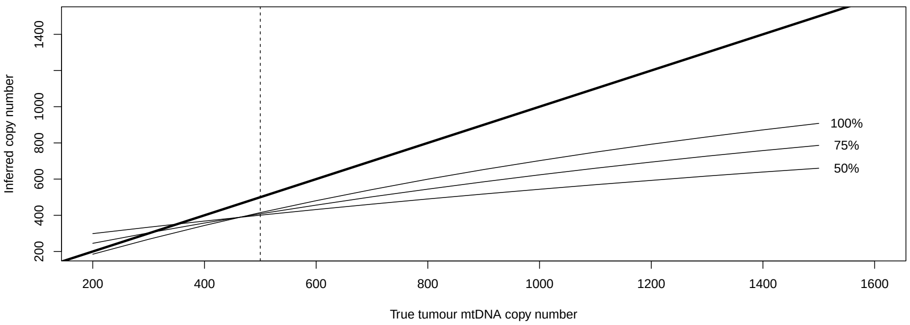

Fig. 6. Showing how the effect of removing duplicates can quickly outstrip the effect of impurity in the sample. For a range of true tumour mtDNA-CN values and a fixed mtDNA-CN of 500 for contaminating benign tissue (indicated with a vertical dotted line) three curves are depicted for the copy number inferred from simulation with three different levels of tumour purity. A line of agreement is shown in bold for contrast.

Consideration of Duplicates It seems clear that in calculating the ratio of coverages in Equation 7, that duplicates should either be retained in both numerator and denominator, or duplicates should be removed from both numerator and denominator. Early examples of research into mitochondria using sequencing technologies retained the duplicates [13, 4, 10], but many more recent investigations have been secondary analyses. It is almost certainly the case that reported values pertaining to the nuclear genome will have been calculated after removing duplicates and revisiting an entire WGS data set to recalculate values for the nuclear genome will be costly. Therefore it may be more convenient to remove duplicates from both numerator and denominator, as indeed appears to have been done in the recent pan-cancer characterization [15] and in other studies.

5.2 An Approach to Correct the Estimate of mtDNA Copy Number

The problem with removing duplicates in both denominator and numerator is that it is only the PCR duplicates that we would wish to remove, and we have seen that while the contribution of fragmentation duplicates will have minimal effect on the nuclear genome calculations, it will have potentially great effect on the mitochondrial calculations (Fig. 6).

Assuming that we do not wish to pay the cost of reanalysing the complete data set, then we are typically in the position of having the nuclear coverage with PCR and fragmentation duplicates removed, and mitochondrial coverage with PCR and fragmentation duplicates removed. For little cost, it is possible to extract and reprocess the mitochondrial-mapping sequences, while applying the methods of this paper to the nuclear genome.

We are then left with observations of the nuclear coverage with PCR and Fragmentation duplicates removed, an estimate of the corresponding fragmentation duplicate proportion, an estimate of the PCR duplicate proportion, and the observed mitochondrial coverage with no duplicates removed. From these it is clearly possible to obtain an estimate either of the ratio of coverages with no duplicates removed, or the ratio of coverages with PCR duplicates removed.

Note that the fragmentation duplicate proportion in the mitochondrial genome can not be estimated directly using our methods due to the lack of heterozygous SNPs in the mtDNA. Note also that in many cases, the correction of the nuclear genome with the nuclear fragmentation duplicate proportion will have minimal effect and might be dropped for even greater computational simplicity.

5.3 Example

We contrast the mtDNA CN calculated with duplicates removed with a corrected estimate for our example data in Fig. 7. In this figure we can see that, not only are the absolute values of mtDNA CN poorly estimated if all duplicates are removed, but the general reduction in copy number will shrink differences between groups, making comparisons less powerful (although a contrast as striking as blood vs tissue still shows a clear difference).

In calculating the coverages, we have assumed that the alignment process was well-behaved and that there are minimal biases - enabling a natural measure of coverage to be used. Note also, that since these are benign samples, we do not need to worry about tumour purity or ploidy in the calculations.

Applying the method to cancer samples is only marginally trickier, but as previously mentioned, the inference of nuclear copy number can often present multiple credible solutions. For example, distinguishing between a copy number of 2 and a high tumour purity and a copy number of 4 and lower tumour purity can sometimes be near impossible (consider sequencing a sample immediately following a nuclear genome doubling event). It should be noted that such uncertainty will naturally lead to a reciprocal change in the inferred mtDNA copy number, and so uncertainty in the nuclear copy number is something of which one should be aware. Although the ranges of mtDNA copy number for a cancer type typically range by more than a factor of two, the inferred mtDNA copy number may be a tool for distinguishing between competing nuclear copy number solutions.

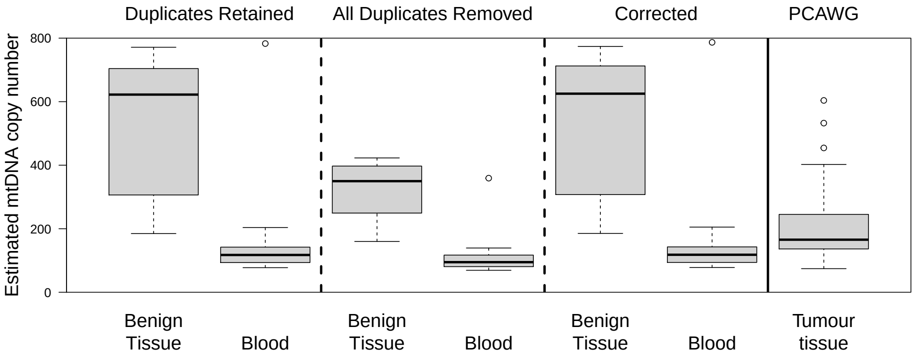

Fig. 7. Showing estimates of mtDNA CN in the gold standard scenario that no duplicates (excepting optical duplicates) are removed from either coverage estimate, the scenario that all duplicates are removed, and then also estimates corrected using the approach described above. For comparison, the estimates for OAC tumour tissue recently reported [15] are also shown.

6 Conclusions

We have set out a framework for estimating the PCR duplicate proportion in a WGS library. This we have argued from basic principles, but have also demonstrated to provide sensible and consistent results. We have updated our software [11], better to make these methods available. Our approach relies on being able to identify a subset of the genome where the PCR duplicate to fragmentation duplicate ratio is constant (i.e. regions of constant copy number state), and we require knowledge of the minor allele fraction (which should preferably be 0.5).

Should we not know the true minor allele fraction or should the number of cells being sequenced be such that removing one DNA fragment greatly changes the minor allele fraction, then these methods will be biased. For all reasonable experimental scenarios, their application would still be an improvement over attributing the basic observed duplicate proportion to be the PCR duplicate proportion.

There are implications too for quality control metrics that rely on the PCR duplicate proportion, e.g. sequencing library complexity. Complexity is an important metric, allowing comparison with previously sequenced libraries in order to detect out-of-control library preparation [7]. It also predicts the value of generating further sequencing from a sample, making it invaluable for experimental design (especially adaptive designs), and will be underestimated if fragmentation duplicates are not corrected for.

We have also shown that the estimation of DNA copy number is affected by the removal of fragmentation duplicates and that regions of high copy number can be severely affected. Additionally, we have demonstrated a computationally effective way to make corrections when the copy number estimation is a secondary analysis on WGS data for which duplicates have been removed.

7 Appendix: Proof that $\boldsymbol{K = (\sum_j \nu_j)!/\prod_j \nu_j!}$

The proof is by induction.

7.1 The Base Case ($\boldsymbol{M = 2}$)

If $M = 2$, we either have the case where two fragments have been observed from the same starting molecule (and so $\nu_2 = 1$, $\nu_j = 0$ for all $j > 2$), or we have the right-hand case where one fragment is observed from each of two original molecules (and so $\nu_1 = 2, \nu_j = 0$ for all $j > 1$). In the first case, $K = 1!/1! = 1$ and in the second $K = 2!/2! = 1$ as required.

7.2 The Assumption ($\boldsymbol{M = G - 1, G > 2}$)

For typographical convenience, we write $\nu_+ = \sum_j \nu_j$ in what follows. We assume that for all partitions where $M = G - 1$, that the relationship

$$a = \nu_+! \,/\, \prod_j \nu_j! \tag{11}$$

holds.

7.3 The Inductive Step ($\boldsymbol{M = G}$)

We assume that our partition of G duplicates is represented by the vector $(\nu_1, \nu_2, \nu_3, ...)$. We view the set of G duplicates as having arisen from a set of $G - 1$ fragments and then sequencing one more. We must distinguish between the two cases: 1) where the new duplicate in the set is the first from a previously unseen molecule (only possible if $\nu_1 > 0$), and 2) where the new duplicate is a further PCR duplicate from a previously seen molecule.

Case 1: A New Molecule If we have observed a new molecule with our Gth fragment then the previous set of $G - 1$ duplicates must have been represented by the vector $(\nu_1 - 1, \nu_2, \nu_3, ...)$. Clearly this is only possible if $\nu_1 > 0$ and, since observing a new molecule in this situation will always result in our observed partition, the full coefficient is inherited from the previous set (there will of course be a factor of F_D as well).

Hence the contribution to a from this case is

$$\frac{\mathbb{I}(\nu_1 > 0)\,(\nu_+ - 1)!}{(\nu_1 - 1)!\,\prod_{j>1} \nu_j!} \tag{12}$$

where $\mathbb{I}()$ is the indicator function.

Case 2: A PCR Duplicate from a Previously Observed Molecule In this case, the previous set of $G-1$ duplicates must have been represented by the vector $(..., \nu_{k-1}+1, \nu_k - 1, ...)$ for some k such that $\nu_k > 0$ and $k > 1$.

The coefficient, a', associated with that vector is

$$a' = \frac{\mathbb{I}(\nu_k > 0)\,\nu_+!}{(\nu_{k-1}+1)!\,(\nu_k - 1)!\,\prod_{j \notin k,(k-1)} \nu_j!}$$

but a new PCR duplicate added to that set might create patterns other than the one in which we are interested, so only a portion of the coefficient makes a contribution to our estimate of a. It would only have led to our observed pattern if the PCR duplicate had been of a molecule of which there previously existed $k-1$ copies. The fraction of the coefficient, a', that contributes to our value of a (not withstanding a factor P_D) is therefore the proportion of molecules of which there were previously $k-1$ copies: $(\nu_{k-1}+1)/\nu_+$.

The additive contribution to a for this value of k is therefore

$$\frac{(\nu_{k-1}+1)}{\nu_+}\frac{\mathbb{I}\,(\nu_k > 0)\nu_+!}{(\nu_{k-1}+1)!\,(\nu_k - 1)!\,\prod_{j \notin k,(k-1)} \nu_j!}$$

and in total the contributions from this second case are

$$\sum_{k>1}\left(\frac{(\nu_{k-1}+1)}{\nu_+}\frac{\mathbb{I}(\nu_k > 0)\nu_+!}{(\nu_{k-1}+1)!\,(\nu_k - 1)!\,\prod_{j \notin k,(k-1)} \nu_j!}\right). \tag{13}$$

Combining The Two Cases. If we combine the terms from the two cases as represented by expressions (12) and (13) then we get

$$a = \frac{\mathbb{I}(\nu_1 > 0)\,(\nu_+ - 1)!}{(\nu_1 - 1)!\,\prod_{j>1} \nu_j!} + \sum_{k>1}\left(\frac{(\nu_{k-1}+1)}{\nu_+}\frac{\mathbb{I}(\nu_k > 0)\,\nu_+!}{(\nu_{k-1}+1)!\,(\nu_k - 1)!\,\prod_{j \notin k,(k-1)} \nu_j!},\right)$$

which we can simplify by removing the terms in the first fraction on the right hand side, to get

$$a = \frac{\mathbb{I}(\nu_1 > 0)\,(\nu_+ - 1)!}{(\nu_1 - 1)!\,\prod_{j>1} \nu_j!} + \sum_{k>1}\left(\frac{\mathbb{I}(\nu_k > 0)\,(\nu_+ - 1)!}{(\nu_{k-1})!\,(\nu_k - 1)!\,\prod_{j \notin k,(k-1)} \nu_j!}\right)$$

and this can be tidied to

$$a = \frac{\mathbb{I}(\nu_1 > 0)\,(\nu_+ - 1)!}{(\nu_1 - 1)!\,\prod_{j>1} \nu_j!} + \sum_{k>1}\left(\frac{\mathbb{I}(\nu_k > 0)\,(\nu_+ - 1)!}{(\nu_k - 1)!\,\prod_{j \neq k} \nu_j!},\right)$$

Adjusting the products to be independent of 1 and k, we get

$$a = \frac{\mathbb{I}(\nu_1 > 0)\,(\nu_+ - 1)!\;\nu_1!}{(\nu_1 - 1)!\prod_j \nu_j!} + \sum_{k>1}\left(\frac{\mathbb{I}(\nu_k > 0)\,(\nu_+ - 1)!\;\nu_k!}{(\nu_k - 1)!\prod_j \nu_j!}\right).$$

Tidying up the other terms,

$$a = \frac{\mathbb{I}(\nu_1 > 0)\,(\nu_+ - 1)!\;\nu_1}{\prod_j \nu_j!} + \sum_{k>1}\left(\frac{\mathbb{I}(\nu_k > 0)\,(\nu_+ - 1)!\;\nu_k}{\prod_j \nu_j!}\right)$$

We can now combine everything into one sum over k:

$$a = \sum_k\left(\frac{\mathbb{I}(\nu_k > 0)\,(\nu_+ - 1)!\;\nu_k}{\prod_j \nu_j!}\right)$$

Moving the terms that are independent of k out of the sum,

$$a = \frac{(\nu_+ - 1)!}{\prod_j \nu_j!}\sum_k\left(\nu_k\mathbb{I}(\nu_k > 0)\right) = \frac{(\nu_+ - 1)!}{\prod_j \nu_j!}\;\nu_+,$$

whence

$$a = \frac{\nu_+!}{\prod_j \nu_j!}$$

as was to be shown.

8 Data Availability

The raw data are archived in the European Genome-Phenome Archive [EGA: EGAD00001000704]. Processed data and code to generate these values, alongside code to reproduce the figures in this text, have been added to the GitHub repository: https://github.com/dralynch/duplicates. DOI:10.5281/zenodo.6257577.

9 OCCAMS Consortium

For membership of the OCCAMS Consortium see [6].

References

1. Bansal, V.: A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. BMC Bioinformatics **18**(Suppl 3), 43 (2017)
2. Bogenhagen, D., Clayton, D.A.: The number of mitochondrial deoxyribonucleic acid genomes in mouse L and human HeLa cells. Quantitative isolation of mitochondrial deoxyribonucleic acid. Journal of Biological Chemistry **249**(24), 7991–7995 (1974)

3. Broad Institute: Picard toolkit. https://broadinstitute.github.io/picard/ (2019)
4. Calvo, S.E., Compton, A.G., Hershman, S.G., Lim, S.C., Lieber, D.S., Tucker, E.J., Laskowski, A., Garone, C., Liu, S., Jaffe, D.B., Christodoulou, J., Fletcher, J.M., Bruno, D.L., Goldblatt, J., Dimauro, S., Thorburn, D.R., Mootha, V.K.: Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. Science Translational Medicine **4**(118), 118ra10 (2012)
5. Castle, J.C., Biery, M., Bouzek, H., Xie, T., Chen, R., Misura, K., Jackson, S., Armour, C.D., Johnson, J.M., Rohl, C.a., Raymond, C.K.: DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing. BMC Genomics **11**, 244 (2010)
6. Frankell, A., Jammula, S., Li, X., Contino, G., Killcoyne, S., Abbas, S., Perner, J., Bower, L., Devonshire, G., Ococks, E., Grehan, N., Mok, J., O'Donovan, M., MacRae, S., Eldridge, M., Tavaré, S., The OCCAMS Consortium, Fitzgerald, R.: The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. Nature Genetics **51**(3) (2019)
7. Guo, Y., Ye, F., Sheng, Q., Clark, T., Samuels, D.C.: Three-stage quality control strategies for DNA re-sequencing data. Briefings in Bioinformatics (2013)
8. Heddi, A., Faure-Vigny, H., Wallace, D.C., Stepien, G.: Coordinate expression of nuclear and mitochondrial genes involved in energy production in carcinoma and oncocytoma. Biochimica et Biophysica Acta - Molecular Basis of Disease **1316**(3), 203–209 (1996)
9. Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**, 1754–60 (2009)
10. Lindberg, J., Mills, I.G., Klevebring, D., Liu, W., Neiman, M., Xu, J., Wikström, P., Wiklund, P., Wiklund, F., Egevad, L., Grönberg, H.: The mitochondrial and autosomal mutation landscapes of prostate cancer. European Urology **63**(4), 702–8 (2013)
11. Lynch, A.G., Smith, M.L., Eldridge, M.D., Tavaré, S.: Duplicates. https://github.com/dralynch/duplicates (2016). DOI DOI:10.5281/zenodo.6257577
12. Meng, S., Han, J.: Mitochondrial DNA copy number alteration in human cancers. North American Journal of Medicine and Science **6**, 22–25 (2013)
13. Vasta V., V., Ng B., S.B., Turner, E.H., Shendure, J., Hahn, S.H.: Next generation sequence analysis for mitochondrial disorders. Genome Medicine **1**(10), 1–10 (2009)
14. Weaver, J.M.J., Ross-Innes, C.S., Shannon, N., Lynch, A.G., Forshew, T., Barbera, M., Murtaza, M., Ong, C.J., Lao-Sirieix, P., Dunning, M.J., Smith, L., Smith, M.L., Anderson, C.L., Carvalho, B., O'Donovan, M., Underwood, T.J., May, A.P., Grehan, N., Hardwick, R., Davies, J., Oloumi, A., Aparicio, S., Caldas, C., Eldridge, M.D., Edwards, P.A.W., Rosenfeld, N., Tavaré, S., Fitzgerald, R.C., the OCCAMS Consortium: Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. Nature Genetics **46**, 837–843 (2014)
15. Yuan, Y., Ju, Y.S., Kim, Y., Li, J., Wang, Y., Yoon, C.J., Yang, Y., Martincorena, I., Creighton, C.J., Weinstein, J.N., Xu, Y., Han, L., Kim, H.L., Nakagawa, H., Park, K., Campbell, P.J., Liang, H.: Comprehensive molecular characterization of mitochondrial genomes in human cancers. Nature Genetics **52**(3), 342–352 (2020)
16. Zhou, W., Chen, T., Zhao, H., Eterovic, A.K., Meric-Bernstam, F., Mills, G.B., Chen, K.: Bias from removing read duplication in ultra-deep sequencing experiments. Bioinformatics **30**, 1073–1080 (2014)