JOHANNES KEPLER
UNIVERSITY LINZ

Author
Monika Heinzl

Submission
Institute of Bioinformatics

Thesis Supervisor
A**ssoz. Univ-.Prof.^{in} Dr.^{in}
Irene Tiemann-Boege**

Assistant Thesis Supervisor
Gundula Povysil, MD, PhD

May 2018

Development of algorithms for the analysis of duplex sequencing data

Bachelor's Thesis

to confer the academic degree of

Bachelor of Science

in the Bachelor's Program

Bioinformatics

**JOHANNES KEPLER
UNIVERSITY LINZ**
Altenberger Str. 69
4040 Linz, Austria
www.jku.at
DVR 0093696

Affidavit

Hereby, I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Linz, May 2018

Table of Contents

1. Abstract

Duplex sequencing overcomes limitations of conventional next-generation sequencing by tagging DNA molecules with double-stranded tags, which allows the detection of ultra-rare mutations that occur in less than 1% of cells. This method creates single-stranded consensus sequences (SSCSs) from the reads, which then form duplex consensus sequences (DCSs) that requires a large sequence coverage. In addition, during consensus assembly we are able to identify that a large part of the sequencing data is lost. We describe a new approach for the analysis of duplex sequencing data that focuses on improving the PE-reads/SSCS/DCS ratios and represents the data in a graphical way. First, we calculated the Hamming distance between the tags to distinguish whether similar tags truly stem from different molecules or occurred due to sequencing or PCR errors and, then we evaluated the effect of mismatch correction in the tag. Additionally, we showed the distribution of family sizes in order to identify any bias between forward and reverse strands during amplification, that could contribute to wasted sequencing capacities. Moreover, we developed a new approach for detecting chimeric reads, which are formed when two or more molecules are joined together during PCR. Our results have shown that allowing mismatches in the tags recovered a large amount of data and improved the SSCS/DCS ratio. Families that were only split due to sequencing errors in their tags were grouped together. Furthermore, our tools identified successfully chimeric reads (~28%) in the sequencing data. Fortunately, 80% of them occurred as singletons and would be filtered out during further data analysis. Analysing only those tags, which form DCSs, suggested that probably most of them originated from different molecules and only few of them occurred as chimeras. Finally, we were able identify different sources of read loss in the formation of DCSs and during trimming which has allowed the recovery of some of the sequencing data. All introduced tools can be used from the command line but can also be found in the Galaxy system.

2. Introduction

Next-generation sequencing (NGS) has become now routine in sequencing millions of nucleotides and in detecting mutations in a whole genome. These methods use data from single-stranded DNA fragments, but the high error rate (one false base call in 100-1,000 nucleotides) limits the detection of rare mutations that occur in fewer than 5% of the cells. Low frequency mutations, which are present in less than 1% of cells, can be identified with a high read depth in the sequencing run, however, this makes it difficult to identify true variations from false-positives. (Kennedy et al., 2014) Therefore, most human studies, limit the genetic information only to a small fraction when using high-throughput sequencing. The mutations may be present in these samples at lower frequencies than the error rate itself. (Lou et al., 2013) To overcome these high error rates, molecular tagging of single-stranded DNA before amplification allows the reduction of false-positives, but errors in the first rounds of PCR cannot be corrected. (Kennedy et al., 2014; Schmitt et al., 2015) A highly sensitive method, like droplet digital PCR, allows to check each polymorphic site at high sensitivity. However, the disadvantage of such technology is, that the sequence information needs to be known *a priori* since this technology does not retrieve sequence information. (Stoler, Arbeithuber, Guiblet, Makova, & Nekrutenko, 2016)

2.1. Concept of duplex sequencing

Duplex sequencing helps to understand biological substructures and to identify generation of mutations and rare variants, and this method plays also a major role in offering diagnostic accuracy required in precision medicine. (Fox, Reid-Bayliss, Emond, & Loeb, 2014) Limitations in sequencing single-stranded molecules can be overcome as DNA is double stranded. Duplex sequencing can identify and correct sequencing errors because it compares the sequence of tagged DNA fragments from one strand with the other part of the double stranded DNA fragment. (Schmitt et al., 2012) Primary advantage of duplex sequencing is the detection of single mutations among $<10^{-7}$ sequenced nucleotides. (Kennedy et al., 2014) In comparison, the probability of detecting a true mutation with NGS is 50% with an error rate of 10^{-2}. Thus, duplex sequencing increases the power and precision of high-throughput sequencing. (Fox et al., 2014) But due to its unique sensitivity, that detects mutations at ultra-low frequencies, the method is very costly. Duplex sequencing requires a much higher sequencing capacity than NGS to produce an appropriate sequencing coverage. (Kennedy et al., 2014)

This method uses unique tags to label each DNA fragment and can trace every PCR product back to its original fragment. Before amplification, each fragment is tagged by a random complementary dsDNA on both ends of the molecule. (Kennedy et al., 2014) Afterwards the reads are sorted by their tag into families, which are then categorized by their direction (forward and reverse reads) and grouped on the basis that the tags of two families match. Next, the reads of a family are aligned to themselves, which results in single-stranded consensus sequences (SSCSs). At this stage, sequence variants can be identified, because they are present in all reads and came from a single molecule; whereas, sequencing and amplification errors result as polymorphism within a family, since they are only to some degree within a family.

Afterwards, the SCSSs from the paired families are compared and a duplex consensus sequence (DCS) is generated and aligned to the reference genome to call the mutations. PCR errors from first rounds of amplification, which are still present in one of the SSCSs, can be identified and removed because they do not appear in both SSCSs. Mutations are considered as true mutations if they are present in both SSCSs; whereas, other sequence variants are probably sequencing or PCR errors. (Stoler et al., 2016)

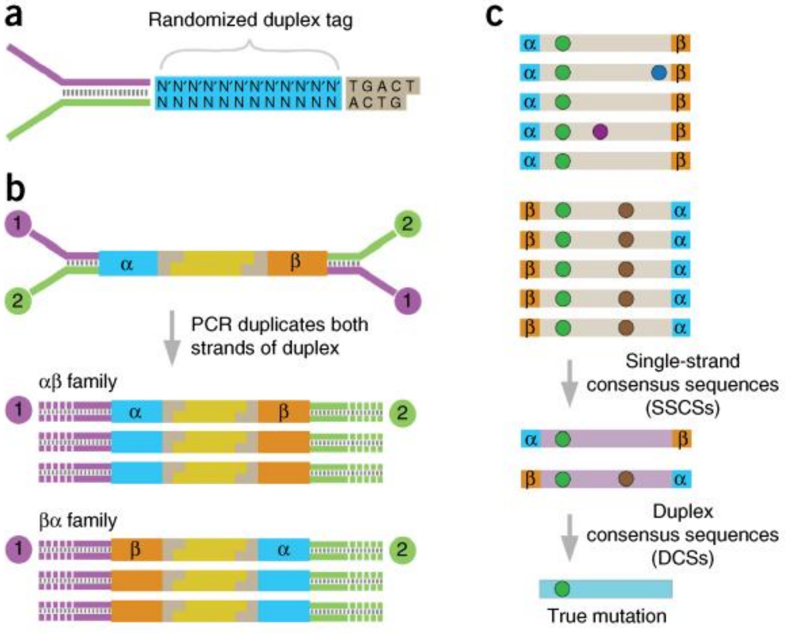

a
Randomized duplex tag
N'N'N'N'N'N'N'N'N'N'N'N' TGACT
NNNNNNNNNNNN ACTG
b
1
2
α
β
2
1
PCR duplicates both strands of duplex
αβ family
1
α
β
2
βα family
1
β
α
2
c
α
β
Single-strand consensus sequences (SSCSs)
Duplex consensus sequences (DCSs)
True mutation

Figure 1: Concept of Duplex Sequencing

(**a**) The duplex sequencing adapter consists of a random double-stranded tag. (**b**) The adapters are ligated on both ends of the DNA fragment and the tagged DNA samples are amplified which gives two related, but distinct PCR products. The reads are then sorted by their unique tags and grouped into αβ or βα families. (**c**) Next, each family is aligned and the SSCSs are formed and mutations (green dots), sequencing (blue or purple dots) or PCR errors (brown dots) can be identified. In this step, only sequencing errors can be removed, but not PCR errors. Finally, the SSCSs of both strands are grouped and the DCS is generated. Only mutations, which are present in both SSCSs, remain until this final step, whereas PCR errors are eliminated. (Kennedy et al., 2014)

Figure is adapted from (Kennedy, Salk, Schmitt, & Loeb, 2013)

2.1. Problems of duplex sequencing data

Duplex sequencing aims to produce an equal ratio of 2:1 between SSCSs and DCS, but more importantly to recover a substantial amount of DCSs from a certain number of reads. However, in reality, it is nearly impossible to produce equal amounts of SSCSs and DCSs. It could happen that not both strands of the molecule are amplified, but only forward or reverse strand. Also, in the final step where the SSCSs are combined to DCSs, the number of useful sequences is reduced. In theory, three reads per family are required to align them, but if there is a mutation present in one of the tags within a family, the read will be discarded. Due to this, it could happen that the family does not have enough reads and no SSCS is formed. Hence, no DCS is formed and the ratio between SSCSs and DCSs will deteriorate. (Kennedy et al., 2014)

Furthermore, a problem of duplex sequencing is the generation of chimeric products during amplification. Chimeras may result in genetic mixing, which then produces diversities new to the original sample and the detection of chimeric reads from real sequences remains a widespread problem. Several studies indicate that 30% of PCR products might occur as artificial chimeras. (Smyth et al., 2010) Highly similar sequences form most likely chimeras, which are highly difficult to remove from the data. (Smyth et al., 2010)

These artifacts can be produced (Kanagawa, 2003)

- **from an incomplete extended primer:**
  A chimeric product is being created, when an incomplete extended molecule acts as a primer and anneals to closely related templates. This kind of mechanism occurs during later stages of PCR. Since the amount of primers, which were incompletely extended, is too high and they can compete against the original primer during annealing. The frequency of chimeric reads can be limited by reducing the number of cycles during PCR.

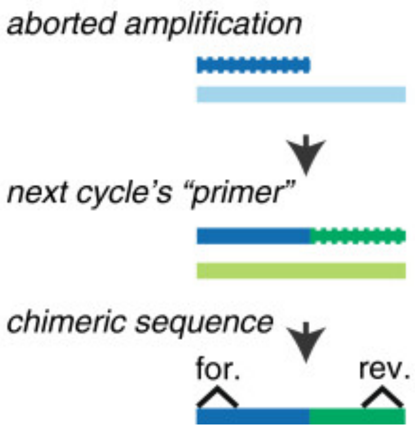


Figure 2: Chimeric reads formation

Chimeras are produced from fragments of various templates. A molecule was not fully extended and will act as a primer for a different template, which results in the formation of a chimeric read. Figure is adapted from (Fichot, 2013)

- **by template switching:**
  During extension, the extended strand switches from the original template to a sequence that has annealed toward the 3' end of the DNA fragment and leads to the formation of a chimeric read. When the extension of the primer gets close to the binding site of the other template, the process may become less efficient in this region and the primer switches the template. Higher temperatures during extension and a higher amount of repeated sequences favours the creation of chimeric reads. The binding of base pairs is less stable at higher temperatures and the 3' end from the extending strand becomes more easily free.

My Bachelor thesis focuses on these major problems of duplex sequencing by helping to understand the sequencing data and to find possible solutions for improving the amplification and sequencing protocol. We have developed new tools, that support the user to improve the ratio between SSCSs and DCSs. The main approach was to infer information about the tags by introducing a similarity measure. The Hamming distance quantifies similarity or dissimilarity between two DNA sequences of equal length by calculating the number of differences between them. (Wang, Kao, & Hsiao, 2015)

$$D_{i,j} = \sum_{k=0}^{n} [X_{ik} \neq X_{jk}]$$

so that $D_{i,j}$ is the number of sites where X_i match and X_j do not match,
k is the index of a particular site from a total number of sites n.
(Pinheiro, de Souza Pinheiro, & Sen, 2005)

This analysis was also performed with various tag lengths and only with tags, that form DCSs, to see if sequencing errors are filtered out during data analysis. In addition, the calculation of the Hamming distance can also be used to identify chimeric reads from real molecules. Since the loss of a huge amount of data is the main problem of duplex sequencing, tags obtained during different stages of the analysis have been represented by their family sizes. Finally, a graphic output was implemented to show the tag's family sizes and their Hamming distances.

3. Materials

The data in this report was obtained during the project by Prof. Tiemann-Boege and her research group at the Institute of Biophysics (Johannes-Kepler University, Linz), which focuses on the detection of selfish mutations in human sperm cells using duplex sequencing. Many mutations in humans are derived from transmissions of parents to their children. But also many affected individuals arise in each generation by inheriting *de novo* mutations. (Arnheim & Calabrese, 2009) Selfish mutations are *de novo* mutations, which have gain-of-function properties, that are under positive selection which leads to the expansion of mutant clones over time. These mutations are associated with paternal age effect (PAE) disorders, which means that the incidence rate of the mutations in the children increases with the age of the father. (Maher, Goriely, & Wilkie, 2014) Those mutations mostly originate in the father's germline, because there are more division after puberty in men than women. (Arnheim & Calabrese, 2009) Some disorders caused by these selfish mutations are achondroplasia, Apert, Noonan and Costello syndrome. The analysis of sperm DNA confirmed that these paternal age effect mutations are present above the mutation rate in most men. Selfish mutations from the male germline arise 1000-fold more frequently than the expected mutation rate. The selfish mutation rate lies at ~10^{-5} which is still too low for conventional NGS. (Maher et al., 2014)

This thesis includes the analysis of one library with tags labelling the sequencing reads for demonstrating the results of our developed algorithms. The tools expect the sequencing data with all tags before the alignment to the SSCSs. The input data should be in tabular format with information about family sizes, the tags itself and directions of the strands in which PCR was performed (ab = forward or ba = reverse strands). For analysing the read loss during data analysis, an additional text file with all tags, that were aligned to the reference genome and the regions of the reference genome, and FASTA files with tags after the formation of DCSs and after trimming are required.

4. Methods

4.1. Distribution of family sizes

At the beginning of the analysis of the sequencing data, the family sizes, which specify the number of reads per family, were graphically represented in a histogram. This family size distribution gives information about three important features of the sequencing data:

- Since by default three reads per family are required for the alignment to SSCSs, families with less than three members are ignored during the formation of the DCSs and useful sequencing data is lost during the data analysis. We show the number of singletons (=families with only one member), which gives an idea of how much data might be lost at the beginning of the analysis.
- Next, the existence of any bias during the formation of the reverse and forward strands and the amount of possible DCSs have been obtained by splitting the tags into three groups:
    - tags of forward reads (ab) with no partner,
    - tags of reverse reads (ba) with no partner and
    - tags of reads which form a DCS.
- Finally, the quantities of large families have been investigated, because they cause a waste of sequencing capacities. Families with too many reads form only few SSCSs and therefore less DCSs are created.

4.2. Calculation of Hamming distances

The second approach of my thesis includes the distinction, whether tags came from different molecules or from the same molecule but with sequencing errors or PCR errors in the tag, by calculating the Hamming distance. Since the whole dataset contained more than one million tags, the comparison of all tags would be computationally too demanding and would take a couple of days. Therefore, we parallelized the algorithms and sampled 1,000 tags and then compared them to the whole dataset to estimate the minimum Hamming distance between tags. We were able to verify that a sample of 1,000 tags gives an estimate for the whole dataset by calculating the Hamming distance for a sample of 10,000 and ~130,000 tags (supplementary material *figure 1*). We know that one duplex tag consists of 24 nucleotides, which gives 4^{24} unique tags for labelling the molecules (f*igure 2*). Since we have more possible tags than molecules, it is very unlikely that multiple molecules will share the same tags. Therefore, small Hamming distances occur very unlikely per chance and indicate sequencing or PCR errors.

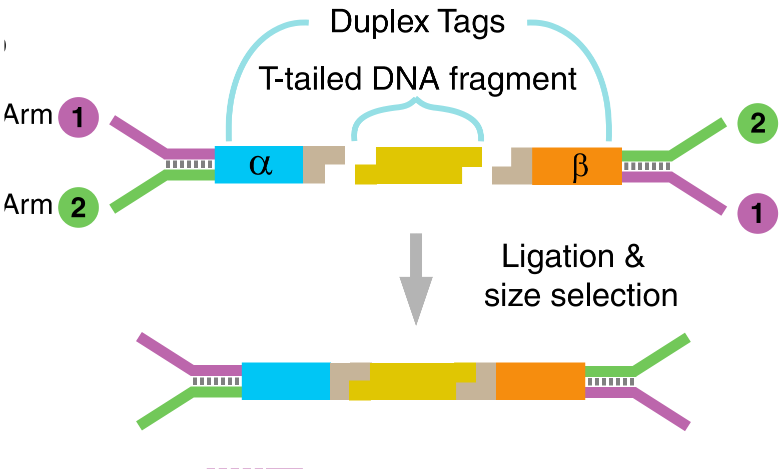


Figure 3: Schematic of a DNA fragment with ligated adapters

The adapters consist of a fixed number of nucleotides and are ligated to the DNA fragment on both ends. In our data, adapters with a length of 12 nucleotides were used, which should result in 4^{24} unique tags to characterize ~1 million DNA fragments.
Figure is adapted from (Schmitt et al., 2012)

4.2.1. Analysis of shorter tag lengths

The effect of shorter tags on the Hamming distance was tested by splitting the tags from its original length in half and shortening the halves to the desired length. The flanking regions may contain many repetitions, therefore the nucleotides at the end of each adapter (‘a’ and ‘b’) were skipped until the specified length of the tag was reached (*figure 4*). In this analysis we have used tag lengths of 13nt+13nt, 12nt+12nt, 10nt+10nt, 8nt+8nt and 6nt+6nt.

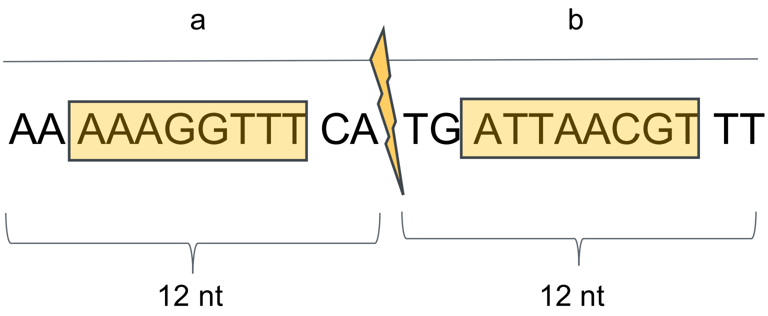


Figure 4: Workflow for shortening the tags

This is an example for shortening a tag with 24 nucleotides to 16 nucleotides. The tags were split into 2 halves (a and b) and the flanking regions of both halves will be skipped until the desired length for the whole tag is reached. Only the part of the tag, which is highlighted in yellow, is kept in the analysis.

4.2.1. Detection of chimeric reads

For identifying chimeric reads in the dataset, the Hamming distances of both halves of the tag were calculated. We have developed a method which is looking at the minimum Hamming distances of the individual parts of the tag and not at the minimum Hamming distance of the whole tag. First, each tag was split into its halves (‘a’ and ‘b’) and then we calculated the Hamming distance for the first part of the tag (‘a’) and looked for the minimum value afterwards. Since one tag can have multiple tags with the same Hamming distance in the dataset, the resulting sample size is much larger than initially selected. The Hamming distance for the second part of the tag (‘b’) was calculated by comparing the ‘b’ part of the sample to all tags with the same minimum Hamming distance of the first part. Finally, the same process was repeated starting with the second part of the tag to identify all possible chimeras.

Chimeric reads have normally very different Hamming distances within the tag and therefore, we suspected, that the absolute difference between those Hamming distances should be very large, which would make it possible to identify chimeras from real molecules. But at this point, we cannot tell for sure if the identified molecules are true chimeras, because we do not know whether the difference originates due to a low and a very large Hamming distance in both parts or one part of the tag is completely identical (HD=0) to a second molecule, which would represent a chimera. Therefore, we calculated the relative difference (=relative delta) by dividing the absolute difference by the sum of the Hamming distances of both halves.

Finally, the data can be grouped into three categories:

I. In theory, a low relative delta means that larger total Hamming distances were almost equal split up into partial Hamming distances. This case would be expected, if all tags have originated from different molecules (*figure 5a*).

II. Higher relative differences occurred either by a low total Hamming distance, which identifies tags that originated due to sequencing errors, and/or large absolute differences, that detects possible artificially introduced chimeras. (*figure 5b*).

III. Finally, true chimeric reads can be distinguished from true molecules since one of the parts is identical to another half, which occurs very unlikely per chance (*figure 5c*)

a small relative delta <= 0.5:	$\frac{small\ avsolute\ delta}{large\ total\ HD} \leq 0.5$ $\frac{2}{8} = 0.2$ e.g. HD_a = 3 and HD_b = 5 in the tag	**molecule**
b large relative delta > 0.5:	$\frac{small\ avsolute\ delta}{small\ \ total\ HD} > 0.5$ $\frac{3}{5} = 0.67$ e.g. HD_a = 1 and HD_b = 4 in the tag $\frac{large\ avsolute\ delta}{large\ total\ HD} > 0.5$ $\frac{10}{14} = 0.71$ e.g. HD_a = 2 and HD_b = 12 in the tag	**molecule** **or** **chimera**
c relative delta = 1	$\frac{absolute\ delta}{total\ HD} = 1$ $\frac{12}{12} = 1$ e.g. HD_a = 12 and HD_b = 0 in the tag	**chimera**

Figure 5: Formula for calculating the relative difference between the Hamming distances

The delta difference is estimated by dividing the absolute difference (=delta) between the HDs by the sum of the partial HDs. A small relative delta (<=0.5) indicates an equal distribution of the total Hamming distance between both parts of the tag, whereas a high relative delta (>0.5) might show chimeras formed during amplification or true molecules. Chimeras can be very likely detected if one half of the tag is identical with another half, but the second half is very different to the rest. This kind of tags can be identified when the relative delta equals exactly to 1.

4.4 Analysis of tags forming DCSs

The Hamming distance analysis can also be applied only on tags, which form DCSs later. The tags were filtered and only those tags, which occur twice in the dataset and have at least a family size of three, were kept. We were able to infer whether the DCSs originated from different molecules or were split into multiple families due to sequencing errors or PCR errors. Furthermore, we wanted to analyze where DCSs are lost during the bioinformatic mutation detection with the duNovo pipeline which was described here **https://github.com/galaxyproject/dunovo/wiki**.

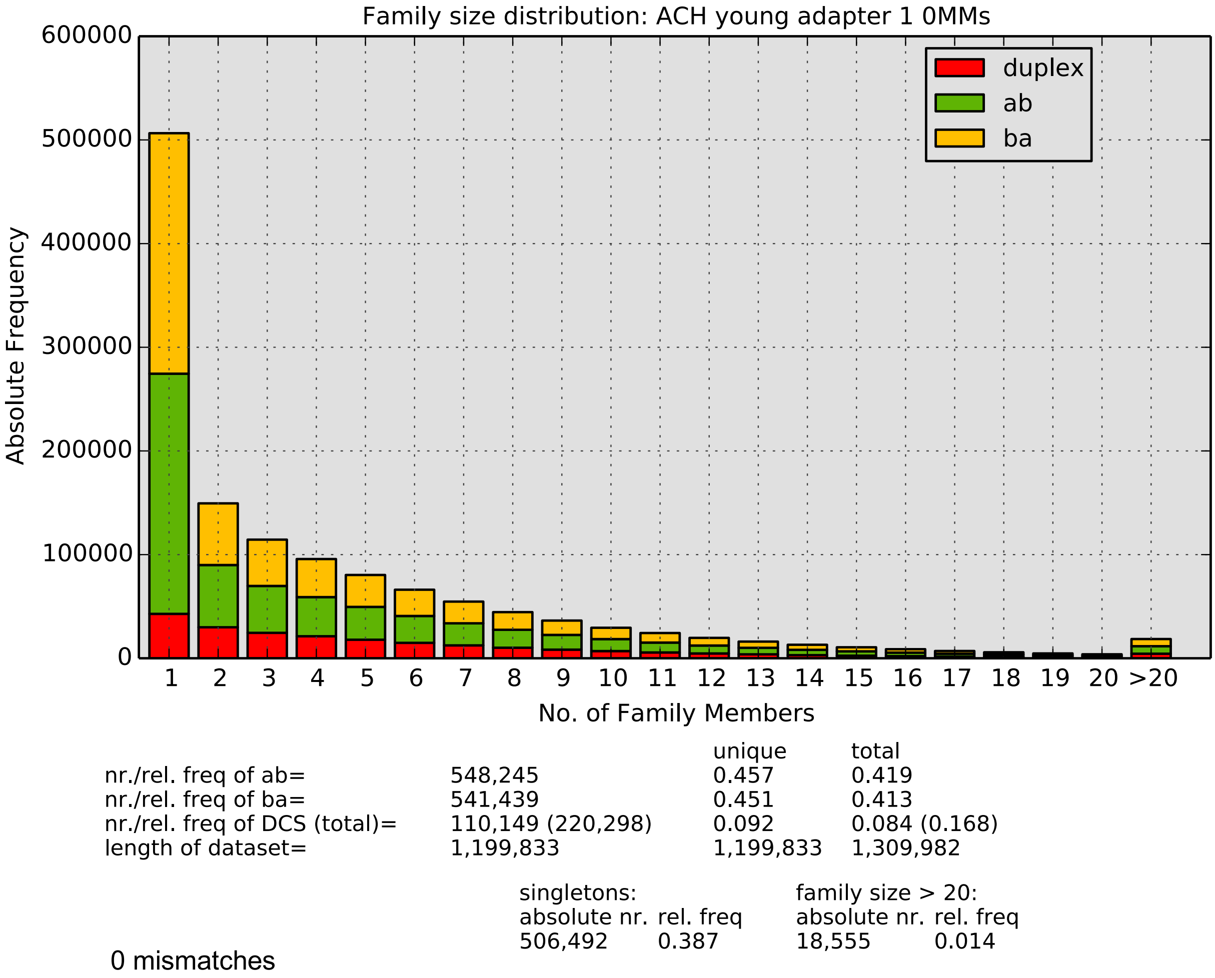
4.5 Detection of read loss during data analysis

We have stated in the beginning, that a lot of tags are lost during data analysis, but until now we were not able to see in which step of the analysis this happens. Possible reasons for data loss are small family sizes, reads with bad quality scores or too short read lengths after trimming. However, it is possible to fetch the data with the tags from various stages of the analysis and those datasets can be plotted in histogram with the read's family sizes. The different datasets were selected from the duNovo Galaxy pipeline which is used for reference free mutation calling of duplex sequencing data and includes the alignment of tags to SSCS, formation of DCSs and the actual mutation calling process. (Stoler et al., 2016) The most important steps, in which a lot of tags might be lost, are the alignment to SSCSs, the alignment to DCSs, trimming and the alignment to the reference genome.

5. Results

5.1. Representation of family sizes

In *figure 6*, the original tags of all SSCSs were grouped by their read family sizes and colored by their direction in which the strands were produced during PCR. We used this kind of distribution to see the grouping of reads into families and if there is any bias between the strands.



		unique	total
nr./rel. freq of ab=	548,245	0.457	0.419
nr./rel. freq of ba=	541,439	0.451	0.413
nr./rel. freq of DCS (total)=	110,149 (220,298)	0.092	0.084 (0.168)
length of dataset=	1,199,833	1,199,833	1,309,982

singletons:		family size > 20:	
absolute nr.	rel. freq	absolute nr.	rel. freq
506,492	0.387	18,555	0.014

0 mismatches

Figure 6: Family size distribution

The family sizes were separated after families that have only a forward (ab), reverse (ba) or both strands (duplex = DCS). Similar amounts of forward and reverse strands were recovered, which means that there was no strand bias in the formation of the SSCSs. In addition, families with large sizes (>20) contributed very little to the data. In theory, high amounts of SSCSs and DCSs should be produced. But in reality, only a small part of SSCSs formed DCSs (~9%) and most of the families are singletons (40%), which will cause a huge loss of useful sequencing data.

The plot suggests, that forward and reverse strands were almost equally amplified across all family sizes. The maximum number of reads within a family was very high (FS=206), but the family size distribution has shown, that the amount of families with more than 20 reads has contributed very little to our data (1.4%). The strands were separated after their direction and formation to DCS, which was expressed in total percentages and unique percentages. Since we were calculating the total numbers by counting both strands of the DCS and the unique numbers consist only of a single forward or reverse strand of the DCS, the percentages of the unique numbers are lower than the total numbers. But in the best case, only 9% of the SSCSs were present in both directions, which represent the DCSs (red part of the bars), and single forward and reverse strands contribute each with ~45% to all tags. As shown below, allowing mismatches in the tags can recover some singletons and merge them with their original family.

5.1. Calculation of Hamming distance of the tags

Next, we calculated the Hamming distance to measure the similarity between the tag of the dataset. First, we verified, that a sample of 1,000 tags is a representative sample for all tags of the data (supplementary material *figure 1*). Then the Hamming distance between 1,000 tags and the whole dataset was calculated to analyze the tags in detail. The minimum Hamming distances for each tag were graphically represented either as a histogram categorized after the family sizes or in a family size distribution separated by the Hamming distances.

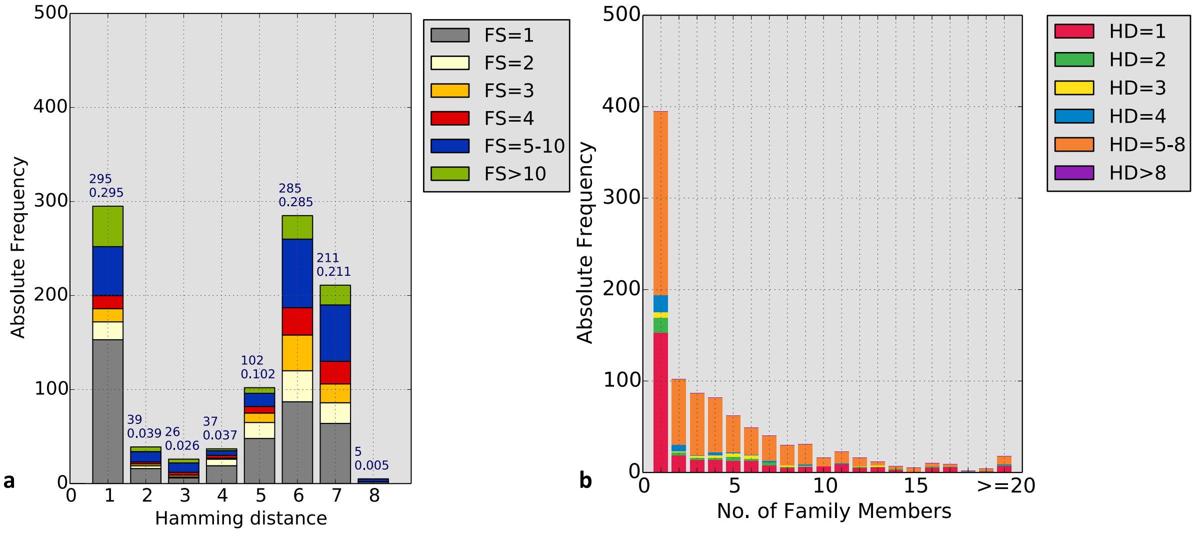


Figure 7: Hamming distance analysis of tags

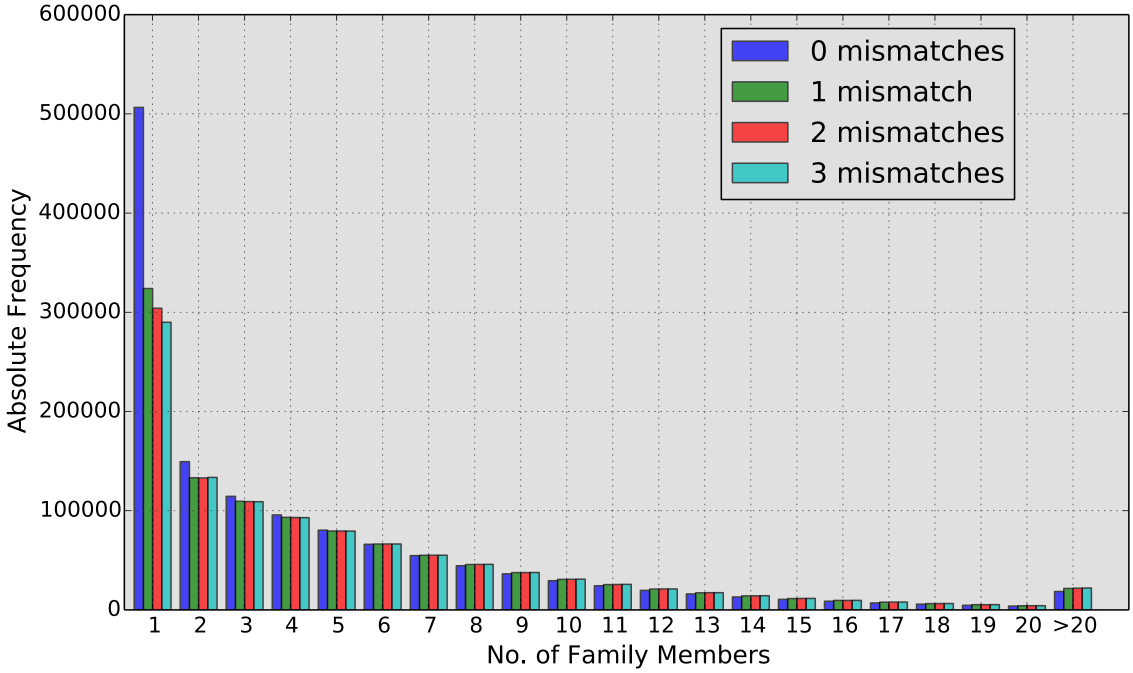
The Hamming distance was calculated for a subset of 1,000 randomly selected tags vs. all tags in the dataset. Afterwards the minimum value for each tag in the sample was calculated. The figure shows two possible ways to represent the Hamming distances: (**a**) Histogram with the Hamming distance, that is separated after the family sizes. 29.5% of the tags differ only by one nucleotide from the rest, which is very unlikely and, therefore, originated due to sequencing errors within the tags. Whereas high Hamming distances indicate, that the tags truly came from different molecules. (**b**) Family size distribution separated after the Hamming distances. Not only small Hamming distances contribute to the singletons, which can be recovered by mismatch correction, but also high Hamming distances (up to 8 mismatches), which indicate that the tags came from different molecules and are not different due to sequencing errors.

Most of our tags differed with at least five to eight nucleotides to the rest, but a big percentage of the tags differed by only one nucleotide (*figure 7a*). In addition, about half of the singletons differed by one to three mismatches, whereas larger family sizes consist of tags with bigger Hamming distances (*figure 7b*). Smaller Hamming distances are indicative of sequencing or PCR errors in the tags that cause families to be split up. These reads can be put into their original families by mismatch correction, which is described in detail in the next section.

5.1.1. Mismatch correction recovers tags with small family sizes

Mismatch correction recovers some of these reads by allowing mismatches in the tags and places the reads in other families, which in turn increases the sizes of some families, but reduces the number of different tags / families in the whole dataset.

In the protocol used for the data, one, two or three mismatches were allowed in the tags and the effects of mismatch correction can be observed in *figure 9* on the Hamming distances of a sample of 1,000 and in *figure 8* on the whole dataset. Allowing mismatches in the tag has reduced the number of singletons that would be lost in further steps of the analysis and has placed them mostly into families with at least seven and more members. The fraction of singletons has dropped over 10 percent, when comparing the dataset before mismatch correction (~39%) with after mismatch correction, where three mismatches have been allowed (~27%).



	singletons:		family size > 20:		
	absolute nr.	rel. freq	absolute nr.	rel. freq	total nr. of families
0 mismatches	506,492	0.387	18,555	0.014	1,309,982
1 mismatch	323,913	0.290	21,629	0.019	1,118,788
2 mismatches	304,054	0.276	21,856	0.020	1,099,732
3 mismatches	289,912	0.267	22,072	0.020	1,086,642

Figure 8: Family size distribution of the whole dataset with mismatch correction

The figure shows the family sizes of the whole dataset with no, one, two or three mismatches allowed in the tags. Allowing mismatches reduced the number of small families, especially the singletons (FS=1, from 38.7% to 26.7%), but increased the number of larger families with at least 7 members (FS>=7).

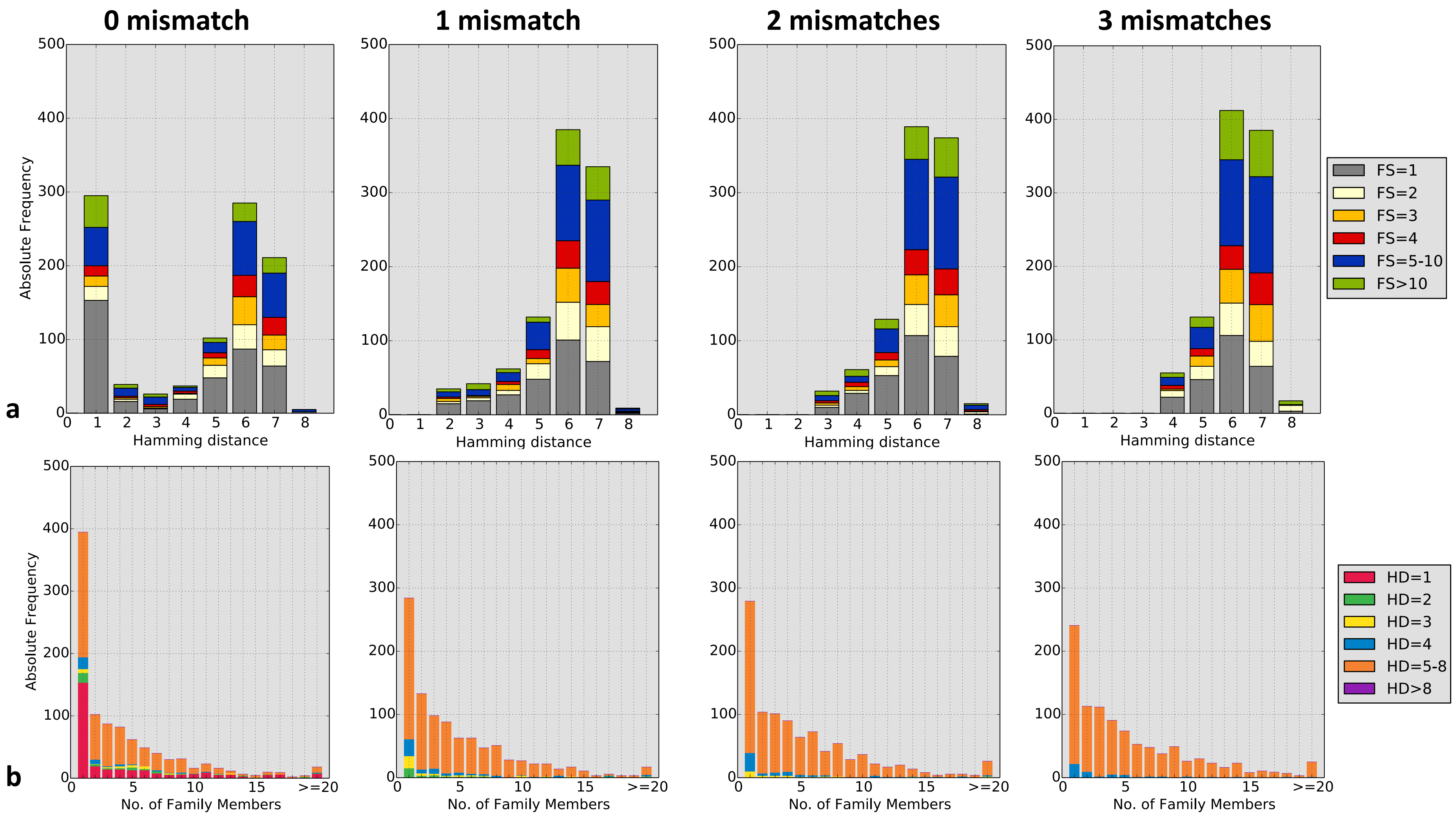


Figure 9: Effect of mismatch correction on Hamming distances of a sample

The effects of mismatch correction with no, one, two or three mismatches in the tags on the Hamming distances are shown. (**a**) 29.5% of the tags differed by only one nucleotide in the tag, and therefore might have occurred due sequencing errors. Those reads can be recovered by allowing up to three mismatches, which has placed the reads into larger families (mainly into families with 5-10 or more than 10 members). (**b**) The family size distribution of mismatch correction shows that a big part of the singletons was recovered. However, most of the singletons still differed with at least 4 mismatches to the rest and they probably originated from different molecules.

Overview of DCSs and SSCSs in the data

DCS (FS >= 1)		
	nr. of duplex	**rel. frequency**
without MM correction	110,149	9.2%
1 MM	116,091	11.6%
2 MMs	124,293	12.7%
3 MMs	136,240	14.4%

SSCS without partner (FS >= 1)		
	nr. of ab strands	**rel. frequency**
without MM correction	548,245	45.7%
1 MM	446,798	44.6%
2 MMs	429,128	44%
3 MMs	410,785	43.2%
	nr. of ba strands	**rel. frequency**
without MM correction	541,439	45.1%
1 MM	439,808	43.9%
2 MMs	422,018	43.3%
3 MMs	403,377	42.4%

Table 1: Overview of DCSs and SSCSs in the data

In this table, the fractions of SSCSs with no partner sequence in both direction of amplification and the DCSs are shown. Before mismatch correction only 9.2% of the data formed DCSs, whereas after allowing three mismatches the DCSs have raised to 14.4%.

The numbers in *table 1* and *table 2* allow all family sizes in the SSCSs, because we wanted to estimate the highest possible amount of DCSs without filtering the SSCSs after their family sizes. Since the number of singletons decreased with mismatch correction, the ratios between SSCSs and DCSs has improved too, when allowing all family sizes (FS>=1) in the alignment. Before mismatch correction only 9.2% of the strands formed DCSs, but if we would allow one, two or three mismatches in the tag, 14% of the SSCSs could form duplex sequences.

Ratios between SSCSs and DCSs

	DCS*2	SSCS	Ratio SSCS/DCS
without MM correction	220,298	1,089,684	11.9
1 MM	232,182	886,606	9.6
2 MMs	248,586	851,147	8.8
3 MMs	272,480	814,162	8
optimum	1	2	2

Table 2: Ratios between SSCSs and DCSs

The ratios in this table were estimated by dividing the SSCSs through the DCSs. The ratio should be around 2 but can only be achieved in a sequencing run with no errors in the tag and successful amplification of both strands. After mismatch correction, the ratio of our data was reduced from 11.9 to 8, but still exceeds the optimal ratio.

Without mismatch correction, the ratios between SSCSs and DCSs were far too high (~12). After allowing three mismatches the ratios decreased from 12 to 8, but still they were four times higher than the optimal ratio (*table 2*).

5.1.1. Effect of tag length on Hamming distance

Usually, tags with a length of 2*12 nucleotides are used in duplex sequencing. Here, we have tested the effect of shorter and longer lengths of tags on the Hamming distance (*figure 10*). The following tag lengths were tested: 2*6nt, 2*8nt, 2*10nt, 2*12nt and 2*13nt resulting in a total length of 12nt, 16nt, 20nt, 24nt, 26nt, respectively. The data did not allow any mismatch in the tag, since this experiment was simulated only *in silico* by selecting a subset from the whole tag and mismatch correction is usually performed on the whole tag. Therefore, we are not interested in the effect of mismatch correction on a shorter tag length. The experiment for testing the tags with 26nt was obtained from the laboratory, therefore, the sizes of the datasets might vary.

The results showed that a shorter tag length (length=12nt, length=16nt, length=20nt) resulted in a much smaller minimal Hamming distance (HD=1-2, HD=4-5, *figure 10*). Whereas bigger Hamming distances (HD=6-7, HD=6-8) were estimated from the experiments with longer tags (length=24nt, length=26nt, *figure 10*). Very short tags with a total length of 12 nucleotides have Hamming distances of one or two (HD=1-2). Therefore, tags with small Hamming distances in one of its halves (if tag length is 2*12nt) are not unusual and can arise by chance.

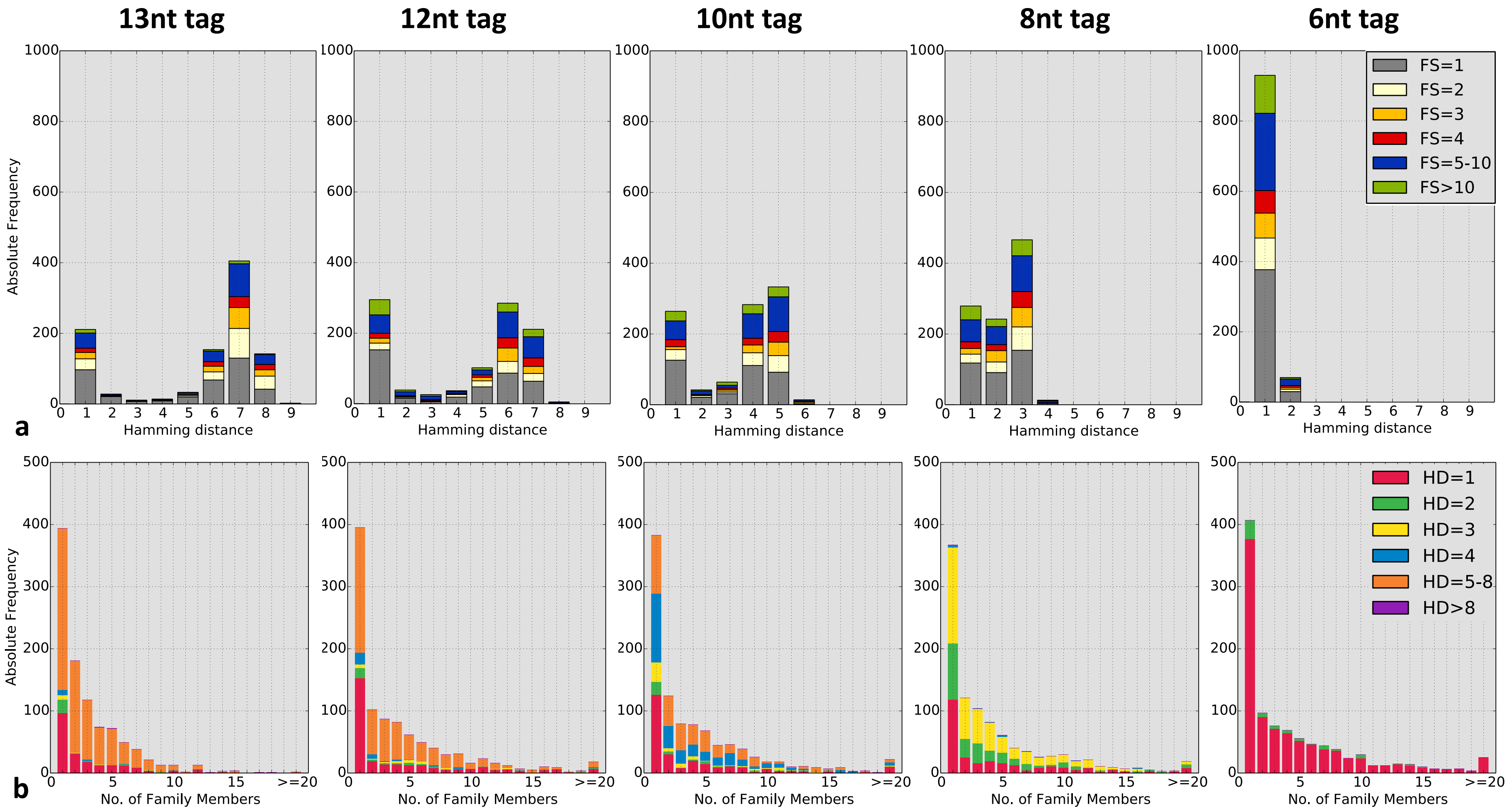


Figure 10: Effect of tag length on Hamming distances of tag

Hamming distances from various tag lengths were calculated (**a**) and their family sizes separated after their Hamming distances (**b**) are shown. The original tag (2*12nt) was shortened to lengths of 2*6nt, 2*8nt and 2*10nt. For longer tags (2*13nt) we had data from the laboratory available. The shorter the tags were, the smaller Hamming distances occurred.

5.1.2. Analysis of chimeric reads

The halves of the tags in the test data consisted of 12nt, therefore, we expected the same results as in *section 5.1.1*, where we used a total tag length of 2*6nt. Reasons for differences between the results are caused due to amplification errors, sequencing errors or the formation of chimeric reads. We developed algorithms for detecting tags of chimeric reads in the data by using the Hamming distance of both halves in the tag (*figure 11*). More tags (n=12,474 and n=13,840) than specified in the beginning (n=1000) resulted from this analysis, because each tag in the sample can have multiple tags with the minimum Hamming distance in the whole dataset.

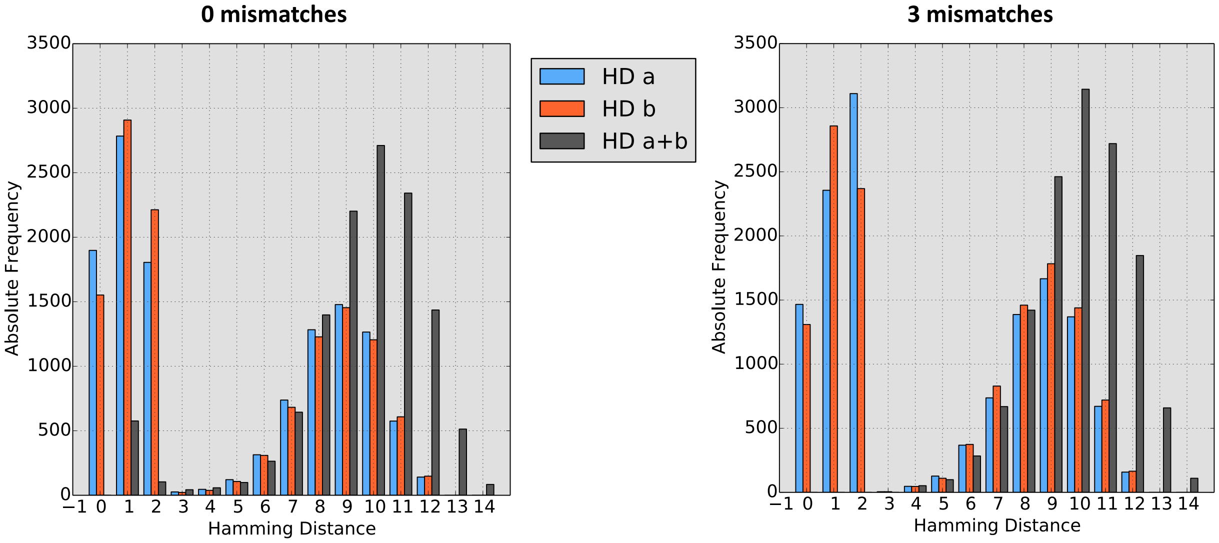


Figure 11: Hamming distances of both parts of the tag

The left plot does not include any mismatch correction, whereas the right plot allows three mismatches. The sample sizes (n=12,474 and n=13,904) are much larger than specified in the beginning (n=1,000), since one tag can have multiple tags with the same minimum Hamming distance. The Hamming distances of the halves (each with 12 nucleotides) of the tags were calculated (blue and orange). One half has a much smaller Hamming distance than the other half, because we were looking for minimum Hamming distance of the half, that we had started the analysis with.

The grey bars ('a' + 'b' halves) in the plot indicate the sum of Hamming distances from both halves of the tag, whereas orange (part 'a') and blue (part 'b') bars show the Hamming distances of the halves. The tags were selected based on the minimum Hamming distance of the individual halves, which in turn gives much smaller Hamming distances (HD=0-2) in one part and very high Hamming distances in the second half (HD=8-10). In other words, the blue bars with the smaller Hamming distance and the red bars with the larger Hamming distances form one whole tag with an overall Hamming distance and vice versa. The results showed that without mismatch correction a lot of tags with Hamming distances of zero in one half resulted from the analysis. Whereas the plots with mismatch correction suggested that the number of halves with zero Hamming distance was reduced but Hamming distances of two and larger was increased in both halves.

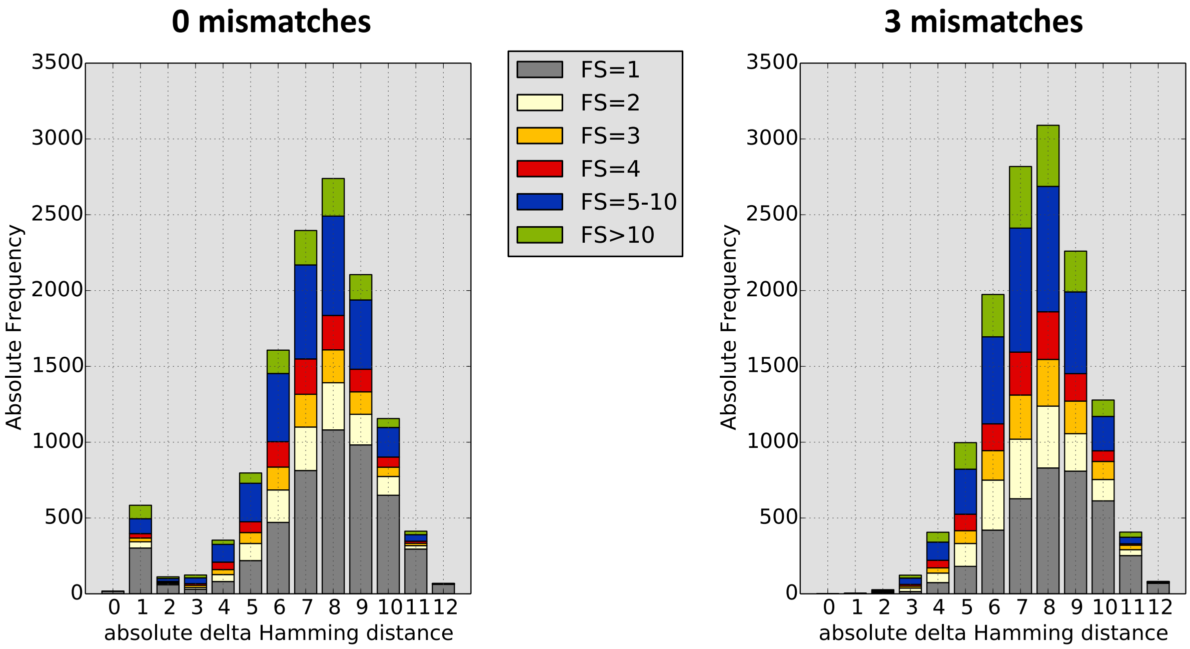


Figure 12: Absolute difference between the Hamming distances within the tag

The histograms contain the absolute differences between the Hamming distances of both halves of the tag. The sample sizes (n=12,474 and n=13,904) are much larger than specified in the beginning (n=1,000), since one tag can have multiple tags with the same minimum Hamming distance. Large differences indicate an unequal splitting of the whole Hamming distance. Differences of 12 are very unlikely to occur per chance and therefore, might be a sign for chimeric reads produced during PCR.

Next, the difference between the Hamming distance of both halves was calculated (=absolute delta HD). In *figure 12*, we can already see, that most of our tags were composed of very different halves (delta=7-9 with 0 and 3 mismatches). We suspect, that tags with differences of 11 or 12 were artificially introduced by the means of chimeric reads. Very large differences indicated completely different halves within one tag. However, this measure was not very informative regarding the distribution of the Hamming distances within the tag. Thus, we calculated the relative difference (relative delta Hamming distance) by dividing the absolute difference by the sum of Hamming distances of both halves (=total Hamming distance).

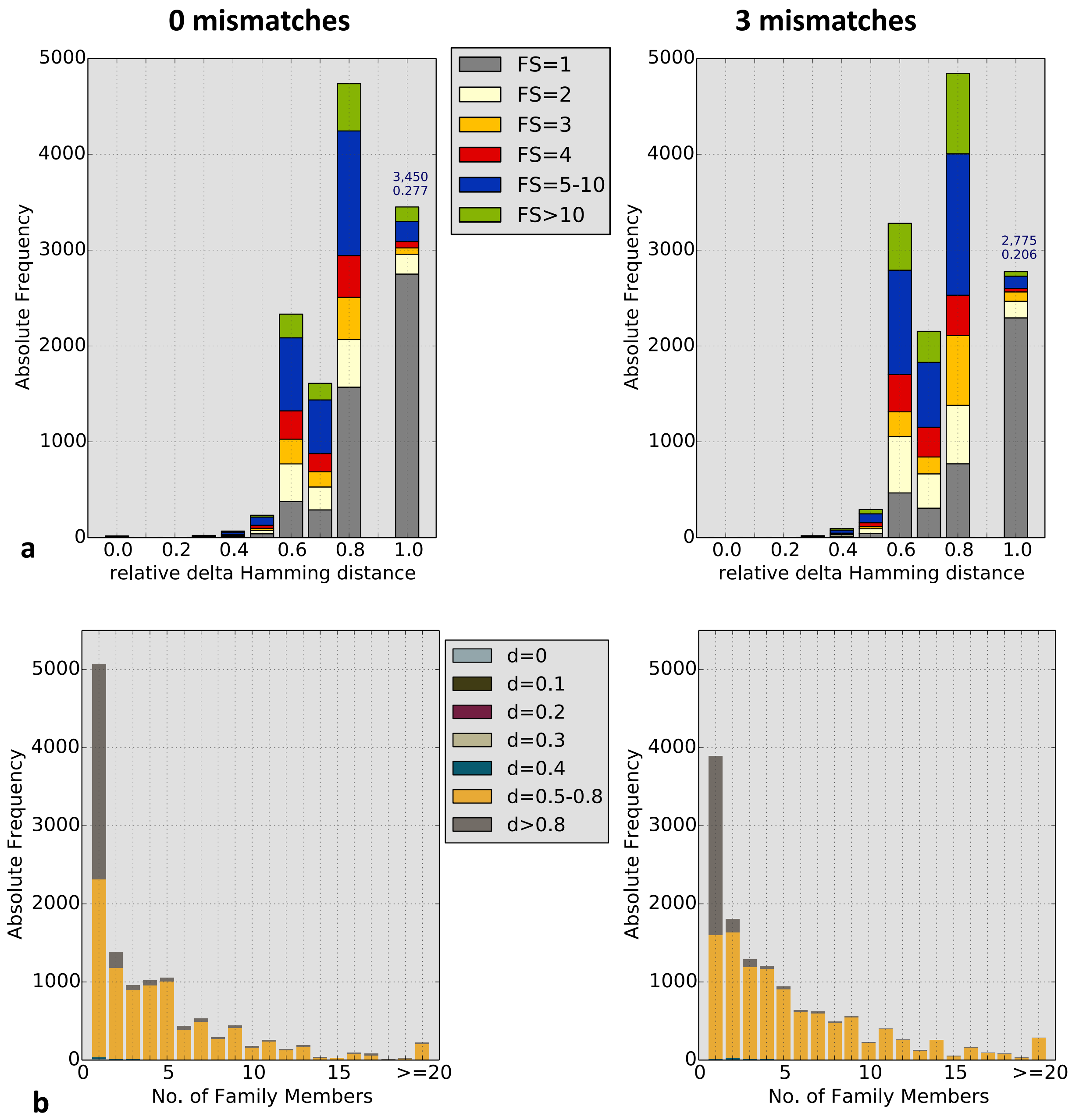


Figure 13: Relative difference between the Hamming distances within the tag

(**a**) Histogram of the relative differences between the Hamming distances. Differences greater or equal than 0.5 might indicate chimeric reads but can also be true molecules. Whereas a relative difference of 1 are very likely chimeric reads. In those tags one half is identical to another half (HD=0) and the other part is completely different to the rest.
(**b**) Family size distribution separated by the relative difference: Most of the singletons have a high relative difference and might indicate chimeric reads, but fortunately they will be filtered out of the dataset later.
The sample sizes (n=12,474 and n=13,904) are much larger than specified in the beginning (n=1,000), since one tag can have multiple tags with the same minimum Hamming distance.

Most of our tags in our sample had a very large relative difference (delta=0.6-0.8). About 28% of our tags without mismatch correction were composed of one completely identical half (HD=0) and one completely different half resulting in a relative difference of one. After allowing three mismatches this percentage drops to ~21% (*figure 13*). In both cases, the singletons, which are skipped later when forming the DCSs, contributed most to the tags with a relative difference of more than 0.5. But still some families with larger sizes have a high relative difference (>0.5).

Finally, we analyzed those tags with a relative difference of one in *figure 14* in more detail by calculating the Hamming distance of the non-identical half, which will represent the total Hamming distance of the whole tag.

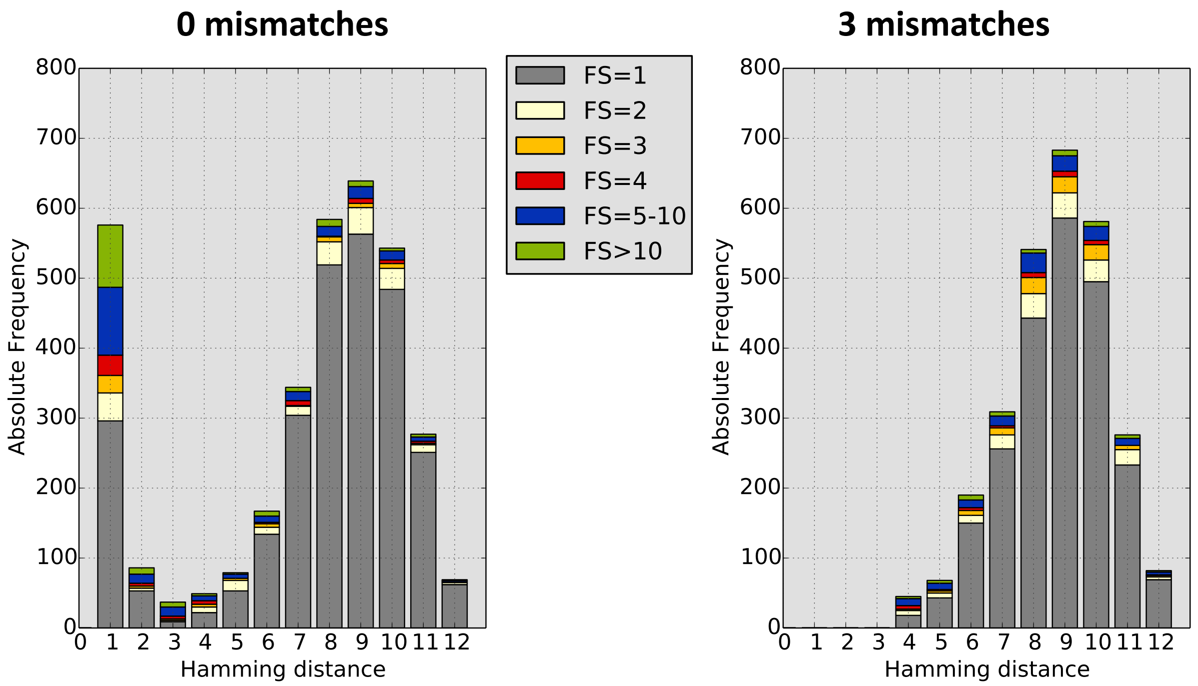


Figure 14: Hamming distances of the chimeric reads

The tags in these plots represent the fraction of tags, where the relative difference is one (0MMs: nr of tag=3,450, 3MMs: nr. of tags=2,775). Tags, which indicate chimeric reads, have one identical half (HD=0), whereas the other half is completely different to the rest (HD=8-11). Mismatch correction reduces the fraction of tags which occurred due to sequencing errors in the tag. Especially the tags with 0-3 mismatches completely vanished from the dataset.

Without mismatch correction a lot of tags had very small Hamming distances (HD=1 in 16.7% of tags with one identical half), which will explain the high fraction of tags with a Hamming distance of one in the data, where we analyzed the whole tag. But still most of the tags had very large Hamming distances. In other words, one half had a Hamming distance of zero whereas the other half had a very large Hamming distance (HD=8-10). Most of those tags (~83%) with one identical half had a family size of one (*figure 15*) and therefore, do not contribute to DCSs when at least three reads per family are required.

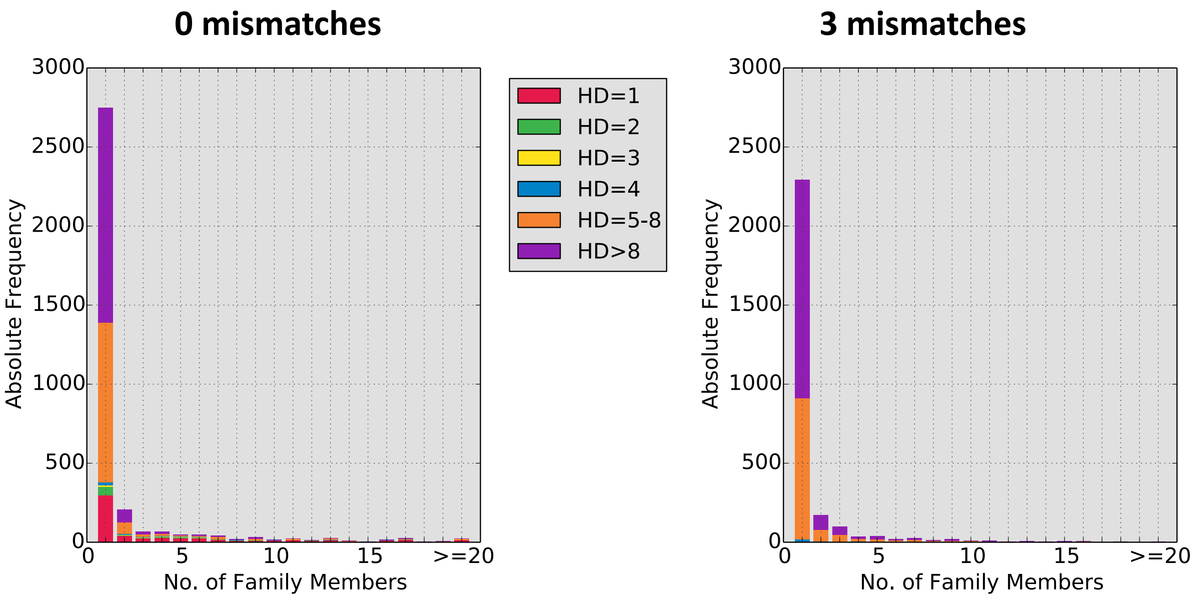


Figure 15: Family size distribution of chimeric reads

The family sizes of the chimeric reads are represented in a histogram. Most of the chimeras are singletons, which are filtered out of the dataset, when forming the SSCSs, because at least tags with family sizes of 3 reads are required by default.

5.3 Analysis of tags forming DCSs

Usually, DCSs are comprised of two identical tags from a forward and reverse strand with a family size of at least three reads. In order to find out whether the errors mentioned above are also present in the final DCSs, we performed the Hamming distance analysis for the whole tag (24nt), its family size distribution and the Hamming distance of the individual halves within the tags only for those tags that can form DCSs. *Figures 16-18* contain the forward and reverse strands of the DCSs. Both strands have equal Hamming distances, since the tag is the same, but the family sizes can differ between the strands. Mismatch correction for three mismatches produced the best results for the whole dataset, therefore the analysis was performed on the dataset, where three mismatches have been allowed.

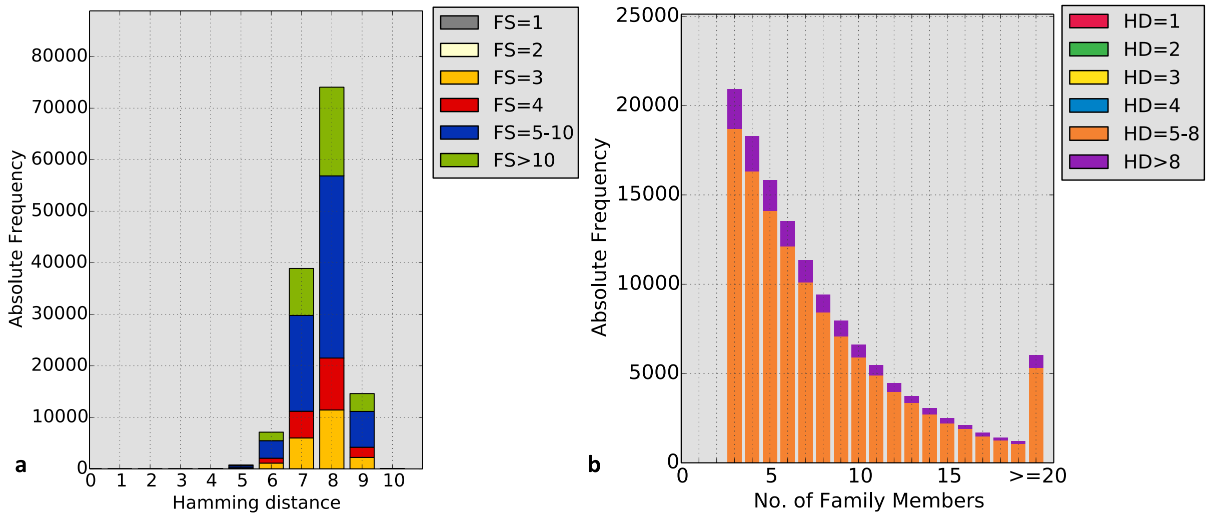


Figure 16: Hamming distance analysis of DCSs

Only tags (n=67,786), which form DCSs, were used for this analysis and were compared against all of them (n=67,786).
(**a**) The Hamming distances are mostly between 7 and 9, they very likely originated from different molecules and are not different due to errors in the tag. (**b**) Since by default a minimal family size of 3 is required, smaller family sizes cannot be observed.

All families with less than three members are not part of the analysis, but most of the tags had family sizes larger than five (*figure 16b*) and the sample tags differed with around seven to nine nucleotides to the rest (HD=7-9, *figure 16a*). Therefore, most of the erroneous tags were filtered out during data analysis and the remaining tags of the DCSs stem from different molecules. Additionally, we were interested if there is still a problem with chimeric reads.

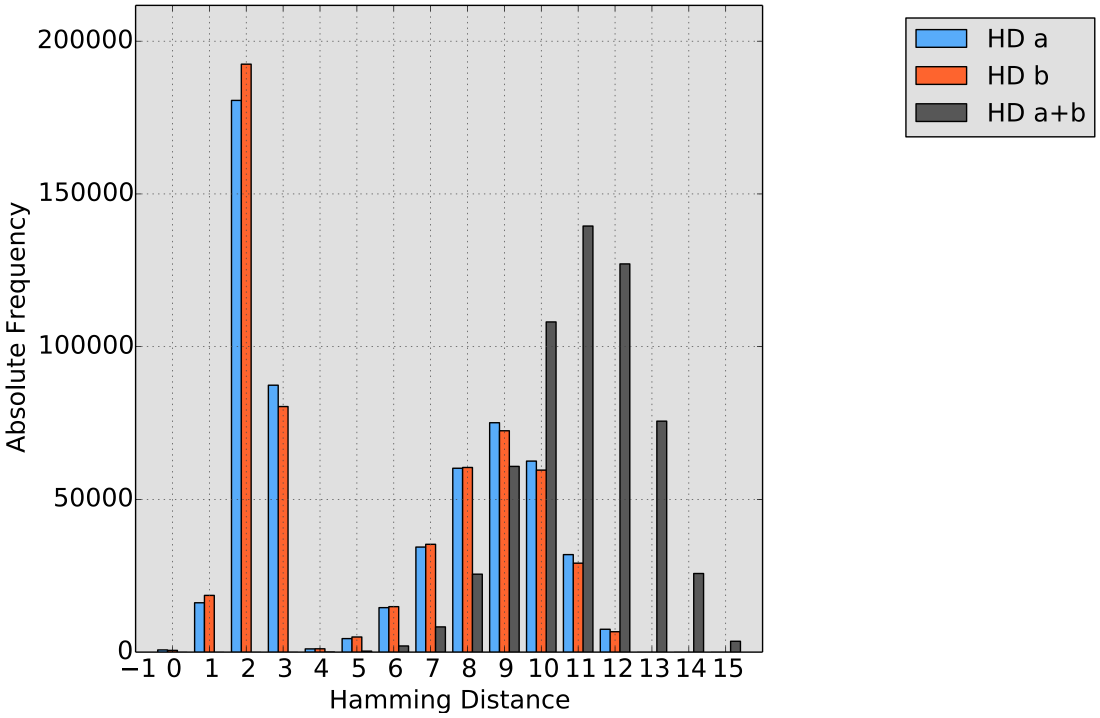


Figure 17: Hamming distances within the tags of DCSs

The Hamming distances of both halves of the tag were individually calculated (blue and orange). The total number of tags (n=576,524) in this analysis is much higher than initially specified (n=67,786), since one tag can have multiple tags with the same minimum Hamming distance. The grey bars represent the sum of the Hamming distances of both halves. Hamming distances of zero are very rare in DCSs. But still, there are a lot of tags with small Hamming distances (HD=1-3) in only one part of the half.

The Hamming distances for both parts of the tags were calculated. Compared to the whole dataset, very few halves of the tag were identical to each other (HD=0), but rather differed with two to three nucleotides (HD=2-3) in one half and eight to ten nucleotides (HD=8-10) in the other part. As expected, the sum of the partial Hamming distances in the *figure 17* shifted to much higher values than the total minimum Hamming distance of the whole tag (*figure 7a*).

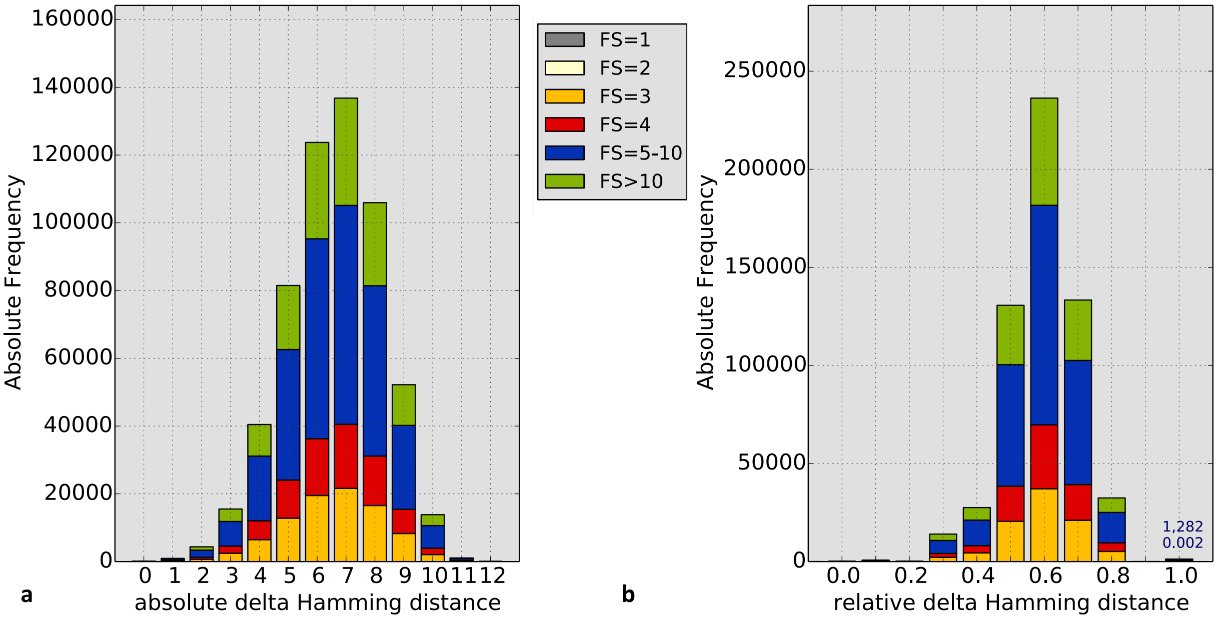


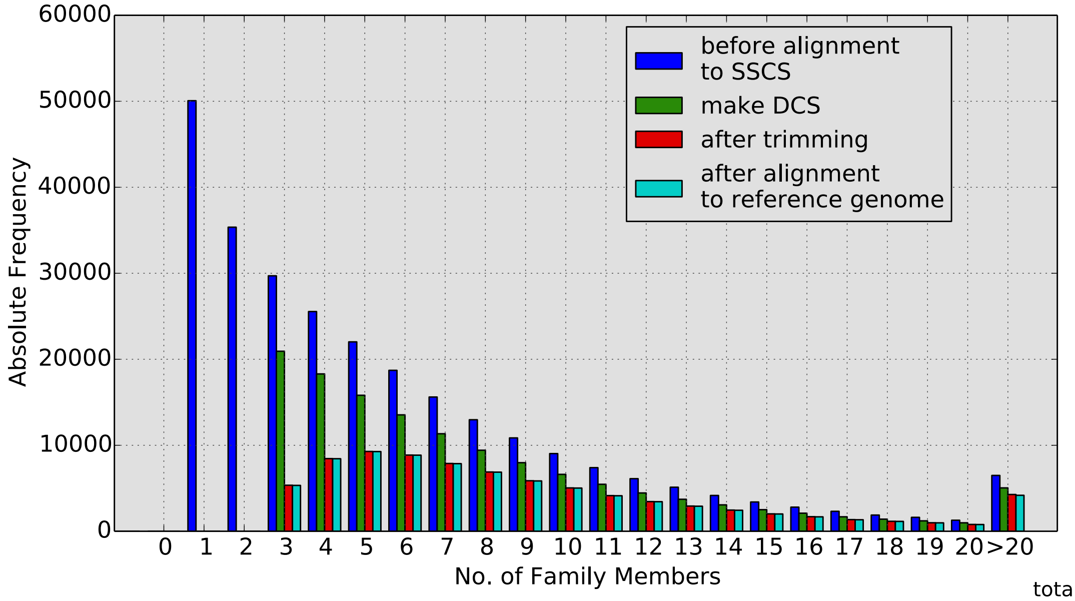
Figure 18: Absolute and relative differences of tags of DCSs

The total number of tags (n=576,524) in this analysis is much higher than initially specified (n=67,786), since one tag can have multiple tags with the same minimum Hamming distance. **a**) Absolute differences between the halves of tags of DCSs: Differences of 11-12, which might indicate chimeric reads, occur very rarely in DCSs. (**b**) For verifying the chimeric reads, the relative differences were calculated. Only 0.2% of the DCSs are chimeric reads, because they have a relative difference of 1, which means one half of the tag has a Hamming distance of zero.

The left plot (*figure 18a*) shows the absolute differences between the Hamming distances of both halves of the tag. We can already see very large differences (HD=11-12), which are a typical sign for chimeric reads, almost disappeared from the data with tags of DCSs. For verifying the identity of chimeric reads, the right plot (*figure 18b*) shows the relative differences where the fraction of chimeras has dropped to 0.2% of the data.

5.4 Analysis of read loss during the Galaxy pipeline

The distribution of family sizes contains all original tags before the alignment to the SSCSs, after the alignment to the DCS, after trimming and after the alignment to the reference genome. The reads were trimmed at their 3' end and based on the frequency of filter bases, which were defined in the experimental protocol. Reads with less than 200 nucleotides were omitted from the further analysis and only kept if both strands of the DCS reached the specified read length. *Figure 19* contains the data, where three mismatches have been allowed, and both family sizes of reverse and forward strand, whereas the total numbers below the plot indicate the single counts of the DCS.



				total numbers
	AB	BA	total nr. of tags (unique, FS>=1):	950,402
max. family size =	86	116	DCS (before alignment to SSCS, FS>=1):	136,240
absolute frequency=	1	1	total nr. of tags (unique, FS>=3):	570,008
relative frequency=	0.00002	0.00002	DCS (before alignment to SSCS, FS>=3):	67,786
			make DCS:	67,786
total nr. of reads (SSCS)	5,764,559		after trimming:	41,460
			after alignment to reference genome:	41,267

Figure 19: Analysis of read loss

Family size distribution with various stages of the analysis with correction for 3 mismatches: The datasets were obtained before the alignment to SSCSs, after the alignment to DCSs, after trimming and after the alignment to the reference genome. The plot contains both counts of forward and reverse strand, whereas the numbers below represent the single count of the DCS. Most of data has been filtered out during the formation of DCSs and during trimming.

The numbers before alignment to SSCSs have been calculated by our programs, which search for all tags that have a duplicate in the dataset and have at least a family size of three. If we compare these values to the numbers obtained after the alignment to the DCSs, then, as expected, both counts agree. However, the numbers of the tags aligned to the reference genome was reduced to ~41,000 tags, which means that 26,519 tags were lost during the trimming step.

Therefore, we adapted the minimal read length parameter of the duNovo Galaxy pipeline (Stoler et al., 2016) to reduce the number of filtered reads.

Comparison of trimming parameters

	read length = 200	**read length = 36**
total nr. of tags	950,402	950,402
before alignment to SSCSs with FS>= 3	67,786	67,786
make DCS	67,786	67,786
after trimming	41,460	67,282
after alignment to reference genome	41,267	67,002

Table 3: Effect of various trimming parameters on data loss

The read length specified during trimming were changed from 200nt to 36nt. Stringent filtering with long reads leads to read loss of ~20,000 families. If we also allow shorter reads in our data, more data can be recovered. The family size distribution with the new trimming parameters can be found in the Supplementary Material Figure 2.

As *table 3* shows, many reads were lost due to a short read length after trimming. When setting the minimal read length to 36nt, only 504 tags were filtered out. In sum, approximately 784 tags instead of more than 26,519 tags were lost during data analysis with the new trimming parameters. However, if we compare the original numbers of tags (~950,000 tags) to the final number of tags, which were aligned to reference genome, only approx. 7% of the tags (~67,000 tags) could be recovered throughout the analysis.

6. Discussion

The calculation of the Hamming distances between the tags which label the DNA fragments and the distribution of their family sizes give various insights into the duplex sequencing data. At first, were able to see in the distribution of family sizes that there was no bias towards any direction of the reads during amplification. Furthermore, the frequency of large families is small enough, so that we do not waste a lot of sequencing capacities.
The family size distribution showed that the majority of read families are singletons. Since by default three reads per family are required to form SSCSs and later on DCSs, a lot of data is lost. In addition to the singletons, unique SSCSs are ignored in the further analysis, because they cannot form a DCS. (Stoler et al., 2016) The optimal ratio between SSCSs and DCSs is 2:1 but in our data we achieved only a 12:1 ratio. This ratio is an indication that a huge amount of data is lost during data analysis because many SSCSs are only present in either forward or reverse direction and therefore cannot form DCSs. We suspect that these unique SSCSs were introduced either due to errors in the tags or failure to amplify the other strand.

Second, we analysed the tags in more detail by calculating the Hamming distances between the tags. Since we have far more unique tags than molecules to label for, it is expected that the tags differ with multiple nucleotides to each other if they originated from different molecules. The Hamming distance analysis verified this conclusion, since mainly larger Hamming distances (HD=5-7) occurred in the sequencing data. But also, about 30% of the data differed by only nucleotide to the rest. We expect that these are due to sequencing errors or PCR errors, therefore mismatch correction for up to three mismatches should be save and can be applied on further data analyses. However, mismatch correction with four or more mismatches should be avoided, because they might already represent different molecules.
The distribution of family sizes has shown that many singletons differed by only one nucleotide but allowing up to three mismatches in the tag moved small families with up to three members into larger families. Although mismatch correction reduced the number of singletons, not all of them have been recovered but we were able to decrease the ratio between SSCSs and DCSs to 8:1.

Third, we investigated the effect of the tag length on the Hamming distance. Shorter tags (2*6nt, 2*8nt) resulted in smaller Hamming distances than longer tags. Therefore, for shorter tags we cannot be sure whether tags with small Hamming distances have originated from sequencing errors in the tags. Although longer tags achieved higher Hamming distances, typically tags with a length of 2*12 nucleotides are used in order to limit expenses.

The formation of chimeric reads is a known problem in sequencing millions of DNA fragments. For detecting chimeric reads, the fourth analysis comprised calculating the Hamming distance in both halves of the tags separately. The Hamming distances showed that many tags differ only in one half to other tags, but the second half is identical, which is very unlikely to occur by chance. Sources for these tags are either chimeric reads or sequencing error in one part of the tag. The data after mismatch correction suggested that only few of those tags have been corrected by allowing three mismatches in the tag because they are a result of artificially produced chimeras. Those chimeric reads usually have tags with very large differences (up to 11 or 12 nucleotides) between the partial Hamming distances. Relative differences of one between the partial Hamming distances indicate that almost 28% of the tags are very likely chimeric reads. This effect was probably introduced due to errors during adapter synthesis or jumping PCR. Additionally, there were a lot of tags with a high relative difference, which indicates tags, where the total Hamming distance was not split up equally into partial Hamming distances. Our analysis of shorter tags showed that minimal Hamming distance of one or two is common for tags with a length of 12 nucleotides. Therefore, we cannot identify those tags as chimeras, because they still might originate from different molecules. However, most of the chimeras in the data were singletons, which might explain the high fraction of singletons in the whole sample.

The original analysis was based on a sample of the whole dataset, not regarding whether the families would finally end up in DCSs. Our results of the fifth analysis showed tags that are part of a DCS have only high Hamming distances which indicates that they all stem from different molecules and were not split up due to sequencing or PCR errors in the tags. In addition, the DCSs were composed of only 0.2% artificially produced chimeras.

Due to high amounts of singletons in the duplex sequencing data, a huge number of SSCSs, which might represent potential sequences for the DCSs, are lost. However, the final analysis showed that also in later stages in the data analysis like trimming or alignment to the reference genome many more tags are filtered out due to low quality reads or too short read lengths. By adapting the trimming parameters, we were able to recover ~26,000 tags which means that 2.8 % more DCSs that can be used for the final mutation detection. From this point of view, we cannot recover more tags of the dataset bioinformatically and the experimental protocol should be revised and adapted accordingly.

Finally, all developed programs were made accessible through the Galaxy system to provide the tools for a high number of potential users, but they can also be used via the command line (**https://github.com/monikaheinzl/galaxyProject**).

7. Conclusion

Duplex sequencing allows the identification of ultra-low frequency mutations, but this method produces a lot of families with only one read. These singletons will not be used in the formation of DCSs and a huge amount of useful data is lost during the analysis. Our developed tools introduce an effective approach for analysing duplex sequencing data. The calculation of the Hamming distance allowed us to identify tags which were produced due to sequencing errors. Although mismatch correction reduced the errors in the tags and increased the tag's family sizes, not all of the tags were recovered. Our new approaches can also be used to identify chimeric reads in the data, which have been introduced due to PCR errors. Fortunately, most of these tags were filtered out during data analysis, therefore probably no sequencing errors and very few chimeric reads might be aligned to DCSs. Our methods helped to identify where the data might be lost in the process of data analysis, which in turn helps to improve the conditions of data analysis. All tools, that have been used in this thesis, can be found in the Galaxy system but can also be used from the command line.

8. List of Figures

9. List of Tables

10. List of Abbreviations

SSCS	single-stranded consensus sequence
DCS	duplex consensus sequence
NGS	Next generation sequencing
PCR	Polymerase chain reaction
DNA	deoxyribonucleic acid
HD	Hamming distance
FS	family size

11. References

Arnheim, N., & Calabrese, P. (2009). Understanding what determines the frequency and pattern of human germline mutations. *Nature Reviews. Genetics*, *10*(7), 478–488. https://doi.org/10.1038/nrg2529

Fox, E. J., Reid-Bayliss, K. S., Emond, M. J., & Loeb, L. A. (2014). Accuracy of Next Generation Sequencing Platforms. *Next Generation, Sequencing & Applications*, *1*. https://doi.org/10.4172/jngsa.1000106

Kanagawa, T. (2003). Bias and Artifacts in Multitemplate Polymerase Chain Reactions(PCR). *Journal of Bioscience and Bioengineering*, *96*(4), 317–323. https://doi.org/10.1263/jbb.96.317

Kennedy, S. R., Schmitt, M. W., Fox, E. J., Kohrn, B. F., Salk, J. J., Ahn, E. H., … Loeb, L. A. (2014). Detecting ultralow-frequency mutations by Duplex Sequencing. *Nature Protocols*, *9*(11), 2586–2606. https://doi.org/10.1038/nprot.2014.170

Lou, D. I., Hussmann, J. A., McBee, R. M., Acevedo, A., Andino, R., Press, W. H., & Sawyer, S. L. (2013). High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(49), 19872–19877. https://doi.org/10.1073/pnas.1319590110

Maher, G. J., Goriely, A., & Wilkie, A. O. M. (2014). Cellular evidence for selfish spermatogonial selection in aged human testes. *Andrology*, *2*(3), 304–314. https://doi.org/10.1111/j.2047-2927.2013.00175.x

Pinheiro, H. P., de Souza Pinheiro, A., & Sen, P. K. (2005). Comparison of genomic sequences using the Hamming distance. *Journal of Statistical Planning and Inference*, *130*(1–2), 325–339. https://doi.org/10.1016/J.JSPI.2003.03.002

Schmitt, M. W., Fox, E. J., Prindle, M. J., Reid-Bayliss, K. S., True, L. D., Radich, J. P., & Loeb, L. A. (2015). Sequencing small genomic targets with high efficiency and extreme accuracy. *Nature Methods*, *12*(5), 423–425. https://doi.org/10.1038/nmeth.3351

Schmitt, M. W., Kennedy, S. R., Salk, J. J., Fox, E. J., Hiatt, J. B., & Loeb, L. A. (2012). Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences*, *109*(36), 14508 LP-14513. Retrieved from http://www.pnas.org/content/109/36/14508.abstract

Smyth, R. P., Schlub, T. E., Grimm, A., Venturi, V., Chopra, A., Mallal, S., … Mak, J. (2010). Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene*, *469*(1–2), 45–51. https://doi.org/10.1016/j.gene.2010.08.009

Stoler, N., Arbeithuber, B., Guiblet, W., Makova, K. D., & Nekrutenko, A. (2016). Streamlined analysis of duplex sequencing data with Du Novo. *Genome Biology*, *17*(1), 180. https://doi.org/10.1186/s13059-016-1039-4

Wang, C., Kao, W.-H., & Hsiao, C. K. (2015). Using Hamming Distance as Information for SNP-Sets Clustering and Testing in Disease Association Studies. *PLoS ONE*, *10*(8), e0135918. https://doi.org/10.1371/journal.pone.0135918

12. Supplementary Material

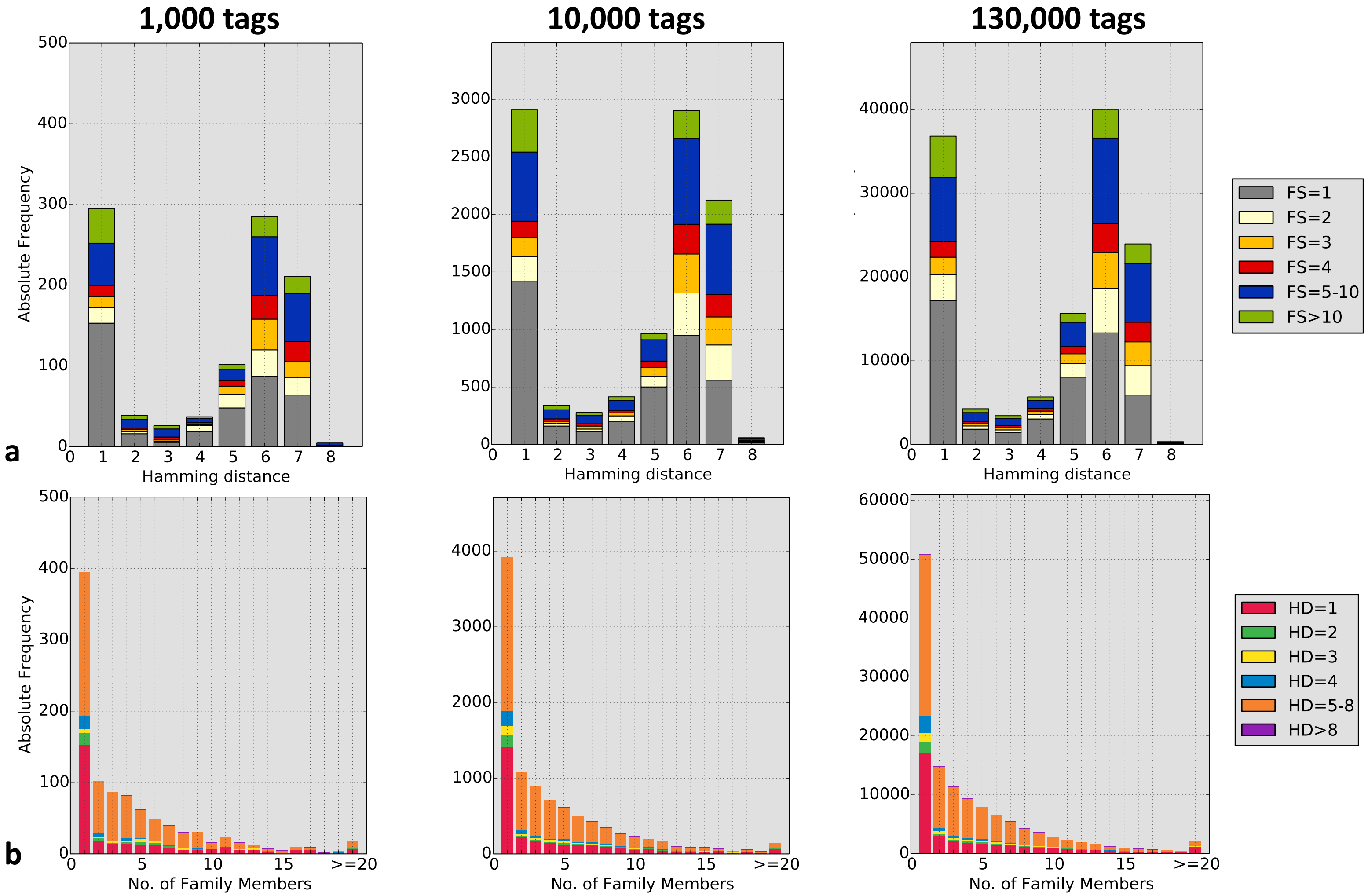
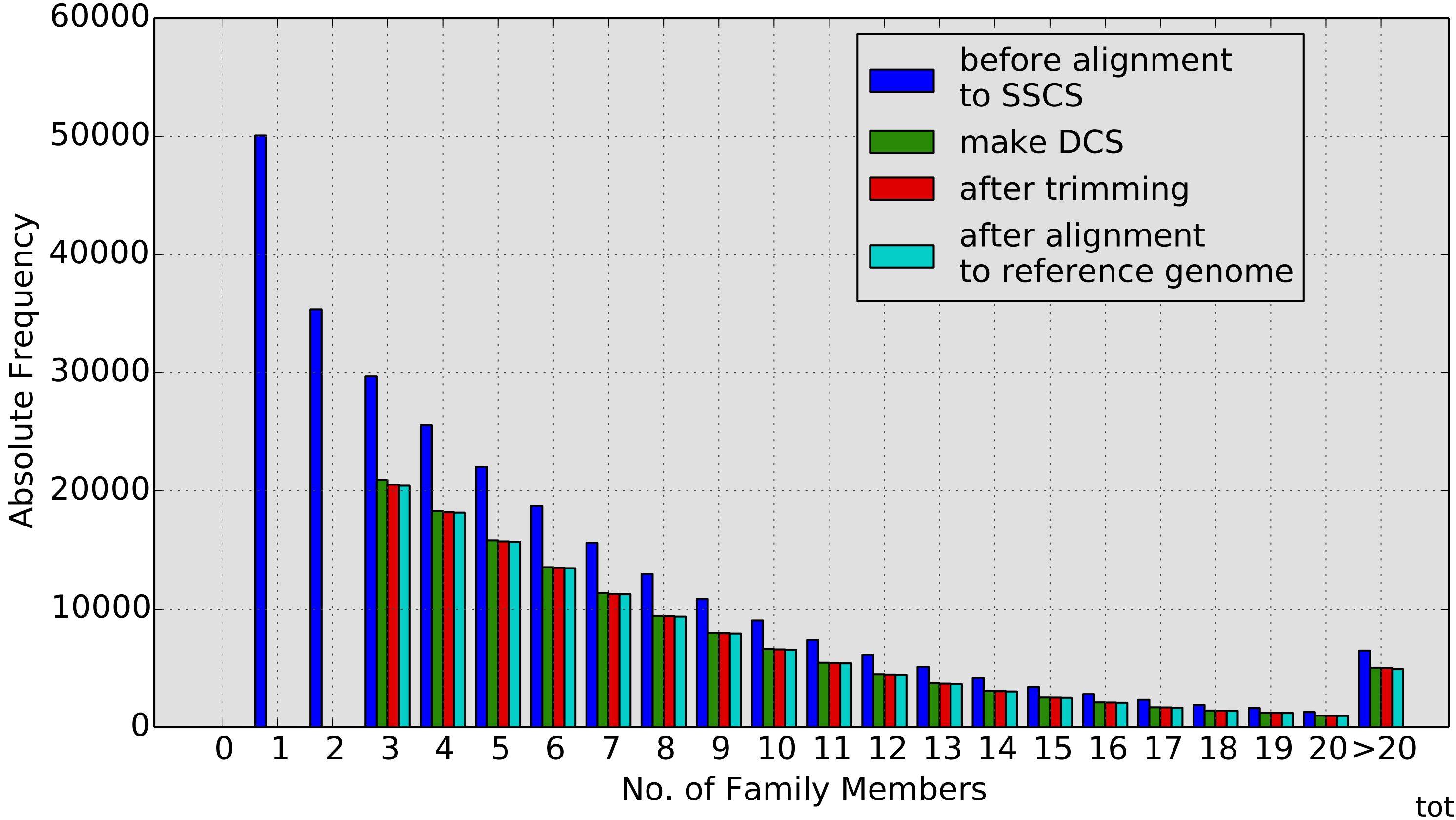


Figure 20: Hamming distance analysis with different sizes of the sample

We have selected the size for the sample with 1,000 tags, since it would computationally be too demanding to analyze more than 1 million tags. Here, we verified with sample sizes of 10,000 tags and ~130,000 tags that 1,000 tags are a representative sample of our whole dataset. Hamming distance analysis (**a**) and family size distribution (**b**) with various sample sizes are shown.



	AB	BA
max. family size =	90	176
absolute frequency=	1	1
relative frequency=	0.00001	0.00001
total nr. of reads (SSCS)	5,764,559	

	total numbers
total nr. of tags (unique, FS>=1):	950,402
DCS (before alignment to SSCS, FS>=1):	136,240
total nr. of tags (unique, FS>=3):	570,008
DCS (before alignment to SSCS, FS>=3):	67,786
make DCS:	67,786
after trimming:	67,282
after alignment to reference genome:	67,002

Figure 21: Analysis of read loss with adapted trimming parameters

Family size distribution with various stages of the analysis with correction for 3 mismatches: The plot contains both counts of forward and reverse strand, whereas the numbers below represent the single count of the DCS. The read length has been changed from 200nt to 36nt to improve the read loss of duplex sequencing. Due to the less stringent filtering of the reads, more read families are recovered throughout the analysis. Only ~500 tags are skipped during trimming with the adapted read length.