Analysis of Min-Hashing for Variant Tolerant DNA Read Mapping*

Jens Quedenfeld[1] and Sven Rahmann[2]

1 **Chair of Theoretical Computer Science, Technical University of Munich, Munich, Germany; and**
Bioinformatics, Computer Science XI, TU Dortmund, Dortmund, Germany
jens.quedenfeld@in.tum.de

2 **Genome Informatics, Institute of Human Genetics, University Hospital Essen, University of Duisburg-Essen, Essen, Germany; and**
Bioinformatics, Computer Science XI, TU Dortmund, Dortmund, Germany
Sven.Rahmann@uni-due.de

Abstract

DNA read mapping has become a ubiquitous task in bioinformatics. New technologies provide ever longer DNA reads (several thousand basepairs), although at comparatively high error rates (up to 15%), and the reference genome is increasingly not considered as a simple string over ACGT anymore, but as a complex object containing known genetic variants in the population. Conventional indexes based on exact seed matches, in particular the suffix array based FM index, struggle with these changing conditions, so other methods are being considered, and one such alternative is locality sensitive hashing.

Here we examine the question whether including single nucleotide polymorphisms (SNPs) in a min-hashing index is beneficial. The answer depends on the population frequency of the SNP, and we analyze several models (from simple to complex) that provide precise answers to this question under various assumptions. Our results also provide sensitivity and specificity values for min-hashing based read mappers and may be used to understand dependencies between the parameters of such methods. We hope that this article will provide a theoretical foundation for a new generation of read mappers.

1998 ACM Subject Classification F.2.2 Pattern Matching

Keywords and phrases read mapping, min-hashing, variant, SNP, analysis of algorithms

Digital Object Identifier 10.4230/LIPIcs.WABI.2017.21

1 Introduction

In bioinformatics, DNA read mapping has become a basic first step of many sequence analysis tasks. Formally, one is given millions of short DNA fragments ("reads"), i.e., strings of typical length 100–300 over the DNA alphabet $\Sigma := \{A,C,G,T\}$, and a reference genome, which is a long DNA sequence (approx. $3 \cdot 10^9$ basepairs for the human genome). For each read, one seeks the (ideally unique) interval of the reference, where the read (or its reverse complement) matches best, i.e., with smallest edit distance. Leaving aside problems such as repetitive regions in the genome, structural rearrangements, split reads, etc., read mapping is thus simply a large-scale approximate string matching problem.

* This work has been supported by the DFG, Collaborative Research Center SFB 876, project C1 (http://sfb876.tu-dortmund.de/).

The most efficient methods today use an index data structure to quickly locate long maximal exact matches (MEMs) between each read and reference. These "seed" matches are then extended to full alignments for verification. Several popular read mappers use the FM index (a compressed representation of a suffix array based on the Burrows Wheeler Transform [5]) because of its small space requirements. Filtering with long MEMs works efficiently if the number of differences between read and reference is small, but breaks down for higher error rates, as one obtains no specific long MEMs, but many unspecific short ones.

Two trends are currently changing the landscape of read mappers. First, there are new technologies on the market that output much longer reads (e.g., up to 60 000 basepairs), but at higher error rates (10%–15%). Second, it is becoming more accepted that the human reference genome is not well represented by a single string over Σ. Each person (even each single cell) has its own individual genome, and the field of *pan-genomics* is exploring how to best represent the entirety of known genomic information of a species [11]. Most (but not all) of the genetic variation between individuals arises from *single nucleotide polymorphisms* (SNPs), which are single positions in the genome at which there exist at least two different basepair choices in the population. Other important types of variations are variable-length insertions and deletions and rearrangements of genomic regions. As index data structures based on suffix arrays struggle with these changing conditions, alternative types of indexes are being explored.

One such alternative indexing method is by *locality sensitive hashing* on q-gram sets of genomic windows. This approach has been already used for the genome assembly tools MHAP [1] and Minimap [8] as well as for the read mappers BALAUR [9] and MashMap [6]. The hashing functions can be constructed to *take variants into account* (see Section 2). We have recently designed and implemented a prototype of a variant tolerant read mapper called VATRAM [10] based on min-hashing [2]. During the design, we noted the following question, which we call the *variant indexing decision problem*: Given a reference genome (string) and a variant with its frequency p in the population, should the variant be included in the index or not? Inclusion of a variant will allow to match reads containing the variant more reliably to the correct region, at the expense of a lower sensitivity of the region for reads that do not contain the variant. The answer therefore depends of the population frequency of the variant.

The purpose of this article is to present a detailed analysis of the variant indexing decision problem for a min-hashing index. In Section 2, we present an overview of read mapping based on min-hashing. We then analyze the variant indexing problem (Section 3), moving from a simple model to a general one, yielding widely applicable results and a limit theorem. Note that our results hold not only for VATRAM, but also for other window-based min-hashing read mappers such as *BALAUR* [9].

2 Background: Read mapping with min-hashing

Let $\mathcal{Q}$ be a set of basic objects (here, the set of all 4^q DNA q-grams). Locality sensitive hashing in general (and min-hashing as a special case) is a method that assigns an integer number (hash value) $h(Q)$ to each object set $Q \subset \mathcal{Q}$, such that the probability that two sets Q, Q' obtain the same hash value $h(Q) = h(Q')$ depends on a well-defined similarity measure if the hash function h is chosen randomly among a well-defined collection of hash functions.

Min-hashing on q-grams in particular works as follows [2]. We identify each q-gram with its base-4 numerical representation after mapping A$\mapsto$0, C$\mapsto$1, G$\mapsto$2, T$\mapsto$3. This provides a

natural (lexicographic) order among the q-grams. Let π be a permutation on $\mathcal{Q}$; it induces a different order on $\mathcal{Q}$. For a q-gram set $Q \subset \mathcal{Q}$, let $\min_\pi(Q)$ be the (numerical representation of the) smallest q-gram in Q according to π. Min-hashing consists of choosing a random permutation π and using $h(Q) := \min_\pi(Q)$. The following lemma states that min-hashing is locality sensitive hashing using the Jaccard coefficient as similarity value.

▶ **Lemma 1** (Min-hash property). *Given two sets $Q \subset \mathcal{Q}$, $Q' \subset \mathcal{Q}$ and the set Π of all permutations on $\mathcal{Q}$, let $\pi \in \Pi$ be a random permutation. For any $Q \subset \mathcal{Q}$, define $h(Q) := \min_\pi(Q)$. The probability that Q and Q' are hashed to the same value $h(Q) = h(Q')$ is equal to the Jaccard coefficient of Q and Q',*

$$P(h(Q) = h(Q')) = \frac{|Q \cap Q'|}{|Q \cup Q'|}. \tag{1}$$

A proof can be found in [2].

The read mapper VATRAM [10] uses min-hashing to create an index on a given reference genome. The genome is divided into overlapping windows of length w; here w should be slightly larger than the typical length n of the reads (e.g. $w = 1.4n$ [10]). The distance between two window start positions is denoted by o; we assume $o < w$, so the windows overlap. (For long reads, the reads may be divided into overlapping windows as well.) Let $a = (a_i)$ be the sequence of genome window sequences. From each window a_i, we obtain the q-grams (substrings of length q) it contains, and we denote the resulting q-gram set by Q_i. We then apply min-hashing to each window and to each read (or read window), choosing the same permutation π. The hash value of a window or read is also called its *signature value* (with respect to π). If a read's signature value is equal to $h(Q_i)$, there is a certain probability that the read originates from window a_i. However, this agreement may also just be a random hit. Therefore, several different independent permutations are used to improve sensitivity and specificity. If the number of common signature values between the read and window a_i is higher than expected by chance, the probability that the read originates from that window is high.

The index data structure efficiently maps signature values to genome windows and may itself be implemented using hashing. Note that the memory usage for the index grows linearly with the number of permutations.

In practice, choosing a random permutation among all $(4^q)!$ ones is impossible even for small q due to limitations of pseudo-random number generation with finite memory. Furthermore it is important that $\min_\pi(Q)$ can be computed efficiently. Therefore in VATRAM both q-grams and permutations are represented by $2q$-bit vectors. Applying a permutation is simulated by combining the q-gram bit-vector and the permutation bit-vector with an *exclusive or* (XOR) operation. The signature value is the smallest of the resulting values,

$$h_\pi(Q) = \min\{x \oplus \pi \mid x \in Q\}. \tag{2}$$

Using $2q$ random bits and the XOR technique instead of a true random permutation means that the pre-conditions of Lemma 1 do not hold and the min-hash property may be violated [3]. However, empirical studies have shown that in practice the XOR technique approximates the desired property well [4].

VATRAM supports single nucleotide polymorphisms where one nucleotide is replaced by another. For each known variant in window a_i of the reference genome, we may add not only the original reference q-grams to set Q_i, but also all q-grams resulting variant. In this case Q_i has a larger cardinality than sets from windows without variants.

3 Analysis of variant-tolerant min-hashing

Adding q-grams resulting from a variant increases the number of common q-grams between the reference window and a read containing the variant, so the *collision probability* that both are mapped to the same signature value becomes larger. However, for reads not containing the variant, adding new q-grams increases the size of the union in Eq. (1), so the collision probability is reduced.

The *variant indexing decision problem* asks how frequently a variant must occur in the population, such that adding the variant is beneficial on average. We provide an answer to this question under certain assumptions in each subsection. These assumptions start out strong (and unrealistic) and are successively relaxed; this organization should allow for an accessible exposition.

We always assume that the variant under consideration has a population frequency of $0 < p \leq 0.5$, i.e., the variant appears with a probability of p in a random read. (If $p > 0.5$, we would define the variant as reference and vice versa.) We use the following notation.

- The window length is w; the number of q-grams in the window is $v = w - q + 1$.
- The read length is n; the number of q-grams in the read is $m = n - q + 1$.

3.1 Basic model

Our initial assumptions are as follows.

- The entire read is contained in a single reference window a_i.
- There is a single SNP inside the window.
- The SNP occurs "far" from the ends of the read and window, such that exactly q of the q-grams are affected.
- There are no sequencing errors in the read, i.e., all other positions correspond exactly to the reference.
- Each q-gram in Q_i (including those stemming from the variant) occurs only once in window a_i.
- Only a single signature value is used.

We compare two strategies: In the DEFAULT strategy we only use the original nucleotides in the reference genome to build the index. In the SNP strategy we use both the original nucleotides and those containing the variant. For both cases, the collision probabilities P_{DEFAULT} and P_{SNP} are easily calculated.

▶ **Lemma 2.** *Under the above assumptions, with $v = w - q + 1$ and $m = n - q + 1$,*

$$P_{\text{DEFAULT}} = (1 - p) \cdot m/v + p \cdot (m - q)/(v + q), \tag{3}$$

$$P_{\text{SNP}} = m/(v + q). \tag{4}$$

Adding the variant to the index is beneficial if and only if

$$p > m/(m + v). \tag{5}$$

Proof. All collision probabilities follow from calculating the cardinalities of intersections and unions of q-gram sets and applying Eq. (1).

For the DEFAULT strategy, with a probability of $1 - p$ (variant does not occur in the read) all q-grams of the read belong to the window's q-gram set, so the collision probability is m/v. With probability p (the variant occurs in the read), the intersection's size is reduced by q and the union's size is enlarged by q, leading to a collision probability of $(m-q)/(v+q)$.

In the SNP strategy, the intersection size is always m and the union size is always $v+q$.

Adding the variant is beneficial if and only if $P_{\text{SNP}} > P_{\text{Default}}$, which is equivalent to inequality (5) by the next lemma, whose proof is elementary algebra. ◀

▶ **Lemma 3.** *If the condition*

$$P_{\text{SNP}} > P_{\text{Default}} \text{ is equivalent to } s > (1-p)\cdot d_0 + p \cdot d_1 \tag{6}$$

for some constants s, d_0, d_1 independent of p with $d_0 > s > 0$ and $d_0 > d_1 > 0$, then the condition is also equivalent to

$$p > (d_0 - s)/(d_0 - d_1). \tag{7}$$

A typical parameter configuration that produces good results for reads with a length of $n = 100$ bases is $w = 140$ and $q = 16$ [10]. For these parameters the frequency threshold is $p > 40.5\%$. Most known variants are less frequent than that; therefore the benefits of the SNP strategy are likely marginal in this simplified model. However, some of the assumptions were unrealistic, and we now consider increasingly realistic models.

3.2 Consideration of errors

So far we did not take sequencing errors or unknown variants into account. Both types of differences between the read and the reference are called *errors*. We assume that

- exactly $e < m - q$ of the q-grams in the read are affected by errors,
- the errors occur far from the SNP position,
- the q-grams produced by the errors are different from the existing q-grams in the read and in the window.

▶ **Lemma 4.** *Under the above assumptions, with $v = w - q + 1$ and $m = n - q + 1$,*

$$P_{\text{Default}} = (1-p)\cdot(m-e)/(v+e) + p\cdot(m-q-e)/(v+q+e), \tag{8}$$

$$P_{\text{SNP}} = (m-e)/(v+q+e). \tag{9}$$

Adding the variant to the index is beneficial if and only if

$$p > (m-e)/(m+v), \tag{10}$$

which is equivalent to (5) for $e = 0$.

Proof. The proof is similar to that of Lemma 2. Intersection sizes are reduced by e, while union sizes grow by e in comparison to Lemma 2. With $d_0 = (m-e)/(v+e)$, $d_1 = (m-q-e)/(v+q+e)$ and $s = (m-e)/(v+q+e)$, the conditions of Lemma 3 are satisfied for $e < m - q$, and (10) follows by elementary algebra. ◀

The result shows that the population frequency threshold decreases when more errors are considered. For example, if we consider the standard configuration ($n = 100$, $w = 140$ and $q = 16$) and $e = 0, q, 2q$, corresponding to no, one and two isolated errors, the thresholds are 40.5%, 32.9% and 25.2%, respectively. (In the extreme case $e = m-q-1$, the frequency threshold becomes 8.1%. Of course, for so many errors, the indexing strategy breaks down as a whole and the variant indexing decision problem is meaningless.)

The above analysis assumed that the errors did non interfere the q-grams generated by the variant, which is unrealistic. Instead, errors will be distributed randomly, so we may assume that they affect (averaging over many reads) q-grams of the variant and q-grams outside the variant with an equal proportion. In this model, we call $\varepsilon := e/m$ the error rate. Using this modified error model has considerable effects on the threshold.

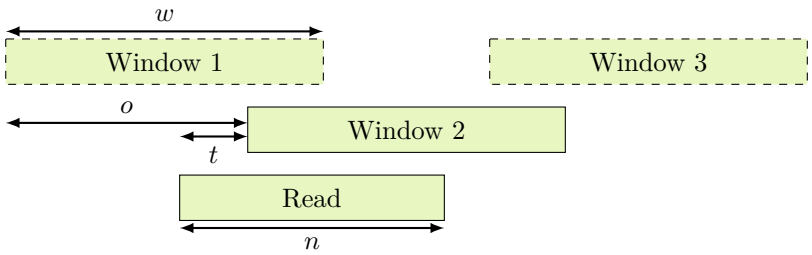

Figure 1 Visualization of the parameters n (read length), w (window length), o (window distance, $o < w$ ensures overlapping windows) and t (the number of read's q-grams located outside of the major window). Note that the number of bases located outside (shown in the visualization) is equal to the number of read's q-grams located outside if and only if $t \leq n-q+1 = m$, so the read and the major window have at least $q-1$ common positions. This conditions is always fulfilled when using realistic parameter configurations. We assume that $o > w - n + 1$, so there exist configurations where the read is not contained in a single window.

▶ **Lemma 5.** *Under the above assumptions,*

$$P_{\text{DEFAULT}} = (1-p) \cdot \frac{m(1-\varepsilon)}{v + m\varepsilon} + p \cdot \frac{(m-q)(1-\varepsilon)}{v + q + (m-q)\varepsilon}, \tag{11}$$

$$P_{\text{SNP}} = \frac{m(1-\varepsilon)}{v + q + m\varepsilon}. \tag{12}$$

Adding the variant to the index is beneficial if and only if

$$p > \frac{m}{v+m} - \frac{eq}{(v+m)(v+e+q)}\,. \tag{13}$$

Proof. The arguments are the same as in the proof of Lemma 4, but it is not a fixed number e that is subtracted from the intersection and added to the union, but the intersection is reduced by a factor of $(1-\varepsilon)$; the corresponding number is added to the union. Note that only the constant factor of p differs between (8) and (11) and (9) is identical to (12). The threshold (13) follows from Lemma 3; a detailed derivation is presented in Appendix A.1. ◀

In this model, the effect of e is marginal for typical parameter configurations, which is different from the model considered in Lemma 4. With the same parameters as above, the behavior for the extreme case $e = m - 1$ yields a threshold of 37.6%, which is close to the threshold of 40.5% for $e = 0$.

3.3 Consideration of partial overlaps between read and windows

So far we assumed that the read is contained entirely in one window. Indeed, this is the case if the distance o between two window starting points satisfies $o \leq w - n + 1$. Now we consider $o > w - n + 1$, where a read may overlap with two windows. We consider the window with the larger overlap (see Fig. 1) and call it the *major window*.

When we slide the read over the reference, there are o different configurations how the read overlaps two windows. For each configuration we track the number t of q-grams of the read that do *not* overlap the major window. For $w-n+1$ configurations, the major window contains the entire read and so $t = 0$. When the read moves further along the reference, t increases until the next window becomes the major window and t decreases towards zero. So there are $o - w + n - 1$ configurations with a non-zero value of t.

If $o-w+n-1$ is even, the maximum value of t that is attained is $t_{\max} = (o-w+n-1)/2$ and each value is attained twice.

If $o-w+n-1$ is odd, the maximum value of t that is attained is $t_{\max} = (o-w+n)/2$, this value is attained once and each lower $t = 1, \ldots, t_{\max} - 1$ is attained twice.

For a unified consideration of intersection and union sizes in Lemma 1, we introduce a collision probability parameterized by t and two additional numbers I and U, which stand for the number of q-grams by which the intersection is reduced and the union is enlarged, respectively,

$$P_{t,I,U,0} := \frac{m-t-I}{v+t+U}.$$

As in Section 3.2, we generalize this quantity by considering q-grams affected by errors, calling again $\varepsilon := e/m$ the q-gram error rate (typical ε of interest are $\varepsilon = 0$, q/m, $2q/m$, etc.) and define

$$P_{t,I,U,\varepsilon} := \frac{(m-t-I)(1-\varepsilon)}{(v+t+U)+\varepsilon(m-t-I)}\,. \tag{14}$$

If we consider each window configuration equally likely, we may obtain average collision probabilities $P^{+}_{I,U,\varepsilon}$ by summing $P_{t,I,U,\varepsilon}$ over the relevant values of $t = 0, 1, \ldots, t_{\max}$ with appropriate weights ω_t:

$$P^{+}_{I,U,\varepsilon} := \sum_{t=0}^{t_{\max}} \omega_t \cdot P_{t,I,U,\varepsilon}, \tag{15}$$

where, according to the discussion above, $t_{\max} = \lfloor (o-w+n)/2 \rfloor$, and the weights are $\omega_0 = (w-n+1)/o$, and, if $o-w+n-1$ is even, $\omega_t = 2/o$ for $1 \le t \le t_{\max}$, but if $o-w+n-1$ is odd, $\omega_t = 2/o$ for $1 \le t < t_{\max}$ and $\omega_{t_{\max}} = 1/o$.

▶ **Lemma 6.** *Under the assumptions and with the notation of this section,*

$$P_{\text{DEFAULT}} = (1-p)\cdot P^{+}_{0,0,\varepsilon} + p \cdot P^{+}_{q,q,\varepsilon},$$
$$P_{\text{SNP}} = P^{+}_{0,q,\varepsilon}.$$

Proof. The lemma is merely re-stating the arguments of the previous sections: For the DEFAULT strategy we have $U = I = 0$ if the read does not contain the variant (an event with probability $1-p$) and $U = I = q$ if the read does contain it (an event with probability p). For the SNP strategy, we always have $I = 0$ and $U = q$. ◀

Before stating a general threshold result (Sec. 3.5), we consider an even more complex model considering more than one signature value.

3.4 Consideration of signature length

So far we considered the collision probabilities for a single signature value. To improve sensitivity and specificity, we may instead use S different random permutations and consider the event where at least s out of S signature values of the read match the corresponding ones of the window. Then the collision probability $P_{t,I,U,\varepsilon}$ in (14) becomes

$$P^{(s,S)}_{t,I,U,\varepsilon} = \sum_{k=s}^{S} \binom{S}{k} \left(P_{t,I,U,\varepsilon}\right)^k \left(1 - P_{t,I,U,\varepsilon}\right)^{S-k}. \tag{16}$$

For $s = 1$ (thereby only increasing sensitivity but not specificity), we obtain

$$P_{t,I,U,\varepsilon}^{(1,S)} = 1 - (1 - P_{t,I,U,\varepsilon})^S \,. \tag{17}$$

Correspondingly, the average (15) over window configurations generalizes to

$$P_{I,U,\varepsilon}^{+(s,S)} := \sum_{t=0}^{t_{\max}} \omega_t \cdot P_{t,I,U,\varepsilon}^{(s,S)} \tag{18}$$

with the same weights ω_t as before, distinguishing the cases when $o = w + n - 1$ is odd resp. even.

3.5 General results

The following result is the most general we present, considering errors, partial overlaps between read and windows and signature length.

▶ **Theorem 7.** *Under the assumptions and with the notation of Sec. 3.4,*

$$P_{\text{DEFAULT}} = (1-p) \cdot P_{0,0,\varepsilon}^{+(s,S)} + p \cdot P_{q,q,\varepsilon}^{+(s,S)},$$
$$P_{\text{SNP}} = P_{0,q,\varepsilon}^{+(s,S)}.$$

Adding the variant to the index is beneficial if and only if

$$p > T := \frac{P_{0,0,\varepsilon}^{+(s,S)} - P_{0,q,\varepsilon}^{+(s,S)}}{P_{0,0,\varepsilon}^{+(s,S)} - P_{q,q,\varepsilon}^{+(s,S)}} \,. \tag{19}$$

Proof. To see that Lemma 3 applies for the threshold, note that for every choice of $(n, w, o, q, \varepsilon, s, S)$, the average collision probability $P_{0,I,U,\varepsilon}^{+(s,S)}$ is a decreasing function of both I and U. ◀

We do not see a way to considerably simplify the threshold T in (19). Before presenting numerical results (Sec. 3.6), we consider a limit result for large signatures.

▶ **Theorem 8.** *Let (n, w, o, q) be given parameters and let $s := 1$. For any variant population frequency $p > 0$ and for each number e of q-grams affected by errors, there exists a signature length S such that $P_{\text{SNP}} > P_{\text{DEFAULT}}$. In other words, for each $\varepsilon := e/m$, the threshold*

$$T^{(1)} := \frac{P_{0,0,\varepsilon}^{+(1,S)} - P_{0,q,\varepsilon}^{+(1,S)}}{P_{0,0,\varepsilon}^{+(1,S)} - P_{q,q,\varepsilon}^{+(1,S)}}$$

converges to zero for increasing signature length S.

Proof. To show $P_{\text{SNP}} > P_{\text{DEFAULT}}$, it is sufficient to show that for large enough S, the inequality

$$P_{t,0,q,\varepsilon}^{(1,S)} > (1-p) \cdot P_{t,0,0,\varepsilon}^{(1,S)} + p \cdot P_{t,q,q,\varepsilon}^{(1,S)} \tag{20}$$

is satisfied for all t because P_{SNP} and P_{DEFAULT} are positively weighted sums of these terms with the same weights (ω_t), cf. (18). By (17), (20) is equivalent to

$$p > \frac{(1 - P_{t,0,0,\varepsilon})^S - (1 - P_{t,0,q,\varepsilon})^S}{(1 - P_{t,0,0,\varepsilon})^S - (1 - P_{t,q,q,\varepsilon})^S} \tag{21}$$

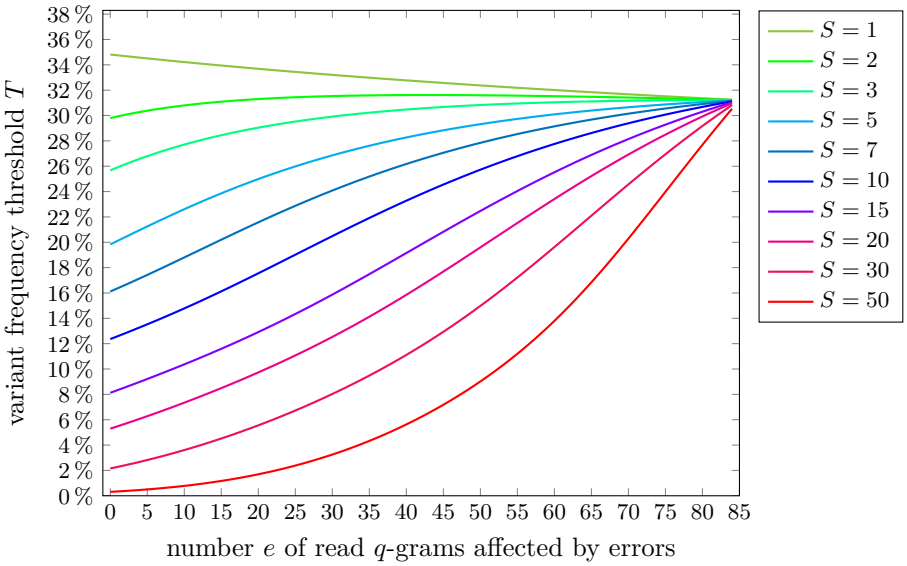

Figure 2 Variant frequency threshold as a function of the number e of erroneous q-grams for different signature lengths S (color-coded). If the probability of a SNP variant is greater than the threshold, the SNP strategy is better than the DEFAULT strategy. Parameters are $n = 100$, $w = 140$, $o = 125$ and $q = 16$, $s = 1$.

for every t. Writing $\ell_{t,I,U} := \ln(1 - P_{t,I,U,\varepsilon})$ (for fixed ε), this is equivalent to

$$p > \frac{\exp(S\ell_{t,0,0}) - \exp(S\ell_{t,0,q})}{\exp(S\ell_{t,0,0}) - \exp(S\ell_{t,q,q})} = \frac{\exp(S(\ell_{t,0,q} - \ell_{t,0,0})) - 1}{\exp(S(\ell_{t,q,q} - \ell_{t,0,0})) - 1} =: T_{t,S} \tag{22}$$

With elementary means, we verify that $P_{t,I,U,\varepsilon}$ is a decreasing function of both I and U, so

$$P_{t,0,0,\varepsilon} > P_{t,0,q,\varepsilon} > P_{t,q,q,\varepsilon} \qquad \text{and} \qquad \ell_{t,q,q} > \ell_{t,0,q} > \ell_{t,0,0}. \tag{23}$$

It follows that both numerator and denominator of $T_{t,S}$ tend towards infinity and to find the limit, we may apply De L'Hôpital's rule and find

$$\lim_{S\to\infty} T_{t,S} = \lim_{S\to\infty} \exp[S((\ell_{t,0,q} - \ell_{t,0,0}) - (\ell_{t,q,q} - \ell_{t,0,0}))] = 0, \tag{24}$$

because $\ell_{t,0,q} - \ell_{t,q,q} < 0$. So for arbitrarily small $p > 0$ and each t there exists $S(t)$ such that $p > T_{t,S(t)}$ is satisfied. Take $S := \max_{t=0,\ldots,t_{\max}} S(t)$; then (20) is satisfied, proving the theorem. ◀

3.6 Numerical results

We show numerical threshold values for a realistic case, where the parameters $n = 100$ (read length), $w = 140$ (window length), $o = 125$ (window distance) and $q = 16$ result from extensive experimentation [10]. Therefore $m = n - q + 1 = 85$ and $v = w - q + 1 = 125$. Figure 2 shows the frequency threshold from (19) as a function of the number e of erroneous q-grams (recall $\varepsilon = e/m$) for different signature lengths S using $s = 1$.

For $S = 1$, the graph looks similar to the results of Lemma 5: The threshold decreased by about 5% points (because we consider the possibility of partial overlaps), and the influence of errors is marginal (about 3% points between $e = 0$ and $e = m - 1$).

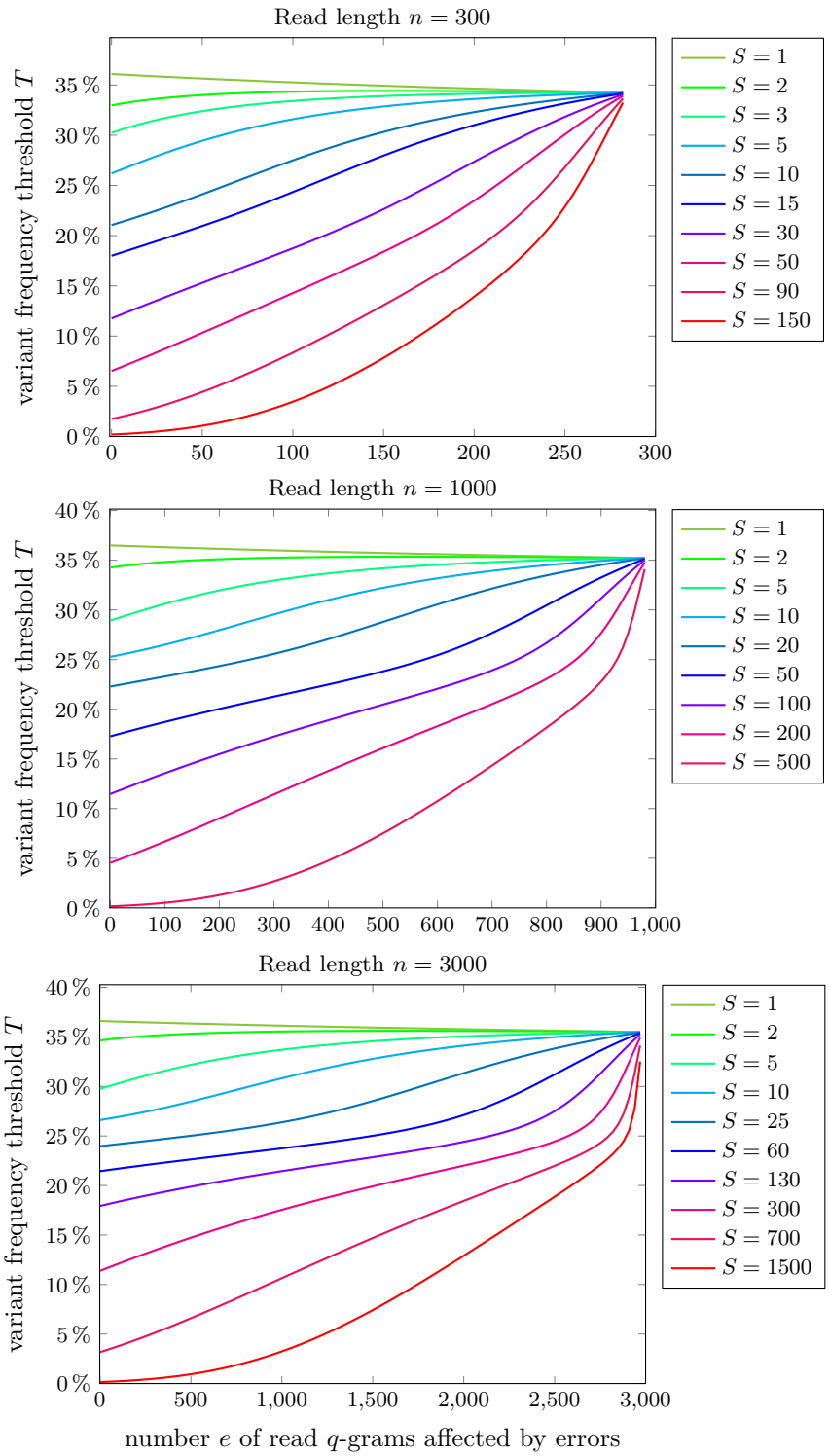

Figure 3 Variant frequency threshold as a function of the number e of erroneous q-grams for different signature lengths S (color-coded) and read lengths n (see heading of the diagrams). If the probability of a SNP variant is greater than the threshold, the SNP strategy is better than the DEFAULT strategy. Parameters are $w = 1.4n$, $o = 1.25n$ and $q = 16$, $s = 1$.

For larger signatures, the picture changes, and the threshold decreases rapidly with S (for fixed e). For error-free reads, the SNP strategy is already better if the variant's population frequency is as small as 0.3%, if use $S = 50$ permutations. As the number of errors increases, the effect of using more signatures is less important, but Theorem 8 shows that enough signature values will always favor the SNP strategy.

Figure 3 shows the frequency threshold for longer reads ($n \in \{300, 1000, 3000\}$). The window length and distance are fit to the read length, i.e. $w = 1.4n$ and $o = 1.25n$, the q-gram length is still 16. The memory consumption increases linearly with S/o, so for longer reads we can increase the signature length S to keep the memory consumption constant. Therefore the maximal signature length plotted in figure 3 is $S = n/2$. If the ratio of q-grams affected by errors is low, then the frequency threshold is almost independent of the read length. If much more than the half of q-grams are affected by errors, then the frequency threshold decreases significantly with the read length. This is especially useful for very long reads with high rates produced by third-generation sequencing machines.

4 Discussion and conclusion

As third-generation sequencing techniques produce ever longer reads with (comparatively) high error rates and the human reference genome is about to replaced by a pan-genome reference, the community is considering alternatives to an FM index for representing the human genome. A fundamental question is whether adding known variants, and in particular SNPs, which represent about 90% of the known variants, should be added to a given index data structure. We investigated this question for min-hashing, a particular form of locality sensitive hashing and found practical decision rules that depend on the population frequency p of a SNP. Theorem 8 shows that (under the right circumstances) it can be beneficial to add even rare variants to a min-hashing index.

Assumptions

Our calculations are based on several assumptions that we did not relax during the development of Theorem 7:

1. All q-grams resulting from a SNP or errors are different from all other q-grams in the read or in the window.
2. We consider each SNP in isolation and assume that exactly q of the q-grams in the read are affected.
3. Errors affect q-grams of the variant and in the remaining part of the read with equal proportion ε.

We consider these assumptions reasonable in practice for the following reasons. While we cannot rule out the possibility that one of the erroneous q-grams is equal to another q-gram in the read, such an event occurs rarely if $v \ll 4^q$, which is always the case in real applications. The second assumption can be relaxed in our framework by choosing different values for U and I in (16) than 0 and q to model specific configurations of SNP positions inside a window. One could even compute weighted averages over all such configurations for a given SNP rate, say 0.5%, across the genome. We here decided to focus on the simple case of a single isolated SNP, which is a case common enough to be relevant. The proportional error distribution may be criticized as averaging early over important effects such as clumping of affected q-grams, but we found that the results obtained with this model match extensive simulations performed in the context of the development of VATRAM (data not shown; [10]).

Sensitivity and specificity analysis

We focused on the population frequency threshold of a variant above which it becomes beneficial to add its q-grams to the index. Of course, our formulas for P_{SNP} and P_{DEFAULT} may also be used to compute sensitivity and specificity values of the min-hashing indexing approach in the first place (see [1] for first arguments). We may define the sensitivity P^* as the desired collision probability (i.e., the probability to find the correct window for a read). Choosing the optimal strategy for each variant depending on its population frequency, we have $P^* = \max\{P_{\text{SNP}}, P_{\text{DEFAULT}}\}$. A sensitivity close to 1 is desirable and we may optimize free parameters (w, o, q, s, S) to achieve a desired sensitivity. On the other hand, it is in our best interest to keep the number of false positive collisions (i.e., of collisions resulting from random coincidence of signature values) as low as possible. By estimating the size distribution of the intersection of the q-gram set of a random read with a random window, we obtain a null collision probability P^0 and the expected number $E = P^0 \cdot G/o$ of colliding windows for a read, where G is the genome size. It follows that Ew basepairs must be aligned to the read for verification. It is mainly the choice of q, S and s that influences P^0, and choosing sufficiently large values will decrease P^0. On the other hand, the memory requirements of the index are proportional to $(4^q + G/o) \cdot S$, so we are interested in keeping this quantity as small as possible. In the future, we aim to use the results presented in this article to optimize the min-hashing parameters to achieve optimum sensitivity with a tolerable false positive rate for a given amount of available memory.

Acknowledgments. This work has been supported by the DFG, Collaborative Research Center SFB 876, project C1 (`http://sfb876.tu-dortmund.de/`). We thank all members and advisors of Project Group 583 [7] for their input.

References

1 Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P. Drake, Jane M. Landolin, and Adam M. Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.*, 33(6):623–630, 2015. Corrigendum in Nat. Biotechnol. 33(10), 1109 (2015).

2 Andrei Z. Broder. On the resemblance and containment of documents. In *Compression and Complexity of Sequences (SEQUENCES'97)*, pages 21–29. IEEE, 1997.

3 Andrei Z. Broder, Moses Charikar, Alan M Frieze, and Michael Mitzenmacher. Min-wise independent permutations. In *Proceedings of the 30th annual ACM symposium on Theory of computing (STOC)*, pages 327–336. ACM, 1998.

4 Matthew Casperson. Minhash for dummies. `http://matthewcasperson.blogspot.de/2013/11/minhash-for-dummies.html`, November 2013.

5 P. Ferragina and G. Manzini. Indexing compressed text. *J. ACM*, 52(4):552—581, 2005.

6 Chirag Jain, Alexander Dilthey, Sergey Koren, Srinivas Aluru, and Adam M. Phillippy. A fast approximate algorithm for mapping long reads to large reference databases. In *International Conference on Research in Computational Molecular Biology*, pages 66–81. Springer, 2017.

7 Benjamin Kramer, Jens Quedenfeld, Sven Schrinner, Marcel Bargull, Kada Benadjemia, Jan Stricker, and David Losch. VATRAM – VAriant Tolerant ReAd Mapper. Technical report, Project Group PG583, Computer Science, TU Dortmund, Germany, 2015.

8 Heng Li. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, page btw152, 2016.

9 Victoria Popic and Serafim Batzoglou. Privacy-preserving read mapping using locality sensitive hashing and secure kmer voting. *bioRxiv*, page 046920, 2016.

10 Jens Quedenfeld and Sven Rahmann. Variant tolerant read mapping using min-hashing. *arXiv*, 1702.01703, 2017.

11 The Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, Oct 2016. Online first,. doi: 10.1093/bib/bbw089.

A Appendix: Proof details

A.1 Detailed calculation for Lemma 5

In Lemma 5,

$$P_{\text{SNP}} \geq P_{\text{DEFAULT}}$$

is by Lemma 3 equivalent to

$$\begin{aligned}
p &> \frac{d_0 - s}{d_0 - d_1} \\
&= \frac{\frac{m-e}{v+e} - \frac{m-e}{v+e+q}}{\frac{m-e}{v+e} - \frac{m-q-e+eq/m}{v+q+e-eq/m}} \\
&= \frac{(m-e)\cdot\frac{(v+e+q)-(v+e)}{(v+e)(v+e+q)}}{\frac{(m-e)(mv+mq+mv-eq)-(m^2-mq-me+eq/m)(v+e)}{(v+e)(mv+mq+me-eq)}} \\
&= \frac{\frac{q(m-e)}{v+e+q}}{\frac{m^2q-meq+qmv-eqv}{mv+mq+me-eq}} \\
&= \frac{(m-e)(mv+mq+me-eq)}{(v+e+q)\,(m^2-me+mv-ev)} \\
&= \frac{m(v+e+q)-eq}{(v+e+q)(m+v)} \\
&= \frac{m}{m+v} - \frac{eq}{(v+m)(v+e+q)}\,,
\end{aligned}$$

which is the final statement of Lemma 5. ◀