Finding all maximal perfect haplotype blocks in linear time

Jarno N Alanko, Hideo Bannai, Bastien Cazaux, Pierre Peterlongo, Jens Stoye

Finding all maximal perfect haplotype blocks in linear time

Jarno Alanko
Department of Computer Science, University of Helsinki, Finland
jarno.alanko@helsinki.fi

Hideo Bannai
Department of Informatics, Kyushu University, Japan
bannai@inf.kyushu-u.ac.jp

Bastien Cazaux
Department of Computer Science, University of Helsinki, Finland
bastien.cazaux@helsinki.fi

Pierre Peterlongo
Univ. Rennes, Inria, CNRS, Irisa, France
pierre.peterlongo@inria.fr

Jens Stoye
Faculty of Technology and Center for Biotechnology (CeBiTec), Bielefeld University, Germany
jens.stoye@uni-bielefeld.de

Abstract

Recent large-scale community sequencing efforts allow at an unprecedented level of detail the identification of genomic regions that show signatures of natural selection. Traditional methods for identifying such regions from individuals' haplotype data, however, require excessive computing times and therefore are not applicable to current datasets. In 2019, Cunha *et al.* (Proceedings of BSB 2019) suggested the *maximal perfect haplotype block* as a very simple combinatorial pattern, forming the basis of a new method to perform rapid genome-wide selection scans. The algorithm they presented for identifying these blocks, however, had a worst-case running time quadratic in the genome length. It was posed as an open problem whether an optimal, linear-time algorithm exists. In this paper we give two algorithms that achieve this time bound, one conceptually very simple one using suffix trees and a second one using the positional Burrows-Wheeler Transform, that is very efficient also in practice.

2012 ACM Subject Classification Theory of computation → Pattern matching; Applied computing → Bioinformatics; Mathematics of computing → Combinatorial algorithms; Applied computing → Computational genomics

Keywords and phrases Population genomics, selection coefficient, haplotype block, positional Burrows-Wheeler Transform

Digital Object Identifier 10.4230/LIPIcs.WABI.2019.8

Funding *Hideo Bannai*: JSPS KAKENHI Grant Number JP16H02783
Pierre Peterlongo: ANR Hydrogen ANR-14-CE23-0001

Acknowledgements We thank the organizers of DSB 2019 (`dsb2019.gitlab.io`) for giving us the opportunity to present earlier work in this area and start a discussion from which the present results originated. We would also like to thank Michel T. Henrichs for providing a script to convert VCF files to haplotype matrices and for assisting with the production of Figure 3.

1 Introduction and Background

As a result of the technological advances that went hand in hand with the genomics efforts of the last decades, today it is possible to experimentally obtain and study the genomes of large numbers of individuals, or even multiple samples from an individual. For instance, the *National Human Genome Research Institute* and the *European Bioinformatics Institute* have collected more than 3500 genome-wide association study publications in their *GWAS Catalog* [3].

Probably the most prominent example of large-scale sequencing projects is the *1000 Genomes Project* (now *International Genome Sample Resource*, IGSR), initiated with the goal of sequencing the genomes of more than one thousand human individuals to identify 95% of all genomic variants in the population with allele frequency of at least 1% (down toward 0.1% in coding regions). The final publications from phase 3 of the project report about genetic variations from more than 2,500 genomes [1, 12].

Recently, several countries announced large-scale national research programs to capture the diversity of their populations, while some of these efforts started already more than 20 years ago. Since 1996 Iceland's deCODE company is mining Icelanders' genetic and medical data for disease genes. In 2015, deCODE published insights gained from sequencing the whole genomes of 2,636 Icelanders [8]. *Genome of the Netherlands* (GoNL) is a whole genome sequencing project aiming to characterize DNA sequence variation in the Dutch population using a representative sample consisting of 250 trio families from all provinces in the Netherlands. In 2016, GoNL analysed whole genome sequencing data of 769 individuals and published a haplotype-resolved map of 1.9 million genome variants [10]. Similar projects have been established in larger scale in the UK: Following the *UK10K* project for identifying rare genetic variants in health and disease (2010–2013), *Genomics England* was set up in late 2012 to deliver the 100,000 Genomes Project [13]. This flagship project has by now sequenced 100,000 whole genomes from patients and their families, focusing on rare diseases, some common types of cancer, and infectious diseases. The scale of these projects is culminating in the US federal *Precision Medicine Initiative*, where the NIH is funding the *All of Us* research program[1] to analyze genetic information from more than 1 million American volunteers. Even more extreme suggestions go as far as to propose "to sequence the DNA of all life on Earth"[2].

The main motivation for the collection of these large and comprehensive data sets is the hope for a better understanding of genomic variation and how variants relate to health and disease, but basic research in evolution, population genetics, functional genomics and studies on demographic history can also profit enormously.

One important approach connecting evolution and functional genomics is the search for genomic regions under natural selection based on population data. The *selection coefficient* [7] is an established parameter quantifying the relative fitness of two genetic variants. Unfortunately, haplotype-based methods for estimating selection coefficients have not been designed with the massive genome data sets available today in mind, and may therefore take prohibitively long when applied to large-scale population data. In view of the large population sequencing efforts described above, methods are needed that – at similar sensitivity – scale to much higher dimensions.

---

[1] `allofus.nih.gov`

[2] Biologists propose to sequence the DNA of all life on Earth, by Elizabeth Pennisi. Science News, Feb. 24, 2017. `https://doi.org/10.1126/science.aal0824`.

Only recently a method for the fast computation of a genome-wide selection scan has been proposed that can be computed quickly even for large datasets [4]. The method is based on a very simple combinatorial string pattern, *maximal perfect haplotype blocks*. Although considerably faster than previous methods, the running time of the algorithm presented in that paper is not optimal, as it takes $O(kn^2)$ time in order to find all maximal perfect haplotype blocks in k genomes of length n each. This is sufficient to analyse individual human chromosomes on a laptop computer, for datasets of the size of the 1000 Genomes Project (1,000s of genomes and 1,000,000s of variations). However, with the larger datasets currently underway and with higher resolution it will not scale favourably. More efficient methods are therefore necessary and it was phrased as an open question whether there exists a linear-time algorithm to find all maximal perfect haplotype blocks.

In this paper we settle this open problem affirmatively. More specifically, after some basic definitions in Section 2 we present in Sections 3 and 4 two new algorithms for finding all maximal perfect haplotype blocks in optimal time. The latter of these two algorithms is then experimentally compared to the one from [4] in Section 5, proving its superiority in running time by a factor of about 5 and memory usage by up to two orders of magnitude for larger data sets. Section 6 concludes the paper.

2 Basic Definitions

The typical input to genome-wide selection studies is a set of haplotype-resolved genomes, or *haplotypes* for short. Clearly, for a given set of haplotypes only those sites are of interest where there is variation in the genomes. Therefore, formally, we consider as input to our methods a $k \times n$ *haplotype matrix* where each of the k rows corresponds to one haplotype and each of the n columns corresponds to one variable genetic site.

Most methods distinguish only between ancestral and derived allele, a consequence of the popular *infinite sites assumption* [11]. Therefore the entries in a haplotype matrix are often considered binary where the ancestral allele is encoded by 0 and the derived allele is encoded by 1. However, the computational problem and its solutions considered in this paper do not depend on this restriction and instead are applicable to any type of sequence over a constant-size alphabet Σ.

The concept of a maximal perfect haplotype block as defined in [4] is the following, where $S|_K$ denotes the elements of an ordered set S restricted to index set K:

▶ **Definition 1.** *Given k sequences $S = (s_1, \ldots, s_k)$ of same length n (representing the rows of a haplotype matrix), a* maximal perfect haplotype block *is a triple (K, i, j) with $K \subseteq \{1, \ldots, k\}$, $|K| \geq 2$ and $1 \leq i \leq j \leq n$ such that*

1. *$s[i, j] = t[i, j]$ for all $s, t \in S|_K$* (equality),
2. *$i = 1$ or $s[i-1] \neq t[i-1]$ for some $s, t \in S|_K$* (left-maximality),
3. *$j = n$ or $s[j+1] \neq t[j+1]$ for some $s, t \in S|_K$* (right-maximality), *and*
4. *$\nexists K' \subseteq \{1, \ldots, k\}$ with $K \subset K'$ such that $s[i, j] = t[i, j]$ for all $s, t \in S|_{K'}$* (row-maximality).

Definition 1 is illustrated in Figure 1.

In [4] it was shown that the number of maximal perfect haplotype blocks is in $O(kn)$, while the algorithm presented there takes $O(kn^2)$ time to find all blocks. It is based on the observation that branching vertices in the trie T_p of the suffixes of the input sequences starting at position p correspond to right-maximal and row-maximal blocks, while left-maximality can be tested by comparing T_p and T_{p-1}. In the next two sections we show how this running time can be improved.

0	1	0	1	0	1	0	0
1	0	1	1	1	1	0	1
0	1	0	1	1	1	0	0

Figure 1 Illustration of Definition 1: a binary 3×8 haplotype matrix with three maximal perfect haplotype blocks $(\{1,3\},1,4)$, $(\{2,3\},4,7)$ and $(\{1,2,3\},6,7)$ highlighted. (The example contains additional maximal perfect haplotype blocks that are not shown.)

3 Linear-Time Method I: Based on Suffix Trees

In this section, we present our first algorithm to find all maximal perfect haplotype blocks in linear time. This solution is purely theoretical, it would likely require large amounts of memory while being slow in practice. However, it demonstrates the connection to the concept of maximal repeats in strings. We recall from [9, Section 7.12] that a *maximal repeat* is a substring occurring at least twice in a string or a set of strings and such that it cannot be extended to the left or to the right without losing occurrences.

Let $\mathbb{S} = s_1\$_1 s_2\$_2 \dots s_k\$_k$, with the $\$_i$ being k different characters absent from the original alphabet Σ. The key point is that any maximal perfect haplotype block in S is a maximal repeat in $\mathbb{S}$. The opposite is not true: In a maximal perfect haplotype block, all occurrences of the repeat are located at the same position of each sequence of S (equality condition in Definition 1), while this constraint does not exist for maximal repeats in $\mathbb{S}$.

Nevertheless, finding all maximal perfect haplotype blocks in S can be performed by computing all maximal repeats in $\mathbb{S}$, while keeping only those whose occurrences are located at the same positions over all s_i in which they occur. This can be done by performing the following procedure:

1. "Decorate" each sequence $s_i \in S$ to create $s_i^+ = \alpha_0 s_i[1]\alpha_1 s_i[2]\alpha_2 \dots s_i[n]\alpha_n$, where the *index characters* $\alpha_0, \alpha_1, \dots, \alpha_n$ are $n+1$ symbols from an alphabet Σ', disjoint from the original alphabet Σ.
2. Find in $\mathbb{S}^+ = s_1^+\$_1 s_2^+\$_2 \dots s_k^+\$_k$ all maximal repeats.
3. Any maximal repeat $r = \alpha_p r_1 \alpha_{p+1} r_2 \alpha_{p+2} \dots r_\ell \alpha_{p+\ell}$ in $\mathbb{S}^+$ with $\ell \geq 1$ corresponds to a maximal perfect haplotype block of length ℓ, starting at position $p+1$ in the input sequences from S.

The key idea here is that the index characters impose that each maximal repeat occurrence starts at the same position in all sequences and, as a consequence, ensure that all occurrences occur in distinct sequences from S.

Hence any maximal repeat $r = \alpha_p r_1 \alpha_{p+1} \dots r_\ell \alpha_{p+\ell}$ defines a unique maximal perfect haplotype block $(K, p+1, p+\ell)$. The value $|K|$ is the number of occurrences of r. Also the set K can be derived from occurrence positions of r in $\mathbb{S}^+$, as any position in r corresponds to a unique position in $\mathbb{S}$. We prefer to omit useless technical details here.

The maximal repeat occurrences in $\mathbb{S}^+$ may be found using a suffix tree, constructed in time linear with respect to the size of the input data $O(kn)$, even for large integer alphabets [6], as we have here. The maximal repeat detection is also linear with the size of the input data [9, Section 7.12.1]. Therefore the overall time complexity is $O(kn)$.

1	2	3	4	5	6	a_6	a_6^{-1}
1	0	0	1	0	0	2	3
1	0	0	0	0	0	4	1
0	0	0	1	1	0	1	5
0	1	0	0	0	0	5	2
0	0	1	1	0	0	3	4

1	2	3	4	5	6	d_6
1	0	0	0	0	0	7
0	1	0	0	0	0	3
1	0	0	1	0	0	5
0	0	1	1	0	0	4
0	0	0	1	1	0	6

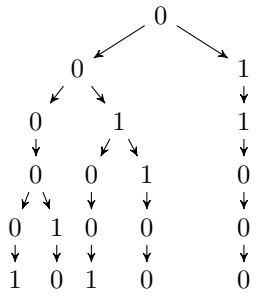

Figure 2 Available blocks. **Left:** an example of a haplotype matrix up to column 6 with the two arrays a_6 and a_6^{-1} on the right. **Center:** the colexicographically sorted rows and the array d_6 listed on the right. **Right:** the trie of the reverses of the rows of the matrix. For example, the block $(\{1,2,4,5\},5,6)$ is available because $a_6^{-1}(1)=3$, $a_6^{-1}(2)=1$, $a_6^{-1}(4)=2$, $a_6^{-1}(5)=4$ is the consecutive range $[x,y]=[1,4]$, we have $d_6[r]\leq 5$ for all $r\in[1+1,4]$ with $d_6[3]=5$, and we have $x=1$ and $d_6[4+1]=6>5$. The repeat in the block is 00, and we see it is a branching node in the trie on the right.

4 Linear-Time Method II: Based on the Positional BWT

Here we present our second algorithm to find all maximal perfect haplotype blocks in linear time. It works by scanning the haplotype matrix column by column while maintaining the positional Burrows-Wheeler Transform (pBWT) [5] of the current column. For simplicity of presentation we assume that all rows of the haplotype matrix S are distinct. Recall that the pBWT of S consists of a pair of arrays for each column of S: For each l, $1\leq l\leq n$, we have arrays a_l and d_l of length k such that the array a_l is a permutation of $[1,k]$ with $S[a_l[1]][1..l]\leq\ldots\leq S[a_l[k]][1..l]$ colexicographically (i.e. right-to-left lexicographically) and the array d_l is such that $d_l[1]=l+1$ and for all r, $1<r\leq k$, we have $d_l[r]=1+\max\{j\in[1,l]:S[a_l[r]][j]\neq S[a_l[r-1]][j]\}$. Further let us denote by a_l^{-1} the inverse permutation of a_l. For readers familiar with string processing terminology, the arrays a_l and a_l^{-1} are analogous to the suffix array and the inverse suffix array, respectively, while the arrays d_l are analogous to the LCP array. We note that the term pBWT is somewhat misleading, since there is no array analogous to the BWT.

Conditions 1, 2 and 4 (equality, left-maximality and row-maximality) of Definition 1 can be stated in terms of the arrays a_l and d_l as follows. Suppose we have a block (K,i,j). If the block is a maximal perfect haplotype block, then the set $\{a_j^{-1}[r]\mid r\in K\}$ must be a contiguous range $[x,y]$ of indices such that the following holds:

- $d_j[r]\leq i$ for all $r\in[x+1,y]$ (equality),
- there exists at least one $r\in[x+1,y]$ such that $d_j[r]=i$ (left-maximality), and
- $(x=1$ or $d_j[x]>i)$ and $(y=k$ or $d_j[y+1]>i)$ (row-maximality).

We call a block satisfying these conditions an *available* block and $[x,y]$ the *colexicographic range* of the block. Let us consider the set B_l of available blocks ending at column l. We have that $|B_l|\leq k$ because each available block corresponds to a distinct branching node in the trie of the reverses of $\{S[1][1..l],\ldots,S[k][1..l]\}$, and the number of branching nodes in the trie is bounded from above by the number of leaves k. The branching nodes of the trie can be enumerated in $O(k)$ time by using a standard algorithm [2] for enumerating LCP intervals of the LCP array of the trie, $LCP_l[r]=l-d_l[r]+1$. This gives us the colexicographic ranges $[x,y]$ of all available blocks in B_l. An example is shown in Figure 2.

The only thing left is to show how to check the right-maximality property of an available

block, i.e., whether $j = n$ or $|\{S[a[r]][j+1] : r \in [x, y]\}| \geq 2$. To check the condition in constant time for $j \neq n$, we build a bit vector V_j such that $V_j[1] = 1$ and $V_j[r] = 1$ if and only if $S[a_j[r]][j+1] \neq S[a_j[r-1]][j+1]$. Now the block is right-maximal if and only if $V_j[x+1..y]$ contains at least one 1-bit. We can build a vector of prefix sums of V_j to answer this question in constant time.

Time and space complexity

We assume the *column stream model*, where we can stream the haplotype matrix column by column. We can thus build the arrays d_l, a_l and a_l^{-1} on the fly column by column [5], and also easily build the required prefix sums of arrays V_l from these. The time is $O(nk)$, since each of the n columns takes $O(k)$ time to process. The algorithm needs to keep in memory only the data for two adjacent columns at a time, so in space $O(k)$ we can report the colexicographic ranges of all maximal blocks ending in each column $l \in [1, n]$. If the colexicographic range of a block at column l is $[x, y]$, then the rows in the original haplotype matrix are $a_l[x], a_l[x+1], \ldots, a_l[y]$. There are $O(nk)$ blocks and $O(k)$ rows per block, so the time to report all rows explicitly is $O(nk^2)$. Alternatively, we can store a complete representation of the answer taking $O(nk)$ space by storing all the a_l arrays and the colexicographic ranges of the maximal perfect blocks for each column, from which we can readily report all rows in any maximal perfect block in constant time per row.

5 Empirical Evaluation

Since the algorithm of Section 3 is mostly of theoretical interest, we evaluate only the pBWT-based algorithm presented in Section 4. The source code is available from `https://gitlab.com/bacazaux/haploblocks`. As a baseline for comparison we use the implementation of the trie-based algorithm by Cunha et al. [4], available from the same gitlab site. The experiments were run on a machine with an Intel Xeon E5-2680 v4 2.4GHz CPU, which has a 35 MB Intel SmartCache. The machine has 256 gigabytes of memory at a speed of 2400MT/s. The code was compiled with `g++` using the `-Ofast` optimization flag.

Our test data consists of chromosomes 2, 6 and 22 from phase three of the 1000 Genomes Project [1], which provides whole-genome sequences of 2,504 individuals from multiple populations worldwide. We preprocessed the data by extracting all biallelic SNPs from the provided VCF files[3] and converting them to a binary haplotype matrix using our own program `vcf2bm`, also available from `https://gitlab.com/bacazaux/haploblocks`.

Our implementation has a user-defined parameter allowing to adjust the minimum size of a reported maximal perfect haplotype block (K, i, j), where *size* is defined as the width $(j - i + 1)$ times the number of rows $(|K|)$ in the block. Table 1 shows the running times and memory usage of our implementation on the different chromosomes and for different settings of the minimum block size parameter. The larger the minimum block size, the faster the algorithm is, because there are less blocks to report. In general, it takes only a few minutes to process a complete human chromosome. Locating all 323,163,970 blocks of minimum size 10^6 in all 22 human autosomes (non-sex chromosomes) took in total 4 hours and 26 minutes with a memory peak of 12.8 MB (data not shown).

Table 2 shows a comparison of our implementation to the trie-based implementation from [4]. Our implementation is about 5 times faster on all datasets, and the memory

[3] `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`

data set	#lines	#columns	min block size	time	memory	#blocks
chr. 22	5,008	1,055,454		4min 54s	12.8MB	148,613,645
chr. 22	5,008	1,055,454	500,000	3min 50s	12.8MB	16,076,453
chr. 22	5,008	1,055,454	1,000,000	3min 40s	12.8MB	2,228,762
chr. 22	5,008	1,055,454	2,000,000	3min 43s	12.8MB	4,779
chr. 6	5,008	4,800,101		19min 42s	12.8MB	624,689,548
chr. 6	5,008	4,800,101	500,000	17min 20s	12.8MB	89,840,467
chr. 6	5,008	4,800,101	1,000,000	16min 30s	12.8MB	11,388,982
chr. 6	5,008	4,800,101	2,000,000	16min 36s	12.8MB	5,585
chr. 2	5,008	6,786,300		31min 57s	12.8MB	946,717,897
chr. 2	5,008	6,786,300	500,000	25min 06s	12.8MB	160,094,115
chr. 2	5,008	6,786,300	1,000,000	23min 24s	12.8MB	25,533,314
chr. 2	5,008	6,786,300	2,000,000	23min 18s	12.8MB	120,243

Table 1 Running times and memory usage of our pBWT-based implementation. Note that in our streaming implementation the memory usage is dominated by the number of haplotypes times the buffer size, and therefore is essentially constant in this study.

	`trie`		`pBWT`	
data set	time	memory	time	memory
chr. 22	17min 08s	927.8MB	3min 40s	12.8MB
chr. 6	1h 34min 34s	3.23GB	16min 30s	12.8MB
chr. 2	2h 07min 21s	4.46GB	23min 24s	12.8MB

Table 2 Comparison of the trie-based implementation from [4] and our pBWT-based implementation with minimum block size 10^6.

consumption is up to 93 times smaller.

Using the method for estimating a local selection coefficient from the size of maximal perfect haplotype blocks covering a certain genomic region presented in [4], it is now possible to generate chromosome-wide selection scans indicating the loci of maximum selection, as shown in Figure 3 for the complete human chromosome 2 (size parameter 10^6), in less than half an hour.

6 Conclusion

In this paper we presented two algorithms that are able to find all maximal perfect haplotype blocks in a haplotype matrix of size $k \times n$ in linear time $O(kn)$. In particular the second method, based on the positional Burrows-Wheeler Transform, performs also extremely well in practice, as it allows for a streaming implementation with extremely low memory footprint.

While an initial implementation of the method is available from `https://gitlab.com/bacazaux/haploblocks`, a user-friendly software combining the algorithm presented here with the computation of the selection coefficient suggested in [4] remains to be developed.

References

1 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A

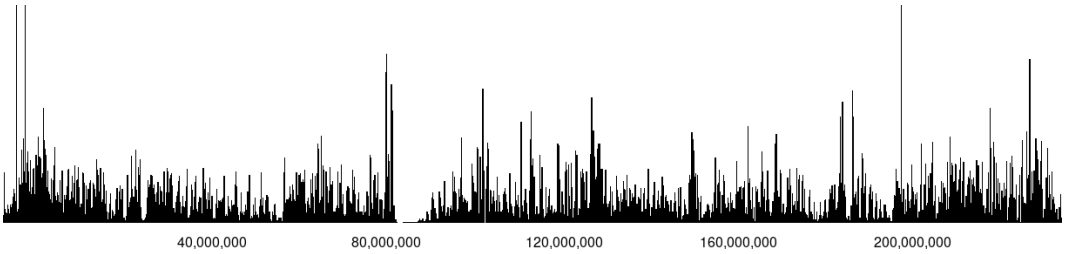

Figure 3 Selection scan for human chromosome 2. Shown is for each position of the chromosome the largest maximum likelihood estimate derived from any maximal perfect haplotype block overlapping that locus. It is easy to spot potential regions of high selection. The centromere, located around 93 Mbp, shows no signal as sequencing coverage is low here and no SNPs could be called.

McVean, and Gonçalo R Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. doi:10.1038/nature15393.

2 Mohamed Ibrahim Abouelhoda, Stefan Kurtz, and Enno Ohlebusch. Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms*, 2(1):53–86, 2004. doi:10.1016/S1570-8667(03)00065-0.

3 Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, Daniel Suveges, Olga Vrousgou, Patricia L Whetzel, Ridwan Amode, Jose A Guillen, Harpreet S Riat, Stephen J Trevanion, Peggy Hall, Heather Junkins, Paul Flicek, Tony Burdett, Lucia A Hindorff, Fiona Cunningham, and Helen Parkinson. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, 2018. doi:10.1093/nar/gky1120.

4 Luís Cunha, Yoan Diekmann, Luis Antonio Brasil Kowada, and Jens Stoye. Identifying maximal perfect haplotype blocks. In *Advances in Bioinformatics and Computational Biology - 11th Brazilian Symposium on Bioinformatics, BSB 2018, Niterói, Brazil, October 30 - November 1, 2018, Proceedings*, pages 26–37, 2018. doi:10.1007/978-3-030-01722-4_3.

5 Richard Durbin. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics*, 30(9):1266–1272, 2014. doi:10.1093/bioinformatics/btu014.

6 Martin Farach. Optimal suffix tree construction with large alphabets. In *Proceedings 38th Annual Symposium on Foundations of Computer Science*, pages 137–143. IEEE, 1997.

7 John H. Gillespie. *Population Genetics – A Concise Guide*. The Johns Hopkins University Press, Baltimore and London, 1998.

8 Daniel F Gudbjartsson, Hannes Helgason, Sigurjon A Gudjonsson, Florian Zink, Asmundur Oddson, Arnaldur Gylfason, Søren Besenbacher, Gisli Magnusson, Bjarni V Halldorsson, Eirikur Hjartarson, Gunnar Th Sigurdsson, Simon N Stacey, Michael L Frigge, Hilma Holm, Jona Saemundsdottir, Hafdis Th Helgadottir, Hrefna Johannsdottir, Gunnlaugur Sigfusson, Gudmundur Thorgeirsson, Jon Th Sverrisson, Solveig Gretarsdottir, G Bragi Walters, Thorunn Rafnar, Bjarni Thjodleifsson, Einar S Bjornsson, Sigurdur Olafsson, Hildur Thorarinsdottir, Thora Steingrimsdottir, Thora S Gudmundsdottir, Asgeir Theodors, Jon G Jonasson, Asgeir Sigurdsson, Gyda Bjornsdottir, Jon J Jonsson, Olafur Thorarensen, Petur Ludvigsson, Hakon Gudbjartsson, Gudmundur I Eyjolfsson, Olof Sigurdardottir, Isleifur Olafsson, David O Arnar, Olafur Th Magnusson, Augustine Kong, Gisli Masson, Unnur Thorsteinsdottir, Agnar Helgason, Patrick Sulem, and Kari Stefansson. Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, 47:435–444, 2015. doi:10.1038/ng.3247.

9 Dan Gusfield. *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press, Cambridge, 1997.

10 Jayne Y Hehir-Kwa, Tobias Marschall, Wigard P Kloosterman, Laurent C Francioli, Jasmijn A Baaijens, Louis J Dijkstra, Abdel Abdellaoui, Vyacheslav Koval, Djie Tjwan Thung, René Wardenaar, Ivo Renkens, Bradley P Coe, Patrick Deelen, Joep de Ligt, Eric-Wubbo Lameijer, Freerk van Dijk, Fereydoun Hormozdiari, The Genome of the Netherlands Consortium, Jasper A Bovenberg, Anton J M de Craen, Marian Beekman, Albert Hofman, Gonneke Willemsen, Bruce Wolffenbuttel, Mathieu Platteel, Yuanping Du, Ruoyan Chen, Hongzhi Cao, Rui Cao, Yushen Sun, Jeremy Sujie Cao, Pieter B T Neerincx, Martijn Dijkstra, George Byelas, Alexandros Kanterakis, Jan Bot, Martijn Vermaat, Jeroen F J Laros, Johan T den Dunnen, Peter de Knijff, Lennart C Karssen, Elisa M van Leeuwen, Najaf Amin, Fernando Rivadeneira, Karol Estrada, Jouke-Jan Hottenga, V Mathijs Kattenberg, David van Enckevort, Hailiang Mei, Mark Santcroos, Barbera D C van Schaik, Robert E Handsaker, Steven A McCarroll, Arthur Ko, Peter Sudmant, Isaac J Nijman, André G Uitterlinden, Cornelia M van Duijn, Evan E Eichler, Paul I W de Bakker, Morris A Swertz, Cisca Wijmenga, Gert-Jan B van Ommen, P Eline Slagboom, Dorret I Boomsma, Alexander Schönhuth, Kai Ye, and Victor Guryev. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nature Communications*, 7:12989, 2016. doi:10.1038/ncomms12989.

11 Motoo Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893, 1969.

12 Peter H Sudmant, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, Miriam K Konkel, Ankit Malhotra, Adrian M Stütz, Xinghua Shi, Francesco Paolo Casale, Jieming Chen, Fereydoun Hormozdiari, Gargi Dayama, Ken Chen, Maika Malig, Mark J P Chaisson, Klaudia Walter, Sascha Meiers, Seva Kashin, Erik Garrison, Adam Auton, Hugo Y K Lam, Xinmeng Jasmine Mu, Can Alkan, Danny Antaki, Taejeong Bae, Eliza Cerveira, Peter Chines, Zechen Chong, Laura Clarke, Elif Dal, Li Ding, Sarah Emery, Xian Fan, Madhusudan Gujral, Fatma Kahveci, Jeffrey M Kidd, Yu Kong, Eric-Wubbo Lameijer, Shane McCarthy, Paul Flicek, Richard A Gibbs, Gabor Marth, Christopher E Mason, Androniki Menelaou, Donna M Muzny, Bradley J Nelson, Amina Noor, Nicholas F Parrish, Matthew Pendleton, Andrew Quitadamo, Benjamin Raeder, Eric E Schadt, Mallory Romanovitch, Andreas Schlattl, Robert Sebra, Andrey A Shabalin, Andreas Untergasser, Jerilyn A Walker, Min Wang, Fuli Yu, Chengsheng Zhang, Jing Zhang, Xiangqun Zheng-Bradley, Wanding Zhou, Thomas Zichner, Jonathan Sebat, Mark A Batzer, Steven A McCarroll, The 1000 Genomes Project Consortium, Ryan E Mills, Mark B Gerstein, Ali Bashir, Oliver Stegle, Scott E Devine, Charles Lee, Evan E Eichler, and Jan O Korbel. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015. doi:10.1038/nature15394.

13 Clare Turnbull, Richard H Scott, Ellen Thomas, Louise Jones, Nirupa Murugaesu, Freya Boardman Pretty, Dina Halai, Emma Baple, Clare Craig, Angela Hamblin, Shirley Henderson, Christine Patch, Amanda O'Neill, Andrew Devereau, Katherine Smith, Antonio Rueda Martin, Alona Sosinsky, Ellen M McDonagh, Razvan Sultana, Michael Mueller, Damian Smedley, Adam Toms, Lisa Dinh, Tom Fowler, Mark Bale, Tim J P Hubbard, Augusto Rendon, Sue Hill, Mark J Caulfield, and 100 000 Genomes Project. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ*, 361:k1687, 2018. doi:10.1136/bmj.k1687.