Microsatellites; genotyping, mutation rate and effect on disease

by

Snædís Kristmundsdóttir

Dissertation submitted to the School of Technology, Department of Engineering
at Reykjavík University in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

June 2022

Thesis Committee:

Bjarni V. Halldórsson, Supervisor
Associate Professor, Reykjavík University, Iceland

Daníel F. Guðbjartsson, Co-advisor
Research professor, University of Iceland, deCODE Genetics, Iceland

Sigrún Helga Lund, Co-advisor
Research Scientist, deCODE Genetics, Iceland

Eyjólfur Ingi Ásgeirsson, Co-advisor
Associate Professor, Reykjavík University, Iceland

Melissa Gymrek, Examiner
Assistant Professor, University of California San Diego, USA

ISBN 978-9935-9655-2-3 electronic version
ISBN 978-9935-9655-3-0 printed version
ORCID: 0000-0001-8981-0883

The undersigned hereby certify that they recommend to the Department of Engineering, School of Technology, Reykjavík University, that this Dissertation entitled "**Microsatellites; genotyping, mutation rate and effect on disease**", submitted by **Snædís Kristmundsdóttir**, be accepted as partial fulfilment of the requirements for the degree of **Doctor of Philosophy (Ph.D.) in Engineering**

June 30th, 2022

date

Bjarni V. Halldórsson, Supervisor
Associate Professor, Reykjavík University, Iceland

Daníel F. Guðbjartsson, Co-advisor
Research professor, University of Iceland, deCODE Genetics, Iceland

Sigrún Helga Lund

Sigrún Helga Lund, Co-advisor
Research Scientist, deCODE Genetics, Iceland

Eyjólfur Ingi Ásgeirsson, Co-advisor
Associate Professor, Reykjavík University, Iceland

Melissa Gymrek, Examiner
Assistant Professor, University of California San Diego, USA

....
date

....
Snædís Kristmundsdóttir
Doctor of Philosophy

Microsatellites; genotyping, mutation rate and effect on disease

Snædís Kristmundsdóttir

June 2022

Abstract

Microsatellites are polymorphic tracts of short tandem repeats (STRs) with one to six base-pair (bp) motifs and account for around 3% of the human genome. Just like copying by hand a text where the same word occurs many times in a row, the replication of microsatellites is error prone and frequently adds or removes one or more copies of the repeat motif. As a result, microsatellites mutate several orders of magnitude faster than unique genomic sequences and for a given microsatellite, a population can have many possible length variations. The first objective of this study was to implement a method to jointly determine the number of repeats present at each microsatellite in the genome for a large number of samples. The second goal was to make the determination of repeat numbers more computationally efficient while simultaneously increasing the detection sensitivity of heavily expanded microsatellite alleles, known as repeat expansions. Last, the software was run on two large sets of whole genome sequenced individuals, one from Iceland and the other from the UK biobank. Using the genealogy information available on the Icelandic set, de novo mutation events were detected and the effects of parental sex, age and genotypes on the types and number of mutations found in their offspring were estimated.

Keywords: Microsatellites, Genotyping, Mutations

Örtungl; Arfgerðarákvörðun, stökkbreytitíðni og áhrif á sjúkdóma

Snædís Kristmundsdóttir

júní 2022

Útdráttur

Um það bil þrjú prósent af erfðamengi mannsins eru örtungl, en þau eru fjölbreytilegar raðir af stuttum samliggjandi endurtekningum þar sem endurtekna röðin er á bilinu einn til sex basar á lengd. Líkt og við afritun á texta þar sem sama orðið er endurtekið oft í röð, þá er villuhættan meiri við afritun örtunglaraða en við aðrar raðir erfðamengisins og afleiðingin er að endurtekningu er bætt við eða hún tapast miðað við upprunalega basaröð. Vegna þessa stökkbreytast örtungl nokkrum stærðargráðum hraðar en aðrar raðir erfðamengisins og fyrir ákveðið örtungl getur hópur af fólki haft margar mismunandi lengdarútgáfur. Fyrsta markmið verkefnisins var að hanna og skrifa hugbúnað sem gæti ákvarðað fjölda endurtekninga fyrir öll örtungl í erfðamenginu hjá mörgun einstaklingum í einu. Næst, var reikniritinu hraðað en það jafnframt gert næmara fyrir stórum útþenslu örtungla samsætum, sem geta valdið mörgum mismunandi heilkennum hjá þeim sem þær bera. Að lokum var hugbúnaðurinn notaður til að meta arfgerð allra einstaklinga í tveimur stórum þýðum, frá Íslandi annars vegar og Bretlandi hins vegar. Ættfræðiupplýsingar um íslenska þýðið voru notaðar til að greina stökkbreytingar í afkvæmum sem ekki fundust í foreldrum og stökkbreytingarnar notaðar til að meta hvernig aldur kyn og arfgerð foreldra hefur áhrif á tegund og fjölda stökkbreytinga sem þeir arfleiða afkvæmi sín að.

Efnisorð: Örtungl, Arfgerðarákvörðun, Stökkbreytingar

I dedicate this to my daughters, my husband, my parents, my siblings, my friends and my colleagues.

Acknowledgements

First of all I would like to thank my sister for being the absolute best person in the world and for her excellent choice of spouse. I would then like to thank said spouse for being the most influential person in my life. I would most definitely not be where I am today without you.

Last, I would like to express my deepest gratitude to my colleagues at deCODE genetics for helping, listening and sharing your endless knowledge with me.

Preface

This dissertation is original work by the author, Snædís Kristmundsdóttir. Two of the papers it consists of have been published in Bioinformatics and the third has been submitted for review.

Publications

Along with the three papers this thesis consists of, I have also contributed to these publications as a part of my thesis work:

Gunnarsson, Bjarni, et al. "A sequence variant associating with educational attainment also affects childhood cognition." Scientific reports 6 (2016): 36189.

Jónsson, Hákon, et al. "Parental influence on human germline de novo mutations in 1,548 trios from Iceland." Nature 549.7673 (2017): 519.

Eggertsson, Hannes P., et al. "Graphtyper enables population-scale genotyping using pangenome graphs." Nature genetics 49.11 (2017): 1654.

Jónsson, Hákon, et al. "Whole genome characterization of sequence diversity of 15,220 Icelanders." Scientific data 4 (2017): 170115.

Kehr, Birte, et al. "Diversity in non-repetitive human sequences not found in the reference genome." Nature genetics 49.4 (2017): 588.

Jónsson, Hákon, et al. "Multiple transmissions of de novo mutations in families." Nature genetics 50.12 (2018): 1674.

Zink, Florian, et al. "Insights into imprinting from parent-of-origin phased methylomes and transcriptomes." Nature Genetics 50.11 (2018): 1542.

Benonisdottir, Stefania, et al. "Sequence variants associating with urinary biomarkers." Human molecular genetics 28.7 (2019): 1199-1211.

Eggertsson, Hannes P., et al. "GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs." Nature communications 10.1 (2019): 1-8.

Gudbjartsson, Daniel F., et al. "Lipoprotein (a) Concentration and Risks of Cardiovascular Disease and Diabetes." Journal of the American College of Cardiology 74.24 (2019): 2982-2994.

Ivarsdottir, Erna V., et al. "Sequence variation at ANAPC1 accounts for 24% of the variability in corneal endothelial cell density." Nature communications 10.1 (2019): 1284.

Holley, Guillaume, et al. "Ratatosk: hybrid error correction of long reads enables accurate variant calling and assembly." Genome Biology 22.1 (2021): 1-22.

Beyter, Doruk, et al. "Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits." Nature Genetics 53.6 (2021): 779-786.

Bell, Steven, et al. "A genome-wide meta-analysis yields 46 new loci associating with biomarkers of iron homeostasis." Communications biology 4.1 (2021): 1-14.

Björnsson, Eythór, et al. "Large-scale screening for monogenic and clinically defined familial hypercholesterolemia in Iceland." Arteriosclerosis, thrombosis, and vascular biology 41.10 (2021): 2616-2628.

Olafsdottir, Thorhildur, et al. "Loss-of-Function Variants in the Tumor-Suppressor Gene PTPN14 Confer Increased Cancer Risk." Cancer Research 81.8 (2021): 1954-1964.

Halldorsson, Bjarni V., et al. "The sequences of 150,119 genomes in the UK biobank." bioRxiv (2021).

Contents

Chapter 1

Introduction

Here I will present relevant concepts and give an outline of the papers my thesis work consists of.

1.1 Background

Microsatellites are polymorphic tandem repeats of DNA sequences, with repeat motif lengths ranging from one to six base pairs (bp). They are abundant in the human genome and have a higher mutation rate than the more standard types of genetic variation, single nucleotide polymorphisms (SNPs) and small insertion/deletions (indels).

Because of their repetitive nature, errors occur at microsatellites when DNA is copied, allowing for mutations to occur from parent to offspring. These errors occur frequently enough that microsatellites are typically highly polymorphic, i.e. different individuals have different copy counts of the microsatellite repeat. Unfortunately, these errors are also frequently introduced during sample preparation, causing complications in determining the copy count carried by the individual.

The goal of my research can be split into distinct sub goals, each contributing to incorporating microsatellite polymorphisms into genetic association studies - studies which correlate the genetic differences of individuals with the diseases and human conditions they exhibit. Genetic association studies are typically only performed for SNPs and indels while microsatellites are often not used, due to their complicated nature, despite the fact that they span 3% of the genome and have a high information content.

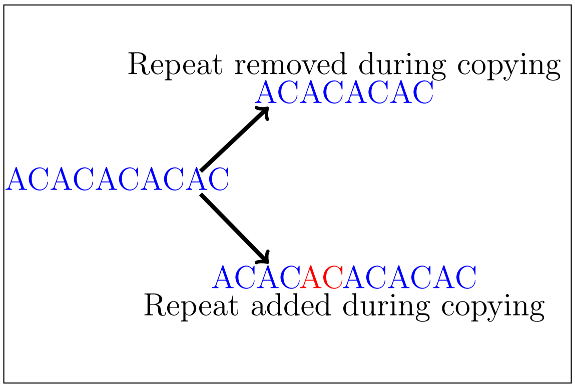

Figure 1.1: Copying microsatellite DNA sequences can result in both addition and removal of repeat motifs.

1.2 Subgoals and their respective papers

My first goal was to efficiently obtain reliable and consistent microsatellite repeat counts across a large set of whole genome sequenced DNA samples. To this end I developed mathematical models and algorithms, implemented in the microsatellite genotyper popSTR. I showed that popSTR outperforms other available microsatellite genotypers in terms of accuracy. I also showed that it is sufficiently efficient to be run on large sample sets, where several petabytes of sequence data need to be analyzed. The method and these results were published in the peer reviewed journal, Bioinformatics in 2017 as "popSTR: population-scale detection of STR variants" which is included as Chapter 2.

After establishing the applicability of the method at a population scale, I decided to broaden the spectrum of popSTR's utilization. To this end I collaborated with the clinical sequencing team at deCODE genetics and created a pipeline for inspecting known pathogenic repeat expansions microsatellites. During this work I also updated the algorithm and software in several ways, both with respect to run time and genotyping quality. As a part of my effort to allow for microsatellites to be processed in the same manner as other genetic variants in downstream association studies binSTR was added to the software package. binSTR groups microsatellite alleles according to user specified constraints to enable testing in an association study a group of alleles against another group, as opposed to testing a single allele against all others. These improvements, extensions and updates were also published in Bioinformatics in 2019 as "popSTR2 enables clinical and population-scale genotyping of microsatellites" which is included as Chapter 3.

My last sub goal was to run the software at scale and to estimate the polymorphism and mutation rate of microsatellites. To this end I genotyped 53,026 Icelandic samples and 150,119 samples from the UK biobank (UKB) at 5,401,401 tandem repeats. popSTR allowed for the calling of these sets, which are probably the world's two largest sets of microsatellites to date. The Icelandic set contained 6,082 parent offspring trios from which I interpreted reliable Mendelian inheritance violations as microsatellite de novo mutations (mDNMs) and used them to estimate the microsatellite mutation rate. I further examined contributing factors to the mutation of microsatellites such as genomic location, motif length, reference repeat tract length, base pair content and age of transmitting parent at conception. These results demonstrated a previously unknown increase in mDNMs with maternal age. Using the mDNM set, I defined parental phenotypes for all parents in the trios quantifying the number of mDNMs transmitted to their offspring and performed a genome wide association. The association returned two genome wide significant variants, both located in coding regions of DNA repair genes and both increasing the number of mDNMs transmitted from parent to offspring. This is the first time that common genetic variants have been found that influence mutation rate. This work has been submitted for review, see Chapter 4.

Results generated by the software this thesis centers around have been reported in several other projects and has contributed to research performed by a number of scientists at deCODE genetics. The list presented on page xv gives an overview of these projects and others I have contributed to applying the expertise I have obtained during my PhD studies.

Chapter 2

popSTR: population-scale detection of STR variants

Bioinformatics, 33(24), 2017, 4041–4048
doi: 10.1093/bioinformatics/btw568
Advance Access Publication Date: 1 September 2016
Original Paper

Sequence analysis

popSTR: population-scale detection of STR variants

Snædís Kristmundsdóttir[1,*], Brynja D. Sigurpálsdóttir[2], Birte Kehr[1] and Bjarni V. Halldórsson[1,2,*]

[1]deCODE genetics/Amgen and [2]School of Science and Engineering, Reykjavík University, Reykjavík, 101, Iceland

*To whom correspondence should be addressed.

Associate Editor: Gunnar Ratsch

Received on April 4, 2016; revised on June 1, 2016; accepted on August 26, 2016

Abstract

Motivation: Microsatellites, also known as short tandem repeats (STRs), are tracts of repetitive DNA sequences containing motifs ranging from two to six bases. Microsatellites are one of the most abundant type of variation in the human genome, after single nucleotide polymorphisms (SNPs) and Indels. Microsatellite analysis has a wide range of applications, including medical genetics, forensics and construction of genetic genealogy. However, microsatellite variations are rarely considered in whole-genome sequencing studies, in large due to a lack of tools capable of analyzing them.

Results: Here we present a microsatellite genotyper, optimized for Illumina WGS data, which is both faster and more accurate than other methods previously presented. There are two main ingredients to our improvements. First we reduce the amount of sequencing data necessary for creating microsatellite profiles by using previously aligned sequencing data. Second, we use population information to train microsatellite and individual specific error profiles. By comparing our genotyping results to genotypes generated by capillary electrophoresis we show that our error rates are 50% lower than those of lobSTR, another program specifically developed to determine microsatellite genotypes.

Availability and Implementation: Source code is available on Github: https://github.com/DecodeGenetics/popSTR

Contact: snaedis.kristmundsdottir@decode.is or bjarni.halldorsson@decode.is

1 Introduction

Microsatellites (a.k.a. short tandem repeats, STRs) are short DNA sequences containing a repeated motif of length 2–6 base pairs. The human reference genome contains approximately 1 million microsatellites, covering almost 1% of the genome (Gymrek *et al.*, 2016). Microsatellites have a mutation rate estimated between 1×10^{-4} and 1×10^{-3} mutations per locus per generation (Sun *et al.*, 2012), much higher than the mutation rate estimated for SNPs (Kong *et al.*, 2012) of 1.2×10^{-8}. Due to their high mutation rate, the alleles of a microsatellite vary greatly between individuals (Sun *et al.*, 2012). Apart from identical twins, no pair of individuals alive today has the same combination of alleles for all microsatellites (Cox and Mays, 2000). Using relatively few microsatellites, it is possible to create a unique genetic profile for every individual (Cox and Mays, 2000), making microsatellites appealing for applications such as forensic analysis (Veselinović, 2006).

Their high mutation rate made microsatellites particularly alluring for genotyping during the linkage era (Gudbjartsson *et al.*, 2000). Despite their abundance and the increasing availability of whole genome sequencing data, microsatellites are however often neglected in GWAS studies (Gudbjartsson *et al.*, 2015), in large due to a lack of tools capable of analyzing them (Duitama *et al.*, 2014).

The high mutation rate can be attributed to the repetitive structure of microsatellites, which causes a secondary DNA conformation that makes replication slippage events more likely than in other locations of the genome (Mirkin, 2007). Replication slippage occurs during DNA replication when the copy strand being created and the original template strand get shifted in their relative positions, causing a part of the template to either be copied twice or not copied at all (cf. Fig. 1), resulting in either an increase or decrease in the number of motif repeats (Brown, 2002).

Replication slippage can occur within individual cells, as well as when the DNA sample is being analyzed. A slippage event that occurs during replication of a sex cell results in a germline mutation and may be passed on to an offspring, while slippage events within other cells of the body lead to somatic mutations. Slippage events also frequently occur in PCR amplification, a pre-processing step often performed prior to sequencing, or during sequencing itself. As a result, the sequence reads of an individual contain both reads from its germline variants and reads resulting from slippage events, complicating the genotyping of microsatellites.

The genotyping of microsatellites is further complicated by the fact that their high mutation rate can make it difficult to align microsatellite reads to the correct location on the genome; most popular read-to-reference aligners trade-off between the tolerance of insertions/deletions and running time. Yet another complication is the length of the microsatellite, as aligners generally require a unique match to the genome to seed their alignment, reads that are fully contained within a microsatellite can often not be placed within the genome. Further, reads that do not fully encompass the microsatellite and only contain a portion of the microsatellite can only give a lower bound on the number of repeats (Gymrek *et al.*, 2012).

A number of methods have been developed to genotype microsatellites (Gelfand *et al.*, 2014; Gymrek *et al.*, 2012; Highnam *et al.*, 2013). We present *popSTR*, a method capable of studying microsatellite (*STR*) variation within all individuals of a *pop*ulation simultaneously. Microsatellite mutation rates have been shown to vary greatly between microsatellites as well as between individuals (Sun *et al.*, 2012). Consequently, our model allows for an error model specific to each microsatellite and individual being studied.

Our results show that popSTR is both faster and more accurate than lobSTR (Gymrek *et al.*, 2012), a previously described method for determining microsatellites. popSTR also finds more microsatellite genotypes than the general purpose genotype caller GATK (McKenna *et al.*, 2010), with the ones found also being more reliable.

2 Methods

popSTR requires three inputs; a reference genome, a list of microsatellite locations (markers) on the reference genome and sequencing data of the set of individuals (population) being studied. We assume that the sequencing data is Illumina whole genome paired-end sequencing data, mapped to the reference genome and stored in BAM-files, with one BAM file per individual. The output of popSTR are for each marker the set of alleles occurring in at least one individual in the population and the genotype likelihoods of all allele pairs of the marker for each individual.

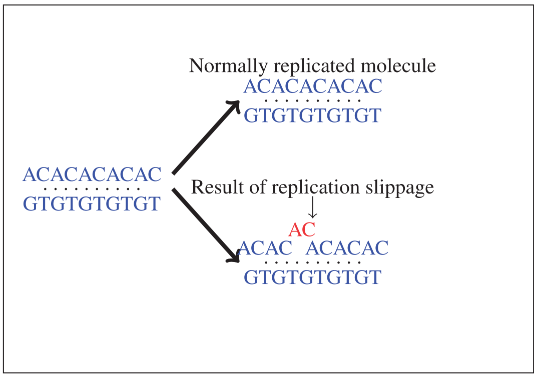

Fig. 1. An extra repeat element added because of replication slippage (Brown, 2002)

popSTR starts by determining a set of informative reads for each marker/individual pair and computing various attributes for the reads. Subsequently, an iterative algorithm is employed to train error models and report genotypes.

2.1 Read selection and processing

The input to the read selection algorithm is a BAM-file, containing the read pairs of a single individual, *j*, the reference genome and a file containing a set, *I*, of microsatellite locations. The algorithm outputs for each microsatellite $i \in I$, a set R_{ij} of reads aligned to the microsatellite and for each read $r \in R_{ij}$ a set of attributes computed for *r*.

The algorithm iterates through the sequencing data and the microsatellite location file in parallel and compares read coordinates to microsatellite coordinates. For each microsatellite, *i*, we determine a set of candidate informative reads as those reads whose alignment intersects the microsatellite location as well as unmapped mates of reads that have been mapped near the microsatellite (within a fixed distance, chosen by default as 1000 bp).

For each candidate informative read we first determine if the read contains the repeat motif of the microsatellite. Those reads that contain the repeat sequence are aligned to the sequences flanking the microsatellite location. The read is split into three parts; the sequence before the microsatellite, the microsatellite repeat sequence and the sequence after the microsatellite. Figure 2 shows how two subsequences are constructed from the read, containing the repeat and the flanking base pairs on either side. Both subsequences are aligned to the reference genome using an overlap alignment and the Needleman-Wunsch algorithm; the first sequence is aligned to the bases preceding the repeat in the reference and the second is aligned to the bases following the repeat in the reference. If the sum of the alignment scores exceeds a minimum threshold the read is considered aligned. The user also specifies a minimum number of flanking bases needed on each side of the repeat. Aligned reads that meet this threshold are added to R_{ij}.

To increase our sensitivity in identifying microsatellite containing reads we also process reads when there is a strong support for the alignment on one side of the microsatellite, while only few bases can be aligned at the other end. We also consider reads to be aligned at both ends if at least four bases can be aligned on each side. Such reads are added to R_{ij} if the sum of the aligned flanking bases is greater than or equal to twice the user specified minimum number of flanking bases on each side.

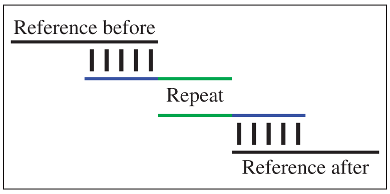

Fig. 2. We split the read into two overlapping parts where the first part has the repeat as a suffix and the second part has the repeat as a prefix. We then align the first part to the reference sequence preceding the microsatellite and the second part to the reference sequence after the microsatellite

We estimate the length, $l^i(r)$ of the microsatellite repeat in r as the number of bases in r between the last base aligned to the sequence preceding i in the reference and the first base aligned to the sequence following i in the reference. We represent alleles of a microsatellite i with the number of times its repeat motif $m(i)$ is repeated. We let $|m(i)|$ represent the length of the repeat motif of microsatellite i, the allele A_r reported by r can be computed as:

$$A_r = \frac{l^i(r)}{|m(i)|} \qquad (1)$$

Some microsatellite alleles are very long, at times longer than the read length used for sequencing. Reads overlapping long microsatellites can only give partial information on the length of the microsatellite allele; the length of the microsatellite allele must be at least as long as the overlap of the read with the microsatellite. To address the challenge presented by reads only able to give a lower bound on the number of repeats, we set a user-specified maximum length of allele, m_l. All alleles longer than m_l are lumped together and reported as the composite allele $\geq m_l$. Reads that contain repeats that span the entire length of the read or occur at either end of a read and the base pair length of the repeat is at least m_l are processed and the number of repeats is reported as:

$$A_r = \frac{m_l}{|m(i)|} \qquad (2)$$

For each read $r \in R_{ij}$ we store a number of attributes relevant to the alignment, summarized and defined in Table 1. These attributes were chosen with the intent of revealing reads that are the result of misclassification events, i.e. sequencing or mapping errors.

2.2 Iterative algorithm

There are multiple sources of error that need to be accounted for in our model. Replication slippage is dependent on the marker being considered as well as the individual. In addition, some of the reads may be the results of sequencing or mapping errors.

Replication slippage has two forms; *full motif slippage* and *stutter noise*. A *full motif slippage* is when the length of the slippage is an integer multiple of the length of the repeat motif of the microsatellite, all other slippages are referred to as *stutter noise*. Following lobSTR (Gymrek *et al.*, 2012) we model these two types of slippage events separately. We assume a Poisson distribution for full motif slippage events and a geometric distribution for stutter noise. In what follows, we will refer to the rate of full motive slippage events as *slippage rate* while we will refer to the rate of stutter noise as *stutter rate*.

Sequencing and mapping errors are accounted for using logistic regression classification of the reads for each microsatellite separately. Based on the attributes computed above and the genotype of an individual at the microsatellite, the classifier assigns a probability to each read of being an error read, i.e. the result of a mapping or sequencing error.

We use an iterative approach to simultaneously train logistic regression classifiers, estimate slippage and stutter rates for each microsatellite and a slippage rate for each individual. We start by describing the individual steps of our algorithm and then show how these are combined into an algorithm.

2.2.1 Read classification

To identify reads resulting from sequencing or mapping errors we train a logistic regression classifier (Hosmer and Lemeshow, 2004) for each microsatellite using the reads of all individuals. At each iteration of the algorithm, each individual has a currently estimated genotype at the microsatellite. This currently estimated genotype allows us to label reads as either TRUE or FALSE. Reads reporting one of the two alleles in the current genotype are labelled as TRUE and reads reporting other alleles that further cannot be explained with a single slippage event, are labelled as FALSE. We use the attributes computed in the read selection step (cf. Table 1) as control variables for the logistic regression classifiers.

The resulting classifier allows us to assign a probability, $p_i(r)$, to each read, r, representing the probability that r is correctly classified as a read from microsatellite i. Reads classified as TRUE are believed to represent the sequence of the individual at the marker being considered. Reads classified as FALSE are believed to be the result of a mapping or a sequencing error.

2.2.2 Slippage rate estimation

The frequency of slippage events varies between microsatellites. To account for this we estimate a marker specific slippage rate.

Assuming we know which reads are the results of a full motif slippage event, we can estimate the slippage rate at microsatellite i by dividing the number of reads resulting from full motif slippage by the total number of reads aligned to the microsatellite. S_i^{M}, the slippage rate at microsatellite i, could be estimated as:

$$S_i^{\mathrm{M}} = \frac{n_i^!}{n_i} \qquad (3)$$

where $n_i^!$ represents the number of reads aligned to microsatellite i that do not support the current genotype and are considered to be results of a full motif slippage and n_i represents the total number of reads aligned to microsatellite i.

The above expression however ignores the fact that individuals may have different slippage rates. We assume that the slippage of

Table 1. The attributes used as control variables in the Logistic regression classification

Attribute	Definition
Quality score	Mapping quality score of the aligned read.
Microsatellite purity	No of base pairs matching microsatellite repeat sequence/No of base pairs in microsatellite sequence
Repeat bases over 20	The number of base pairs with a PHRED-scaled quality over 20 in the microsatellite sequence.
Flanking bases on right over 20	The number of base pairs with a PHRED-scaled quality over 20 in flanking bases after the repeat.
Edit distance of mate	Edit distance of aligned base pairs of the mate sequence to the reference.
Left side alignment score	Alignment score of sequence before the microsatellite to the reference.
Right side alignment score	Alignment score of sequence after the microsatellite to the reference.
Was unaligned	Boolean value indicating if the read was unaligned by BWA.
Alignment shift	Measures changes from original alignment during the realignment of flanking sequences.
Read length	Total length of the read.

marker i in individual j, S_{ij}, is a composite of a marker specific slippage rate, S_i^{M}, and an individual specific slippage rate S_j^{P}.

$$S_{ij} = S_i^{\mathrm{M}} + S_j^{\mathrm{P}} \tag{4}$$

Given the current genotype of individual j at marker i we construct the set $R^{!}_{ij}$ of those reads that do not agree with either of the alleles of the current genotype and are considered to be the result of full motif slippage events.

We can then estimate S_{ij} as:

$$S_{ij} = \frac{\sum_{r\in R^{!}_{ij}} p_i(r)}{\sum_{r\in R_{ij}} p_i(r)} \tag{5}$$

Consequently, we can estimate S_i^{M} and S_j^{P} as

$$S_i^{\mathrm{M}} = \frac{\sum_{r\in R^{!}_{ij}} p_i(r)}{\sum_{r\in R_{ij}} p_i(r)} - S_j^{\mathrm{P}} \tag{6}$$

$$S_j^{\mathrm{P}} = \frac{\sum_{r\in R^{!}_{ij}} p_i(r)}{\sum_{r\in R_{ij}} p_i(r)} - S_i^{\mathrm{M}} \tag{7}$$

Giving us multiple estimates for each S_i^{M} and S_j^{P}. We weight these estimates by the inverse variance of S_{ij} and the number of correctly classified reads at microsatellite i in individual j. The variance of S_{ij} is $S_{ij}(1-S_{ij})$, assuming S_{ij} obeys a binomial distribution. The weight, w_{ij} of microsatellite i in individual j is then:

$$w_{ij} = \frac{\sum_{r\in R_{ij}} p_i(r)}{S_{ij}(1-S_{ij})} \tag{8}$$

Allowing us to estimate S_i^{M} and S_j^{P} as:

$$S_i^{\mathrm{M}} = \sum\nolimits_j \frac{w_{ij}}{\sum_j w_{ij}} \cdot \left(\frac{\sum_{r\in R^{!}_{ij}} p_i(r)}{\sum_{r\in R_{ij}} p_i(r)} - S_j^{\mathrm{P}} \right) \tag{9}$$

$$S_j^{\mathrm{P}} = \sum\nolimits_i \frac{w_{ij}}{\sum_i w_{ij}} \cdot \left(\frac{\sum_{r\in R^{!}_{ij}} p_i(r)}{\sum_{r\in R_{ij}} p_i(r)} - S_i^{\mathrm{M}} \right) \tag{10}$$

2.2.3 Stutter rate estimation

Following the model presented in lobSTR (Gymrek *et al.*, 2012), we estimate a microsatellite specific parameter t_i, for the geometric distribution assumed for stutter noise as:

$$t_i = \frac{1}{1+\bar{x}_i} \tag{11}$$

Where $\bar{x}_i$ is an estimate of the fraction of reads at microsatellite i that are results of stutter noise events.

To estimate $\bar{x}_i$ we start by computing the absolute value of the minimum base pair distance to the current genotype for all reads covering microsatellite i. A read from individual j, supporting either allele of the individual's current genotype (A,B) has a distance of zero but reads not supporting the current genotype have a distance of:

$$\mathrm{dist}(r) = \min(|l(A) - l^i(r)|, |l(B) - l^i(r)|) \tag{12}$$

where $l(A)$ and $l(B)$ represent the base pair length of alleles A and B, respectively and $l^i(r)$ represents the base pair length of the allele reported by the read. We then estimate $\bar{x}_i$ as the average of this number modulo the length of the repeat motif at microsatellite i.

2.2.4 Computing genotype likelihoods

We focus our attention on determining the likelihood of a genotype, gt. We are given a set R of reads, which we assume are independent observations of the microsatellite i, allowing us conclude that:

$$L(R|\mathrm{gt}) = \prod\nolimits_{r\in R} L(r|\mathrm{gt}) \tag{13}$$

We now show how to compute $L(r|\mathrm{gt})$, adding terms for each source of error successively to our model. We first consider the case when the only sources of error are full motif slippage events and read misclassification events. Recall, that A_r represents the number times the repeat motif of i is repeated in r and that the alleles of a genotype are represented with the number of times the repeat motif, m_i, is repeated. Given an allele, A, we compute $x_r(A)$ as the number of slippage events needed to explain r with A as $x_r(A) = |A - A_r|$. We assume that the number of slippage events follows a Poisson distribution with $\lambda = S_{ij}$. This gives the following expression for a homozygous genotype gt = (A,A).

$$L(r|A,A) = p_i(r)\cdot \mathrm{pois}(x_r(A); S_{ij}) \tag{14}$$

For a heterozygous genotype (A, B) we assume that each allele is drawn with equal probability:

$$L(r|A,B) = \\ p_i(r)\cdot\left(\frac{1}{2}\cdot \mathrm{pois}(x_r(A); S_{ij}) + \frac{1}{2}\cdot \mathrm{pois}(x_r(B); S_{ij})\right) \tag{15}$$

The above expression assigns a very small likelihood for reads that are not the results of slippage events. With probability $1 - p_i(r)$ the read being considered is an error read, in this case we assume that each of the other reported alleles is equally likely. We let n^i be the number of alleles present in the population for microsatellite i and refine our expression for $L(r|A,B)$ as follows:

$$L(r|A,B) = p_i(r)\cdot\left(\frac{1}{2}\cdot \mathrm{pois}(x_r(A); S_{ij}\right) + \frac{1}{2}\cdot \mathrm{pois}(x_r(B); S_{ij})) + \frac{1-p_i(r)}{n^i} \tag{16}$$

Slippage events are more likely to delete repeat units than insert. To account for this, we further refine our model and add a parameter, p_{d}, representing the probability that if a slippage event occurs, this event results in a deletion of a motif. Given an allele A and a read r we compute $a_{\mathrm{r}}^{\mathrm{A}}$ as p_{d} if $A - A_{\mathrm{r}} \le 0$ and $1 - p_{\mathrm{d}}$ if $A - A_{\mathrm{r}} > 0$. Our refined model then becomes:

$$L(r|A,B) = p_i(r)\cdot \\ \left(\frac{1}{2}\cdot \mathrm{pois}(x_{\mathrm{r}}(A); S_{ij})\cdot a_{\mathrm{r}}^{\mathrm{A}} + \frac{1}{2}\cdot \mathrm{pois}(x_{\mathrm{r}}(B); S_{ij})\cdot a_{\mathrm{r}}^{\mathrm{B}}\right) + \frac{1-p_i(r)}{n^i} \tag{17}$$

Finally, we account for stutter noise, for which we assume a geometric distribution and use the marker specific t_is estimated using Equation (11). To reflect this in our model we split $x_r(A)$ and $x_r(B)$ into their integer and decimal portions. We let $x_{\mathrm{r}}^k(A)$ denote the integer portion and $x_{\mathrm{r}}^{\mathrm{d}}(A)$ the decimal portion of $x_{\mathrm{r}}(A)$. Similarly we split $x_{\mathrm{r}}(B)$ into $x_{\mathrm{r}}^{\mathrm{k}}(B)$ and $x_{\mathrm{r}}^{\mathrm{d}}(B)$ and our final model becomes:

$$L(r|A,B) = p_i(r)\cdot \\ \left(\frac{1}{2}\cdot \mathrm{pois}(x_{\mathrm{r}}^{\mathrm{k}}(A); S_{ij})\cdot \mathrm{geom}(x_{\mathrm{r}}^{\mathrm{d}}(A); t_i\right)\cdot a_{\mathrm{r}}^{\mathrm{A}} \\ + \frac{1}{2}\cdot \mathrm{pois}(x_{\mathrm{r}}^{\mathrm{k}}(B); S_{ij})\cdot \mathrm{geom}(x_{\mathrm{r}}^{\mathrm{d}}(B); t_i)\cdot a_{\mathrm{r}}^{\mathrm{B}}) \\ + \frac{1-p_i(r)}{n^i} \tag{18}$$

Given a set of reads $R_{i,j}$ for a microsatellite i and individual j we compute this genotype likelihood for all genotypes A, B present in the population. The *current genotype* is the A, B with the highest $\prod_{r\in R_{i,j}} L(r|A,B)$.

2.2.5 Algorithm pseudocode

The algorithm can now be described with the following pseudocode:

```
• Select and process reads
• Initialize genotypes.
• Initialize all $S_i^{\mathrm{M}}, S_j^{\mathrm{P}}, p_i, t_i$.
• While genotypes have not converged:
    • Use $S_i^{\mathrm{M}}, p_i, t_i, S_j^{\mathrm{P}}$ to compute genotypes.
    • Update $S_j^{\mathrm{P}}$s using $S_i^{\mathrm{M}}$s,$p_i$s and $t_i$s.
    • From the current genotypes determine the probability of each
      read being TRUE and FALSE.
    • Update $p_i$s using read classification.
    • Update $t_i$s using current genotypes.
    • Update $S_i^{\mathrm{M}}$s using current genotypes, $S_j^{\mathrm{P}}$s and $p_i$s.
• Compute genotype likelihoods and exit.
```

2.3 Kernelization of iterative algorithm

Our iterative algorithm can be too memory and time intensive for large data sets. In order to make our time and memory requirements more manageable we can kernelize our algorithm. We select a small set of well behaving microsatellites and individuals with high quality sequencing data for our initial training, a set we refer to as a *kernel*. Within this kernel we apply the full algorithm described in section 2.2.5.

Once this kernel has been trained we estimate individual specific slippage rates, S_j^{P}s, using only the markers within the kernel, keeping the marker slippage rates, S_i^{M}s, the stutter rates, t_is, and the marker classification models ($p_i(r)$s) fixed.

Once the S_j^{P}s have been trained for all individuals, j, we train S_i^{M}, t_i and $p_i(r)$ for all markers i keeping the S_j^{P}s fixed, allowing us to compute a final set of genotype likelihoods.

3 Implementation

PopSTR was written in C++using the sequence analysis library SeqAn (Döring *et al.*, 2008) which allows for easy reading and manipulation of data stored in BAM-files.

The implementation of popSTR has four steps. In the first step, we identify the reads useful for genotyping, compute their attributes and initialize genotypes. In the second step, we estimate S_j^{P}s, S_i^{M}s, t_is and p_is on a kernel of markers. In the third step, we use results from the kernelization to compute S_j^{P}s. In the final step we train S_i^{M}s, t_is and p_is and finally perform genotyping.

3.1 Read selection and processing

We use the fact that the sequencing data has already been aligned (in a BAM file), allowing us to limit the number of reads that we consider. We can however not limit our search only to reads that have been aligned to a microsatellite, as alignment to microsatellites by general purpose aligners, such as BWA (Li and Durbin, 2009), is not reliable. General purpose aligners trade accuracy and speed in their implementation and do not account for the high mutation rate of microsatellites. We limit our search to reads that have been aligned to microsatellites and reads that are unaligned but have a mate that is aligned near the microsatellite being considered. Sequences already aligned to non-microsatellite sequences are unlikely to be useful while sequences that are unaligned may in fact contain a microsatellite but have not been aligned because they are too different from the reference.

When selecting reads and in order to perform the read classification, we compute a number of attributes related to the reads' alignment and their sequencing quality. As previously mentioned, candidate microsatellite reads are processed by first identifying the repeat sequence within the read. Subsequently, the sequences flanking the repeat are aligned to the sequences flanking the microsatellite in the reference genome. The quality of this alignment is one of the attributes used as a control variable in the logistic regression classification. We define *purity* of an alignment as the number matching base pairs divided by the total number of base pairs in the alignment. The purity of a microsatellite repeat sequence is the number of base pairs matching the repeat divided by the total number of base pairs in the repeat. The purity of the repeat sequence in the read is another control variable. All attributes computed, used as control variables in the logistic regression, are summarized in Table 1.

Further, some attributes are required to reach a minimum value for the read to be used. The minimum microsatellite purity required is relative to the purity of the microsatellite sequence in the reference and also depends on the number of flanking bases available in the read. Table 2 summarizes these filters used.

Finally, we do not consider low quality reads, i.e. the ones that fail platform or vendor quality checks nor reads that are PCR or optical duplicates.

3.2 Kernelization

Convergence has been reached in the kernelization when <0.5% of the genotypes are updated between iterations.

We initialize the slippage rate for individual j, using the following expression

$$S_j^{\mathrm{P}} = \frac{n_j^!}{n_j} \tag{19}$$

where $n_j^!$ represents the number of reads from individual j not supporting the initialized genotype and n_j represents the total number of reads from individual j.

3.3 Individual slippage rate computation

The marker slippage and stutter rates estimated (S_i^{M}s and t_is) and the logistic regression classifiers ($p_i(r)$s) trained during the kernelization are used to directly estimate the individual specific slippage rates (S_j^{P}s). First, we compute the attributes of reads aligned to the microsatellites in the kernel. Next, we assign misclassification probabilities, $p_i(r)$s to the reads using the logistic regression classifiers from the kernel and we update the genotypes, with marker slippage and stutter rates from the kernel using the expression given in Equation (18) to determine the most likely genotype. Finally, we use the expression given in Equation (10) to estimate an individual slippage rate, S_j^{P}s. We iterate this process, keeping the marker specific properties from the kernel constant, until the individual slippage rates, (S_j^{P}s), have reached convergence.

Table 2. Minimum numeric values when identifying useful reads

Name	Condition	Minimum value
Microsatellite purity	both flanking	0.75*(ref. purity)
	one sided flanking	0.8*(ref. purity)
	no flanking	0.85*(ref. purity)
Alignment purity	one sided flanking	0.7
No. repeats	motiflength = 2	4
	motiflength = 3	3
	motiflength $\in 4, 5, 6$	2

3.4 Marker slippage and stutter rate computation, logistic regression and genotyping

We fix the probability of deletion Equation (18) as $p_d = 0.85$ and consequently $1 - p_d = 0.15$. We iterate between updating genotypes and updating the microsatellite slippage and stutter rates, (S_i^Ms and t_is), and logistic regression classifiers ($p_i(r)$s), while keeping individual slippage rates, (S_j^Ps), constant, until convergence has been reached.

4 Results

4.1 Data set

We analyzed microsatellites for 15 220 whole genome sequenced individuals, sequenced using Illumina sequencers. Sequencing reads had previously been mapped to GRCh38 using BWA (Li and Durbin, 2009).

We ran Tandem Repeat Finder on 1 Mb non-overlapping intervals using the parameters and options suggested in (Willems *et al.*, 2014) to determine the microsatellite locations (Benson, 1999). We filtered the output in order to include only repeats with motif-length between 2 and 6 bp. We removed repeats with alignment scores below thresholds suggested in (Willems *et al.*, 2014), repeats closer than 100 bp to each other and repeats longer that 100 bp. This resulted in a set of 880 355 microsatellite locations found on GRCh38 autosomes after excluding microsatellites located in high coverage regions.

We initially ran popSTR on a set of 8453 individuals and 880 355 microsatellites. We chose a kernel set of 703 individuals with high-quality sequencing data and 8303 microsatellites on chromosome 1, based on their imputation info (Gudbjartsson *et al.*, 2015) when imputed into the Icelandic population.

Our comparison set is based on the genotypes of 15 220 individuals on 880 355 microsatellites. Out of these a total of 380 261 microsatellites were found to be polymorphic and were subsequently imputed into the Icelandic population (Gudbjartsson *et al.*, 2015).

For comparison purposes, we chose 141 markers where capillary electrophoresis benchmark genotypes were available, sequenced as parts of various disease association efforts at deCODE genetics (Sun *et al.*, 2012).

Comparisons to lobSTR were done by choosing 10 individuals from the 15 220 sequenced individuals. The 10 individuals were chosen to have a large number of electrophoresis genotypes available.

The 15 220 samples were also genotyped using the GATK (McKenna *et al.*, 2010) genotype caller and imputed into the Icelandic population. GATK is a general purpose genotype caller that does not distinguish between indels and microsatellites. To further investigate the quality of our genotypes we matched our microsatellites coordinates to output coordinates of indel alleles from the GATK genotype caller where the indel allele matched the microsatellite repeat motif. We then compared the imputation results into the Icelandic population for markers where a match was found.

4.2 Comparison to lobSTR

We compared the popSTR and lobSTR genotypes to inhouse benchmark data obtained through capillary electrophoresis. The capillary electrophoresis genotypes are represented as base pair distances from a reference individual, while genotypes reported from sequencing (by popSTR or lobSTR) are presented as lengths of the microsatellite alleles. As we did not have the length of the microsatellite alleles of the reference individual, we considered the genotypes reported from capillary electrophoresis and sequencing to agree if an identical difference in lengths between the two alleles carried by the individual was reported by the two methods.

Both lobSTR and popSTR can be expected to report more accurate genotypes when more reads overlapping the microsatellite are used in the genotyping. We therefore condition our results on the number of reads used in the genotyping. The number of reads used for genotyping at a particular location is upper bounded by the sequencing coverage at the given location. Not all reads overlapping the location can however be used as both algorithms require the reads to fully overlap the microsatellite and sequences flanking the microsatellite on both sides. Figure 3 shows the accuracy of lobSTR and popSTR as a function of the number of reads used by lobSTR. The figure clearly shows that, as expected, the accuracy of both methods increases when more reads overlapping the microsatellite are used. The figure also shows that popSTR consistently has higher genotyping accuracy than lobSTR.

Table 3 summarizes the comparison between the two methods. We observe that, when we restrict our analysis to microsatellites and individuals where there are at least 10 reads overlapping the microsatellite, popSTR has a 96% agreement with the capilllary electrophoresis genotypes while lobSTR has a 92% agreement. Consistently, over all coverage thresholds the number of genotypes that are in disagreement with the capillary electrophoresis genotypes is approximately two times higher for lobSTR than for popSTR, i.e. the error rate of popSTR is 50% lower than that of lobSTR.

To further confirm the accuracy of our method we compared the popSTR genotypes of 409 individuals to the benchmark genotypes, considering the same 141 markers as in the comparison to lobSTR. Figure 4 shows how the accuracy of popSTR increases with the number of reads used in the genotyping.

We compared the running times of popSTR and lobSTR and found an average speed-up provided by popSTR of 74.7%. The average running time of lobSTR from start to finish was 39.2 h per individual (SD 6.7 h). This includes the time of the alignment (39.1 h) and allelotyping (0.1 h) step of lobSTR.

The running time for the steps of popSTR we ran jointly was divided by the number of genotyped individuals (15 220) and then added to the average running time of individually run steps.

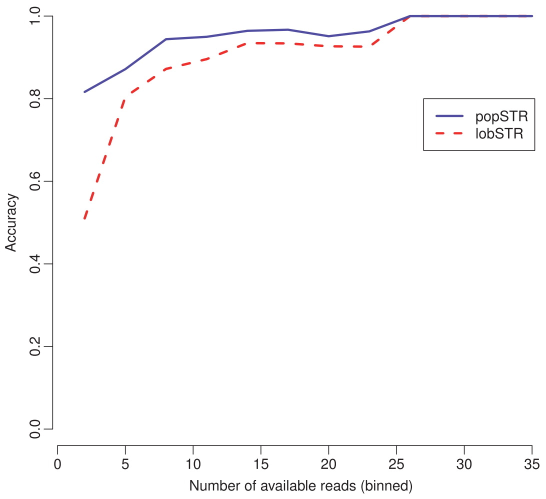

Fig. 3. The accuracy of the lobSTR and popSTR genotypers as a function of the number of reads used by lobSTR overlapping the microsatellite, binned with bin size = 3. Results are averages over 10 individuals and 141 microsatellites

Table 3. Genotyping accuracy of lobSTR and popSTR compared with capilllary electrophoresis genotypes

Coverage filter	lobSTR	popSTR
≥1	87.3%	93.5%
≥5	89.5%	94.3%
≥10	92.0%	96.0%
≥15	93.5%	96.4%
≥20	94.4%	97.2%

The results are thresholded on the number of reads available to lobSTR.

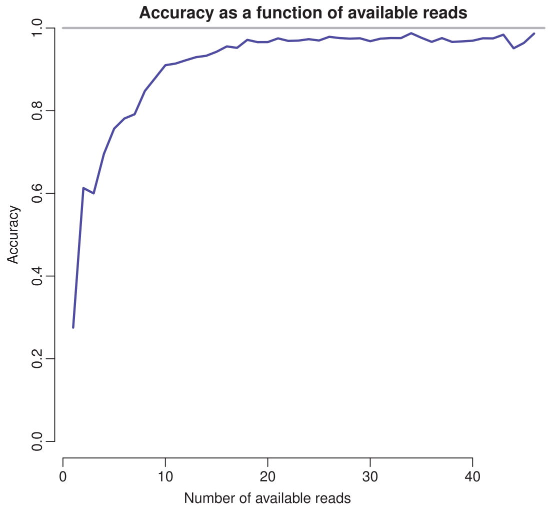

Fig. 4. The accuracy of the popSTR genotyper as a function of the number of reads used by popSTR overlapping a microsatellite. Results are averages over 409 individuals and 141 microsatellites

Summing up the running time of all steps of popSTR we get a total of 9.9 hours per individual. Table 4 breaks the total running time of popSTR into its components.

Table 4. Running times of popSTR steps

Step	Run individually	Run jointly	Time
Kernelization		✓	0.1 h
Read selection and processing	✓		9.2 h
Individual slippage estimation	✓		0.25 h
Genotyping		✓	0.35 h

4.3 Comparisons to GATK

We use imputation info (Gudbjartsson *et al.*, 2015) to compare the quality of genotypes reported by GATK and popSTR. Imputation info is a measure between 0 and 1, representing confidence in genotype assignment reported by the imputation software (Gudbjartsson *et al.*, 2015), with larger values of imputation info representing higher confidence. We have previously determined imputation info of greater than or equal to 0.9 as a threshold for which we believe that the genotypes are highly reliable (Gudbjartsson *et al.*, 2015).

GATK is a general purpose tool for determining genotypes and does not have a specific model for microsatellites, but rather lumps them in a category with indels. We compared the imputation info of popSTR microsatellites to the imputation info of indel alleles from GATK in cases where alleles reported by GATK were located within a microsatellite sequence. At microsatellite locations, some of the indels reported by GATK contain the microsatellite motif, while others do not. We condition our comparison to GATK on whether the microsatellite motif is found in the indel reported by GATK.

For a judicious comparison, we construct a single number for each microsatellite by summing the info of each allele weighted by frequency. This is shown in Equation (20) where i_w represents the weighted info value and f_a and i_a represent the frequency and imputation info of allele a, respectively.

Table 5. Imputation info comparison of popSTR and GATK

	Matching coordinates	Matching motifs
Total	152 152	75,057
popSTR info > GATK info	107 104 (70.4%)	56 521 (75.3%)
GATK info > popSTR info	45 048 (29.6%)	18 536 (24.7%)
popSTR info > 0.9	120 317	62 962
GATK info > 0.9	92 854	49 684
Either info > 0.9	133 366	68 216
popSTR info > GATK info	99 787 (74.8%)	53 812 (78.9%)
GATK info > popSTR info	33 579 (25.2%)	14 404 (21.1%)
Mean popSTR info	0.95 (SD 0.16)	0.96 (SD 0.12)
Mean GATK info	0.90 (SD 0.16)	0.93 (SD 0.08)

$$i_{\mathrm{w}} = \frac{\sum_{\mathrm{a}} f_{\mathrm{a}} * i_{\mathrm{a}}}{\sum_{\mathrm{a}} f_{\mathrm{a}}} \qquad (20)$$

For a total of 152 152 microsatellites found by popSTR an indel was reported by GATK within the microsatellite. We compared the imputation info for these popSTR-microsatellite/GATK-indel pairs in several different ways; these are summarized in Table 5 (Matching coordinates).

In 75 057 of the microsatellites found by popSTR the indel reported by GATK contained the microsatellite motif. The same comparison was performed for these pairs as the previous ones and is also summarized in Table 5 (Matching motifs).

5 Conclusion

Here we have shown that, by creating a microsatellite profile for an individual using previously aligned data, it is possible to significantly decrease the running time of microsatellite genotyping by considering only reads that are either aligned to a known microsatellite location or not aligned at all. The filtering dismisses a large portion of the data immediately while minimally effecting the microsatellite profile. Our results also show that the genotyping accuracy of our program is higher than for the general purpose genotype caller GATK as well as lobSTR, a program specifically designed for calling of microsatellites.

Several improvements could still be made to our model and method. Our method does not consider reads where neither the read nor its mate align to the reference genome. Our method also assumes that the mate of the read containing the microsatellite is correctly mapped. If the read pair were to be mapped to a graph reference (a reference genome containing all variants) it is possible that a joint alignment of both the read containing the microsatellite and its mate would reveal the correct location for the read pair. We do not account for possible sampling biases, i.e. it may be more likely that we observe reads that are similar to the reference than those that are highly divergent from the reference. Similarly, there may be biases introduced by our alignment algorithm or filtering steps not accounted for in our model. Finally, our implementation is optimized for Illumina paired-end sequencing data. Although we believe that our algorithm could be used for other types of sequencing data the method would need to be tuned to the error models of those data.

Conflict of Interest: none declared.

References

Benson,G. (1999) Tandem repeats finder: a program to analyze dna sequences. *Nucleic Acids Res*., **27**, 573–580.

Brown,T.A. (2002). *Genomes*, 2nd edn. Wiley-Liss, Oxford.

Cox,M., and Mays,S. (2000). *Human Osteology: In Archaeology and Forensic Science*. Cambridge University Press, Cambridge.

Döring,A. *et al*. (2008) SeqAn an efficient, generic c++ library for sequence analysis. *BMC Bioinformatics*, **9**, 11.

Duitama,J. *et al*. (2014) Large-scale analysis of tandem repeat variability in the human genome. *Nucleic Acids Res*., **42**, 5728–5741.

Gelfand,Y. *et al*. (2014) VNTRseek-a computational tool to detect tandem repeat variants in high-throughput sequencing data. *Nucleic Acids Res*., **42**, 8884–8894.

Gudbjartsson,D.F. *et al*. (2000) Allegro, a new computer program for multipoint linkage analysis. *Nat Genet*., **25**, 12–13.

Gudbjartsson,D.F. *et al*. (2015) Large-scale whole-genome sequencing of the icelandic population. *Nat Genet*., **47**, 435–444.

Gymrek,M. *et al*. (2012) lobstr: A short tandem repeat profiler for personal genomes. *Genome Res*., **22**, 1154–1162.

Gymrek,M. *et al*. (2016) Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet*., **48**, 22–29.

Highnam,G. *et al*. (2013) Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res*., **41**, e32.

Hosmer, D.W., Jr, and Lemeshow,S. (2004). *Applied Logistic Regression*. John Wiley & Sons, New York.

Kong,A. *et al*. (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, **488**, 471–475.

Li,H., and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–1760.

McKenna,A. *et al*. (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res*., **20**, 1297–1303.

Mirkin,S.M. (2007) Expandable dna repeats and human disease. *Nature*, **447**, 932–940.

Sun,J.X. *et al*. (2012) A direct characterization of human mutation based on microsatellites. *Nat. Genet*., **44**, 1161–1165.

Veselinović,I. (2006) Microsatellite DNA analysis as a tool for forensic paternity testing (DNA paternity testing). *Med. Pregl*., **59**, 241–243.

Willems,T. *et al*. (2014) The landscape of human str variation. *Genome Res*., **24**, 1894–1904.

Chapter 3

popSTR2 enables clinical and population-scale genotyping of microsatellites

Bioinformatics, 36(7), 2020, 2269–2271
doi: 10.1093/bioinformatics/btz913
Advance Access Publication Date: 5 December 2019
Applications Note

Sequence analysis

popSTR2 enables clinical and population-scale genotyping of microsatellites

Snædis Kristmundsdottir[1,2,*], **Hannes P. Eggertsson**[1], **Gudny A. Arnadottir**[2] **and Bjarni V. Halldorsson**[1,2,*]

[1]deCODE genetics/Amgen, Reykjavík 102, Iceland and [2]School of Science and Engineering, Reykjavík University, Reykjavík 102, Iceland

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on June 28, 2019; revised on November 27, 2019; editorial decision on December 2, 2019; accepted on December 3, 2019

Abstract

Summary: popSTR2 is an update and augmentation of our previous work 'popSTR: a population-based microsatellite genotyper'. To make genotyping sensitive to inter-sample differences, we supply a kernel to estimate sample-specific slippage rates. For clinical sequencing purposes, a panel of known pathogenic repeat expansions is provided along with a script that scans and flags for manual inspection markers indicative of a pathogenic expansion. Like its predecessor, popSTR2 allows for joint genotyping of samples at a population scale. We now provide a binning method that makes the microsatellite genotypes more amenable to analysis within standard association pipelines and can increase association power.

Availability and implementation: https://github.com/DecodeGenetics/popSTR.

Contact: snaedisk@decode.is or bjarni.halldorsson@decode.is

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Microsatellites, a.k.a. short tandem repeats (STRs), are tandem repeats with repeat motif lengths between one and six base pairs. They are one of the most frequent types of variation in the human genome, surpassed only by single nucleotide polymorphisms (SNPs) and indels and have a mutation rate estimated to be three to five orders of magnitude higher than for other types of genetic variation (Jónsson *et al.*, 2017; Sun *et al.*, 2012). Genotyping microsatellites from whole-genome sequence (WGS) data is challenging since they are highly polymorphic and library preparation methods may modify the true number of repeats in the sequence (Gymrek *et al.*, 2012). WGS-based association and clinical analysis commonly do not consider microsatellites, partially due to a lack of tools capable of analyzing them.

Tandem repeat expansions occur when microsatellites expand beyond a certain length threshold, making them unstable and thus more likely to expand further. A number of repeat expansions are known to be disease-causing (Gatchel and Zoghbi, 2005) and an increase in the use of WGS-technologies for genetic diagnostics has created a need for fast estimation of the repeat number at disease-associated loci.

Here, we present extensions to our previously published software popSTR and improvements of its previous implementation, both with respect to runtime and accuracy. We increased our expansion detection sensitivity, updated our sample specific slippage estimation kernel, reduced the dimensions of our logistic regression model and updated external libraries to decrease I/O time and handle both BAM and CRAM files. We further created a panel of known repeat expansion markers and a pipeline to determine at each loci whether read support for a pathogenic expansion is present. Last, we provide a method to bin genotypes into user specified bins to increase power of downstream association analysis. By combining this set of functionalities, we hope to make popSTR2 applicable in a wide range of situations. Both when analyzing large cohorts to make population inferences and disease associations as well as analyzing small sets or single samples in a clinical context.

2 Materials and methods

Figure 1 gives a high level description of the algorithm's workflow, a more detailed description is given in Supplementary Section S1.1 and a full description is given in Kristmundsdóttir *et al.* (2017). To summarize, we start by computing various quality-indicating attributes for all reads encompassing each of the microsatellites being considered, i.e. overlapping its coordinates and containing repeats of the relevant motif. We also look for repeats in unaligned reads with mates aligned close to the repeat region. An update of our read selection step is to also look for repeats of the relevant motif in reads aligned to longer repeats of the same motif in other locations of the genome that have mates aligned close to the repeat region. This can happen when a repeat has expanded considerably and the read reporting it is thus highly divergent from the reference sequence. After the set of informative reads has been created, the algorithm iterates between genotyping and assigning to each read a probability

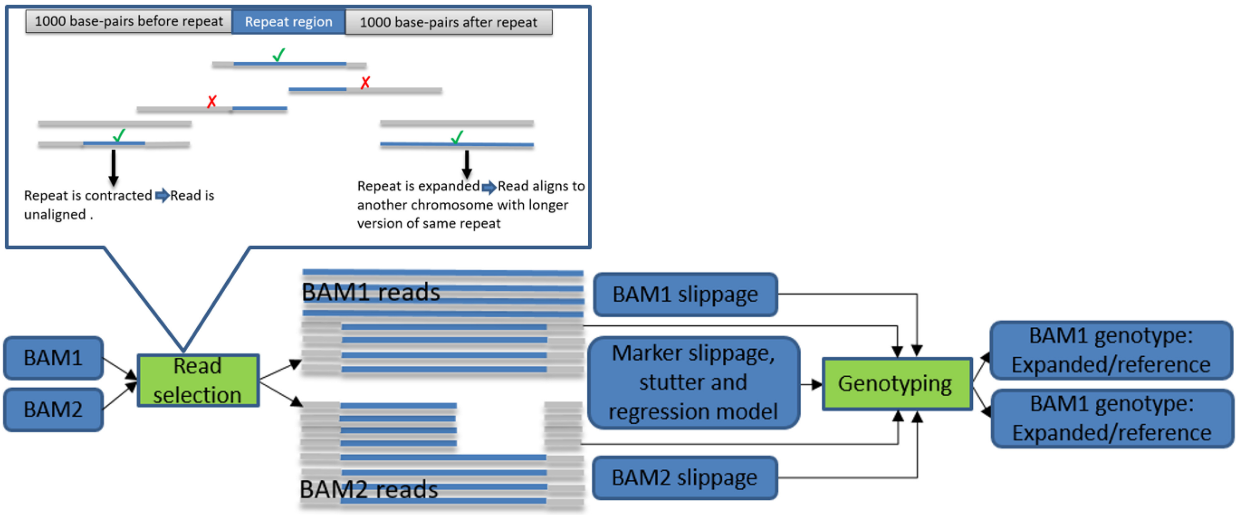
1000 base-pairs before repeat
Repeat region
1000 base-pairs after repeat
Repeat is contracted ➡ Read is unaligned .
Repeat is expanded ➡ Read aligns to another chromosome with longer version of same repeat
BAM1
BAM2
Read selection
BAM1 reads
BAM2 reads
BAM1 slippage
Marker slippage, stutter and regression model
BAM2 slippage
Genotyping
BAM1 genotype: Expanded/reference
BAM1 genotype: Expanded/reference

Fig. 1. Results of read selection are passed into genotyping model along with sample and marker-specific parameters

of reporting a true allele. Since this type of iterative parameter estimation is time and resource intensive, we supply a kernel of reliable markers to efficiently estimate these parameters. For details on kernel construction see Supplementary Section S1.2. We replaced the SeqAn BAM I/O module (Reinert *et al.*, 2017) with the one from htslib (Li *et al.*, 2009; https://github.com/DecodeGenetics/SeqAnHTS). The update provides CRAM file support, decreases I/O demands and runtime. Algorithmic improvements reduced runtime from 11.25 to 2.17 CPU hours/million markers per sample. See Supplementary Table S1 for a breakdown of our runtime analysis.

2.1 Application to population-based genotyping

Useful reads and their attributes are used along with marker and sample specific parameters to perform genotyping. The marker-specific parameters can be estimated by popSTR2, but we also provide a default set of parameters. By default we require 20 samples for the parameters to be estimated since estimation with fewer samples would not yield reliable results. The sample-specific slippage parameter is estimated using a kernel of reliable markers described above and supplied with the software. Our genotyping model (Supplementary Equation S1 in Supplementary Material) computes the likelihood of observing a read, r, given genotypes A and B and selects the genotype pair that maximizes this likelihood over the set of reads being considered. The model previously assumed constant probabilities of adding and removing repeats across all markers, fixing a_r^A in Supplementary Equation S1 from Supplementary Material to 0.85 if whole repeats were removed and consequently to 0.15 if whole repeats were added. It has however been shown that microsatellites have very different mutation profiles depending on their various properties, e.g. repeat motif, repeat purity, reference allele length, etc. (Brinkmann *et al.*, 2002). To reflect this we have replaced the hard coded values with marker-specific estimates, computed as follows. Assuming that we know which reads result from whole motif slippage events, we can estimate the fraction of slippage events that added whole repeats at microsatellite i:

$$p_i^u = \frac{\sum_{r\in R_i^{!u}} p_i(r)}{\sum_{r\in R_i^{!}} p_i(r)} \tag{1}$$

where $R_i^{!u}$ is the set of reads at microsatellite i, considered to be results of slippage events that add whole motifs and $R_i^{!}$ is the set of all reads at microsatellite i reporting whole motif slippage events, regardless of their direction. The probability of removing repeats is then trivially computed as $p_i^d = 1 - p_i^u$.

Our previous version created one output file per sample and computed nine attributes from each read used for genotyping. Due to increased data quality and consistency we were able to reduce the number of attributes to six, which simplified and sped up the logistic regression analysis. To make population scale inferences and genotyping easier we now write one output file per marker, i.e. all alleles discovered in a population accessible in the same file.

Association pipelines commonly assume biallelic variants or multi allelic variants where only a single allele is tested for association with a phenotype, rather than associating a subset of the alleles with it (Gudbjartsson *et al.*, 2015; Purcell *et al.*, 2007). This is not optimal for microsatellites where alleles above or below a certain length threshold may be pathogenic (Lee and McMurray, 2014). In an effort to increase association power we provide binSTR, a software for grouping alleles as a preprocessing step for association analysis. To allow for various patterns of allele groups, binSTR enables not only binarizing but also binning into a user determined number of groups where each group is defined by a list of allele indices passed as a parameter.

2.2 Application to clinical genetics

We have, through literature review, assembled a panel containing 31 STR markers, each associated with a disease or syndrome when the number of repeats passes a certain threshold, hereafter referred to as pathogenicity threshold. We provide a script which reports which of these markers, if any, contain evidence of a repeat expansion. The script runs the read selection step described above to scan a given BAM file at all panel locations and extracts for each of them all reads containing information on the number of repeats present. Expanded alleles have often undergone a dramatic increase in length, decreasing the odds of finding informative reads supporting them. Genotyping models assuming equal probabilities of drawing reads from each haplotype are thus not reliable in these cases. To account for this, our script scans the informative reads for any repeat tracts longer than the given threshold for each marker and flags locations harboring such reads for further manual inspection. Since many of the pathogenicity thresholds exceed the current read lengths by a considerable number of base pairs the scripts also counts and reports all fully repetitive reads, i.e. reads containing only repeats of the relevant motif. See Supplementary Table S4 for a table summarizing the markers included in the panel along with a pathogenicity threshold for each of them. As the set of pathogenic variants and our understanding of them grows the panel can easily be extended and thresholds for existing markers updated.

3 Experiments

We compared popSTR2 to HipSTR (Willems *et al.*, 2017), a commonly used microsatellite genotyper on chr21 of the CEU trio consisting of NA12878, NA12891 and NA12892 and on chr21 of 10 trios sequenced at deCODE genetics.

The runtime reduction was 40% for the CEU trio and 26% for the deCODE trios. To compare the accuracy of these two methods we extracted markers where both methods had high confidence genotypes for all members of at least one trio and at least one trio member had a non-homozygous-reference genotype and recorded the number of trios where the offspring genotype did not match the parental ones. The deCODE trios had slightly more accurate genotypes from popSTR2 than HipSTR (99.8% versus 99.6%) but for the CEU trio hipSTR had a single trio inconsistency in 250 markers while popSTR had 2. For a more detailed comparison of these runs see Supplementary Table S3. To examine the sensitivity of our expansion detection script we ran it on ten samples with a known expanded allele in the 3′-flanking region of the DMPK gene which causes myotonic dystrophy 1 when exceeding 50 copies (Musova *et al.*, 2009) and ten healthy control samples. The expanded samples were sequenced for clinical sequencing analysis at deCODE genetics and the healthy ones as parts of various other projects, also at deCODE genetics. The script flagged the DMPK locus in all expanded individuals and none of the control samples.

Last, we genotyped 49 962 Icelandic samples to examine the allelic spectrum of this repeat in the Icelandic population. The resulting distribution was in concordance with ones previously published for European populations with a bimodal distribution consisting of a peak at 5 repeats and another one between 11 and 13 repeats (Dean *et al.*, 2006; Magaña *et al.*, 2011) (see Supplementary Fig. S1).

4 Conclusion

We updated the microsatellite genotyper popSTR to decrease runtime and increase genotype quality and accuracy. This was done by replacing external libraries, re-training the data provided with the software and decreasing the number of variables in our logistic regression analysis. To expand the application range we extended the software to provide both a clinical sequencing analysis script for quickly estimating expansion status at known disease loci and a binning software for grouping genotypes by allele length range before performing disease association on them. It is our hope that these updates and extensions will make popSTR2 applicable in a broader spectrum of situations, i.e. for single sample clinical sequencing analysis as well as large scale association efforts. Analysis methods (Dashnow *et al.*, 2018; Dolzhenko *et al.*, 2017; Tang *et al.*, 2017; Tankard *et al.*, 2018) sensitive to detecting expanded repeats are not explicitly intended for population scale analysis of STRs at a genome wide scale. Conversely, other methods which aim at population and genome scale analysis (Gymrek *et al.*, 2012; Willems *et al.*, 2017) do not focus on and reporting of expanded repeats. GangSTR (Mousavi *et al.*, 2019) is, to our knowledge, the only method intended to perform accurate genotyping of both short and expanded microsatellites. It however does not mark known pathogenic variants in its output nor flags those expansions passing pathogenicity thresholds. By supplying a panel of known expansions along with an easily executable and fast script to flag potentially expanded repeats for further manual inspection we aim to direct users to the correct putative expansion as quickly as possible.

Conflict of Interest: none declared.

References

Brinkmann,B. *et al.* (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am. J. Hum. Genet.*, **62**, 1408–1415.

Dashnow,H. *et al.* (2018) Stretch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.*, **19**, 121.

Dean,N. *et al.* (2006) Transmission ratio distortion in the myotonic dystrophy locus in human preimplantation embryos. *Eur. J. Hum. Genet.*, **14**, 299–306.

Dolzhenko,E. *et al.* (2017) Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.*, **27**, 1895–1903.

Gatchel,J.R. and Zoghbi,H.Y. (2005) Diseases of unstable repeat expansion: mechanisms and common principles. *Nat. Rev. Genet.*, **6**, 743.

Gudbjartsson,D.F. *et al.* (2015) Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.*, **47**, 435.

Gymrek,M. *et al.* (2012) lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res.*, **22**, 1154.

Jónsson,H. *et al.* (2017) Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*, **549**, 519.

Kristmundsdóttir,S. *et al.* (2016) popstr: population-scale detection of STR variants. *Bioinformatics*, **33**, 4041–4048.

Lee,D.-Y. and McMurray,C.T. (2014) Trinucleotide expansion in disease: why is there a length threshold? *Curr. Opin. Genet. Dev.*, **26**, 131–140.

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

Magaña,J.J. *et al.* (2011) Distribution of CTG repeats at the DMPK gene in myotonic distrophy patients and healthy individuals from the Mexican population. *Mol. Biol. Rep.*, 38, 1341–1346.

Mousavi,N. *et al.* (2019) Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.*, **47**, e90.

Musova,Z. *et al.* (2009) Highly unstable sequence interruptions of the CTG repeat in the myotonic dystrophy gene. *Am. J. Med. Genet. A*, **149**, 1365–1374.

Purcell,S. *et al.* (2007) Plink: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.*, **81**, 559–575.

Reinert,K. *et al.* (2017) The SeqAn C++ template library for efficient sequence analysis: a resource for programmers. *J. Biotechnol.*, **261**, 157.

Sun,J.X. *et al.* (2012) A direct characterization of human mutation based on microsatellites. *Nat. Genet.*, **44**, 1161.

Tang,H. *et al.* (2017) Profiling of short-tandem-repeat disease alleles in 12,632 human whole genomes. *Am. J. Hum. Genet.*, **101**, 700–715.

Tankard,R.M. *et al.* (2018) Detecting expansions of tandem repeats in cohorts sequenced with short-read sequencing data. *Am. J. Hum. Genet.*, **103**, 858–873.

Willems,T. *et al.* (2017) Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods*, **14**, 590.

1 Supplementary materials

1.1 Algorithm

The algorithm takes a BAM/CRAM file and a list of microsatellite intervals to determine a set of reads informative of a sample's genotype. It considers reads encompassing each interval and reads with mates close to the repeat region that are either aligned to longer repeats of the same motif or could not be aligned at all. These misaligned and unaligned reads are considered to retrieve heavily contracted or expanded alleles that are highly divergent from the reference, see Figure 1. For each chosen read, a set of attributes, used to estimate read reliability, are computed and used as parameters when training logistic regression models for each marker. After the read-selection step, the algorithm iterates until convergence between genotyping according to latest parameter estimations and updating parameters using the latest genotypes. Since estimating parameters with few samples is not likely to yield reliable results, by default, we require 20 samples for the parameters to be estimated. A set of default parameters is supplied with the software for analysis of smaller sample sets. Comparing an allele reported by a read to a genotype, allows the algorithm to assign a label to the read. These labels, along with the read attributes, enable training of a per marker logistic regression model that assigns to each read, a probability $p_i(r)$ of reporting a true germline allele. The observed sequence may differ from the germline variant due to somatic mutations and sequencing problems which cause the addition or removal of full and partial motifs. The labels allow us to estimate marker-specific slippage and stutter rates S_i, t_i representing the probability of adding or removing a full or a partial motif, respectively. When determining the most likely genotype (A, B) given a read r, we use i as a marker identifier and j as a sample identifier in the following model.

$$
\begin{aligned}
L(r|A, B) = p_i(r)\cdot &\\
&\left(\frac{1}{2}\cdot pois(x_r^k(A); S_{ij})\cdot geom(x_r^d(A); t_i)\cdot a_r^A\right.\\
+\frac{1}{2}\cdot pois(x_r^k(B); S_{ij})\cdot geom(x_r^d(B); t_i)\cdot a_r^B&\left.\right) + \frac{1-p_i(r)}{n^i}
\end{aligned}
\tag{1}
$$

We let n^i be the number of alleles present in the population for microsatellite i. With probability $1 - p_i(r)$ the read being considered is an error read, in which case we assume that each allele is equally likely. We compute $x_r(A)$ as the number of slippage events needed to explain r with A. To separate whole and partial motif slippage events, we split $x_r(A)$ into its integer and decimal portions. We let $x_r^k(A)$ denote the integer portion which follows a Poisson distribution with $\lambda = S_{ij}$, a combination of the marker and sample-specific slippage rate estimates. $x_r^d(A)$ denotes the decimal portion of $x_r(A)$ which follows a geometric distribution with parameter t_i. We use Equation 1 in 2.1 of the main text to estimate and represent the different probabilities at each marker of adding and removing motifs. The result of this estimation is then used as a_r^A and a_r^B in Equation 1. Last, we select the genotype pair that maximizes the likelihood function in Equation 1 over the set of reads, R, being considered. Figure 1 in the main text shows a schematic view of the workflow described above.

1.2 Kernel construction

Our kernel was selected by genotyping all microsatellites on chr21 in a set of 6086 trios and choosing markers with heterozygous transmission rates between 30 and 70% and Mendelian error rates below 0.1% which resulted in 8779 markers. To verify that this subset accurately represented the entire marker set, we examined whether kernel markers were more likely to have larger repeat units and found that this was not the case, see Supplementary table 2. All our kernel training, parameter training and development was performed using BAM files aligned with the BWA-MEM aligner Li (2013) and generated using the sequencing processing pipeline described in Jónsson *et al.* (2017).

1.3 Repeat list construction

We created our list of repeat coordinates by running the Tandem repeats finder(trf) Benson (1999) on GRCh38 for each chromosome and retaining resulting repeats with motif lengths between one and six base pairs, (command: ./trf409.linux64 chr$num.fa 2 7 7 80 10 22 7 -d -h -ngs > trf.$num).

References

Benson, G. (1999). Tandem repeats finder: a program to analyze dna sequences. *Nucleic acids research*, **27**(2), 573–580.

Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M. T., Hjorleifsson, K. E., Eggertsson, H. P., Gudjonsson, S. A., *et al.* (2017). Whole genome characterization of sequence diversity of 15,220 icelanders. *Scientific data*, **4**, 170115.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*.

Table 1. Breakdown of runtime analysis comparison, all times are given per sample.

	popSTR	**popSTR2**
samples	15,220	40,121
markers	880,355	5,401,401
polymorphic markers	380,261	1,636,021
Read selection	9.2 h	9.45 h
Sample slippage	0.25 h	0.065 h
Genotyping	0.35 h	2.19 h
CPU hours/1,000,000 markers	11.25	2.17

Table 2. Motif length distribution of kernel vs entire marker set

Motif length	**Fraction of kernel**	**Fraction overall**
1bp	48.2%	44.4%
2bp	23%	18.5%
3bp	4.5%	5.5%
4bp	12.8%	15.3%
5bp	6%	7.9%
6bp	5.5%	8.4%

Table 3. Comparison of hipSTR and popSTR2 on runtime and accuracy for chr21.

	CEU trio		**deCODE trios**	
	hipSTR	popSTR2	hipSTR	popSTR2
Runtime	57.48m	34.75m	9.70h	7.15h
Accuracy	99.6%	99.2%	99.6%	99.8%

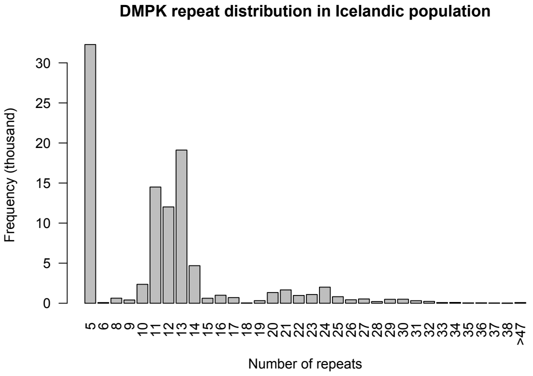

Fig. 1. Distribution of repeat numbers in Icelandic population at DMPK locus. Peaks at five repeats and 11-13 repeats match previously reported distributions in European populations.

Table 4. Disease-associated markers in panel and corresponding pathogenicity thresholds.

Location (HG38)	**Gene**	**Disease**	**Pathogenicity threshold**	**OMIM link**
chr1:149390803-149390842	NBPF19	Neuronal intranuclear inclusion disease	90	603472
chr2:176093058-176093099	HOXD13	Synpolydactyly 1	21	186000
chr2:190880873-190880920	GLS	Global developmental delay, progressive ataxia, and elevated glutamine	90	618412
chr3:63912685-63912716	ATXN7	Spinocerebellar ataxia 7	37	164500
chr3:129172577-129172659	CNBP	Myotonic dystrophy 2	50	602668
chr4:3074877-3074940	HTT	Huntington disease	36	143100
chr4:41745976-41746022	PHOX2B	Congenital central hypoventilation	25	209880
chr5:146878728-146878759	PPP2R2B	Spinocerebellar ataxia 12	55	604326
chr6:16327634-16327724	ATXN1	Spinocerebellar ataxia 1	39	164400
chr6:45422750-45422794	RUNX2	Cleidocranial dysplasia	27	119600
chr6:170561907-170562017	TBP	Spinocerebellar ataxia 17	47	607136
chr7:27199680-27199729	HOXA13	Hand-foot-genital syndrome	18	140000
chr8:104588965-104588999	LRP12	Oculopharyngodistal myopathy	93	164310
chr9:27573485-27573546	C9orf72	Amyotrophic lateral sclerosis	21	105550
chr9:69037285-69037304	FXN	Friedreich ataxia 1	200	229300
chr10:79826377-79826404	LOC642361/ NUTM2B-AS1	Oculopharyngeal myopathy with leukoencephalopathy	-	NatGen
chr12:6936717-6936775	ATN1	Dentatorubral pallidoluysian atrophy	49	125370
chr12:111598950-111599019	ATXN2	Spinocerebellar ataxia 2	35	183090
chr13:70139384-70139429	ATXN8OS	Spinocerebellar ataxia 8	111	608768
chr13:99985449-99985494	ZIC2	Holoprosencephaly-5	25	609637
chr14:23321472-23321492	PABPN1	Oculopharyngeal muscular dystrophy	12	164300
chr14:92071011-92071052	ATXN3	Spinocerebellar ataxia 3	55	109150
chr16:87604283-87604329	JPH3	Huntington disease-like-2	50	606438
chr18:55586154-55586229	TCF4	Corneal dystrophy	40	613267
chr19:13207859-13207898	CACNA1A	Spinocerebellar ataxia 6	20	183086
chr19:18786035-18786050	COMP	Multiple epiphyseal dysplasia	6	132400
chr19:45770205-45770266	DMPK	Myotonic dystrophy 1	50	160900
chr20:2652733-2652775	NOP56	Spinocerebellar ataxia 36	650	614153
chr21:43776445-43776479	CSTB	Myoclonic epilepsy of Unverricht and Lundborg	30	254800
chr22:45795355-45795424	ATXN10	Spinocerebellar ataxia 10	800	603516
chrX:67545318-67545383	AR	Kennedy disease	37	313200

Chapter 4

Microsatellite de novo mutation rate associates with parental age and germline variants in MSH2 and NEIL2

Microsatellite de novo mutation rate associates with parental age and germline variants in *MSH2* and *NEIL2*

Written with: Hakon Jonsson, Marteinn T. Hardarson, Gunnar Palsson, Doruk Beyter, Hannes P. Eggertsson, Arnaldur Gylfason, Gardar Sveinbjornsson, Guillaume Holley, Olafur A. Stefansson, Sigurgeir Olafsson, Gudny. A. Arnadottir, Pall I. Olason, Ogmundur Eiriksson, Gisli Masson, Unnur Thorsteinsdottir, Thorunn Rafnar, Patrick Sulem, Agnar Helgason, Daniel F. Gudbjartsson, Bjarni V. Halldorsson and Kari Stefansson

Abstract

Microsatellites are polymorphic tracts of short tandem repeats (STRs) with one to six base-pair (bp) motifs and are some of the most polymorphic markers in the genome. We describe microsatellites in 53,026 and 150,119 whole genome sequenced samples from Iceland and the UK Biobank (UKB), respectively. Using 6,084 Icelandic parent-offspring trios we find 76,987 microsatellite de novo mutations (mDNMs) and estimate the germline mDNM rate as $4.95 \cdot 10^{-5}$ (95% CI: $4.88\text{-}5.02 \cdot 10^{-5}$) mutations per microsatellite per generation (MMG), corresponding to 63.7 (95% CI: 61.9-65.4) mDNMs per proband per generation. Paternal mDNMs account for 76.9% of the total and occur at longer repeats, while maternal mDNMs affect more bp. mDNMs increase by 0.97 (95% CI: 0.90–1.04) and 0.31 (95% CI: 0.25-0.37) per year of father‘s and mother‘s age, respectively. We found two independent variants correlating with increased number of transmitted mDNMs. A missense variant (1.9% allele frequency) in *MSH2*, a mismatch repair gene, correlates with an increase of mDNMs transmitted from both parents (effect: 13.1 paternal and 7.8 maternal mDNMs). A synonymous variant (20.3% allele frequency) in *NEIL2*, a DNA damage repair gene, correlates with an increase of paternally transmitted mDNMs (effect: 4.4 paternal mDNMs). Thus, the microsatellite mutation rate in humans is in part under genetic control.

Introduction

Mutations enable life to evolve and adapt. Accurate estimates of the rate of mutations and the processes behind them are therefore imperative for understanding evolution, making inferences about population history(*1–6*) and understanding the genetics of disease and other phenotypes(*7–11*).

Around 3% of the human genome are short tandem repeats (STRs)(*12*), some of which are polymorphic, i.e. microsatellites, that mutate several orders of magnitude faster than unique sequences(*13*). Because of their high mutation rate and abundance in the human genome microsatellites have proven useful in a wide range of research(*14*). Genotyping microsatellites became standard practice when PCR-based methods emerged in the late eighties(*14*) and they were the main form of sequence variation studied until the advent of single nucleotide polymorphism (SNP) arrays(*15*).

Hypermutability of microsatellites located within or close to genes can cause diseases and syndromes, collectively referred to as repeat expansion disorders, which are caused by a dramatic microsatellite expansion after the repeat length exceeds a stability threshold, specific to each microsatellite(*16–19*). These include fragile X syndrome (FRAXA)(*20*), myotonic dystrophy type 1 (DM1)(*21*), and spinocerebellar ataxias (SCAs)*(22).*

The rate of mDNMs in humans has been shown to vary by microsatellite motif, motif length, allele length, GC content (*23–25*), parent of origin, and paternal(*26*) age, but researchers have to date been unable to detect a relationship with maternal age. mDNM rates for di- and tetranucleotide repeats were estimated as $1.0 \cdot 10^{-3}$ and $2.73 \cdot 10^{-4}$ mutations per microsatellite per generation (MMG), respectively(*1*) using a set of microsatellites chosen because of their high mutability. An mDNM rate of $5.6 \cdot 10^{-5}$ MMG was reported in a recent study on the contribution of mDNMs to autism spectrum disorder(*26*) which considered tandem repeats with motif lengths between one and twenty bp. To date, datasets used to study mDNMs have for the most part been confined to small sets of well defined microsatellite or focused on specific diseases.

mDNMs are believed to mostly occur through replication slippage caused by a failure in the processes responsible for sequence fidelity before, during, and after DNA replication. These processes are proofreading, which verifies the correct pairing of the most recently added bp, and mismatch/damage repair mechanisms, which detect and replace incorrectly paired or damaged DNA bases(*27*, *28*). Loss of function mutations affecting genes responsible for mismatch/damage repair or proofreading are known to cause somatic microsatellite instability, which in turn can result in increased risk of colorectal, gastric, endometrial and other types of cancer(*29*). However, apart from handful of clinical cases(*30*, *31*), no variants associated with the germline de novo mutation rate in humans have been reported.

Recombination, DNA damage repair and nonhomologous end joining (NHEJ) have also been implicated determinants of mDNMs(*32*, *33*) and since the two germlines have different exposure to these processes it is logical to assume that the mDNM accumulation is different between the sexes. The recombination rate is higher in females(*34*) and oocytes wait in a largely dormant state from the parent‘s birth to reproduction, exposed to DNA damage to be repaired by NHEJ or homologous recombination(*35*). Spermatogonia undergo mitosis continuously, increasing the risk of replication slippage events.

In addition to sex differences between the germlines, genetic factors responsible for genome integrity can be expected to play a role in DNM accumulation. Sequence variants that increase sDNM and mDNM rates are known in somatic tissues in humans(*36*). Sequence variants that increase the sDNM rate are known in animal and yeast models(*37*, *38*), but the detection of germline mutators has been elusive.

We identified and genotyped microsatellites in two large sequencing cohorts. Using these sets we estimated the mDNM rate and identified environmental and genetic determinants of mDNMs.

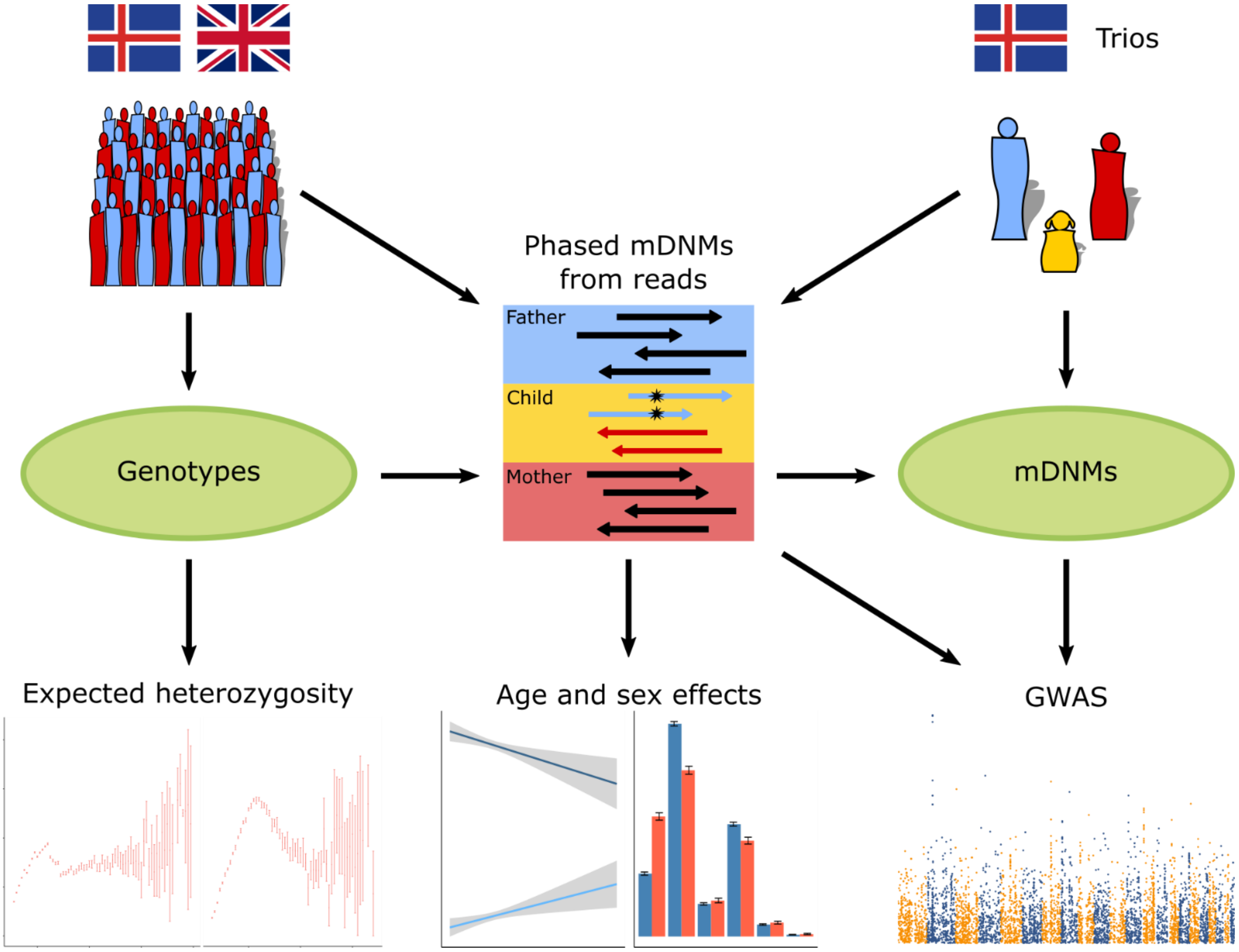

Fig. 1 | Overview of the analysis. We use WGS data from Iceland and UKB and genealogy data from Iceland. From the WGS data we generate microsatellite genotypes. Using trios in the genealogy and the genotypes we detect and phase mDNMs and count the number of mDNMs per trio. We associate the individual mDNM counts with genotypes of the parents in the trios. We compute population wide expected heterozygosity based on the genotypes and observe how it is affected by sequence context. From the phased mDNMs we also estimate age and sex effects on the mDNM rate and to create parental phenotypes based on number of mDNMs found in offspring from the phased mDNMs.

Results

STR identification and genotyping

We used popSTR(*39*) to call genotypes for 5,401,401 autosomal short tandem repeats (STRs), identified by Tandem repeats finder(*40*), in two large WGS datasets, 53,026 Icelanders and 150,119 participants in the UKB, sequenced to an average coverage of 39.2x (min: 19.7, max: 608.3) and 32.5x (min: 23.6, max: 128.1), respectively.

Each STR has a repeat motif along with start and end positions in the reference and we refer to the sequence between these positions as the reference repeat tract (RRT) and its length as RRT length. We define repeat purity as the ratio between the number of times the STR‘s repeat motif is observed in its RRT and the maximum number of repeat motifs if the sequence contained no interruptions (where the highest possible repeat purity is 1). For motif lengths between one and three bp, we compare the behavior of motif equivalence classes where all members in a class are either a circular shift or a reverse complement of each other. For example, the members of the AAT motif equivalence class are: AAT, ATA, TAA, ATT, TAT and TTA (Table 1).

Class representative	Other members
A	T
C	G
AC	CA, GT, TG
AG	GA, CT, TC
AT	TA
CG	GC
AAC	ACA, CAA, GTT, TGT, TTG
AAG	AGA, GAA, CTT, TCT, TTC
AAT	ATA, TAA, ATT, TAT, TTA
ACC	CAC, CCA, GGT, GTG, TGG
ACG	CGA, GAC, CGT, TCG, GTC
ACT	CTA, TAC, AGT, TAG, GTA
AGC	GCA, CAG, GCT, TGC, CTG
AGG	GAG, GGA, CCT, CTC, TCC
ATC	TCA, CAT, GAT, TGA, ATG
CCG	CGC, GCC, CGG, GCG, GGC

Table 1 : Motif equivalence classes and their members.

We found 1,394,292 (25.8%) and 2,393,292 (44.3%) of the STRs to be polymorphic in the Icelandic and UK datasets, respectively and will refer to these polymorphic STRs as microsatellites. We describe microsatellite diversity through polymorphism rate (the fraction of STRs that are polymorphic) and expected heterozygosity*(41)* (Fig. 4, Fig. *5*, Table 3, Table 4, Table 5, Table 6). Our microsatellite genotyping is limited by read length (151 bp for most samples), resulting in a decrease in accuracy when detecting and determining genotypes of alleles with long RRTs. The average expected heterozygosity in our datasets decreases at RRT lengths exceeding 80 bp (Fig. 4, Fig. 5), so we conclude that this is the length where the effect of the RRT length on the genotyping accuracy becomes pronounced.

A number of STR properties correlate with the polymorphism rate and expected heterozygosity(*41*). Both the motif length and how many times the motif is repeated in its RRT (repeat number) affect the polymorphism rate and the expected heterozygosity. Almost all homopolymers are polymorphic and both the fraction of polymorphic STRs and expected heterozygosity decrease with motif length (linear regression $P < 1 \cdot 10^{-320}$ for both datasets, Fig. 4, Fig. 5). For all motif lengths, an increase in repeat number results in both a higher polymorphism rate and expected heterozygosity (linear regression $P < 1 \cdot 10^{-320}$ for both datasets, Fig. 4, Fig. 5).

The fraction of G/C bases in an STR‘s motif (motif GC content) is negatively correlated with polymorphism rate (Table 6) while negative correlation with expected heterozygosity is only observed for motif lengths above two bp (Table 5). Homopolymers from the C motif class have higher expected heterozygosity than A class homopolymers (Table 5) but account for only 0.8% of all homopolymers (Table 2).

CpG microsatellites (CG motif class) also have higher expected heterozygosity on average than the other dinucleotide motif classes but account for only 0.4% of all dinucleotide repeats (Table 2, Table 5). Thus, although C class homopolymers and CG class dinucleotide microsatellites have higher expected heterozygosity values than other classes with equal motif lengths, their overall effect on microsatellite diversity is small since they are so rare.

Enrichment of A motif class homopolymers in the human genome is thought to be a result of the microsatellite-like structure often found at 3‘ ends of reverse transcribed RNA sequences, i.e. poly-A tails(*42*, *43*) and the high expected heterozygosity rate at CG class microsatellites is consistent with CpG sites acting as mutational hot spots(*44*). Negative correlation of GC content to polymorphism rates and expected heterozygosity can likely be explained by the three hydrogen bonds between paired G/C bp, compared to two between A/T bp, making the slippage-causing disassociation from the template strand during replication less likely.

Stratified on RRT length, all correlations between repeat purity and both expected heterozygosity and polymorphism rate were positive (Table 3, Table 4, Fig. 2, Fig. 3). This is consistent with(*14*, *45–47*), reporting a positive correlation between repeat purity and mDNM rate. Interruptions of the repeat sequence decrease the number of locations where replication slippage can occur and thus the mDNM rate(*24*).

Motif class	Maternal	Paternal	Full marker set	TRF output
A	97.5%	97.7%	99.2%	99.2%
C	2.5%	2.3%	0.8%	0.8%
AC	67.6%	84.3%	49.1%	54.4%
AG	8.9%	5.6%	19.9%	23.4%
AT	23.2%	10.0%	30.5%	21.7%
CG	0.4%	0.1%	0.4%	0.4%
AAC	13.0%	13.3%	27.1%	20.8%
AAG	3.8%	2.0%	15.0%	10.0%
AAT	75.3%	77.6%	34.9%	32.9%
ACC	0.8%	0.4%	4.0%	7.4%
ACG	0.0%	0.0%	0.1%	0.1%
ACT	1.2%	0.9%	0.9%	1.9%
AGC	2.1%	1.2%	3.0%	7.8%
AGG	0.8%	0.6%	7.5%	11.3%
ATC	2.4%	3.9%	4.3%	6.1%
CCG	0.6%	0.2%	3.1%	1.6%

Table 2 Fraction of each motif equivalence class within homopolymers, di- and trinucleotide repeats split on maternal and paternal mDNMs, fraction of each motif equivalence class in full marker set and in the Tandem Repeats Finder (TRF) output.

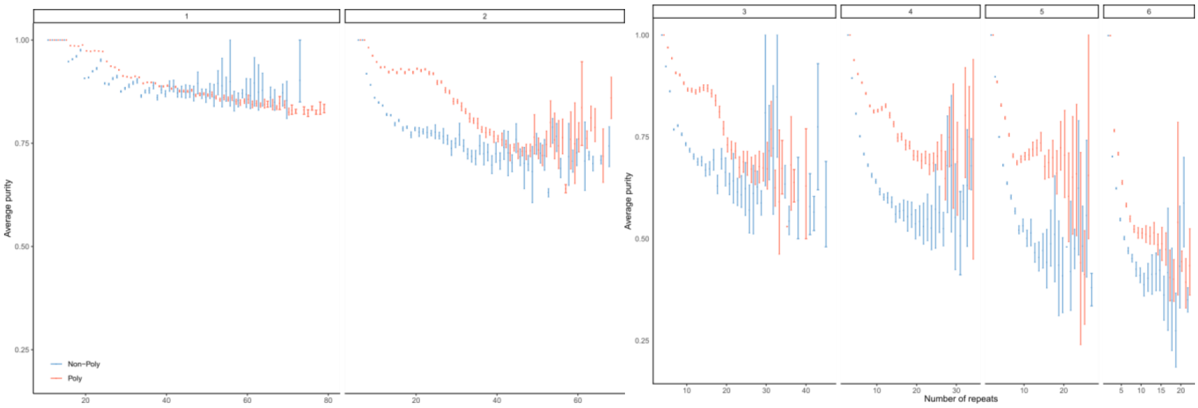

***Fig. 2* | Average repeat purity as a function of repeat number stratified on motif length for the Icelandic data set.** Red marks polymorphic STRs (microsatellites) and blue non-polymorphic STRs.

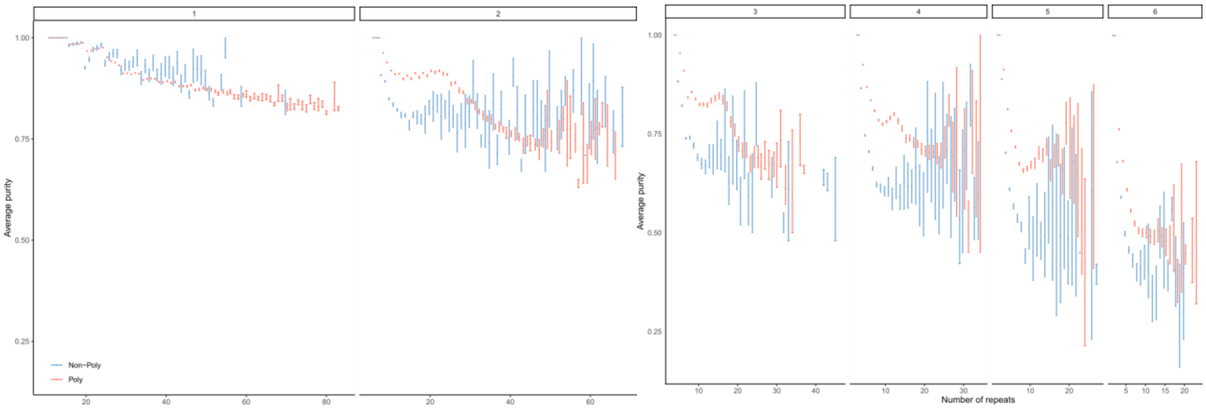

Fig. 3 | Average repeat purity as a function of repeat number stratified on motif length for the UKB data set. Red marks polymorphic STRs (microsatellites) and blue non-polymorphic STRs.

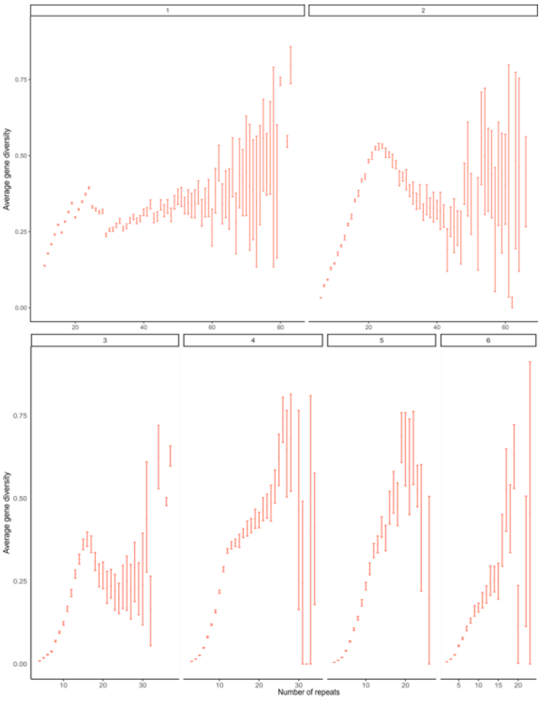

Fig. 4 | Average expected heterozygosity in UKB as a function of repeat number stratified on motif length. The drop in all motif lengths is most likely due our inability to reliably detect long alleles from short reads, causing underestimation of expected heterozygosity values at microsatellite with long reference alleles.

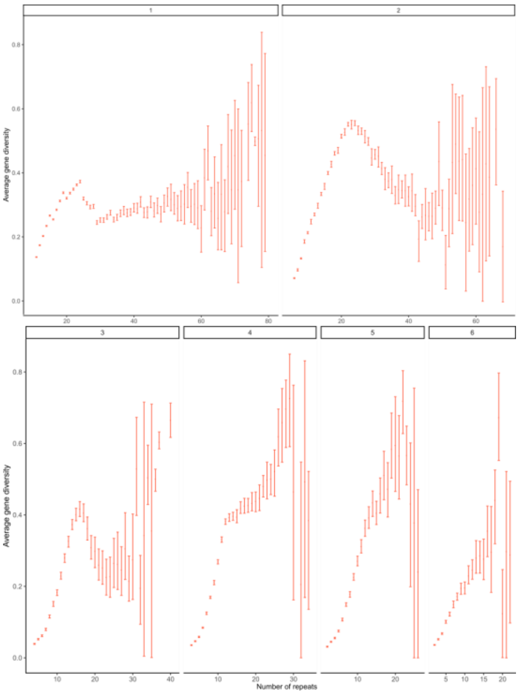

Fig. 5 | Average expected heterozygosity in Icelandic data set as a function of repeat number stratified on motif length.

RRT length	Purity effect (Ice/UKBB)	*P* (Ice/UKBB)
11-20	0.39/0.23	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
21-30	0.71/0.67	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
31-40	0.87/0.85	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
41-50	1.17/1.15	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
51-60	1.17/1.18	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
61-70	1.10/1.13	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
71-80	1.05/1.08	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
81-90	0.96/1.07	$3.9 \cdot 10^{-141}$/$6.8 \cdot 10^{-183}$
>90	0.76/0.83	$5.4 \cdot 10^{-56}$/$6.1 \cdot 10^{-57}$

Table 3 Linear regression coefficients and p-values for the effect of repeat purity on expected heterozygosity within ten base pair bins of RRT length for both the Icelandic and UKB datasets.

RRT length	Purity effect (Ice/UKBB)	*P* (Ice/UKBB)
11-20	7.86/5.31	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
21-30	7.84/6.09	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
31-40	6.11/5.47	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
41-50	5.94/4.96	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
51-60	4.78/3.668	$< 1.0 \cdot 10^{-320}$/$2.4 \cdot 10^{-83}$
61-70	4.42/2.72	$2.8 \cdot 10^{-164}$/$6.0 \cdot 10^{-17}$
71-80	4.22/1.95	$2.0 \cdot 10^{-61}$/$7.1 \cdot 10^{-6}$
81-90	4.03/-0.09	$6.2 \cdot 10^{-26}$/0.9
>90	2.72/0.61	$1.0 \cdot 10^{-16}$/0.1

Table 4 Logistic regression coefficients and p-values for the effect of repeat purity on polymorphism rate within ten base pair bins of RRT length for both the Icelandic and UKB datasets.

Motif length	Effect (Ice/UKBB)	*P* (Ice/UKBB)
1	0.15/0.17	$4.5 \cdot 10^{-312}$/$< 1.0 \cdot 10^{-320}$
2	0.14/0.10	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
3	-0.12/-0.05	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
4	-0.07/-0.04	$1.1 \cdot 10^{-155}$/$< 1.0 \cdot 10^{-320}$
5	-0.06/-0.03	$8.0 \cdot 10^{-141}$/$< 1.0 \cdot 10^{-320}$
6	-0.02/-0.01	$3.9 \cdot 10^{-18}$/$1.52 \cdot 10^{-95}$

Table 5 Linear regression coefficients and p-values for the effect of GC motif content on expected heterozygosity within motif lengths for both the Icelandic and UKB datasets.

Motif length	Effect (Ice/UKBB)	*P* (Ice/UKBB)
1	-0.20/-1.33	$4.9 \cdot 10^{-3}$/$4.6 \cdot 10^{-41}$
2	-0.29/-0.21	$1.3 \cdot 10^{-79}$/$1.6 \cdot 10^{-15}$
3	-0.79/-1.12	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
4	-1.47/-1.97	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
5	-2.02/-1.67	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$
6	-1.55/-1.34	$< 1.0 \cdot 10^{-320}$/$< 1.0 \cdot 10^{-320}$

Table 6 Logistic regression coefficients and p-values for the effect of GC motif content on polymorphism rate within motif lengths for both the Icelandic and UKB datasets.

mDNM rate

We detected 76,987 mDNMs in 6,084 Icelandic trios (mean: 12.7 mDNMs per trio) in 634,406 high quality microsatellites for offspring genotypes that were inconsistent with the genotypes of the parents. Each trio had on average 256,066 microsatellites (40.4%) where all members‘ genotypes passed quality filters and we were able to test for mDNMs.

To estimate the false positive rate of our method we used two methods; PacBio CCS sequence data, available for four of our trios, and mDNM sharing in nine Icelandic monozygotic twin pairs in our set. We used haplotype resolved assemblies of the PacBio data and were able to verify 26 out of the 27 mDNMs present in the PacBio sequenced trios, giving a false positive rate estimate of 3.7% (Table 7).

Chrom	Pos	Motif	Ref. (No repeats)	Offspring gt	Father gt	Mother gt	Verified
chr11	80254522	CTAT	15	15/15	11/16	15/15	yes
chr12	100919737	TATC	15	15/18	15/17	15/17	yes
chr2	199703012	GTTT	8	9/9	8/10	8/9	yes
chr4	115864579	TG	11	11/13	11/12	11/11	no
chr7	86946160	CA	16	14/18	16/18	15/18	yes
chr11	43435069	AGAT	16	17/17	15/18	15/17	yes
chr13	54157465	TA	20	16/19	18/21	16/22	yes
chr14	82746964	AATA	8	8/10	8/9	8/9	yes
chr17	54067449	TTTTA	7	7/8	7/7	7/7	yes
chr18	3742950	CTTC	11	11/14	11/12	11/13	yes
chr18	63468251	AAC	8	10/12	8/10	10/10	yes
chr1	10498281	AAAAC	8	8/7	9/9	7/9	yes
chr4	140179641	TGTTT	8	7/9	8/5	5/7	yes
chr6	44850746	ATCT	13	16/16	15/15	10/16	yes
chr16	69147179	AAT	15	14/16	15/15	15/14	yes
chr1	180853627	AC	21	19/25	19/26	19/20	yes
chr20	34877898	AGAT	12	15/16	13/16	16/16	yes
chr3	85784531	TATC	13	13/12	10/14	11/12	yes
chr4	169443052	TATC	13	12/15	14/14	13/12	yes
chr6	38464486	TTTTC	6	6/5	6/7	6/7	yes
chr7	29184144	AGAT	14	14/15	11/16	14/13	yes
chr8	110980554	ATCT	10	12/12	11/11	12/13	yes
chr11	22272303	CA	22	22/19	23/26	22/25	yes
chr12	29413662	AC	15	16/20	14/16	14/19	yes
chr13	80326450	TG	9	9/8	8/8	8/8	yes
chr1	172791800	AT	7	7/14	7/7	7/13	yes
chr8	19119757	TA	9	9/13	9/12	9/12	yes

Table 7 mDNMs verified by haplotype resolved assemblies generated from pacBio HiFi sequencing data. We were unable to verify homopolymer mDNMs due to high error rates.

The monozygotic twin pairs shared 217 of the 230 mDNMs observed which gives a false positive rate estimate of 5.6%. We note that this is likely to be an overestimate, as some of the differences between the twin pairs could be due to result of post zygotic mutations, representing true differences between twins(*48*).

Using haplotype sharing across 540 three-generation families (795 trios), we counted how many times an mDNM was transmitted from a proband to offspring and estimated the transmission rate. The expected value of the transmission rate is 0.50 and deviations from it quantify false positive mDNM detection rates. For example, if the observed mutations were somatic and thus false positive as mDNMs, we would not observe transmission from the probands to their offspring. We observe a transmission rate of 0.49 (N = 11,228, 95% CI: 0.48-0.50) which gives an estimated false positive rate of 2%, although transmission rates vary between motif lengths and thus the error rate estimates as well (Table 8). Notably, the transmission rate for homopolymers is only 0.4 while the other motif lengths have transmission rates much closer or equal to 0.5.

Motif length (bp)	Transmission rate	Error estimate
1	0.40 (861/2,156)	20%
2	0.50 (2,558/5,083)	0%
3	0.51 (433/857)	2%
4	0.53 (1,458/2,764)	6%
5	0.55 (169/310)	10%
6	0.47 (27/58)	6%
Total	0.49 (5,506/11,228)	2%

Table 8 Transmission rate by motif length and combined for all mDNMs phased using haplotype sharing in three-generation families.

We estimated the average mDNM rate over all motif lengths as $4.95 \cdot 10^{-5}$ MMG and observed an order of magnitude difference in mDNM rates between motif lengths, with rates ranging from $1.0 \cdot 10^{-5}$ MMG for hexanucleotide repeats to $1.1 \cdot 10^{-4}$ MMG for dinucleotide repeats (Table 9, Fig. 6).

Motif (bp)	mDNM rate (95% CI)	#microsatellites	#DNMs
1	$2.15 \cdot 10^{-5}$ ($2.11 \cdot 10^{-5}$-$2.19 \cdot 10^{-5}$)	399,087 (62.9%)	17,194 (22.3%)
2	$1.07 \cdot 10^{-4}$ ($1.06 \cdot 10^{-4}$-$1.09 \cdot 10^{-4}$)	93,653(14.8%)	33,025 (42.9%)
3	$4.92 \cdot 10^{-5}$ ($4.70 \cdot 10^{-4}$-$5.21 \cdot 10^{-4}$)	29,869 (4.7%)	5,469 (7.1%)
4	$8.57 \cdot 10^{-5}$ ($8.28 \cdot 10^{-4}$-$8.86 \cdot 10^{-4}$)	66,541 (10.5%)	18,795 (24.4%)
5	$2.58 \cdot 10^{-5}$ ($2.37 \cdot 10^{-5}$-$2.86 \cdot 10^{-5}$)	29,992 (4.7%)	2,132 (2.8%)
6	$1.04 \cdot 10^{-5}$ ($9.07 \cdot 10^{-6}$-$1.23 \cdot 10^{-5}$)	15,264 (2.4%)	372 (0.5%)
Total	$4.95 \cdot 10^{-5}$ ($4.88 \cdot 10^{-5}$-$5.02 \cdot 10^{-5}$)	634,406	76,987

Table 9: mDNM rate, number of microsatellites in high quality set and number of mDNMs, all by motif lengths. Dinucleotide repeats have the highest mDNM rate and represent almost 43% of our mDNMs.

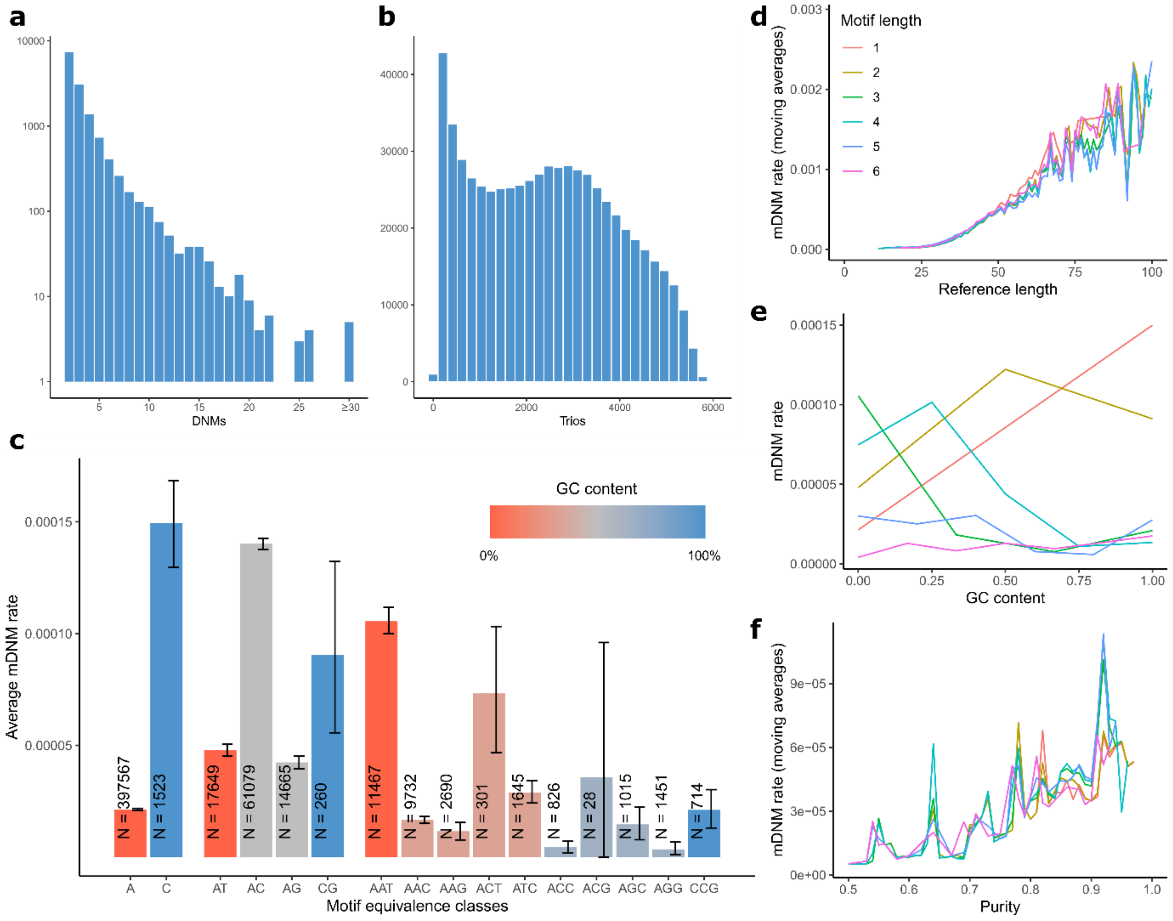

Fig. 6 | mDNM rate a) Histogram showing the number of mDNMs per microsatellite. Only microsatellites with at least one mDNM are shown and the counts are on a log-scale. b) Histogram showing the number of trios available to check for mDNMs per microsatellite with at least one mDNM. c) mDNM rates of motif equivalence classes for motif lengths between one and three bp with error bars representing 95% confidence intervals. d) mDNM rate as function of RRT length stratified on motif length. The RRT length affects mDNM rates in a similar way for all motif lengths. e) mDNM rate as function of GC content in motif. f) mDNM rate as a function of repeat purity stratified on motif length. The mDNM rate increases with purity for all motif lengths.

Using motif length specific mutation rates and the average number of markers available at each trio to extrapolate to the full set of 1,394,292 microsatellites gives an expected number of 63.7 mDNMs (95% CI: 61.9-65.4) per proband per generation. This extrapolated number is comparable with the de novo mutational load of SNPs and indels accessible by short read sequencing(*49*, *50*).
Attributes known to correlate with polymorphism rate were also correlated with the mDNM rate. As the RRT length increased from ten to 100 bp the mDNM rate also increased for all motif lengths (Fig. 6), consistent with findings from previous studies(*23*, *51–53*) and, intuitively, the opportunities for errors during replication increase with a microsatellite‘s RRT length.

Microsatellites with longer RRTs have been shown to be more likely to contract and shorter ones to expand(*23*, *54–59*). We replicate this. For all RRT length thresholds, a higher fraction of mDNMs had a gain of repeat motifs below the threshold than above it, i.e., the microsatellite group with shorter overall RRTs were more likely to expand than the group with longer RRTs (Table 10).

Length threshold (bp)	percentage adding bp below	percentage adding bp above
20	59.4%	52.1%
30	59.6%	50.7%
40	56.9%	49.5%
50	54.2%	49.9%
60	53.5%	51.2%
70	53.5%	50.2%
80	53.4%	46.2%
90	53.4%	50.0%
100	53.4%	42.9%

Table 10 Fraction of mDNMs adding bp above and below different repeat tract length thresholds. The fraction of mDNMs adding bp is higher below the threshold in all cases.

Our mDNM rate estimate is nominally lower than a previous estimate(*26*) of $5.6 \cdot 10^{-5}$ and lower than the two estimates of $10.0 \cdot 10^{-4}$ and $2.7 \cdot 10^{-4}$ for tetra- and dinucleotide repeats, respectively(*1*). This apparent discrepancy could be a result of a more conservative filtering in our study, a younger set of parents, a generally healthier cohort and a different range of motif lengths considered. However, the most likely reason for the apparent discrepancy is the sample size difference between the studies. Our set contains 53,026 individuals while the set analyzed by Mitra et al. (*26*) contained 6,548 individuals. Thus, our minimum detection frequency is $1/(2 \cdot 53{,}026) = 9.0 \cdot 10^{-6}$

compared to the minimum detection frequency(*26*) of $1/(2 \cdot 6{,}548) = 7.6 \cdot 10^{-5}$ enforced by the smaller sample size. We recomputed our mutation rate estimate conditioning on microsatellite frequency (Table 11) and confirmed that at a detection frequency cutoff of $7.6 \cdot 10^{-5}$ our estimate becomes $5.6 \cdot 10^{-5}$ and matches the one presented by Mitra et al.(*26*). Similarly, at a minimum frequency of 10% our estimate is comparable to the one from Sun et al.(*1*) Based on this we conclude that a mDNM rate estimate depends on the size of the sample set studied.

Detection frequency	DNM rate estimate
$9.0 \cdot 10^{-6}$	$5.0 \cdot 10^{-5}$
$7.6 \cdot 10^{-5}$	$5.6 \cdot 10^{-5}$
$1.0 \cdot 10^{-2}$	$8.1 \cdot 10^{-5}$
$5.0 \cdot 10^{-2}$	$9.3 \cdot 10^{-5}$
0.1	$1.0 \cdot 10^{-4}$

Table 11 Microsatellite mDNM rate estimates for different detection frequencies, the number of markers included in the analysis depends on the detection frequency defined by the sample size.

The mDNM rate for homopolymers was positively correlated with the motif G/C content, while di-, tri-, tetra- and pentanucleotide repeats had a negative correlation with the mDNM rate and for hexanucleotide repeats we lacked power to detect a correlation with the mDNM rate (Fig. 6,Table 12).

Motif	Repeat tract length	Purity	GC content	Motif length
All	0.10 ($<1 \cdot 10^{-320}$)	0.12($<1 \cdot 10^{-320}$)	$9.47 \cdot 10^{-5}$(0.58)	-0.15 ($<1 \cdot 10^{-320}$)
1bp	0.16($<1 \cdot 10^{-320}$)	0.28($<1 \cdot 10^{-320}$)	0.02($<1 \cdot 10^{-320}$)	X
2 bp	0.10($<1 \cdot 10^{-320}$)	0.17($<1 \cdot 10^{-320}$)	-0.01($8.6 \cdot 10^{-87}$)	X
3 bp	0.12 ($<1 \cdot 10^{-320}$)	0.14 ($<1 \cdot 10^{-320}$)	-0.03($3.0 \cdot 10^{-235}$)	X
4 bp	0.10($<1 \cdot 10^{-320}$)	0.10 ($<1 \cdot 10^{-320}$)	-0.003($2.2 \cdot 10^{-10}$)	X
5 bp	0.1 ($<1 \cdot 10^{-320}$)	0.05($4.4 \cdot 10^{-121}$)	-0.003($2.4 \cdot 10^{-2}$)	X
6 bp	0.09($<1 \cdot 10^{-320}$)	0.02($8.9 \cdot 10^{-18}$)	$7.5 \cdot 10^{-4}$ (0.76)	X

Table 12 Poisson multiple regression coefficients and p-values for the effect on the mDNM rate by RRT length, repeat purity, GC content and motif length for all markers and stratified on motif length. Repeat tract length and repeat purity remain significant and consistent in their effect directionality across the full data set and all motif length subsets. GC-content is positively correlated to the mDNM rate for homopolymers but for di-, tri-, tetra-, and pentanucleotide repeats the motif GC-content has an inverse correlation to the mDNM rate.

Repeat purity correlated positively with the mDNM rate for all motif lengths (effect = 0.12, $P < 1 \cdot 10^{-320}$, Fig. 6, Table 12), consistent with previously published results(*53*). The correlation remained positive in most cases after conditioning on RRT lengths (Table 13).

Bp	1 bp motif (*P*)	2 bp motif (*P*)	3 bp motif (*P*)	4 bp motif (*P*)	5 bp motif (*P*)	6 bp motif (*P*)
11-20	0.13($3.2 \cdot 10^{-93}$)	0.04($1.2 \cdot 10^{-9}$)	0.05($5.7 \cdot 10^{-4}$)	-0.005(0.36)	-0.006(0.25)	-0.01(0.17)
21-30	0.21($< 1.0 \cdot 10^{-320}$)	0.16($< 1.0 \cdot 10^{-320}$)	0.13($3.4 \cdot 10^{-38}$)	0.05($2.0 \cdot 10^{-35}$)	-0.002(0.56)	0.002(0.65)
31-40	0.20($2.0 \cdot 10^{-36}$)	0.17($< 1.0 \cdot 10^{-320}$)	0.18($7.3 \cdot 10^{-136}$)	0.11($2.2 \cdot 10^{-211}$)	0.06($1.1 \cdot 10^{-28}$)	0.01(0.01)
41-50	0.06(0.23)	0.16($< 1.0 \cdot 10^{-320}$)	0.12($1.4 \cdot 10^{-95}$)	0.09($1.3 \cdot 10^{-293}$)	0.08($4.4 \cdot 10^{-40}$)	0.03($1.0 \cdot 10^{-6}$)
51-60	-0.50($3.7 \cdot 10^{-4}$)	0.15($3.0 \cdot 10^{-252}$)	0.11($5.1 \cdot 10^{-40}$)	0.07($5.8 \cdot 10^{-216}$)	0.06($3.5 \cdot 10^{-24}$)	0.06($1.1 \cdot 10^{-14}$)
61-70	-1.13($5.9 \cdot 10^{-5}$)	0.12($1.6 \cdot 10^{-36}$)	0.12($1.8 \cdot 10^{-12}$)	0.06($3.9 \cdot 10^{-115}$)	0.06($1.4 \cdot 10^{-18}$)	0.06($2.2 \cdot 10^{-9}$)
71-80	0.65(0.48)	0.14($9.3 \cdot 10^{-10}$)	0.03(0.34)	0.04($4.4 \cdot 10^{-42}$)	0.03($6.1 \cdot 10^{-6}$)	0.02(0.15)
81-90	X	0.05(0.58)	0.47(0.31)	0.04($7.6 \cdot 10^{-12}$)	0.05($5.0 \cdot 10^{-5}$)	0.06(0.06)
>90	X	X	5.43(1.00)	0.03($4.3 \cdot 10^{-5}$)	0.008(0.42)	0.04(0.34)

Table 13 Poisson regression coefficients and p-values for effect of repeat purity on mDNM rate split into ten bp RRT length bins stratified on motif length. Data for regression was not available for homopolymers above 80 bp and dinucleotide microsatellites above 90 bp.

The mDNM rate is higher for C class homopolymers than for A class ones (Mann-Whitney U test $P < 1 \cdot 10^{-230}$, Fig. 6), but C homopolymers are much rarer and represent only 0.8% of all homopolymers in our set (Table 2). The AC motif class has the highest mDNM rate of the dinucleotide microsatellites (Table 14, Fig. 6).

Motif equivalence class	Regression α and (*P*) for comparison with AC-class mDNM rate
AG	-1.20 ($<1 \cdot 10^{-320}$)
AT	-1.08 ($<1 \cdot 10^{-320}$)
CG	-0.43 ($7.7 \cdot 10^{-3}$)

Table 14 Regression coefficients and p-values from a Poisson regression comparing the mDNM rate of the other motif equivalence classes to the AC class using available trios per marker as exposure but without correcting for RRT length. All motif classes have significantly lower mDNM rates.

However, the average RRT length of the AC motif class is longest among dinucleotide classes. Including RRT length as a covariate the CG motif class has a higher mDNM rate than all other dinucleotide classes (Table 15), in line with the fact that CpG sites have been shown to act as mutational hot spots(*44*).

Motif equivalence class	Regression α and (*P*) for comparison with CG-class mDNM rate
AC	-0.81 ($5.7 \cdot 10^{-7}$)
AG	-1.48 ($1.7 \cdot 10^{-19}$)
AT	-1.18 ($5.8 \cdot 10^{-13}$)

Table 15 Regression coefficients and p-values from a Poisson regression comparing the mDNM rate of the other motif equivalence classes to the CG class after correcting for RRT length and using available trios per marker as exposure. All motif classes have significantly lower mDNM rates.

The AAT motif class had a higher mDNM rate than eight of the other nine trinucleotide repeat motif classes. Only the rarest class (ACG) did not show a significantly different mDNM rate (Table 16). The AAT motif class accounts for 38.4% of all trinucleotide microsatellites and 79.0% of trinucleotide mDNMs and has an mDNM rate 1.5 times higher than the second highest class.

Motif equivalence class	Regression α (*P*) for comparison with AAT mDNM rate
AAC	-1.35 (2.7 · 10^{-217})
AAG	-1.67 (5.3 · 10^{-75})
ACC	-2.99 (5.6 · 10^{-29})
ACG	-0.06 (0.92)
ACT	-2.49 (3.3 · 10^{-88})
AGC	-1.8 (8.6 · 10^{-41})
AGG	-3.09(4.9 · 10^{-37})
ATC	-1.78 (1.2 · 10^{-120})
CCG	-1.14(1.7 · 10^{-11})

Table 16 Regression coefficients and p-values from a Poisson regression comparing the mDNM rate of the other motif equivalence classes to the AAT class, correcting for RRT length and using available trios per marker as exposure. All motif classes expect for the rarest one (ACG) have significantly lower mDNM rates.

A higher mDNM rate for AAT class motifs has been previously reported for other organisms(*60*, *61*) but not, to our knowledge for, humans.

Previous studies have reported increased efficiency of mismatch repair (MMR) in early-replicating regions of the human genome(*62*). Our results are in line with this since we see 1.28 (95% CI: 1.25-1.31) fold depletion of mDNMs in early replicating regions of the genome(*63*).

The mDNM rate is lower in exons than in other parts of the genome

Exonic mDNMs are rarer than their intergenic and intronic counterparts. In 2,568,858 transmissions of microsatellites intersecting exons by one or more bp, we observed 33 mDNMs.

We estimated the exonic mDNM rate as 1.3 · 10^{-5} MMG, which is 3.9 (95% CI: 2.8-5.6) times lower than the genome-wide estimate. mDNMs are further 1.7 (95% CI: 1.3-2.1), 1.4 (95% CI: 1.3-1.5) and 4.2 (95% CI: 1.2-34.4)-fold depleted in 5‘UTR and 3‘UTR and splice regions, respectively. The 33 exonic mDNMs occurred at 21 unique microsatellites, of which 19 had motif lengths that were multiples of three and since amino acids are coded with three bp codons, mutations at microsatellites with multiple of three motif lengths are unlikely to cause a frameshift but rather an in-frame alteration of a gene. Sixteen of the exonic mDNMs were trinucleotide microsatellites, three were hexanucleotide microsatellites and the remaining two were homopolymers.

Tri- and hexanucleotide repeats were enriched in coding exons (chi squared test $P < 1 \cdot 10^{-320}$) compared with the rest of genome. Microsatellites with motif lengths that are multiples of three accounted for 93.3% of exon intersecting microsatellites and 70.9% of all exon intersecting non polymorphic STRs (Table 17). In contrast, 12.3% of all microsatellites had motifs that are multiples of three (Table 17) and 44.5% of non-polymorphic STRs.

Motif length (bp)	All microsatellites	Microsatellites in coding exons	Non polymorphic STRs in coding exons
1	50.0%	1.8%	0.003%
2	16.7%	0.5%	0.6%
3	5.8%	76.2%	17.5%
4	13.2%	3.0%	6.4%
5	7.1%	1.5%	22.1%
6	6.5%	17.1%	53.4%

Table 17 Motif length composition for all microsatellites and exon intersecting microsatellites.

The average purity of microsatellites was 0.94, while among microsatellites in exons the purity was notably lower (0.87, Mann-Whitney $P = 1 \cdot 10^{-153}$). Purity was positively correlated with the mDNM rate, so decreased purity in exons may decrease occurrences of possibly pathogenic mDNMs. This indicates that there is a possible positive selection for point mutations that reduce the purity of exonic microsatellites or a possible negative selection for point mutations that increase in their purity. The purity difference is largest for trinucleotide repeats, the most common motif length for exon intersecting microsatellites (Table 18). Non-polymorphic coding STRs do not have decreased purity compared to their intergenic counterparts, so point mutations are less likely to be the mechanism preventing mutations at these exonic STRs (Table 18).

Motif length	All microsatellites	Coding microsatellites	Non-polymorphic STRs	Coding non-polymorphic STRs
1	0.98	0.97 (0.25)	0.97	0.95 (0.05)
2	0.95	0.96 (0.14)	0.93	0.97 ($5.9 \cdot 10^{-12}$)
3	0.93	0.89 ($9.0 \cdot 10^{-66}$)	0.94	0.93 ($1.3 \cdot 10^{-8}$)
4	0.89	0.88 (0.87)	0.93	0.96 ($5.7 \cdot 10^{-43}$)
5	0.84	0.89 ($3.0 \cdot 10^{-4}$)	0.96	0.97 ($2.3 \cdot 10^{-26}$)
6	0.81	0.80 (0.12)	0.94	0.94 ($3.8 \cdot 10^{-4}$)
Overall	0.94	0.87 ($1 \cdot 10^{-153}$)	0.95	0.95 ($9.6 \cdot 10^{-14}$)

Table 18 Average purity values per motif length for all microsatellites, coding microsatellites, all non-polymorphic STRs and coding non-polymorphic STRs. Mann-Whitney U-test p-values for significant difference in purity values between exonic and non-exonic microsatellites in each motif length in brackets.

Parent of origin effects

To determine the sex differences in mDNMs formation, we assigned a parent of origin to 46,171 (60.0%) of the mDNMs using a combination of three methods; read pair tracing, allele sharing and haplotype sharing in three-generation families. The concordance was above 93% between all three methods (Table 19, Table 20, Table 21).

		Read pair tracing		
		Paternal	Maternal	Total
3 generation	Paternal	2,199	70	2,269(75.5%)
	Maternal	118	619	737(24.5%)
	Total	2,317(77.1%)	689(22.9%)	

Table 19 Comparison of phasing results for mDNMs phased using both three generation and read pair phasing. The concordance between the methods is 93.7% and the ratio between maternal and paternal mDNMs is similar in both sets.

		Allele based		
		Paternal	Maternal	Total
3 generation	Paternal	2,390	39	2,429(78.2%)
	Maternal	103	576	679(21.8%)
	Total	2,493(80.2%)	615(19.8%)	

Table 20 Comparison of phasing results for mDNMs phased using both three generation and allele based phasing. The concordance between the methods is 95.4% and the ratio between maternal and paternal mDNMs is similar in both sets.

		Read pair tracing		
		Paternal	Maternal	Total
Allele based	Paternal	8,865	9	8,874(79.6%)
	Maternal	21	2,258	2,279(20.4%)
	Total	8,886 (79.7%)	2,267(20.3%)	

Table 21 Comparison of phasing results for mDNMs phased using both read pair and allele based phasing. The concordance between the methods is 99.7%.

We found mDNMs from fathers (N=35,501, 76.9%) to be 3.3 (95% CI: 3.2-3.4), chi squared test $P < 1 \cdot 10^{-320}$) times more common than from mothers (N=10,670, 23.1%) (Table 22).

Motif length (bp)	Maternal	Paternal	Maternal percentage
1	2,962	5,141	36.6%
2	4,119	17,388	19.2%
3	887	2,663	25.0%
4	2,365	9,184	20.5%
5	343	977	26.0%
6	62	148	29.5%
Total	10,738	35,501	23.2%

Table 22 Parental ratios of mDNMs stratified on motif length and on the full set.

Maternal and paternal mDNMs occurred with different probabilities at different RRT lengths, motif lengths, and motif equivalence classes. First, while dinucleotide mDNMs were most common for both parents (Table 23), a higher fraction of maternal mDNMs occurred at homopolymers, tri-, penta-, and hexanucleotide microsatellites, while a larger fraction of paternal mDNMs occurred at di- and tetranucleotide microsatellites (Table 24, Fig. 7).

Motif (bp)	Paternal	Motifs affected	Maternal	Motifs affected	Mann-Whitney *P*
1	5,141(14.5%)	1.67	2,962(27.6%)	2.13	**$1.9 \cdot 10^{-23}$**
2	17,388(49.0 %)	1.38	4,119(38.4%)	1.71	**$6.0 \cdot 10^{-13}$**
3	2,663(7.5%)	1.22	887(8.2%)	1.33	$5.5 \cdot 10^{-2}$
4	9,184(25.9%)	1.06	2,365(22.0%)	1.12	**$9.6 \cdot 10^{-5}$**
5	977(2.8%)	1.08	343(3.2%)	1.21	0.38
6	148(0.4%)	0.95	62(0.6%)	0.77	0.26

Table 23 Motif length composition of paternal and maternal mDNMs, mean number of motifs added/removed for each motif length in each parent and Mann-Whitney U test p-values for different step sizes between parents. Bold represents significant difference in step size between parents (p<0.05).

Motif length (bp)	Odds ratio maternal/paternal (95% CI)	Fisher exact test *P*
1	2.25 (2.14-2.37)	$2.0 \cdot 10^{-198}$
2	0.65 (0.62-0.68)	$6.3 \cdot 10^{-84}$
3	1.11 (1.03-1.20)	$1.0 \cdot 10^{-2}$
4	0.81 (0.77-0.85)	$4.2 \cdot 10^{-16}$
5	1.17 (1.03-1.32)	$1.7 \cdot 10^{-2}$
6	1.39 (1.03-1.86)	$3.0 \cdot 10^{-2}$
Motif length (> 1bp)	Odds ratio maternal/paternal (95% CI) (no 1 bp)	Fisher exact test *P* (no 1 bp)
2	0.84 (0.80-0.88)	$1.0 \cdot 10^{-11}$
3	1.34 (1.24-1.45)	$3.3 \cdot 10^{-12}$
4	1.01 (0.95-1.06)	0.78
5	1.39 (1.22-1.57)	$6.3 \cdot 10^{-7}$
6	1.64 (1.21-2.20)	$1.9 \cdot 10^{-3}$

Table 24 Per motif length enrichment of mDNMs between maternally and paternally phased mDNMs. Top half shows enrichment including homopolymers and bottom half without them. Tri-, penta- and hexanucleotide repeats are enriched in maternal mDNMs while dinucleotide microsatellites are paternally enriched.

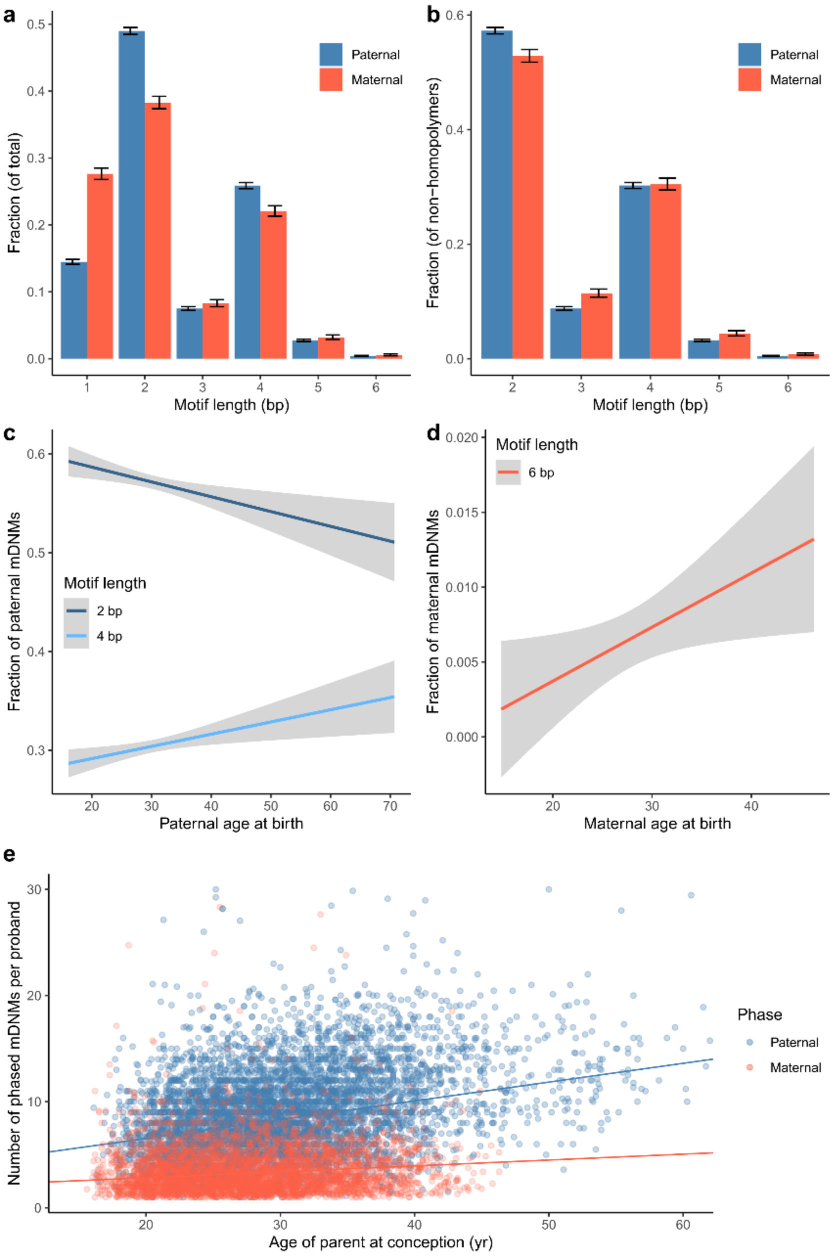

Fig. 7|Sex differences and age effects a) Relative frequencies of paternal and maternal mDNMs at different motif lengths. b) Relative frequencies of paternal and maternal mDNMs at different motif lengths without homopolymer mDNMs. c) Fraction of di- and tetranucleotide paternal mDNMs as a function of paternal age. The tetranucleotide fraction increases while the dinucleotide fraction decreases. d) The fraction of hexanucleotide maternal mDNMs increases with maternal age. e) Paternal and maternal age effect regression lines within our mDNM set. The age effect reported is interpolated to a genome-wide value using the fraction between average number of available microsatellites and total number of microsatellites. Fractions in c) and d) are computed after excluding homopolymers.

Second, the average number of bp affected by each mDNM was greater from mothers (3.4 bp, 95% CI: 3.3-3.4) than fathers (3.1 bp, 95% CI: 3.1-3.1) (Mann-Whitney U test *P* = 5.0 · 10^{-8}, Table 25), consistent with previous results(*26*).

Mutation size (In motifs)						
Motif	<-2 (p/m)	-2 (p/m)	-1 (p/m)	1 (p/m)	2 (p/m)	>2 (p/m)
1	312(168/144)	140(69/71)	2,782(1,671/1,111)	3,962(2,777/1,185)	214(121/93)	679(335/344)
2	610(399/211)	2,482(2,101/381)	7,124(6,192/932)	8,933(6,936/1,997)	1,849(1,490/359)	474(270/204)
3	90(52/38)	160(117/43)	1,516(1,053/463)	1,524(1,251/273)	228(171/57)	28(19/9)
4	134(90/44)	237(170/67)	5,229(3,972/1,257)	5,798(4,861/937)	107(73/34)	33(18/15)
5	14(11/3)	44(30/14)	583(392/191)	651(533/118)	13(7/6)	13(4/9)
6	0(0/0)	4(3/1)	79(50/29)	117(90/27)	6(3/3)	2(2/0)
Total	1,160(720/440)	3,067(2,490/577)	17,313(13,330/3,983)	20,985(16,448/4,573)	2,417(1,865/552)	1,229(648/576)

Table 25: Counts of mutation sizes of phased mDNMs per motif length and for all motif lengths combined. Maternal and paternal counts are given in brackets. Maternal mDNMs affect on average more bp than paternal ones and we see the ratio of maternal mDNMs increases with their size.

Stratifying by motif length revealed that maternal mDNMs affected more bp on average than paternal mDNMs at homopolymer, di- and tetranucleotide microsatellites (Table 23). Overall, paternal mDNMs occurred at microsatellites with greater RRT lengths and stratifying by motif length, we observed significant RRT length differences between the sexes at di-, tetra-, penta- and hexanucleotide microsatellites (Table 26).

Motif length (bp)	Mean RRT length(bp) (pat/mat)	*P*
1	17.0/16.9	0.12
2	39.2/37.2	**1.9 · 10^{-24}**
3	40.8/40.4	0.72
4	52.2/50.1	**5.6 · 10^{-9}**
5	54.7/51.1	**1.6 · 10^{-3}**
6	49.8/42.6	**2.2 · 10^{-2}**

Table 26 Mean RRT lengths for paternal and maternal mDNMs. The paternal mDNMs have longer RRT at di-, tetra-, penta- and hexanucleotide microsatellites. Bold represents significant difference in RRT length between paternal and maternal mDNMs ($p<0.05$).

Because of potential inaccuracy in our homopolymer genotypes, we also computed the results above without homopolymer mDNMs, yielding similar results. A higher fraction of maternal mDNMs occur at tri-, penta- and hexanucleotide microsatellites while dinucleotide microsatellites represent a larger fraction of paternal mDNMs (Table 24, Fig. 7).

The average number of bp involved without homopolymers is larger in maternal mDNMs than in paternal mDNMs (3.9 vs 3.4 bp, Mann-Whitney U test *P* = 2.6 · 10^{-23}). Considering mDNMs with motif lengths above one, the number of repeats in the reference is higher for paternal mDNMs (16.8 vs 15.7 repeats, Mann-Whitney U test $P = 2.8 \cdot 10^{-51}$).

For di- and trinucleotide microsatellites, the relative frequency of maternal and paternal mDNMs differed by the motif classes. The AG and AT motif classes were more commonly affected in maternal mDNMs, while the AC motif class was more commonly affected in paternal mDNMs (Table 27).

Motif equivalence class	% of paternal	% of maternal	Odds ratio maternal/paternal (*P*)
AAC	10.7%	7.6%	**0.68[0.52;0.90] (7.6 · 10^{-3})**
AAG	2.1%	2.7%	1.28[0.77;2.05] (0.36)
AAT	80.9%	84.2%	**1.26[1.03;1.55] (0.02)**
ACC	0.3%	0.0%	No maternal mDNMs.
ACG	0.0%	0.0%	No mDNMs
ACT	1.6%	0.8%	0.49[0.20;1.04] (7.2 · 10^{-2})
AGC	0.6%	1.4%	2.14[0.99;4.50] (5.1 · 10^{-2})
AGG	0.2%	0.2%	1.56[0.19;8.48] (0.64)
ATC	3.4%	2.1%	0.63[0.37;1.02] (7.2 · 10^{-2})
CCG	0.3%	1.0%	**3.39[1.28;9.20] (2.0 · 10^{-2})**
Motif equivalence class	% of paternal	% of maternal	Enrichment
AC	88.7%	79.8%	**0.50[0.46;0.55] (2.9 · 10^{-48})**
AG	4.6%	7.4%	**1.65[1.43;1.88] (4.5 · 10^{-12})**
AT	6.6%	12.7%	**2.05[1.84;2.29] (3.6 · 10^{-35})**
CG	0.04%	0.1%	3.04[0.88;9.76] (6.2 · 10^{-2})

Table 27 Motif equivalence class odds ratios for di- and trinucleotide microsatellites. Maternal mDNMs are more common at AAT, CCG, AG and AT motif class microsatellites while paternal mDNMs are more common at AAC and AC motif class microsatellites. Bold represents significant enrichment (p<0.05).

In trinucleotide repeats, the AAT and CCG motifs were enriched in maternal mDNMs and the AAC motif was enriched in paternal mDNMs (Table 27). The CCG enrichment in maternal events may be the result of the vulnerability of GC rich sequences to alkylation(*64*) or oxidative damage(*65*) and the long time oocytes spend in dictyate arrest before meiosis. Building on this, we considered all non-trinucleotide mDNMs in microsatellites whose motif contained only G or C and found a 2.3 (95% CI: 1.7-3.7) fold enrichment of maternal mDNMs.

Because of a known trend towards paternal expansion at an ATTCT repeat associated with spinocerebellar ataxia 10 (*SCA10*)(*66*, *67*), we looked at motifs from this class specifically to determine if the bias observed at *SCA10* is present genome-wide. The maternal contribution at pentanucleotide mDNMs was 25.9% overall, but for the ATTCT motif equivalence class, the ratio was only 18.9% (chi squared test $P = 0.037$). On average, we see an addition of 0.3 bp in paternal mDNMs at microsatellites with ATTCT motifs, compared to a loss of 3.2 bp in maternal mDNMs (Mann-Whitney U test $P = 5.0 \cdot 10^{-4}$). This indicates that the kind of paternal expansion bias observed at the *SCA10* microsatellite also affects other ATTCT microsatellites in the genome.

Parental age influences number of mDNMs

The number of mDNMs transmitted to the offspring was affected by both paternal ($P = 5.4 \cdot 10^{-176}$) and maternal ($P = 7.2 \cdot 10^{-24}$) age at birth of proband. The increase was 0.97 mDNMs per year in fathers (95% CI: 0.90-1.04) and 0.31 mDNMs per year in mothers (95% CI: 0.25-0.37), resulting in an expected number of 51.2 (35.7 paternal and 15.5 maternal) mDNMs and 77.0 (55.1 paternal and 21.8 maternal) mDNMs extrapolated genome-wide in probands with 20- and 40-year-old parents, respectively. An increase of mDNMs with paternal age is consistent with previous studies(*26*), but, to our knowledge, this is the first demonstration that mDNMs also increase with maternal age.

mDNMs at mono-, di-, tri- and tetranucleotide microsatellites increased significantly with paternal age and mDNMs at di-, tetra- and hexanucleotide repeats increase with maternal age (Table 28). An age effect is likely to be present for all motif lengths in both sexes but we lack power to detect it in the less frequent motif lengths.

Data set	Paternal age effect (95% CI)	Maternal age effect (95% CI)	#paternal/#maternal
All markers	**0.178(0.165-0.190)**	**0.058(0.047-0.069)**	**35,501/10,670**
Motif length 1	**0.013(0.006-0.020)**	0.007(-0.001-0.014)	5,141/2,948
Motif length 2	**0.078(0.069-0.087)**	**0.029(0.022-0.037)**	**17,388/4,084**
Motif length 3	**0.011(0.008-0.014)**	0.002(-0.001-0.004)	2,663/883
Motif length 4	**0.056(0.050-0.062)**	**0.013(0.008-0.017)**	**9,184/2,354**
Motif length 5	0.004(-0.003-0.011)	-0.002(-0.007-0.003)	977/341
Motif length 6	-0.008(-0.021-0.005)	**0.016(0.003-0.029)**	**148/60**

Table 28: Maternal and paternal age effect for all motif lengths and conditioning on motif length. mDNMs at mono-, di-, tri- and tetranucleotide repeats increase with paternal age while mDNMs at di- and tetra- and hexanucleotide repeats increase with maternal age. Bold represents significant association ($P < 0.05$).

The mDNM fractions of each motif length changed with age, indicating that the increase of DNMs with parental age was stronger for some motif lengths. However, the motif lengths changing were not the same for paternal and maternal mDNMs. The fraction of homopolymers decreased with parental age for both sexes, dinucleotide paternal mDNMs decreased while the fraction of tetranucleotides increased with paternal age (Fig. 7). The fraction of hexanucleotide maternal mDNMs increased with maternal age (Fig. 7). Excluding homopolymers resulted in similar results, tetranucleotide mDNMs increased their fraction with paternal age (Linear regression $P = 1.5 \cdot 10^{-6}$) and the fraction of di- and hexanucleotide mDNMs increases with maternal age.

Parental genotypes associate with number of mDNMs in offspring

We constructed parental phenotypes, using counts of mDNMs originating from each parent, and performed a genome-wide association study (GWAS). We considered the mDNM counts per parent with and without homopolymer mDNMs and corrected for age, sex, sequencing method and sample type.

Three correlated SNPs rs4987188, rs112587140 and rs113983130 (Table 29) were associated with an increase in mDNMs for all motif lengths.

Marker 1	Marker 2	Conditional *P* 1-2	Conditional *P* 2-1	R^2
chr2:47416318	chr2:47494068	$8.9 \cdot 10^{-2}$	$5.1 \cdot 10^{-2}$	0.74
chr2:47416318	chr2:47491330:0	$3.6 \cdot 10^{-2}$	$2.3 \cdot 10^{-2}$	0.60
chr2:47416318	chr2:47491330:1	$3.6 \cdot 10^{-2}$	$2.3 \cdot 10^{-2}$	0.60
chr2:47494068	chr2:47491330:0	0.17	0.18	0.80
chr2:47494068	chr2:47491330:1	0.17	0.18	0.80
chr2:47491330:0	chr2:47491330:1	1.00	1.00	1.00

Table 29 Conditional association p-values for markers significantly associating with increased mDNM rate (*MSH2* missense marker in bold). All markers are correlated but a residual signal remains after conditioning the intergenic signal at chr2:47491330 for the missense marker

Using sequence annotation weighted Bonferroni corrected significance(*68*), we selected rs4987188[A], a missense variant (p.Gly322Asp) with a 1.9% allele frequency in the *MSH2* gene, a mismatch repair gene(*69*, *70*), as the lead marker for the association (Fig. 8, Table 30).

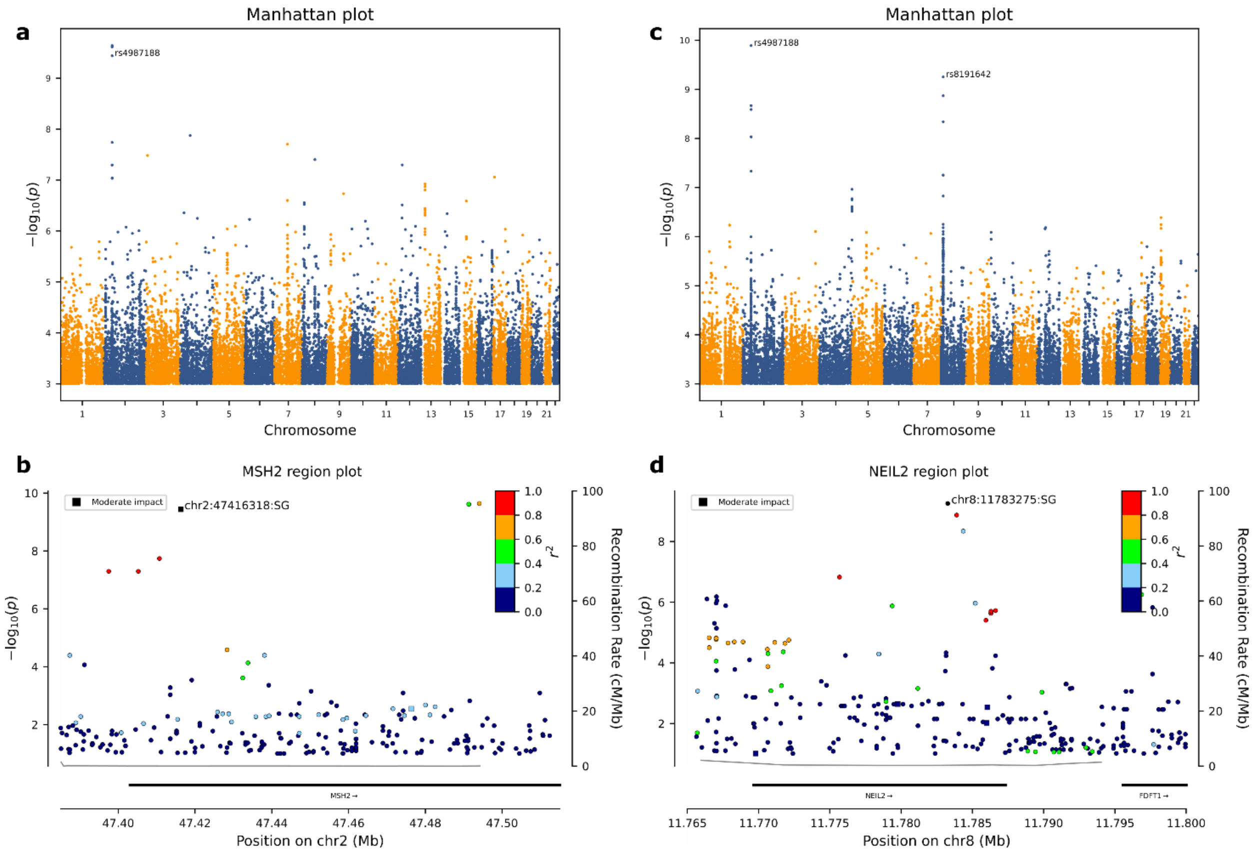

Fig. 8|Genome-wide association a) Manhattan plot showing a missense variant in the MSH2 gene associating with an increased number of mDNMs transmitted to offspring. b) Region plot for the missense variant in MSH2. Two correlated markers also reach genome-wide significance. c) Manhattan plot showing a synonymous variant in the NEIL2 gene which associates with an increased number of mDNMs with motif lengths greater than one bp transmitted to offspring. The plot also shows the p-value for the MSH2 missense variant considering mDNMs with motif lengths greater than one bp. d) Region plot for a synonymous variant in the NEIL2 gene. No other markers at the locus reach genome wide significance.

Chromosomes on Manhattan plots are marked with alternating colors and the threshold for genome wide significance was set as 1 · 10^{-9}.

Population	Allele frequency of rs4987188
Iceland	1.9%
UKB	1.5%
Fingen	3.0%
Africa	0.2%
Asia	0.0.%

Table 30 Allele frequency of rs4987188 in different populations, it is most frequent in the Finnish population and not found in Asians.

Each copy of rs4987188[A] is associated (P = 3.6 · 10^{-10}) with a 0.37 s.d. (95% CI: 0.26-0.48) increase in the number of transmitted mDNMs, corresponding to 13.1 and 7.8 mDNMs genome-wide per paternal and maternal copy, respectively. We tested each parental sex separately and rs4987188 was not genome-wide significant in either one and no significant difference in effect was found in the increase between the two parental sexes (P = 0.14). However, after correcting for parental age and carrier status we see a nominal association (P = 0.019) between paternal age and mutation count in offspring of male carriers, suggesting a possible interaction between paternal age and the effect of rs4987188[A].

The protein encoded by *MSH2* forms two heterodimeric mismatch repair complexes. One predominantly required for repairing mismatched bp and the other mainly responsible for repairing insertion/deletion loops between one and twelve bp(*71*, *72*). Multiple variants in *MSH2* have been reported to cause Lynch syndrome (also known as hereditary non-polyposis colorectal cancer, HNPCC), which results in increased risk of endometrial, colorectal and other cancers(*73*). We tested rs4987188[A] for association with an increased risk of endometrial and colorectal cancer in both the Icelandic and combined meta-analysis data sets available at deCODE genetics, but found no such association. Functional studies of the yeast homologue of rs4987188 indicate that it results in a modest decrease in mismatch repair efficiency(*74*, *75*).

Homopolymers mDNM have a higher false positive rate than other mDNMs categories, to assess the robustness of our results, we reran the GWAS on mDNM counts per parent without homopolymer mDNMs. We recover our *MSH2* and p.Gly322Asp association with a similar effect 0.39 s.d. (P = 1.3 · 10^{-10}). Furthermore, we find a novel association between mDNM counts without homopolymers and two correlated (r^2=0.988) SNPs, rs8191642 (P =5.6 · 10^{-10}) a synonymous variant (Pro188), and rs8191649 (1.4 · 10^{-9}), an intronic variant, in *NEIL2*, a glycolase involved in both transcription and replication associated base excision repair of DNA damaged by oxidation or by mutagenic agents(*76*). Using sequence annotation weighted Bonferroni corrected significance(*68*), we selected rs8191642[G], which has a 20.3% frequency as the lead marker for the association.

Testing rs8191642[G] on each parent separately revealed heterogeneity between the sexes (*P* = $1.2 \cdot 10^{-3}$). Each copy of rs8191642[G] is associated with (*P* = $7.2 \cdot 10^{-12}$) a 0.21 s.d. (95% CI: 0.15-0.27) increase in the number of paternally transmitted mDNMs with motif lengths greater than one bp, corresponding to 4.4 mDNMs genome-wide per copy, but it did not associate with the number of maternally transmitted mDNMs (*P* = 0.0149, effect=0.072).

Somatic and germline mutations in mismatch-repair genes are known to cause microsatellite instability and hypermutator phenotypes in tumors of the colon and endometrium. Sequencing of large cohorts of tumors has revealed several mutational signatures associated with mismatch repair deficiency(*77*).

These signatures show a strong correlation with somatic microsatellite instability. Mutations in components of the base-excision repair pathway, most notably *NTHL1* and *MUTYH,* also cause distinct mutational signatures in tumors with these mutations (COSMIC signatures SBS30 and SBS36(*77*), respectively). Given the effects of rs4987188[A] and rs8191642[G] on the rate of mDNMs, we were interested in knowing if they also affect the mutational spectra of single-base-substitution DNMs, for example by altering the MMR/BER pathways in the testis.

Using only phased DNMs, we compared the 96-class trinucleotide spectra of DNMs transmitted from carrier and non-carrier mothers and fathers. All spectra were highly similar (pairwise cosine similarities >0.95) and showed no hint of MMR deficient related mutational signatures (Fig. 9). This suggests that the effects of rs4987188[A] and rs8191642[G] are confined to microsatellite sites and they do not otherwise affect the fidelity of the mismatch or base-excision repair pathways.

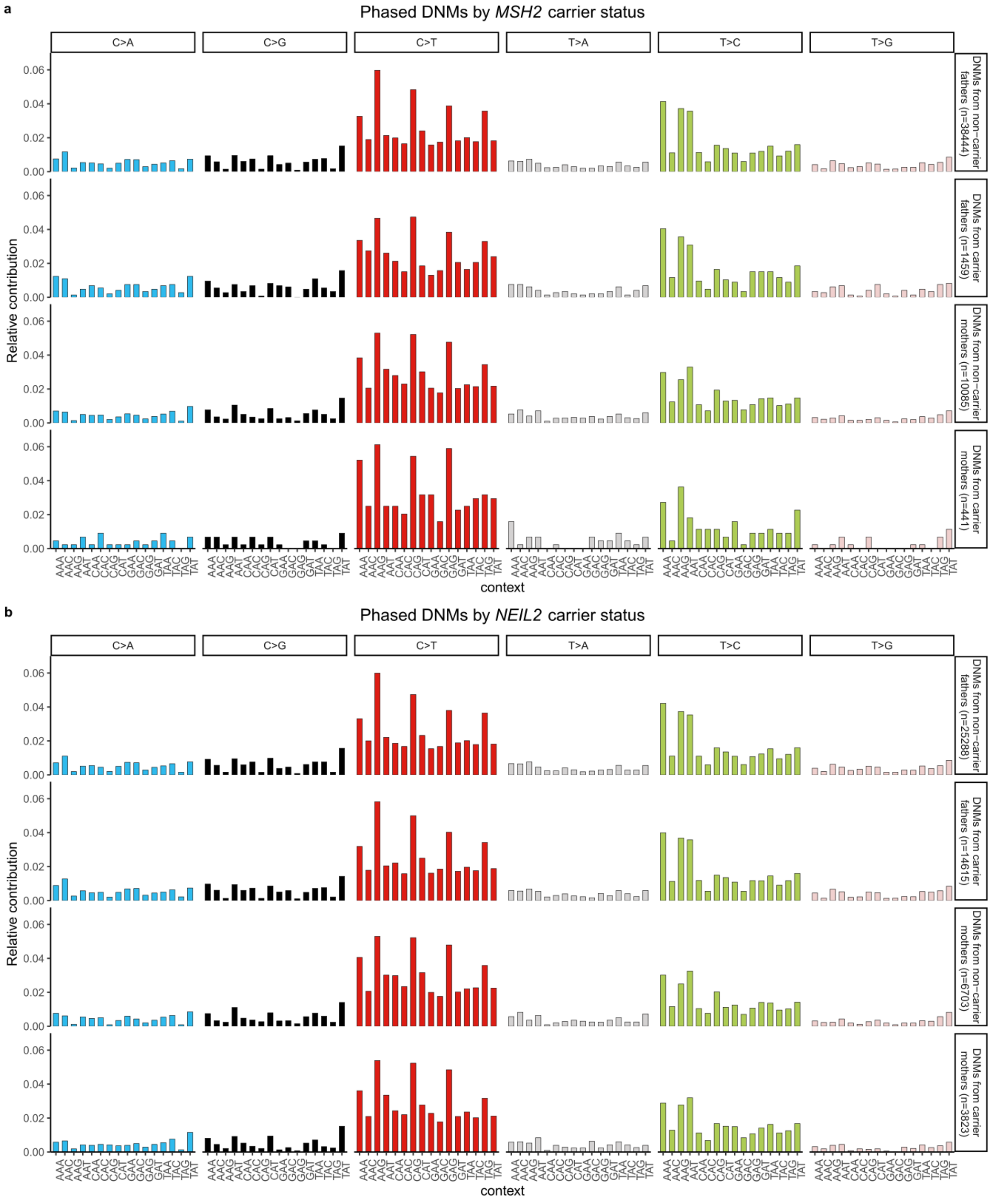

Fig. 9 | sDNM mutation signatures for offspring of paternal and maternal carriers and non-carriers of both rs4987188 and rs8191642. The signatures are not different in any of the four comparisons so we assume that the variants do not affect the generation of sDNMs.

Methods

Identification of tandem repeats

To generate a set of STR locations we ran version 4.09 of Tandem repeats finder(*40*) on the GRCh38(*86*) human reference genome with the following parameters: ./trf409.linux64 genome.fa 2 7 7 80 10 22 100 –d –h –ngs > trf_out_100

Purity computation

We define the ratio between the observed repeat units in a STR and the number of expected repeat units in a perfectly pure STR as the repeat purity.

$$purity = \frac{observed\ repeats}{E[repeats \mid 100\%\ pure]} \qquad \text{Eq. 1}$$

Expected heterozygosity computation

We compute expected heterozygosity using the following formula(*41*)

$$gene\ diversity = \left[\frac{n}{(n-1)}\right]\left(1 - \sum_{i=1}^{k} p_i^2\right) \qquad \text{Eq. 2}$$

where *n* is the total number of sequences, *k* is the number of alleles at the marker and p_i is the frequency of each allele.

Genotype and marker filtering

For a trio comparison we require all members of the trio to have a phred scaled genotype quality value higher than 60. We removed microsatellites which imputed to a 0% frequency and ones that failed our best practices filters. These filters consider coverage, genotype quality, number of samples genotyped and fraction of reads not supporting the reported genotype and are described in(*78*).

We further removed microsatellites outside the Tier 1 regions defined by GIAB(*87*) and microsatellites lying within CNV regions annotated by CNVnator(*88*).

After using long range phased SNP genotypes to phase and impute our microsatellite genotypes into the Icelandic population(*89*), microsatellites with alleles showing a strong deviation from the Hardy Weinberg equilibrium ($P < 1 \cdot 10^{-6}$) or a frequency weighted imputation information lower than 0.9 were removed(*90*). We also required the frequency weighted phasing information to be greater than 0.6 and the r^2 for leave one out cross validation of phased genotypes to be greater than 0.5.

We looked for probands with homozygous mDNMs less than 1Mbp apart, implying a haploid state, and checked them for large CNVs covering the region, causing autozygosity and spurious mDNM calls. Last, we crossed all mDNMs with deletions and duplications imputed into the proband(*91*) and removed the ones intersecting.

De novo detection

For marker-trio combinations where all conditions described above are met, we define a mDNM as a trio genotype combination satisfying neither of the following logical statements:

(1) $proband_{A1} \in M \; and \; proband_{A2} \in F$

(2) $proband_{A1} \in F \; and \; proband_{A2} \in M$

where $proband_{A1}$ and $proband_{A2}$ refer to the two alleles carried by the proband and M and F define the allele pairs carried by the mother and father, respectively.

Microsatellite attribute regression on mDNM rate

We performed a Poisson regression using the number of available markers per trio as an offset on the full data set and stratified by motif length to examine if the effects of other attributes on the mDNM rate remained consistent across motif lengths.

The direction and statistical significance for both RRT length and repeat purity remain consistent for all motif lengths but the effect of GC content in repeat motif is positive for homopolymer repeats (Table 12).

Obtaining confidence intervals for mDNM rate estimates

We used the boot package for R(*92*, *93*) to obtain confidence intervals for both our genome wide mDNM rate estimate and the motif length specific estimates. 100 replicates were used in all cases and 95% confidence intervals extracted using the resulting quantiles.

Extrapolation from mDNM rate to expected number of mDNMs

The per motif mDNM rate was determined from our set of high-quality microsatellite calls. We estimated the expected number of per motif mDNMs among all microsatellites by multiplying the mDNM rate and the number of microsatellites per motif length. The total expected mDNMs is the sum over all motif lengths

$$E[mDNMs] = \sum_{i=1}^{6} r_i \cdot n_i$$

where r_i is the mDNM rate at motif length *i* and n_i is the genome-wide number of microsatellites with motif length *i*.

Construction of mDNM phenotypes

We used maternal and paternal counts at each proband to construct phenotypes for the parents quantifying the number of mDNMs passed on to the proband. In addition to correcting for the parental gender (mother/father), we corrected for parental age at birth of proband, sequencing method, sample type and number of microsatellites available for *de novo* detection in the trio. For parents with more than one proband in our trio set, we used the average of all their offspring. After regressing out the coefficients, we used rank inverse normalization to normalize the phenotype.

Microsatellite mDNM phasing

We determined the parent of origin of mDNMs, using three distinct methods; read pair tracing, allele sharing and haplotype sharing in three-generation families. First, we used long range phased(*89*) SNP and indel genotypes available at deCODE genetics(*89*) to phase reads reporting mDNMs when possible. Read phases enabled us to assign a parent of origin to mDNMs according to the phase of the reads reporting the event since a read phased to one parent and reporting a mDNM indicates that the mDNM was transmitted from that parent. Not all mDNMs had supporting reads containing phased markers and for these, assigning a parent of origin was not possible (Fig. 10)

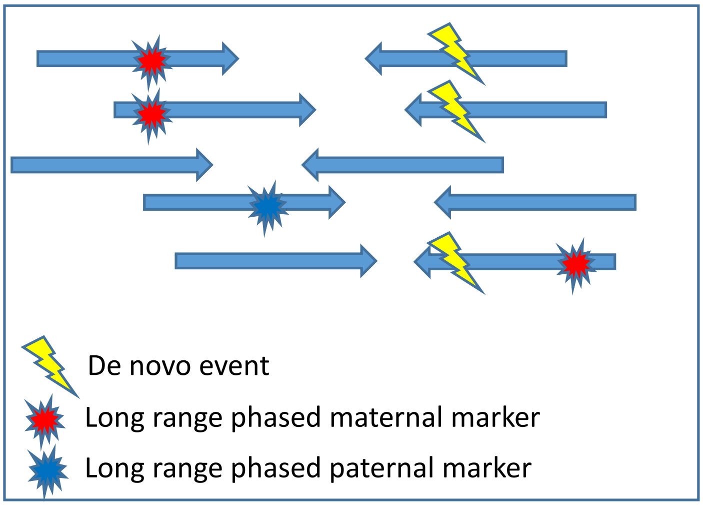
De novo event
Long range phased maternal marker
Long range phased paternal marker

Fig. 10 | Read pair phasing. We use read pairs which contain a long range phased marker and report an mDNM. Reads with long range phased markers covering the de novo site but supporting the other allele can also further give information on the parent of origin. An example of a maternal mDNM where three read pairs report a long range phased maternal marker and a de novo allele, one pair is not informative and one contains a long range phased paternal marker and not the de novo allele.

Second, for trios where the de novo allele was seen in neither parent and the other proband allele was only seen in one parent we phased the mDNM to the other parent (Fig. 11).

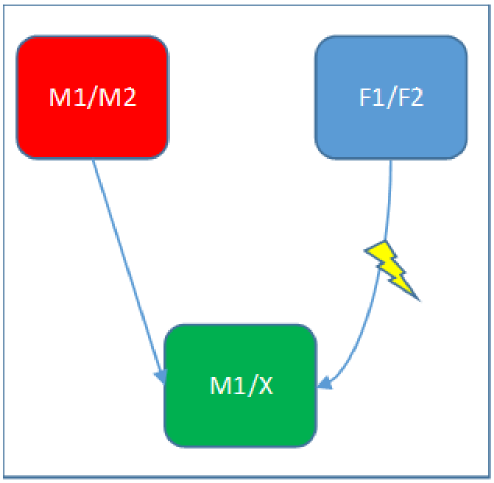

Fig. 11 | Allele based phasing. If the de novo allele is present in neither parent and the other allele is present only in one parent we phase the de novo event to the other parent. Here we phase the de novo event to the father since a maternal allele is seen in the proband but neither of the paternal ones.

Last, we used haplotype sharing in three generation families such that if the mDNM was transmitted from proband to its offspring, we phased the de novo to the parent sharing a haplotype with the offspring, and to the parent not sharing a haplotype if de novo was not transmitted (Fig. 12).

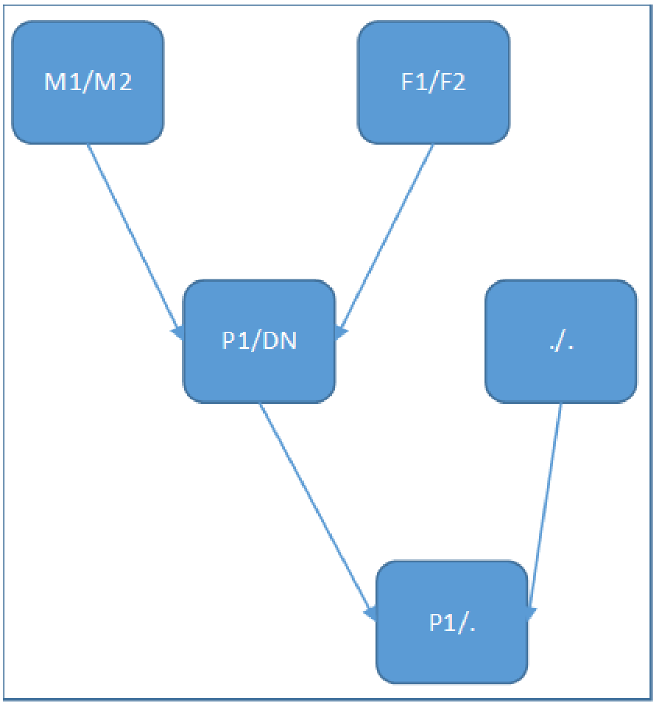

Fig. 12 | Haplotype sharing in three generation families. If the de novo is transmitted from the proband to its offspring we phase the de novo to the parent sharing a haplotype with the offspring, and to the parent not sharing a haplotype if de novo is not transmitted.

Parental age effect regression

We applied a Poisson regression model described in(*49*) on our data with the number of available markers in each trio as a covariate to determine the effect of parental age at childbirth on the number of transmitted mDNMs.

To determine whether the coefficients were robust to the phasing method, we repeated the analysis(*49*) and split the phased markers by phasing method and performed the regression on each subset. Both maternal and paternal age remained significant in all subsets (Table 31).

Phasing method	Paternal effect	*P*	Maternal effect	*P*
Allele based	0.17 [0.15;0.18]	$1.2 \cdot 10^{-102}$	0.07 [0.05;0.08]	$5.1 \cdot 10^{-16}$
Read back	0.16 [0.15;0.18]	$3.6 \cdot 10^{-101}$	0.07 [0.06;0.09]	$7.1 \cdot 10^{-21}$
3 generation	0.18 [0.15;0.21]	$9.8 \cdot 10^{-34}$	0.05 [0.03;0.07]	$1.7 \cdot 10^{-6}$

Table 31 Poisson regression coefficients and p-values for subsets of phased markers by each method applied. The effects for allele based phasing and read back phasing match but the three generation phasing subset gives a stronger paternal effect and a weaker maternal one.

We compute the total predicted number of de novo mutations in a proband with an X year old father and a Y year old mother from the coefficients from our regression model:

$$(I_F + \beta_F \cdot X + I_M + \beta_M \cdot Y)/(\frac{\mu_{markers}}{1{,}394{,}292}) \qquad \text{Eq. 3}$$

Where I_F is the paternal intercept, β_F is the paternal age effect, I_M is the maternal intercept, β_M is the maternal age effect, $\mu_{markers}$ is the mean value of available markers across our trios and 1,394,292 is the total number of microsatellites we detect.

Verification

PacBio HiFi data was available for four trios in our set. We were unable to verify homopolymer mDNMs as the PacBio sequencing error rate was too high(*94*) but we were able to verify the existence of 27 mDNMs with motif length >1. Out of these, 26 were true positives and one was a false positive at a dinucleotide repeat, giving an expected false positive rate of 3.7% for motif lengths greater than 1.

For mDNMs observed in probands with a monozygotic twin also present in our set, we checked whether the mDNM genotypes were concordant between both twins. We compared mDNM calls where the genotype was present in the monozygotic twin of the proband. We treated genotype calls of the other twin present if the genotype quality was higher than or equal to 30, which is half of the value we require for trio mDNM detection. Out of the 230 comparable MZ-twin mDNMs, 217 were found in both twins and 13 were discordant (Table 32).

Motif length (bp)	Shared	Not shared	Error rate
1	47	6	11.3%
2	108	6	5.3%
3	16	1	5.9%
4	44	0	0%
5	2	0	0%
6	0	0	0%

Table 32 Sharing of mDNMs between monozygotic twins stratified on motif length.

Discussion

We generated two large microsatellite genotype sets. The microsatellite genotypes for the UKB samples are publicly available, they have been tested for association with various phenotypes and we have previously reported associations with disease (repeat expansions at *DMPK* and *CACNA1A*)(*78*). It is likely that these genotypes will be a valuable resource for further mining of associations with phenotypes.

In the Icelandic set, we identified 76,987 mDNMs and found a correlation between the number of mDNMs and age at birth of both mothers and fathers and parental genotypes. We observed a previously unreported increase in the number of maternal mDNMs transmitted to offspring with maternal age, consistent with the increase of both SNP/indel DNMs(*49*) and recombination with maternal age(*34*, *79*). mDNMs are often associated with replication adducts(*23*, *24*, *80–83*), however the maternal effect gives novel insight into the formation of mDNMs, as oocytes are in dictyate arrest compared to the actively dividing spermatogonia in aging fathers. The maternal effect indicates that, mDNMs can also occur outside of DNA synthesis during S-phase replication since DNA polymerases operate during most types of DNA repair on long tracts and in homologous recombination pathways(*23*).

The observation of different frequencies of mDNM motif classes transmitted by older mothers and fathers allows us to deconvolve the mutational processes acting in the germlines. Since spermatogonia undergo a greater number of mitotic cell divisions than oocytes, we expect the enrichment of AC motif class mutations observed in paternal mDNMs to be due to out-of-register realignments during replication, resulting in mDNMs(*82*), whereas the maternal enrichment of mDNMs at pure GC repeats is likely to be a result of damage accumulated during the dictyate arrest of oocytes.

In vitro study of how repeat motifs affect the frequency of polymerase slippage during replication reported that motifs less likely to stall replication were more likely to mutate during replication. Of the dinucleotide repeat classes, microsatellites from the AC class had the lowest replication stall affinity(*82*). The higher mDNM rate of microsatellites in the AAT motif class could be a result of its low replication stall affinity(*82*). The two hydrogen bonds between A/T base-pairs compared to the three between G/C base-pairs also makes A/T pairs more likely to disassociate from each other, enabling the formation of secondary structures and possible mDNMs. Finally, repeats with a high A/T-content also have a sequence composition similar to elements involved in DNA unwinding at replication origins(*84*) during mitosis. These repeats could therefore function as aberrant replication origins and cause a higher mDNM rate during replication in S phase(*84*).

Sequence variants detected in humans that decrease sequence stability have for the most part been very deleterious and under strong negative selection(*85*). Interestingly, the variants presented here are both present in high frequencies and have large effects, thereby indicating that an increase in the mDNM rate across the genome is not sufficiently deleterious to sieve out the mutators at *MSH2* and *NEIL2*. A missense variant, rs4987188, in *MSH2* is associated with the number of mDNMs transmitted from parent to offspring with no significant difference in effect between mothers and fathers. The similar effect of rs4987188 across the sexes, indicates that gametes from both sexes are subject to the same sequence fidelity maintenance process. A synonymous variant, rs8191642, in *NEIL2* is associated with the number of mDNMs transmitted from fathers to their offspring. *NEIL2* has been reported to function in both transcription and replication coupled repair(*76*). Thus, it is likely that the association between the variant in *NEIL2* and the number of paternally transmitted mDNMs is due to the more frequent replication of spermatogonia.

We have identified the first germline variants, segregating at high frequencies, that directly affect the mDNM rate in humans. We have also demonstrated for the first time that the number of maternally transmitted mDNMs increases with maternal age. Last, we have generated a publicly available microsatellite genotype set for 150,119 samples, a valuable resource for the scientific community in its efforts to better understand and define the many ways that microsatellites affect human phenotypes.

Bibliography

1. J. X. Sun, A. Helgason, G. Masson, S. S. Ebenesersdóttir, H. Li, S. Mallick, S. Gnerre, N. Patterson, A. Kong, D. Reich, K. Stefansson, A direct characterization of human mutation based on microsatellites. *Nat. Genet.* (2012), doi:10.1038/ng.2398.

2. M. W. Nachman, S. L. Crowell, Estimate of the mutation rate per nucleotide in humans. *Genetics*. **156** (2000), doi:10.1093/genetics/156.1.297.

3. H. Ellegren, Microsatellite mutations in the germline: Implications for evolutionary inference. *Trends Genet.* (2000), doi:10.1016/S0168-9525(00)02139-9.

4. B. Walsh, Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. *Genetics*. **158** (2001), doi:10.1093/genetics/158.2.897.

5. K. S. Kim, T. W. Sappington, Microsatellite data analysis for population genetics. *Methods Mol. Biol.* **1006** (2013), doi:10.1007/978-1-62703-389-3_19.

6. A. I. Putman, I. Carbone, Challenges in analysis and interpretation of microsatellite data for population genetic studies. *Ecol. Evol.* **4** (2014), , doi:10.1002/ece3.1305.

7. J. R. Brouwer, R. Willemsen, B. A. Oostra, Microsatellite repeat instability and neurological disease. *BioEssays*. **31** (2009), , doi:10.1002/bies.080122.

8. Y. D. Kelkar, S. Tyekucheva, F. Chiaromonte, K. D. Makova, The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res.* **18** (2008), doi:10.1101/gr.7113408.

9. U. Polak, E. Mcivor, S. Y. R. Dent, R. D. Wells, M. Napierala, Expanded complexity of unstable repeat diseases. *BioFactors*. **39** (2013), , doi:10.1002/biof.1060.

10. S. N. Shah, S. E. Hile, K. A. Eckert, Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. *Cancer Res.* **70** (2010), , doi:10.1158/0008-5472.CAN-09-3049.

11. C. D. Campbell, E. E. Eichler, Properties and rates of germline mutations in humans. *Trends Genet.* **29** (2013), , doi:10.1016/j.tig.2013.04.005.

12. E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. Fitzhugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. Levine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, M. L. Hong, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P.

Abola, M. J. Proctor, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. De La Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, A. Patrinos, M. J. Morgan, Initial sequencing and analysis of the human genome. *Nature*. **409** (2001), doi:10.1038/35057062.

13. T. Willems, M. Gymrek, G. D. Poznik, C. Tyler-Smith, Y. Erlich, Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates. *Am. J. Hum. Genet.* **98** (2016), doi:10.1016/j.ajhg.2016.04.001.

14. H. Ellegren, Microsatellites: Simple sequences with complex evolution. *Nat. Rev. Genet.* **5** (2004), , doi:10.1038/nrg1348.

15. T. LaFramboise, Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic Acids Res.* **37** (2009), , doi:10.1093/nar/gkp552.

16. T. H. Massey, L. Jones, The central role of DNA damage and repair in CAG repeat diseases. *DMM Dis. Model. Mech.* **11** (2018), , doi:10.1242/dmm.031930.

17. K. J. Rohilla, K. N. Ovington, A. A. Pater, M. Barton, A. J. Henke, K. T. Gagnon, Systematic microsatellite repeat expansion cloning and validation. *Hum. Genet.* **139** (2020), doi:10.1007/s00439-020-02165-z.

18. E. L. Van Der Ende, J. L. Jackson, A. White, H. Seelaar, M. Van Blitterswijk, J. C. Van Swieten, Unravelling the clinical spectrum and the role of repeat length in C9ORF72 repeat expansions. *J. Neurol. Neurosurg. Psychiatry*. **92** (2021), , doi:10.1136/jnnp-2020-325377.

19. D. Y. Lee, C. T. McMurray, Trinucleotide expansion in disease: Why is there a length threshold? *Curr. Opin. Genet. Dev.* (2014), , doi:10.1016/j.gde.2014.07.003.

20. M. J. Salcedo-Arellano, R. J. Hagerman, V. Martínez-Cerdeño, Fragile x syndrome: Clinical presentation, pathology and treatment. *Gac. Med. Mex.* **156** (2020), doi:10.24875/GMM.M19000322.

21. A. Kumar, S. Agarwal, D. Agarwal, S. R. Phadke, Myotonic dystrophy type 1 (DM1): A triplet repeat expansion disorder. *Gene*. **522** (2013), , doi:10.1016/j.gene.2013.03.059.

22. R. A. M. Buijsen, L. J. A. Toonen, S. L. Gardiner, W. M. C. van Roon-Mom, Genetics, Mechanisms, and Therapeutic Progress in Polyglutamine Spinocerebellar Ataxias. *Neurotherapeutics*. **16** (2019), , doi:10.1007/s13311-018-00696-y.

23. H. Ellegren, Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat. Genet.* **24**, 400–402 (2000).

24. K. A. Eckert, S. E. Hile, Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Mol. Carcinog.* **48** (2009), doi:10.1002/mc.20499.

25. T. V. Nikitina, S. A. Nazarenko, Human microsatellites: Mutation and evolution. *Russ. J. Genet.* **40** (2004), , doi:10.1023/B:RUGE.0000044750.21421.65.

26. I. Mitra, B. Huang, N. Mousavi, N. Ma, M. Lamkin, R. Yanicky, S. Shleizer-Burko, K. E. Lohmueller, M. Gymrek, Patterns of de novo tandem repeat mutations and their role in autism. *Nature*. **589** (2021), doi:10.1038/s41586-020-03078-7.

27. N. J. Haradhvala, J. Kim, Y. E. Maruvka, P. Polak, D. Rosebrock, D. Livitz, J. M. Hess, I. Leshchiner, A. Kamburov, K. W. Mouw, M. S. Lawrence, G. Getz, Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat. Commun.* **9** (2018), doi:10.1038/s41467-018-04002-4.

28. G. M. Li, Mechanisms and functions of DNA mismatch repair. *Cell Res.* **18** (2008), , doi:10.1038/cr.2007.115.

29. R. J. Hause, C. C. Pritchard, J. Shendure, S. J. Salipante, Classification and characterization of microsatellite instability across 18 cancer types. *Nat. Med.* **22** (2016), doi:10.1038/nm.4191.

30. A. Minelli, E. Maserati, G. Giudici, S. Tosi, C. Olivieri, L. Bonvini, P. De Filippi, A. Biondi, F. Lo Curto, F. Pasquali, C. Danesino, Familial partial monosomy 7 and myelodysplasia. *Cancer Genet. Cytogenet.* **124**, 147–151 (2001).

31. J. Kaplanis, B. Ide, R. Sanghvi, M. Neville, P. Danecek, T. Coorens, E. Prigmore, P. Short, G. Gallone, J. McRae, L. Moutsianas, C. Odhams, J. Carmichael, A. Barnicoat, H. Firth, P. O'Brien, R. Rahbari, M. Hurles, G. E. R. Consortium, Genetic and chemotherapeutic influences on germline hypermutation. *Nature* (2022), doi:10.1038/s41586-022-04712-2.

32. S. B. Leclercq, E. Rivals, P. Jarne, DNA slippage occurs at microsatellite loci without minimal threshold length in humans: A comparative genomic approach. *Genome Biol. Evol.* **2** (2010), doi:10.1093/gbe/evq023.

33. N. Chatterjee, G. C. Walker, Mechanisms of DNA damage, repair, and mutagenesis. *Environ. Mol. Mutagen.* **58** (2017), , doi:10.1002/em.22087.

34. B. V. Halldorsson, G. Palsson, O. A. Stefansson, H. Jonsson, M. T. Hardarson, H. P. Eggertsson, B. Gunnarsson, A. Oddsson, G. H. Halldorsson, F. Zink, S. A. Gudjonsson, M. L. Frigge, G. Thorleifsson, A. Sigurdsson, S. N. Stacey, P. Sulem, G. Masson, A. Helgason, D. F. Gudbjartsson, U. Thorsteinsdottir, K. Stefansson, Human genetics: Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science (80-.).* **363** (2019), doi:10.1126/science.aau1043.

35. J. K. Collins, K. T. Jones, DNA damage responses in mammalian oocytes. *Reproduction*. **152** (2016), , doi:10.1530/REP-16-0069.

36. P. S. Robinson, T. H. H. Coorens, C. Palles, E. Mitchell, F. Abascal, S. Olafsson, B. C. H. Lee, A. R. J. Lawson, H. Lee-Six, L. Moore, M. A. Sanders, J. Hewinson, L. Martin, C. M. A. Pinna, S. Galavotti, R. Rahbari, P. J. Campbell, I. Martincorena, I. Tomlinson, M. R. Stratton, Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nat. Genet.* **53** (2021), doi:10.1038/s41588-021-00930-y.

37. G. I. Lang, L. Parsons, A. E. Gammie, Mutation rates, spectra, and genome-wide distribution of spontaneous mutations in mismatch repair deficient yeast. *G3 (Bethesda).* **3** (2013), doi:10.1534/g3.113.006429.

38. T. A. Sasani, D. G. Ashbrook, A. C. Beichman, L. Lu, A. A. Palmer, R. W. Williams, J. K. Pritchard, K. Harris, *bioRxiv*, in press, doi:10.1101/2021.03.12.435196.

39. S. Kristmundsdóttir, B. D. Sigurpálsdóttir, B. Kehr, B. V. Halldórsson, popSTR: population-scale detection of STR variants. *Bioinformatics* (2017), doi:10.1093/bioinformatics/btw568.

40. G. Benson, Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27** (1999), doi:10.1093/nar/27.2.573.

41. A. M. Harris, M. DeGiorgio, An unbiased estimator of gene diversity with improved variance for samples containing related and inbred individuals of any ploidy. *G3 Genes, Genomes, Genet.* **7** (2017), doi:10.1534/g3.116.037168.

42. S. S. Arcot, Z. Wang, J. L. Weber, P. L. Deininger, M. A. Batzer, Alu repeats: A source for the genesis of primate microsatellites. *Genomics*. **29** (1995), doi:10.1006/geno.1995.1224.

43. F. Grandi, W. An, Non-LTR retrotransposons and microsatellites: Partners in genomic variation. *Mob. Genet. Elements*. **3** (2013), doi:10.4161/mge.25674.

44. K. D. Robertson, P. A. Jones, DNA methylation: Past, present and future directions. *Carcinogenesis*. **21** (2000), , doi:10.1093/carcin/21.3.461.

45. E. Buschiazzo, N. J. Gemmell, The rise, fall and renaissance of microsatellites in eukaryotic genomes. *BioEssays*. **28** (2006), , doi:10.1002/bies.20470.

46. M. Legendre, N. Pochet, T. Pak, K. J. Verstrepen, Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res.* **17** (2007), doi:10.1101/gr.6554007.

47. R. Sainudiin, R. T. Durrett, C. F. Aquadro, R. Nielsen, Microsatellite mutation models: Insights from a comparison of humans and chimpanzees. *Genetics*. **168** (2004), doi:10.1534/genetics.103.022665.

48. H. Jonsson, E. Magnusdottir, H. P. Eggertsson, O. A. Stefansson, G. A. Arnadottir, O. Eiriksson, F. Zink, E. A. Helgason, I. Jonsdottir, A. Gylfason, A. Jonasdottir, A. Jonasdottir, D. Beyter, T. Steingrimsdottir, G. L. Norddahl, O. T. Magnusson, G. Masson, B. V. Halldorsson, U. Thorsteinsdottir, A. Helgason, P. Sulem, D. F. Gudbjartsson, K. Stefansson, Differences between germline genomes of monozygotic twins. *Nat. Genet.* **53** (2021), doi:10.1038/s41588-020-00755-1.

49. H. Jónsson, P. Sulem, B. Kehr, S. Kristmundsdottir, F. Zink, E. Hjartarson, M. T. Hardarson, K. E. Hjorleifsson, H. P. Eggertsson, S. A. Gudjonsson, L. D. Ward, G. A. Arnadottir, E. A. Helgason, H. Helgason, A. Gylfason, A. Jonasdottir, A. Jonasdottir, T. Rafnar, M. Frigge, S. N. Stacey, O. Th Magnusson, U. Thorsteinsdottir, G. Masson, A. Kong, B. V. Halldorsson, A. Helgason, D. F. Gudbjartsson, K. Stefansson, Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* (2017), doi:10.1038/nature24018.

50. L. A. Bergeron, S. Besenbacher, T. N. Turner, C. J. Versoza, R. J. Wang, A. L. Price, E. Armstrong, M. Riera, J. Carlson, H. Y. Chen, M. W. Hahn, K. Harris, A. S. Kleppe, E. H. López-Nandam, P. Moorjani, S. P. Pfeifer, G. P. Tiley, A. D. Yoder, G. Zhang, M. H. Schierup, The mutationathon highlights the importance of reaching standardization in estimates of pedigree-based germline mutation rates. *Elife*. **11** (2022), doi:10.7554/eLife.73577.

51. H. Fan, J. Y. Chu, A Brief Review of Short Tandem Repeat Mutation. *Genomics. Proteomics Bioinformatics*. **5**, 7–14 (2007).

52. Y. Lai, F. Sun, The Relationship between Microsatellite Slippage Mutation Rate and the Number of Repeat Units. *Mol. Biol. Evol.* **20** (2003), doi:10.1093/molbev/msg228.

53. B. Brinkmann, M. Klintschar, F. Neuhuber, J. Hühne, B. Rolf, Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat. *Am. J. Hum. Genet.* (2002), doi:10.1086/301869.

54. X. Xu, M. Peng, Z. Fang, X. Xu, The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.* **24** (2000), doi:10.1038/74238.

55. Q. Y. Huang, F. H. Xu, H. Shen, H. Y. Deng, Y. J. Liu, Y. Z. Liu, J. L. Li, R. R. Recker, H. W. Deng, Mutation patterns at dinucleotide microsatellite loci in humans. *Am. J. Hum. Genet.* **70** (2002), doi:10.1086/338997.

56. M. G. Gardner, C. M. Bull, S. J. B. Cooper, G. A. Duffield, Microsatellite mutations in litters of the Australian lizard Egernia stokesii. *J. Evol. Biol.* **13** (2000), doi:10.1046/j.1420-9101.2000.00189.x.

57. A. G. Jones, G. Rosenqvist, A. Berglund, J. C. Avise, Clustered microsatellite mutations in the pipefish Syngnathus typhle. *Genetics*. **152** (1999), doi:10.1093/genetics/152.3.1057.

58. C. R. Primmer, N. Saino, A. P. Møller, H. Ellegren, Unraveling the processes of microsatellite evolution through analysis of germ line mutations in barn swallows Hirundo rustica. *Mol. Biol. Evol.* **15** (1998), doi:10.1093/oxfordjournals.molbev.a026003.

59. B. Harr, C. Schlötterer, Long microsatellite alleles in Drosaphila melanogaster have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. *Genetics*. **155** (2000), doi:10.1093/genetics/155.3.1213.

60. S. M. Udupa, M. Baum, High mutation rate and mutational bias at (TAA)n microsatellite loci in chickpea (Cicer arietinum L.). *Mol. Genet. Genomics*. **265** (2001), doi:10.1007/s004380100508.

61. S. Kruglyak, R. Durrett, M. D. Schug, C. F. Aquadro, Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. *Mol. Biol. Evol.* **17** (2000), doi:10.1093/oxfordjournals.molbev.a026404.

62. F. Supek, B. Lehner, Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature*. **521** (2015), doi:10.1038/nature14173.

63. A. Koren, P. Polak, J. Nemesh, J. J. Michaelson, J. Sebat, S. R. Sunyaev, S. A. McCarroll, Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91** (2012), doi:10.1016/j.ajhg.2012.10.018.

64. W. B. Mattes, J. A. Hartley, K. W. Kohn, D. W. Matheson, Gc-rich regions in genomes as targets for dna alkylation. *Carcinogenesis*. **9** (1988), doi:10.1093/carcin/9.11.2065.

65. A. R. Poetsch, S. J. Boulton, N. M. Luscombe, Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis 06 Biological Sciences 0604 Genetics. *Genome Biol.* **19** (2018), doi:10.1186/s13059-018-1582-2.

66. R. P. Grewal, M. Achari, T. Matsuura, A. Durazo, E. Tayag, L. Zu, S. M. Pulst, T. Ashizawa, Clinical features and ATTCT repeat expansion in spinocerebellar ataxia type 10. *Arch. Neurol.* **59** (2002), doi:10.1001/archneur.59.8.1285.

67. T. Matsuura, P. Fang, X. Lin, M. Khajavi, K. Tsuji, A. Rasmussen, R. P. Grewal, M. Achari, M. F. Alonso, S. M. Pulst, H. Y. Zoghbi, D. L. Nelson, B. B. Roa, T. Ashizawa, Somatic and germline instability of the ATTCT repeat in spinocerebellar ataxia type 10. *Am. J. Hum. Genet.* **74** (2004), doi:10.1086/421526.

68. G. Sveinbjornsson, A. Albrechtsen, F. Zink, S. A. Gudjonsson, A. Oddson, G. Másson, H. Holm, A. Kong, U. Thorsteinsdottir, P. Sulem, D. F. Gudbjartsson, K. Stefansson, Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat. Genet.* **48** (2016), doi:10.1038/ng.3507.

69. K. Tamura, M. Kaneda, M. Futagawa, M. Takeshita, S. Kim, M. Nakama, N. Kawashita, J. Tatsumi-Miyajima, Genetic and genomic basis of the mismatch repair system involved in Lynch syndrome. *Int. J. Clin. Oncol.* **24** (2019), , doi:10.1007/s10147-019-01494-y.

70. N. de Wind, M. Dekker, A. Berns, M. Radman, H. te Riele, Inactivation of the mouse Msh2 gene results in mismatch repair deficiency, methylation tolerance, hyperrecombination, and predisposition to cancer. *Cell*. **82** (1995), doi:10.1016/0092-8674(95)90319-4.

71. J. Gorman, A. Chowdhury, J. A. Surtees, J. Shimada, D. R. Reichman, E. Alani, E. C. Greene, Dynamic Basis for One-Dimensional DNA Scanning by the Mismatch Repair Complex Msh2-Msh6. *Mol. Cell*. **28** (2007), doi:10.1016/j.molcel.2007.09.008.

72. S. Tomé, K. Manley, J. P. Simard, G. W. Clark, M. M. Slean, M. Swami, P. F. Shelbourne, E. R. M. Tillier, D. G. Monckton, A. Messer, C. E. Pearson, MSH3 Polymorphisms and Protein Levels Affect CAG Repeat Instability in Huntington's Disease Mice. *PLoS Genet.* **9** (2013), doi:10.1371/journal.pgen.1003280.

73. X. Jia, B. B. Burugula, V. Chen, R. M. Lemons, S. Jayakody, M. Maksutova, J. O. Kitzman, Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. *Am. J. Hum. Genet.* **108** (2021), doi:10.1016/j.ajhg.2020.12.003.

74. K. Drotschmann, A. B. Clark, T. A. Kunkel, Mutator phenotypes of common polymorphisms and missense mutations in MSH2. *Curr. Biol.* **9** (1999), doi:10.1016/S0960-9822(99)80396-0.

75. A. R. Ellison, J. Lofing, G. A. Bitter, Functional analysis of human MLH1 and MSH2 missense variants and hybrid human-yeast MLH1 proteins in Saccharomyces cerevisiae. *Hum. Mol. Genet.* **10** (2001), doi:10.1093/hmg/10.18.1889.

76. S. Rangaswamy, A. Pandey, S. Mitra, M. L. Hegde, Pre-replicative repair of oxidized bases maintains fidelity in mammalian genomes: The cowcatcher role of NEIL1 DNA glycosylase. *Genes (Basel).* **8** (2017), , doi:10.3390/genes8070175.

77. L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, P. Boutros, K. Chan, A. Fujimoto, G. Getz, M. N. Huang, M. Kazanov, M. Lawrence, I. Martincorena, S. Morganella, H. Nakagawa, P. Polak, S. Prokopec, S. A. Roberts, S. G. Rozen, N. Saini, T. Shibata, Y. Shiraishi, M. R. Stratton, B. T. Teh, I. Vázquez-García, F. Yousif, W. Yu, The repertoire of mutational signatures in human cancer. *Nature*. **578** (2020), doi:10.1038/s41586-020-1943-3.

78. B. V. Halldorsson, H. P. Eggertsson, K. H. S. Moore, H. Hauswedell, O. Eiriksson, M. O. Ulfarsson, G. Palsson, M. T. Hardarson, A. Oddsson, B. O. Jensson, S. Kristmundsdottir, B. D. Sigurpalsdottir, O. A. Stefansson, D. Beyter, G. Holley, V. Tragante, A. Gylfason, P. I. Olason, F. Zink, M. Asgeirsdottir, S. T. Sverrisson, B. Sigurdsson, S. A. Gudjonsson, G. T. Sigurdsson, G. H. Halldorsson, G. Sveinbjornsson, K. Norland, U. Styrkarsdottir, D. N. Magnusdottir, S. Snorradottir, K. Kristinsson, E. Sobech, G. Thorleifsson, F. Jonsson, P. Melsted, I. Jonsdottir, T. Rafnar, H. Holm, H. Stefansson, J. Saemundsdottir, D. F. Gudbjartsson, O. T. Magnusson, G. Masson, U. Thorsteinsdottir, A. Helgason, H. Jonsson, P. Sulem, K. Stefansson, The sequences of 150,119 genomes in the UK biobank. *bioRxiv* (2021).

79. B. V. Halldorsson, M. T. Hardarson, B. Kehr, U. Styrkarsdottir, A. Gylfason, G. Thorleifsson, F. Zink, A. Jonasdottir, A. Jonasdottir, P. Sulem, G. Masson, U. Thorsteinsdottir, A. Helgason, A. Kong, D. F. Gudbjartsson, K. Stefansson, The rate of meiotic gene conversion varies by sex and age. *Nat. Genet.* **48** (2016), doi:10.1038/ng.3669.

80. D. Bachtrog, M. Agis, M. Imhof, C. Schlötterer, Microsatellite variability differs between dinucleotide repeat motifs - Evidence from Drosophila melanogaster. *Mol. Biol. Evol.* **17** (2000), doi:10.1093/oxfordjournals.molbev.a026411.

81. D. Dieringer, C. Schlötterer, Two distinct modes of microsatellite mutation processes: Evidence from the complete genomic sequences of nine species. *Genome Res.* **13** (2003), doi:10.1101/gr.1416703.

82. P. Murat, G. Guilbaud, J. E. Sale, DNA polymerase stalling at structured DNA constrains the expansion of short tandem repeats. *Genome Biol.* **21** (2020), doi:10.1186/s13059-020-02124-x.

83. S. T. Lovett, Polymerase Switching in DNA Replication. *Mol. Cell*. **27** (2007), , doi:10.1016/j.molcel.2007.08.003.

84. A. N. Khristich, S. M. Mirkin, On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *J. Biol. Chem.* **295** (2020), , doi:10.1074/jbc.REV119.007678.

85. F. Duraturo, R. Liccardo, M. De Rosa, P. Izzo, Genetics, diagnosis and treatment of lynch syndrome: Old lessons and current challenges (Review). *Oncol. Lett.* **17** (2019), , doi:10.3892/ol.2019.9945.

86. V. A. Schneider, T. Graves-Lindsay, K. Howe, N. Bouk, H. C. Chen, P. A. Kitts, T. D. Murphy, K. D. Pruitt, F. Thibaud-Nissen, D. Albracht, R. S. Fulton, M. Kremitzki, V. Magrini, C. Markovic, S. McGrath, K. M. Steinberg, K. Auger, W. Chow, J. Collins, G. Harden, T. Hubbard, S. Pelan, J. T. Simpson, G. Threadgold, J. Torrance, J. M. Wood, L. Clarke, S. Koren, M. Boitano, P. Peluso, H. Li, C. S. Chin, A. M. Phillippy, R. Durbin, R. K. Wilson, P. Flicek, E. E. Eichler, D. M. Church, Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27** (2017), doi:10.1101/gr.213611.116.

87. J. M. Zook, N. F. Hansen, N. D. Olson, L. Chapman, J. C. Mullikin, C. Xiao, S. Sherry, S. Koren, A. M. Phillippy, P. C. Boutros, S. M. E. Sahraeian, V. Huang, A. Rouette, N. Alexander, C. E. Mason, I. Hajirasouliha, C. Ricketts, J. Lee, R. Tearle, I. T. Fiddes, A. M. Barrio, J. Wala, A. Carroll, N. Ghaffari, O. L. Rodriguez, A. Bashir, S. Jackman, J. J. Farrell, A. M. Wenger, C. Alkan, A. Soylev, M. C. Schatz, S. Garg, G. Church, T. Marschall, K. Chen, X. Fan, A. C. English, J. A. Rosenfeld, W. Zhou, R. E. Mills, J. M. Sage, J. R. Davis, M. D. Kaiser, J. S. Oliver, A. P. Catalano, M. J. P. Chaisson, N. Spies, F. J. Sedlazeck, M. Salit, A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38** (2020), doi:10.1038/s41587-020-0538-8.

88. A. Abyzov, A. E. Urban, M. Snyder, M. Gerstein, CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21** (2011), doi:10.1101/gr.114876.110.

89. A. Kong, G. Masson, M. L. Frigge, A. Gylfason, P. Zusmanovich, G. Thorleifsson, P. I. Olason, A. Ingason, S. Steinberg, T. Rafnar, P. Sulem, M. Mouy, F. Jonsson, U. Thorsteinsdottir, D. F. Gudbjartsson, H. Stefansson, K. Stefansson, Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40** (2008), doi:10.1038/ng.216.

90. D. F. Gudbjartsson, H. Helgason, S. A. Gudjonsson, F. Zink, A. Oddson, A. Gylfason, S. Besenbacher, G. Magnusson, B. V. Halldorsson, E. Hjartarson, G. T. Sigurdsson, S. N. Stacey, M. L. Frigge, H. Holm, J. Saemundsdottir, H. T. Helgadottir, H. Johannsdottir, G. Sigfusson, G. Thorgeirsson, J. T. Sverrisson, S. Gretarsdottir, G. B. Walters, T. Rafnar, B. Thjodleifsson, E. S. Bjornsson, S. Olafsson, H. Thorarinsdottir, T. Steingrimsdottir, T. S. Gudmundsdottir, A. Theodors, J. G. Jonasson, A. Sigurdsson, G. Bjornsdottir, J. J. Jonsson, O. Thorarensen, P. Ludvigsson, H. Gudbjartsson, G. I. Eyjolfsson, O. Sigurdardottir, I. Olafsson, D. O. Arnar, O. T. Magnusson, A. Kong, G. Masson, U. Thorsteinsdottir, A. Helgason, P. Sulem, K. Stefansson, Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* (2015), doi:10.1038/ng.3247.

91. D. Beyter, H. Ingimundardottir, A. Oddsson, H. P. Eggertsson, E. Bjornsson, H. Jonsson, B. A. Atlason, S. Kristmundsdottir, S. Mehringer, M. T. Hardarson, S. A. Gudjonsson, D. N. Magnusdottir, A. Jonasdottir, A. Jonasdottir, R. P. Kristjansson, S. T. Sverrisson, G. Holley, G. Palsson, O. A. Stefansson, G. Eyjolfsson, I. Olafsson, O. Sigurdardottir, B. Torfason, G. Masson, A. Helgason, U. Thorsteinsdottir, H. Holm, D. F. Gudbjartsson, P. Sulem, O. T. Magnusson, B. V. Halldorsson, K. Stefansson, Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53** (2021), doi:10.1038/s41588-021-00865-4.

92. S. T. Buckland, A. C. Davison, D. V. Hinkley, Bootstrap Methods and Their Application. *Biometrics*. **54** (1998), doi:10.2307/3109789.

93. A. Canty, B. . Ripley, boot: Bootstrap R (S-Plus) Functions. R package version 1.3-28. *Http://Cran.R-Project.Org/Doc/Packages/.* (2021).

94. A. Lal, M. Brown, R. Mohan, J. Daw, J. Drake, J. Israeli, Improving long-read consensus sequencing accuracy with deep learning. *bioRxiv* (2021).

Chapter 5

Conclusions and Future Directions

5.1 Summary

During my work I have approached microsatellites from three different directions. First, the focus was strictly on obtaining reliable microsatellite genotypes in a large set of samples. The number of samples and consequently the amount of data available for analysis presented an extra level of complexity on top of the error and mutation rates. As a result, computational improvements were one of the two main objectives in the next publication. The other objective represented the other end of the sample size range and aimed at broadening the software's application spectrum to accurate calling of individual clinical samples. Clinical sequencing samples require an increased level of sensitivity when looking for possible genetic causes of the diseases and syndromes the individuals suffer from. Finally, combining the computationally efficient software with the experience gained using the clinical approach of looking for something in an offspring that is not present in its parents, we genotyped microsatellites in two of the largest sample sets to date and shifted our scope to the source of variability in microsatellites, de novo mutations.

5.2 Conclusion

A different set of conclusions can be drawn from each section of work presented here but perhaps the most important one is how valuable microsatellites are to genetic research and how looking at problems from different perspectives can increase the overall quality of the solutions generated. By focusing first solely on determining repeat counts while accounting for different sources of errors and noise in the data I was able to lay the groundwork for algorithmic improvements. These improvements then made it feasible to generate data sets on a scale larger than previously possible. Collaboration with clinical sequencing experts allowed for a better understanding of the requirements and the best course of action when confirming or rejecting the presence of expanded repeats in affected individuals. No matter how deleterious and pathogenic most mutations in humans are, we cannot deny the fact that they are also what drives the evolution of all living things. Thus, it is imperative to increase our understanding of the factors influencing their occurrence, and detecting mutations in offspring that are not present in their parents is arguably the most powerful way to reach this goal. The high mutation rate of microsatellites made it possible to apply the most stringent filters to the mutations while still keeping the set large enough for drawing important and valuable conclusions from its elements. These results demonstrated for the first time that the rate of mutations in the reproductive cells of humans is directly affected by their genetic sequence, something that has been hypothesized but proven difficult to show with mutations occurring at other types of genetic variants.

5.3 Future Directions

Long read sequencing methods are constantly increasing in quality and will likely soon rival second generation sequencing with respect to error rates and cost. I would like to extend my microsatellite genotyping to long read sequencing data and use the power obtained by longer reads to increase sensitivity when identifying long microsatellite alleles, such as repeat expansions and more accurately estimate their size. Further, my aim is to increase the sensitivity and specificity of my mDNM detection to identify both more differences between the mDNMs originating from each parent and other genetic variants affecting the number of mDNMs transmitted from parent to offspring. Building on the phenotypes I defined for the de novo mutation set I would like to define more specific phenotypes, based on the number of bp affected by the mutations and their direction among other things. As a surrogate phenotype for de novo mutation counts I would also like to compare imputed and sequenced genotypes at high quality, reliable microsatellites and count per individual how many times these disagree. These counts could then be generated for all samples, not just parents in trios and would hopefully replicate the previous associations while also greatly increasing power to detect new associations.

School of Technology, Department of Engineering
Reykjavík University
Menntavegur 1
101 Reykjavík, Iceland
Tel. +354 599 6200
Fax +354 599 6201
www.ru.is