STRetch: detecting and discovering pathogenic short tandem repeats expansions

Harriet Dashnow[1,2], Monkol Lek[3,4], Belinda Phipson[1], Andreas Halman[1,5], Mark Davis[6], Phillipa Lamont[7], Nigel Laing[8], Daniel G. MacArthur[3,4] and Alicia Oshlack[1,2]

1. Murdoch Childrens Research Institute, Royal Children's Hospital, Parkville, VIC, Australia
2. School of Biosciences, The University of Melbourne, Parkville, VIC, Australia
3. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
4. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA
5. Florey Institute of Neuroscience and Mental Health, University of Melbourne, Parkville, VIC, Australia
6. Department of Diagnostic Genomics, PathWest Laboratory Medicine, QEII Medical Centre, Nedlands, WA, Australia
7. Neurogenetic Unit, Royal Perth Hospital, Perth, WA, Australia
8. Harry Perkins Institute of Medical Research, Centre for Medical Research, University of Western Australia, Nedlands, WA, Australia.

Abstract

Short tandem repeat (STR) expansions have been identified as the causal DNA mutation in dozens of Mendelian human diseases. Traditionally, pathogenic STR expansions could only be detected by single locus techniques, such as PCR and electrophoresis. The ability to genotype STRs directly from next-generation sequencing data has the potential to reduce both the time and cost to reaching diagnosis and to discover new causal STR loci. Most existing tools detect STR variation within the read length, and so are unable to detect the majority of pathogenic expansions.

Here we present STRetch, a new genome-wide method to detect pathogenic STR expansions and estimate their approximate size directly from short read sequencing. We show that STRetch can detect pathogenic STR expansions in short-read whole genome sequencing data. We apply STRetch to the analysis of 97 whole genomes to reveal variation at STR loci. Finally, we demonstrate the application of STRetch to solve cases of patients with undiagnosed disease, where STR expansions are a likely cause. A key advantage of STRetch over other tools is that it assesses expansions at all STR loci in the genome and so can be used to detect novel disease-causing STR loci.

STRetch is open source software, available from github.com/Oshlack/STRetch.

Background

Short tandem repeats (STRs), also known as microsatellites, are short (1-6bp) DNA sequences repeated consecutively. Approximately 3% of the human genome consists of STRs [1]. These loci are prone to frequent mutations and high polymorphism, with mutation rates 10 to 100,000 times higher than average mutation rates in other parts of the genome [2]. Dozens of neurological and developmental disorders have been attributed to STR expansions [3]. STRs have also been associated with a range of functions such as DNA replication and repair, chromatin organization, and regulation of gene expression [2, 4, 5].

STR expansions have been identified as the causal DNA mutation in almost 30 Mendelian human diseases [6]. Many of these conditions affect the nervous system, for example Huntington’s disease, spinocerebellar ataxias, spinobulbar muscular atrophy, Friedreich’s ataxia, fragile X syndrome and polyalanine disorders [7]. Most tandem repeat expansion disorders show dominant inheritance, with disease mechanisms varying from expansion of a peptide repeat, disrupting protein function or stability, to causing the aberrant regulation of gene expression [8].

STR expansion diseases typically show genetic anticipation, characterized by greater severity and earlier age of onset as the tandem repeat expands through the generations [9]. In many STR diseases, the probability that a given individual is affected increases with the repeat length. In some cases severity also depends on the gender of the parent who transmitted the repeat expansion [10]. The number and position of imperfect repeat units also influences the stability of the allele through generations [9]. Together these features can be used to identify patients with a disease of unknown genetic basis that might be caused by an STR expansion.

Traditionally, STRs have been genotyped using gel electrophoresis. Polymerase chain reaction (PCR) is performed using primers, which are complementary to unique sequences flanking the STR. The PCR product is then run on a capillary electrophoresis gel to determine its size. Although the method has been scaled to handle dozens of samples, it is still labor-intensive and costly. Each new STR locus to be genotyped requires the design and testing of a new set of PCR primers, along with control samples.

A number of diseases are known to be caused by any one of multiple variants, including STR expansions, single nucleotide variants (SNVs) or short indels. For example there are more than ten STR loci in as many genes known to cause ataxia [11], in addition to SNVs and indels in dozens of genes [12]. For such diseases, this can mean hundreds of dollars spent per STR locus, alongside SNV and short indel testing. For such conditions there is a clear need for a single genomic test that can detect all relevant disease variants including SNVs, indels and STRs.

The ability to genotype STRs directly from next-generation sequencing data has the potential to reduce both the time and cost to reaching diagnosis and to discover new causal STR loci. It is becoming increasingly common to sequence the genomes or exomes of patients with undiagnosed genetic disorders. Currently the analysis of this data is generally restricted to SNVs and short indels, with STRs generally only investigated in an ad-hoc manner if they are a common cause of the disease. The ability to detect STR expansions in next-generation sequencing data gives the potential to perform disease variant discovery in those patients for which no known pathogenic variants are found.

The vast majority of current STR genotyping tools for short-read sequencing data (most notably LobSTR [13], HipSTR [14] and RepeatSeq [15]) are designed to look at normal population variation by looking for insertions and deletions within reads that completely span the STR. These tools are limited to genotyping alleles that are less than the read length, with sufficient unique flanking sequence to allow them to be mapped correctly. However, for most STR loci causing Mendelian disease in humans, pathogenic alleles typically exceed 100bp, with pathogenic alleles at some loci in the range 1,000-10,000bp [16], far exceeding the sizes that can be detected using these algorithms.

One STR genotyper, STRViper [17], detects alleles exceeding the read length by looking for shifts in the insert size distribution. This method requires that the insert size distribution has a relatively low standard deviation and is limited to repeats smaller than the insert size. The mean insert size can be as low as 300-400bp, meaning that for a very large pathogenic repeat, there may be very few or no spanning read pairs. This strategy is therefore unsuitable to detect many of the STR expansions known to cause disease in humans. Another tool, ExpansionHunter [18], uses read pair information and recovery of mismapped reads to detect large STR expansions. However this tool only works on specific loci as defined by the user, and so is not a genome-wide method.

Long read sequencing technologies can potentially sequence right through larger repeat loci [19] however, this sequencing is currently far too expensive for the clinic. Also, high error rates and low throughput in these technologies make them less accurate for genotyping SNVs and short indels and so a poor alternative to short-read sequencing in a clinical setting. Clearly there is still a great need to be able to detect pathogenic expansions from short read data.

Here we present STRetch, a new method to detect pathogenic STR expansions and estimate their approximate size directly from short read sequencing. We show that STRetch can detect pathogenic STR expansions in short-read whole genome sequencing data. We also demonstrate the application of STRetch to solve cases of patients with undiagnosed disease, where STR expansions are a likely cause.

STRetch is open source software, available from github.com/Oshlack/STRetch.

Results

STRetch detects large STR expansions

The STRetch method has been designed to identify expanded STRs from short-read sequencing data and give approximate sizes for these alleles. Briefly, the idea behind

STRetch is to first construct a set of additional STR decoy chromosomes that are added to the reference genome. By mapping to this modified genome, STRetch identifies reads that originated from large STR expansions that now map to the STR decoys. These reads are then allocated back to the genome using read pair information, and the source locus is assessed for an expansion using a statistical test based on coverage of the STR. Figure 1 outlines the STRetch method, and the details are expanded below.

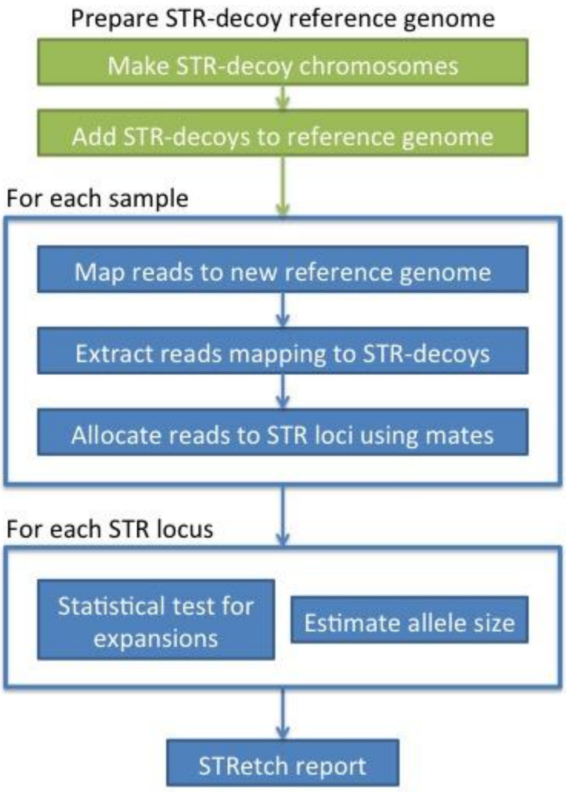

Prepare STR-decoy reference genome
Make STR-decoy chromosomes
Add STR-decoys to reference genome
For each sample
Map reads to new reference genome
Extract reads mapping to STR-decoys
Allocate reads to STR loci using mates
For each STR locus
Statistical test for expansions
Estimate allele size
STRetch report

Figure 1: Summary of the STRetch method

STR decoy chromosomes: generating an STR-aware reference genome

A key feature of STRetch is the generation of STR decoy chromosomes to produce a custom STR-aware reference genome.

Most aligners have difficulty accurately mapping reads containing long STRs. For example, although the BWA MEM algorithm has superior performance for mapping reads containing STRs [20] sometimes reads containing long STRs map to other STR loci with the same repeat unit, or completely fail to map [18]. The systematic mis-mapping of STR reads is unsurprising considering that BWA MEM is optimized to find the longest exact match [21]. For a read made up primarily of STR sequence, the best match is likely to be the longest STR locus in the reference genome with the same repeat unit.

STRetch takes the issue of systematically mis-mapped, or unmapped STR reads and uses this as a way to identify reads that contain long STR sequence. To achieve this, we introduce the concept of STR decoy chromosomes. These chromosomes are sequences that consist of 2000bp of pure STR repeat units that can be added to any reference genome. STR decoy chromosomes for all possible STR repeat units in the range 1-6bp are generated, resulting in 501 new chromosomes added to the reference genome (STRetch provides hg19 with STR decoys, see methods). While reads from STRs similar to the allele in the reference genome will map to their original locus, reads containing large STR expansions will preferentially align to the STR decoy chromosomes. These reads can then be further examined for evidence of a pathogenic expansion.

Mapping to STR decoys to identify reads containing STRs

Once the new STR-decoy reference genome is created, the first step maps reads against the new reference genome using BWA MEM. If the data has already been mapped STRetch can optionally just map a subset of reads likely to contain STRs: those aligning to known STR loci, and unmapped reads (see methods). Any reads mapping to the STR decoy chromosomes are inferred to have originated from an STR.

Determining the origin of STR reads

Next the reads that map to the decoys are assigned to genomic STR positions. STRetch uses the mapping position of the mate to infer which STR locus each read originates from. Known STR loci are obtained from a Tandem Repeats Finder (TRF) [22] annotation of the reference genome. For a given read, if the mate maps within 500bp of a known STR locus with the same repeat unit, then the read is assigned to that genomic STR locus (or the closest matching locus if multiple loci are present).

After all possible reads are assigned, there may be a difference between the number of reads mapping to a given STR decoy chromosome and the number of reads assigned to all STR loci with that same repeat unit. Unassigned STR reads can occur for a variety of reasons, such as their mates map too far from a known locus, their mate also maps to the STR decoy chromosome, or their mate is unmapped. This number will increase in samples with very large and/or multiple STR expansions, as these result in more read pairs where both are contained within the STR. This will result in STRetch underestimating the size of large alleles.

Detecting outlier STR loci

STRetch next uses a statistical test to identify loci where an individual has an unusually large STR. Specifically, STRetch compares the number of STR decoy reads assigned to each locus for a test sample with STR reads from a control set. At each locus the reads are normalized by the average coverage of the sample. A robust z-score (“outlier score”) is then used to test if the log-normalized number of reads is an outlier compared with the controls (see methods). A variety of control samples can be used. Firstly, STRetch can be run on a set of controls and then the estimated median and variance parameters from these controls can be used for subsequent test samples. Secondly, STRetch supplies median and variance parameters estimated from a reference set of PCR-free whole genomes that can be used (see “STRetch reveals STR expansions in 97 whole genomes”). Thirdly, a set of samples that are all being tested can be used and compared with each other. The advantage of the first and last options is that the sequencing is usually run at the same center with the same library preparation protocols, however the data sets might be smaller. If the third option is used it may not detect outliers if the same expansion occurs in many of the samples, although the statistic is robust to up to approximately 50% outliers.

Estimating the size of STR alleles

STRetch works on the assumption that, for a given locus, the number of reads containing the STR repeat unit is proportional to the length of the repeat in the genome being sequenced. Therefore STRetch estimates the size of any detected expansion using the normalized counts allocated to that STR locus. Through simulation we found that the counts are linearly related to the length and we used the simulated results for estimating the allele size (see methods).

To estimate the size of the STR expansion, we performed a simulation with various STR lengths (see methods). Specifically we simulated reads from 100 individuals with the genotype 16xCAG/NxCAG at the SCA8 locus, where N was randomly selected in the range 0 to 500. Our simulated data exhibits a linear relationship between allele size and the number of reads mapping to the STR decoy chromosome. This relationship is used to estimate the size of alleles (Supplementary Figure 2).

Output files

On completion, STRetch generates a tab-delimited output file that contains STR loci for which STR-decoy reads were detected for each sample, along with p-values for statistical significance of an expansion and details of the data including the STR locus (position, repeat unit, size in reference), robust outlier z-score, locus read count and the allele length estimate. By default this file is sorted such that the most significant expansions are ranked at the top.

Stretch is able to recover true pathogenic expansions

In order to test STRetch we generated PCR-free whole genome sequencing on 10 individuals, nine with known pathogenic STR expansions and one unaffected family member. Samples were sequenced to a mean coverage of 41.74x (range 38.35-49.57x), then processed using the Broad GATK pipeline (mapped to hg38 with BWA-MEM, then processed using the GATK best practices).

For analysis with STRetch, we first extracted reads overlapping all known STR loci annotated by Tandem Repeats Finder (see methods), then processed these reads through the STRetch pipeline, using the hg19 reference genome. The STRetch statistics were calculated twice, first using just these 10 samples as controls for each other ("internal control") and then using the 97 WGS described below as controls ("reference control"). We also ran LobSTR/HipSTR to estimate the size of the short allele in each case.

For six of the ten samples we had information about the disease and the estimated allele size by PCR. For the other four samples (Sample 7, Sample 8, Sample 9 and Sample 10) we were initially completely blinded to all patient information, including phenotype. Disease and allele size estimates were only revealed to us after we had correctly identified the causal STR expansion in each case. Table 1 summarises the results of this analysis.

Table 1: Summary of the 10 individuals with known STR alleles. * indicates results after manual addition of the FRDA STR to the reference data.

Sample	disease	gene	repeat unit	type	position	allele ref	allele PCR	allele STRetch	rank (reference control)	p-val (reference control)	rank (internal control)	p-val (internal control)	allele LobSTR	allele HipSTR
Sample 1	SCA1	ATXN1	CAG	coding	chr6:16327865-16327955	30.3	51	50.5	1	3.21E-15	1	4.71E-25	31.3/31.3	36.3/39.3
Sample 2	Unaffected relative of sample 1	ATXN1	CAG	coding	chr6:16327865-16327955	30.3	29/32	32.2	-	-	-	-	30.3/35.3	30.3/35.3
Sample 3	SCA3	ATXN3	CAG	coding	chr14:92,537,355-92,537,396	14	73	65.4	3	5.41E-10	793	0.22	28/28	no call
Sample 4	SCA6	CACNA1A	CAG	coding	chr19:13318673-13318712	13.3	22	no call- 0 STR reads in all samples	-	-	-	-	11.3/24.3	11.3/22.3
Sample 5	SBMA	AR	CAG	coding	chrX:66765159-66765261	33.3	41	43.3	11	6.77E-06	9	6.90E-09	no call	no call
Sample 6	SBMA	AR	CAG	coding	chrX:66765159-66765261	33.3	47	35.1 (0 STR reads)	-	-	-	-	no call	no call
Sample 7	FTDALS1	C9orf72	GGGGCC	intronic	chr9:27573482-27573544	10.8	>50	41.5	5	1.57E-07	959	0.6	no call	no call
Sample 8	DM2	ZNF9	CCTG	intronic	chr3:128891419-128891502	20.8	>75	38.4	1	5.28E-17	1	3.84E-23	14.8/14.8	no call
Sample 9	DM1	DMPK	CAG	coding	chr19:46273462-46273524	20.7	>150	79.3	1	1.48E-33	1	3.24E-48	5.7/5.7	20.7/5.7
Sample 10	FRDA	FXN	GAA	intronic	chr9:71652203-71652205	6	~850	17.7*	NA	NA	2*	1.83e-08*	No call	No call

For the six samples with known information STRetch correctly identified three true positive expansions, and for the true negative (Sample 2) STRetch correctly failed to detect an expansion. STRetch failed to identify the causal locus in two cases (Sample 4 and Sample 6). However, for Sample 4 the PCR allele is only 26bp larger than the reference, making the entire allele 66bp, so well within the read length of 150bp. STRetch only detects STR expansions that are sufficiently large that the repeat maps to the STR decoys instead of the locus. Indeed we see three reads mapped to the genomic locus with 27bp insertion, and no reads from this locus mapping to the STR decoy. This allele can be detected by tools that look for indels within the read, and indeed both LobSTR and HipSTR are able to correctly call this expansion. Sample 6 was found to have lower coverage over the STR region compared with other samples however, some evidence of the repeat was observed and may be detected in future iterations of the software.

For the blinded samples Sample 7, Sample 8 and Sample 9 we were able to correctly determine the causal STR locus simply by ranking variants by their outlier scores then looking for known pathogenic STR loci (supplementary table 1) at the top of the list.

For Sample 10 we were initially unable to identify a significant expansion in a known pathogenic gene. After the variant was revealed to be a large GAA expansion at the FRDA locus we investigated further and discovered that although the reference genome has 6xGAA at this position, the STR is missing from the TRF genome annotation. This locus is positioned in the middle of an Alu element. We hypothesis that TRF fails to annotate this STR due to its relatively small size and the flanking repetitive sequences. After manually adding this locus to the genome annotation and rerunning the analysis on the 10 samples in Table 1, STRetch was able to detect an expansion at this locus.

Most of the true positives that were detected had significant expansions when using both the 97 reference controls and the 10 internal controls. However there were two samples (Sample 3 and Sample 7) that were only significant when using the much larger set of reference controls.

STRetch underestimates the allele size for many of the true positives (figure 2). One likely explanation for this is that the current implementation of STRetch is limited by the insert size of the sequencing data. Some alleles will be so large (e.g. DM1 in

Sample 9 is >450bp) that there will be read pairs where both are completely contained within the STR. So both will map to the STR decoy and will not be assigned to an STR locus. STRetch does not try to use these reads to estimate the allele length, and so will systematically underestimate large alleles. Future implementations of STRetch will assign these reads to loci based on the proportion of previously assigned reads, thereby making estimated lengths more accurate.

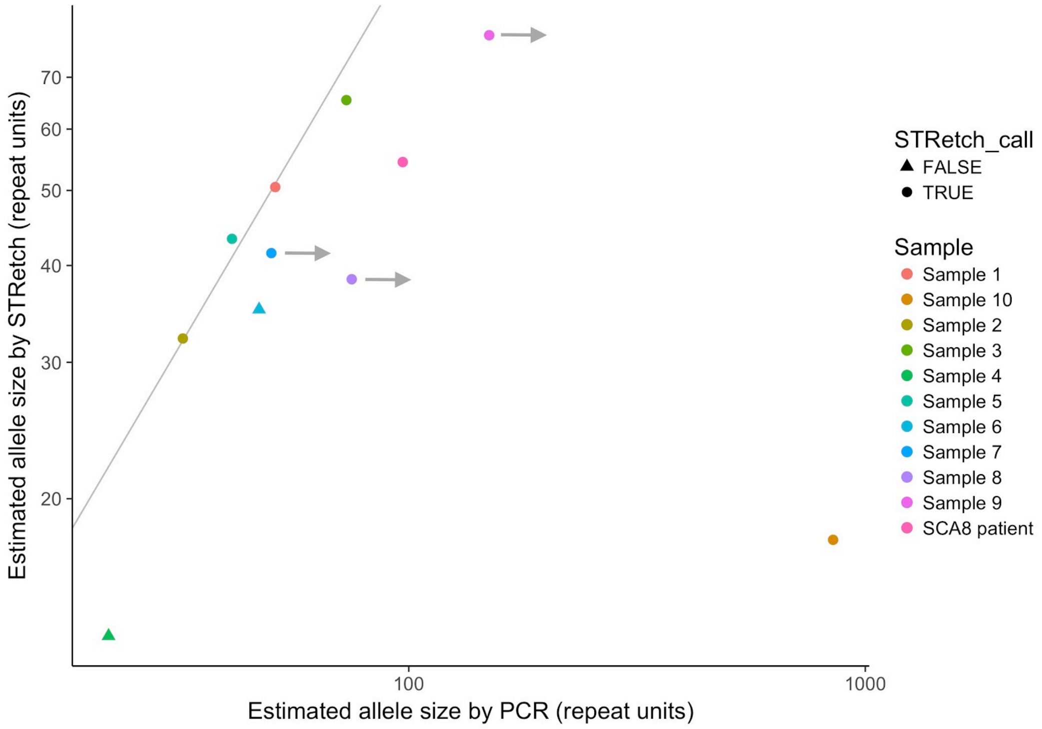


Figure 2: Relationship between the allele sizes estimated by PCR and STRetch for the true positive samples. The triangles indicate where STRetch failed to make a call because no reads from the STR decoy were assigned to that locus, so the hg19 reference allele has been used in place of a STRetch call. Grey arrows indicate where the PCR allele is known to be a lower bound. The SCA8 patient is not part of this analysis, but comes from the case study below.

STRetch reveals STR expansions in 97 whole genomes

We performed PCR-free whole genome sequencing on a set of 97 individuals, most of whom were being investigated for the cause of their Mendelian disease, or are immediate family members of such patients. Many of these cases have inherited neuromuscular disorders, approximately half of which have previously been identified as being caused by a SNV or indel. The remaining individuals are unsolved cases. The set also includes eight individuals from ataxia families comprised of four affected singletons, and a family consisting of two affected children and two unaffected parents. In addition to patient samples and relatives, there are seven unaffected samples including NA12878. Given its enrichment for individuals with Mendelian disease, this control set may contain pathogenic STR expansions. We therefore use robust estimators of the mean and variance to generate z-scores and p-values (see methods).

All samples had previously been mapped with BWA-MEM. We used STRetch to extract reads from all annotated STR loci, as well as unmapped reads, and re-mapped these against the hg19 STR-decoy genome (see methods) and then proceeded with the rest of the STRetch pipeline. We recorded the median and standard deviation for each locus across all individuals for use as a control set for subsequent analyses. Homopolymer loci are the most variable between individuals, showing dramatically higher standard deviations, followed by STRs with 5, 6, 3, 4 and then 2 bp repeat units (supplementary figure 1).

To assess the frequency of expansions at known pathogenetic STR loci, we filtered the STRetch results to those significant expansions intersecting with a set of 23 known pathogenic STR loci (see supplementary table 1). We observed 29 significant STR expansions in seven pathogenic loci (summarised in table 2). Although these are significantly expanded compared to the rest of the control set, their pathogenicity is uncertain as the allele size estimates are only approximate and are often well below the defined pathogenic range.

Table 2: Summary of significant expansions in STR disease loci in 97 WGS samples.

Disease	**Gene**	**Number of patients**
SCA8	ATXN8/ATXN8OS	2
DM2	ZNF9	1
SBMA	AR	2
SCA36	NOP56	1
FXTAS	FMR1	1
SCA3/MJD	ATXN3	11
FTDALS1	C9orf72	11

Nonetheless, a number of the STR expansions are potential candidates for follow-up if the individuals sequenced have a relevant phenotype. STRetch detected a large SCA8 expansion in two individuals, one with unknown phenotype and the second was in one of the ataxia patients (see below). We also detected a DM2 expansion. All three variants were highly significantly expanded compared to the other control samples (p=4.2e-24, 8.2e-23 and 1.5e-10 respectively), and each was ranked as the most significant for that individual. We have referred these variants back to the originating laboratories to determine if the variants fit the phenotype and can be validated.

STRetch also identified STR expansions in a number of other pathogenic loci in these samples, however many of these are unlikely to be sufficiently large to cause disease. Two SBMA expansions were on the limit of detection and significance (ranked 109 and 476, p=0.001 and 0.04 respectively). We detected SCA36 and FXTAS expansions, which likely reflect sub-clinical variation at these loci, with size estimates of 13.5xAGGCCC and 34.6xCCG respectively. We detected a surprising number of SCA3/MJD_ATXN3 expansions: 11 samples with estimated allele sizes in the range 30.5x to 74.3xAGC. As ≥52xAGC is considered pathogenic, with 45x-51x showing incomplete penetrance, many of these may be asymptomatic. However, we have observed STRetch to underestimate allele sizes at this locus so these variants could be a little larger than predicted. We also detected 11 FTDALS1 expansions in the range 20.9-32.9xCCCCGG, all within the normal size range for this locus. Generally affected individuals have greater than 250 repeats so these results likely indicate individuals with pre-mutations.

These likely non-pathogenic variants at known pathogenic loci highlight the ability of STRetch to detect STR expansions at the genome-wide scale. This gives us the power to explore population variation at these loci, and so to better determine the true non-pathogenic range of these known pathogenic loci.

Genetic diagnosis and validation of an ataxia patient

As noted above, six ataxia patients (and two unaffected parents) were included in our 97 WGS cohort. These patients had previously been tested with a panel of the five most common ataxia STR expansions: SCA1, SCA2, SCA3, SCA6 and SCA7. The whole genome data had also been examined for causative SNVs and indels in candidate ataxia genes using the GATK Best Practices recommendations [23], and copy number variations using BreakDancer [24] and Genome STRiP [25] without success in diagnosis.

In one of these patients STRetch identified a highly significant expansion of SCA8, a known but rare disease locus in ATXN8 with an outlier z-score of 11.11 (p=3.55e-24) and an estimated allele size of 54.4x.

As a result this SCA8 expansion was validated using a PCR assay, which confirmed a pathogenic CAG expansion with an allele size of 97x. In addition, an affected sibling, not sequenced using WGS, was tested for an expansion at the same locus and was similarly determined to have a pathogenic allele of ~96x.

We ran LobSTR and HipSTR on the WGS data of this same patient. At the expanded pathogenic locus LobSTR called a homozygous 6xTGC insertion, while HipSTR called a homozygous indel deleting one TAC repeat unit upstream of SCA8 and an insertion of seven TGC repeat units for a net six repeat unit insertion. Both tools report a reference allele size of +15x, so the total allele size is 21x in both cases. PCR analysis of the proband and sibling indicated the expansions were heterozygous both with a short allele size of ~29x, however these sizes may not be directly comparable due to potential variation in the definition of the reference locus size. We configured ExpansionHunter to estimate the allele size for the SCA8 locus detected using STRetch and a 96x allele was estimated (confidence interval 76-113).

The LobSTR and HipSTR analysis reveals the danger of relying solely on within-read signals of STR expansion, as this will miss any insertions larger than the read length. This work demonstrates the ability of STRetch to discover pathogenic STR expansions from WGS data.

Discussion

We have demonstrated the ability of STRetch to detect pathogenic STR expansions in short-read WGS data. STRetch was able to detect most of the known true positive variants where the variant size exceeded the read-length and only missed 1 out of 9 true positives. We further applied STRetch to 97 WGS samples to detect both potentially pathogenic STR expansions and expansions of moderate length in STR disease loci, where the allele size is likely below the threshold for disease. Importantly this set of analysed genomes can act as a control set, and statistical parameters for the STR loci are provided to use in testing for expansions with STRetch. Within this cohort STRetch revealed a previously undetected pathogenic STR expansion in SCA8 that was validated by PCR in the proband and an affected sibling.

The main limitation of STRetch is its tendency to underestimate the allele size of STR expansions, especially variants larger than the insert size. This is because STRetch only assigns STR reads to a genomic locus when there is a uniquely mapping read-pair. Future implementations of STRetch will use reads where both reads in the pair map to the STR decoy chromosome to improve the allele size estimates. It should be noted, however, that the statistical approach for testing for outliers against a set of controls is not as severely affected by this limitation.

As expected, we found that STR genotypers such as LobSTR and HipSTR, that are designed to genotype short STR variation, were unable to detect large pathogenic variants in WGS data. These tools instead called a homozygous genotype, corresponding to the size of the non-expanded allele or called a heterozygous with slight variation in the small allele, or failed to make a call. Using these tools in conjunction with STRetch allows the estimation of the short allele in cases where STRetch has detected an expansion, allowing us to obtain a fuller picture of the genotype from short read sequencing data.

Finally, ExpansionHunter performed accurately on the one sample and locus that we tested. However a key limitation is that ExpansionHunter is not currently configured as a genome-wide tool. Each locus of interest must be defined in a separate configuration file.

Conclusions

Here we have introduced STRetch, a method to test for STR expansions using whole genome sequencing data. We have shown that STRetch can detect pathogenic STR expansions relevant to Mendelian disease. Although the emphasis has been on STRs known to cause Mendelian disease, a key advantage of STRetch over other methods is that it is a genome-wide approach. STRetch tests for expansions at every STR locus annotated in the reference genome, and so has potential to be used for not only diagnostic applications, but also in research to discover new pathogenic STR expansions.

Methods

The STRetch pipeline

The STRetch pipeline is implemented using the Bpipe [26] pipeline framework (v0.9.9.3). This allows for a pipeline combining standard bioinformatics tools with novel scripts, and is compatible with many high performance computing environments, allowing large-scale parallelization over multiple samples.

To summarize the pipeline and components in brief:

Reads are mapped to the reference genome with STR decoy chromosomes using BWA MEM [21] (v0.7.12) and Samtools [27] (v1.3.1). STRetch then counts the number of reads mapping to each STR decoy chromosome using bedtools [28] (v2.26.0). Reads mapping to the STR decoy chromosomes are allocated to an STR locus using paired information (Python v3.5.2 script: identify_locus.py). Median coverage over the whole genome or exome target region is calculated using goleft covmed [29] (v0.1.8), which is later used to normalize the counts. STRetch then predicts the size of the expansion using the number of reads allocated to the locus (R v3.3.1 script: estimateSTR.R).

The STRetch pipeline is freely available under an MIT license from github.com/Oshlack/STRetch.

Generating STR decoy chromosomes

To produce STR decoy chromosomes STRetch generates the set of all possible STR repeat units in the range 1-6bp. These are then grouped by those repeat units that are equivalent as a circular permutation of each other or the reverse complement. For each group the first repeat unit lexicographically was taken to represent that group. For example CAG = AGC = GCA = CTG = TGC = GCT and the group is represented by AGC. STRetch filters out repeat units that could be represented by multiples of a shorter repeat unit. For example ATAT would be filtered out as it is already represented by AT. This resulted in 501 unique repeat units. STRetch then uses an "STR decoy chromosome" for each repeat unit, which consists of the repeat unit repeated in tandem for 2000bp (script: decoy_STR.py). These additional chromosomes can be added to any reference genome (hg19 with STR decoy chromosomes was used for all analyses).

Extracting likely STR reads pairs from aligned bams

In the case where reads have already been mapped to a reference genome, STRetch provides the option of extracting likely STR reads pairs from the bam file for analysis, rather than remapping all reads in the sample.

In this case STRetch defines a region where STR reads are likely to align by taking the Tandem Repeats Finder [22] annotation of the reference genome and expanding the region to include 800bp of flanking sequence on each side. Reads aligned to this region, and all unmapped reads are extracted using samtools view. These are sorted to place together read pairs using samtools collate and then are extracted in fastq format using bedtools bamtofastq. Unpaired reads are discarded.

Aligning and allocating reads to STR loci

Stretch uses BWA MEM to align reads to the custom reference genome. Any read mapping to the STR decoy chromosomes is presumed to have originated from an STR locus.

To determine which STR locus the reads originated from, the mates of the reads mapping to a given STR decoy chromosome are collected. If the mate maps within 500bp of a known STR locus with the same repeat unit, it is assigned to that locus. If multiple loci fall in this range, it is assigned to the closest. This distance was chosen because the average insert size of WGS data was 500bp, so we expect the mate to map within 500bp or less of the STR locus. This value can be configured in the pipeline if required.

To correct for library size (total number of aligned reads) the counts for each STR locus are normalised against the median coverage for that sample within the target region (in the case of exome capture), or across the entire genome. Counts are normalised to a median coverage of 100x, by calculating: normalised counts = (raw counts/median sample coverage)*100. Where log counts are required, log counts = $\log_2$(100*(raw counts + 1)/median sample coverage)). Log_2 normalized counts are used in subsequent statistical analyses.

Detecting outliers

To detect individuals with unusually large STRs, STRetch calculates an "outlier score" for each individual at each locus. The outlier score is a z-score calculated using robust estimates, with a positive score indicating the STR is larger than the median.

A robust z-score and p-value is calculated for each locus *l* using the normalized log counts. First the median and variance across all samples for locus *l* is estimated using Huber's M-estimator [30, 31]. This calculation can be performed over all samples in a batch, or a set of control samples (estimates from the set of 97 PCR-free whole genomes are provided with STRetch). We test the null hypothesis that the log counts, y_{il}, at locus *l* for sample *i* is equal to the median log counts at locus *l* for the control samples. The alternative hypothesis is that the median log counts for locus *l* are greater for sample *i* compared to the control samples. Hence for each sample i and locus *l* we obtain a robust z-score

$$z_{il} = \frac{y_{il}-M}{\sigma_M},$$

where M is the median and σ_M is the square-rooted M-estimator of the variance. One-sided p-values are then obtained from the standard normal distribution and adjusted for multiple testing across the loci using the Benjamini-Hochberg method [32]. A locus is called significant if the adjusted p-value is < 0.05.

Estimating allele sizes

We reasoned that the size of an expanded allele would be proportional to the number of reads containing STR sequence and hence the number of reads allocated to the STR locus. In order to estimate allele sizes we performed simulations of a single locus at a range of allele sizes. Specifically, reads were simulated from the SCA8 locus in ATXN8. One allele was held constant at 16xAGC and then we simulated repeat lengths in the other allele in the range 0-500 repeat units (selected at random from a uniform distribution). Alternate versions of the hg19 reference genome with these alleles were produced using GATK v3.6 FastaAlternateReferenceMaker. Reads were simulated from 10,000bp either side of the SCA8 locus using ART MountRainier-2016-06-05 [33] Reads were 150bp paired end, with insert sizes sampled from a normal distribution (mean 500bp, sd 50bp) and 30X coverage (proportional coverage sampled from each haplotype). The Illumina error profile was used. Simulation code is available at github.com/hdashnow/STR-pipelines.

A plot of the number of reads mapping to the AGC decoy chromosome against the number of AGC repeat units inserted into the ATXN8 locus shows a clear linear relationship between these two variables (Supplementary Figure 2).

We can use this information from the simulated data to provide a point estimate of the allele size of any new sample we analyse with STRetch in the following manner. We fit a linear regression between the number of reads mapping to the STR decoy and the size of the allele from the simulated data (both $\log_2$ transformed), in order to obtain estimates of the intercept and slope parameters, $\beta_0$and β,

$$y = \beta_0 + \beta x + \varepsilon , \varepsilon \sim N(0, \sigma^2).$$

Here y is $\log_2$(coverage) and x is $\log_2$(allele size). Given a new data point from a real sample, the $\log_2$(coverage) for an STR locus of interest, the point estimate of the allele size is thus

$$\log_2\left(\widehat{allelesize}\right) = \frac{\log_2(\mathrm{cov}\,erage) - \hat{\beta}_0}{\hat{\beta}}\ .$$

where allele size is the number of base pairs inserted relative to the reference and coverage is the normalised number of reads allocated to the locus.

Reference data

Reference genome: ucsc.hg19.fasta, with STR decoy chromosomes added as described above.

STR positions in genome annotated bed file: hg19.simpleRepeat.txt.gz. Source: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/

Known STR loci are obtained by performing a Tandem Repeats Finder (TRF) [22] annotation of the reference genome. Pre-computed annotations of many genomes are available from the UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables)

[34]. TRF annotations are converted to bed files annotated with two additional columns: the repeat unit/motif and the number of repeat units in the reference.

Running other STR genotypers

To estimate the size of the shorter allele, LobSTR and HipSTR were run on bam files containing the locus of interest and 1000bp of flanking sequence on either side. We used LobSTR version 4.0.6 with default settings and the LobSTR reference genome and annotation hg19_v3.0.2. HipSTR version 0.4 was used --min-reads 2 and otherwise default settings, with with the provided hg19 reference genome and annotation. In some cases the tools make a different call as to the reference allele in hg19. To make variant calls comparable to STRetch calls we converted genotypes to number of repeat units inserted relative to the reference defined by that tool, then applied that to the reference allele given by STRetch. For example if the STRetch reference allele is 20 repeat units, while in HipSTR it is 10 repeat units, a HipSTR genotype of 10/15 would be reported as 20/25. All imperfect repeat units or other variation was ignored and only the total size of alleles taken.

ExpansionHunter version 2.0.9 was run on the entire mapped bam file using the GRCh37 reference genome. Each locus to be tested must be manually defined (3 loci are provided with the tool: ALS, FMR1 and HD). The configuration file for SCA8 was defined as follows:

```
{
  "RepeatId": "ATXN8",
  "RepeatUnit": "CTG",
  "TargetRegion": "chr13:70713516-70713561"
}
```

References

1. The 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73. doi:10.1038/nature09534.

2. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet. 2010;44:445–77. doi:10.1146/annurev-genet-072610-155046.

3. Orr HT, Zoghbi HY. Trinucleotide repeat disorders. Annu Rev Neurosci. 2007;30:575–621. doi:10.1146/annurev.neuro.29.051605.113042.

4. Hamada H, Seidman M, Howard BH, Gorman CM. Enhanced gene expression by the poly (dT-dG). poly (dC-dA) sequence. Mol Cell Biol. 1984;4:2622–30.

5. Li YY-CC, Korol AB, Fahima T, Beiles A, Nevo E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol. 2002;11:2453–65. doi:10.1046/j.1365-294X.2002.01643.x.

6. Gatchel JR, Zoghbi HY. Diseases of unstable repeat expansion: mechanisms and common principles. Nat Rev Genet. 2005;6:743–55.

7. van Eyk CL, Richards RI. Dynamic Mutations. In: Tandem Repeat Polymorphisms. Springer; 2012. p. 55–77.

8. Hannan AJ, editor. Tandem Repeat Polymorphisms: Genetic Plasticity, Neural Diversity and Disease. Austin/New York: Landes Bioscience/Springer Science+Business Media; 2012.

9. Mirkin SM. Expandable DNA repeats and human disease. Nature. 2007;447:932–40. doi:10.1038/nature05977.

10. Sherman SL, Jacobs PA, Morton NE, Froster-Iskenius U, Howard-Peebles PN, Nielsen KB, et al. Further segregation analysis of the fragile X syndrome with special reference to transmitting males. Hum Genet. 1985;69:289–99.

11. Margolis RL. The spinocerebellar ataxias: Order emerges from chaos. Curr Neurol Neurosci Rep. 2002;2:447–56. doi:10.1007/s11910-002-0072-8.

12. BL F, Lee H, JL D, al et. Exome sequencing in the clinical diagnosis of sporadic or familial cerebellar ataxia. JAMA Neurol. 2014;71:1237–46. http://dx.doi.org/10.1001/jamaneurol.2014.1944.

13. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler

for personal genomes. Genome Res. 2012;22:1154–62. doi:10.1101/gr.135780.111.

14. Willems T, Zielinski D, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nat Publ Gr. 2016; October 2016. doi:10.1101/077727.

15. Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. Nucleic Acids Res. 2013;41:e32. doi:10.1093/nar/gks981.

16. Ashley EA. Towards precision medicine.

17. Cao MD, Tasker E, Willadsen K, Imelfort M, Vishwanathan S, Sureshkumar S, et al. Inferring short tandem repeat variation from paired-end short reads. Nucleic Acids Res. 2014;42:e16. doi:10.1093/nar/gkt1313.

18. Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, van Blitterswijk M, Kingsbury Z, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. bioRxiv. 2016. http://biorxiv.org/content/early/2016/12/19/093831.abstract.

19. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015;517:608–11.

20. Gymrek M. ScienceDirect A genomic view of short tandem repeats. Curr Opin Genet Dev. 2017;44:9–16. doi:10.1016/j.gde.2017.01.012.

21. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Prepr arXiv. 2013;0:3. doi:arXiv:1303.3997 [q-bio.GN].

22. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573.

23. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In: Current Protocols in Bioinformatics. John Wiley & Sons, Inc.; 2002. doi:10.1002/0471250953.bi1110s43.

24. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Meth. 2009;6:677–81. http://dx.doi.org/10.1038/nmeth.1363.

25. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, et al. Large multiallelic copy number variations in humans. Nat Genet. 2015;47:296–303. http://dx.doi.org/10.1038/ng.3200.

26. Sadedin SP, Pope B, Oshlack A. Bpipe: a tool for running and managing bioinformatics pipelines. Bioinformatics. 2012;28:1525–6. doi:10.1093/bioinformatics/bts167.

27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9. doi:10.1093/bioinformatics/btp352.

28. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

29. Pedersen B. goleft. 2016. github.com/brentp/goleft.

30. Ripley BD. Modern applied statistics with S. Springer; 2002.

31. Huber PJ. Robust statistics. 1981.

32. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;:289–300.

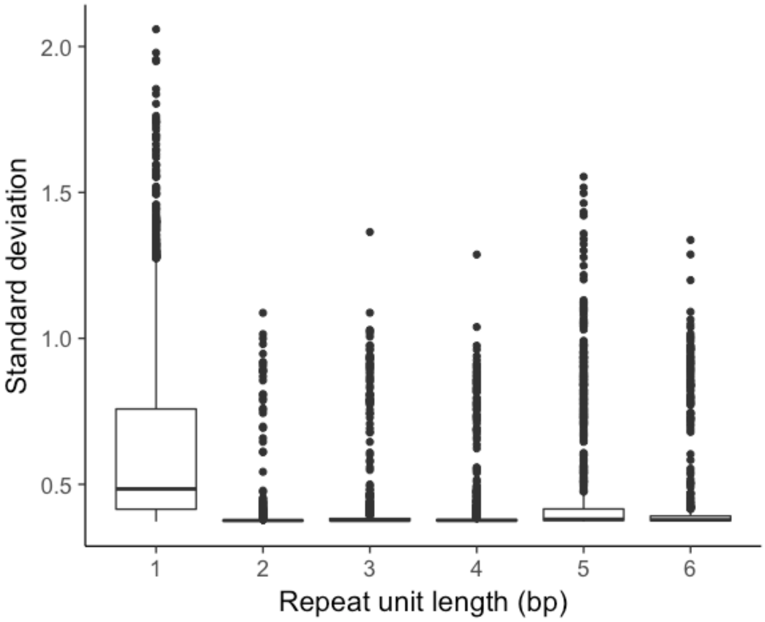
33. Huang W, Li L, Myers JR, Marth GT. ART: A next-generation sequencing read simulator. Bioinformatics. 2012;28:593–4.

34. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004;32 Database issue:D493-6.

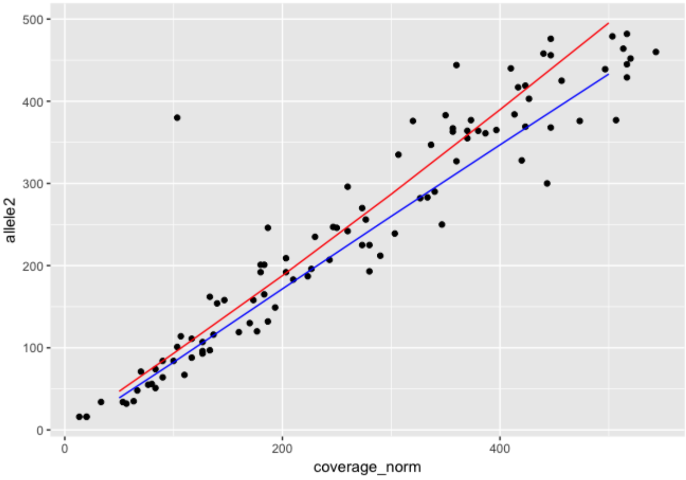
Supplementary tables and figures

Supplementary table 1: Pathogenic STR loci, positions in hg19. Also available as the bed file hg19.STR_disease_loci.bed on Figshare along with the other reference data at https://figshare.com/s/1a39be9282c90c4860cd.

Chromosome	Start	End	Disease	Gene
chr3	63898361	63898392	SCA7	ATXN7
chr3	128891419	128891502	DM2	ZNF9
chr4	3076604	3076695	HD	HTT
chr4	41747993	41748039	CCHS	PHOX2B
chr5	146258291	146258322	SCA12	PPP2R2B
chr6	16327865	16327955	SCA1	ATXN1
chr6	170870995	170871105	SCA17	TBP
chr9	27573482	27573544	FTDALS1	C9orf72
chr9	71652203	71652205	FRDA	FXN
chr12	7045880	7045938	DRPLA	ATN1
chr12	50898785	50898805	FRA12A	DIP2B
chr12	112036754	112036823	SCA2	ATXN2
chr13	70713484	70713561	SCA8	ATXN8/ATXN8OS
chr14	23790681	23790701	OPMD	PAPBN1
chr14	92537355	92537397	SCA3/MJD	ATXN3
chr16	87637889	87637935	HDL2	JPH3
chr19	13318673	13318712	SCA6	CACNA1A
chr19	46273462	46273524	DM1	DMPK
chr20	2633379	2633421	SCA36	NOP56
chr22	46191235	46191304	SCA10	ATXN10
chrX	66765159	66765261	SBMA	AR
chrX	146993555	146993629	FXTAS	FMR1
chrX	147582125	147582273	FRAXE	AFF2



Supplementary figure 1: Robust standard deviation of STR reads assigned to each locus across 97 WGS samples. Homopolymer loci are the most variable between individuals.



Supplementary figure 2: The plot shows the simulated data (black) with the upper (red) and lower (blue) bounds for the linear fit indicated. A plot of the number of reads mapping to the AGC decoy chromosome against the number of AGC repeat units inserted into the ATXN8 locus shows a clear linear relationship between these two variables.