Detecting disease-causing repeat expansions in next-generation sequencing data

Rick M. Tankard[1,2], Mark F Bennett[2], Peter Degorski[2], Martin B. Delatycki[3,4,5], Paul J. Lockhart[3,4] and Melanie Bahlo[2,3]

[1]Murdoch University, [2]The Walter and Eliza Hall Institute of Medical Research, Melbourne, [3]The University of Melbourne,
[4]Murdoch Childrens Research Institute, Melbourne, [5]The Royal Children's Hospital, Melbourne

Introduction

Repeat expansion disorders are responsible for over 20 human neurological disorders, including Huntington disease, spinocerebellar ataxias and intellectual disabilities. These are caused by expansions in DNA of short tandem repeats (STRs). Genetic diagnosis of repeat expansion disorders is slow and costly, as each disorder requires a specific test.

Next generation sequencing (NGS) is now common in the diagnosis and screening of genetic diseases. NGS allows detection of a wide-range of genetic mutations, but methods to detect repeat expansions are in their infancy. The ability to diagnose repeat expansions with NGS has the potential to reduce costs and provide faster results to patients.

Examples of repeat expansion disorders

- Fragile X site A
  - » Intellectual disability
  - » Most common: 1 of 4,000–6,000 males
- Huntington disease
  - » 1 of 10,000–20,000 affected
- Spinocerebellar ataxias (dominant)
  - » 1 of 15,000–100,000 affected
  - » Coordination difficulties
  - » More than 10 caused by repeat expansions
  - » Approximately 30 other rarer genetic causes

Short-tandem repeat (STR) and expansions

CAG motif:

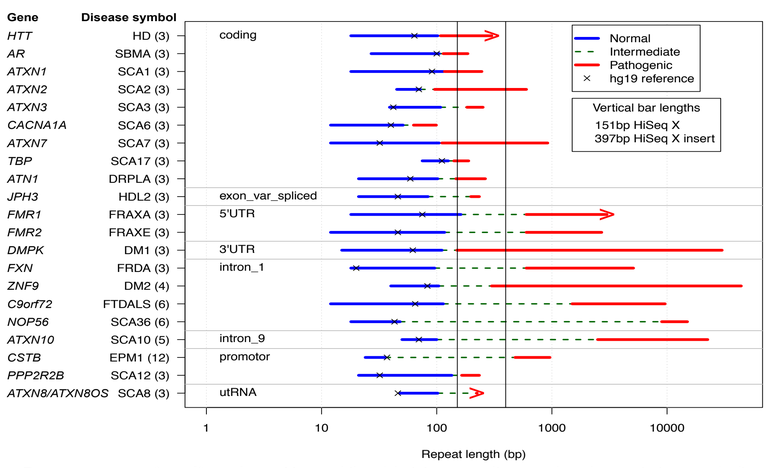
...CTCAAGTCCTTC**CAGCAGCAGCAGCAGCA**ACAGCCGCCACCGC...
...CTCAAGTCCTTC**CAGCAGCAGCAGCAGCAGCAGCAGCAGCA**ACAGCCGCCACCGC...

AC motif:

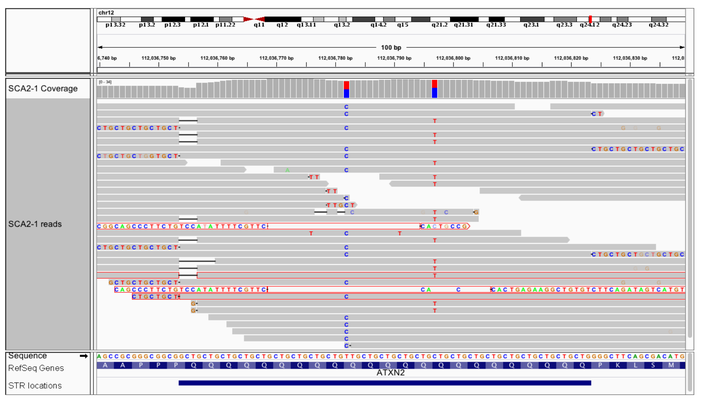
...AGAGATAG**ACACACACACACACACA**AACAAGCAT...
...AGAGATAG**ACACACACACACACACACACACACACACA**AACAAGCAT...

GGGCCT motif:

...GCCGCAGACA**GGGCCTGGGCCTGGGCCTG**CCCTGGGAA...
...GCCGCAGACAG**GGCCTGGGCCTGGGCCTGGGCCTGGGCCTG**CCCTGGGAA...



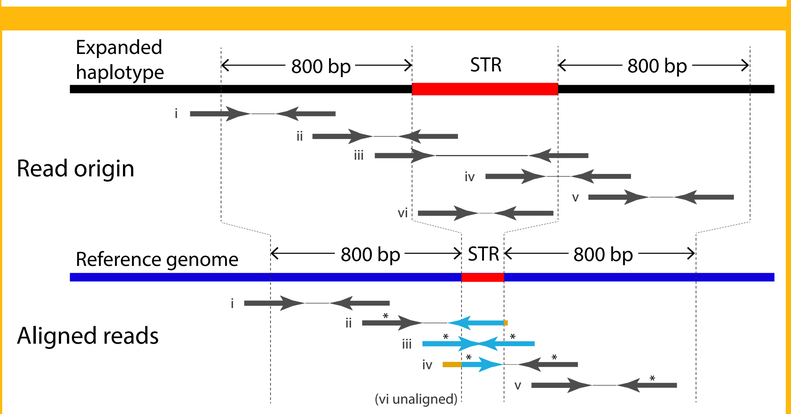
Repeat expansion disorders. *Normal (healthy), intermediate and pathogenic (expanded) allele size ranges in bp on logarithmic scale. Diseases are HD: Huntington disease, SBMA: Kennedy disease, SCA#: Spinocerebellar type #, DRPLA: Dentatorubral-pallidoluysian atrophy, HDL2: Huntington disease-like 2, FRAX*: Fragile-X site *, DM#: Myotonic dystrophy #, FRDA: Friedreich ataxia, EPM1: Myoclonic epilepsy of Unverricht and Lundborg. The number in brackets after each disease indicates the number of bases in each repeat unit of the locus.*[1,2]



IGV screenshot of the SCA2 repeat expansion locus of a WGS sample known to have this disease. The T>C SNP suggests the expanded allele is a pure repeat unlike the reference.

exSTRa (expanded STR algorithm)

We created an R package exSTRa[3] to visualise and classify repeat expansions. An accompanying Perl module is first used to count the repeat content in NGS data, focusing on known locations of repeat expansions.



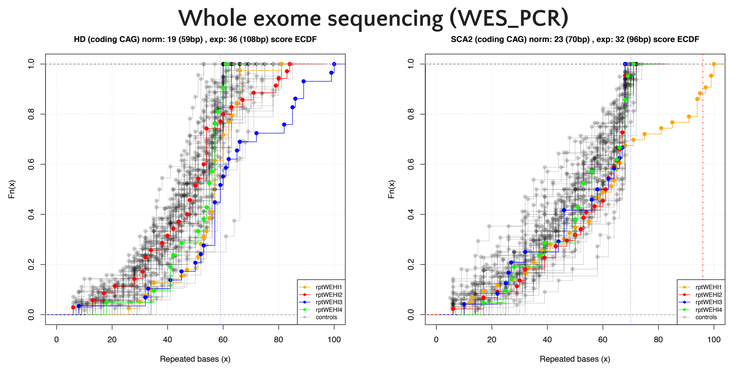
Identifying reads for repeat scoring. *The expanded haplotype represents an actual DNA sequence, with reads labeled from i to v. Reads are aligned by software to a reference genome. Reads marked with * are further analysed due to their proximity and direction to the STR; when their pair appears to overlap the STR (blue/orange), its repeat content is counted.*

Each read overlapping the STR gives one observation. Longer repeats typically have more reads with larger repeat scores. Repeat expansions are classified using an outlier detection method.

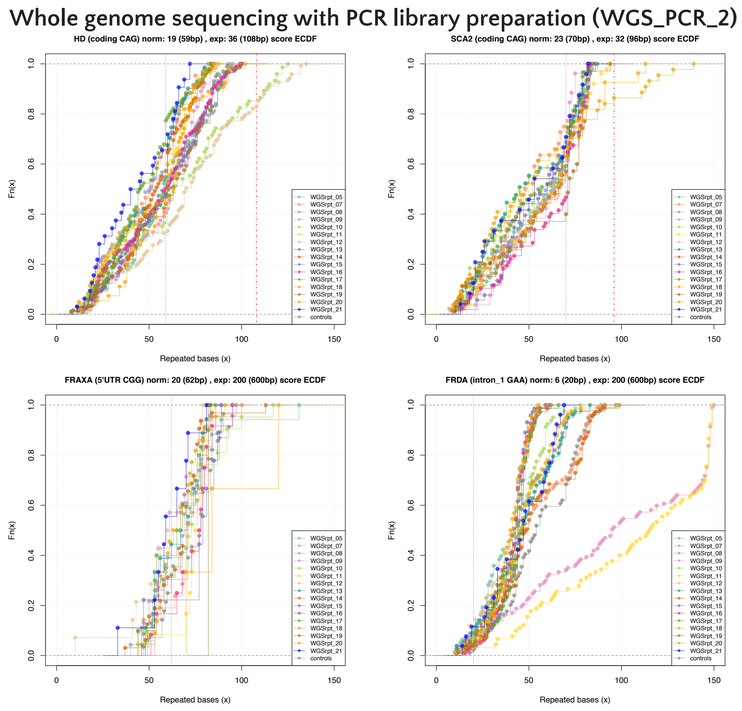
ECDFs of repeat score

ECDF visualisation of repeat score (x) for several disease loci. Expansions can be observed by a shift to the right on larger quantiles.

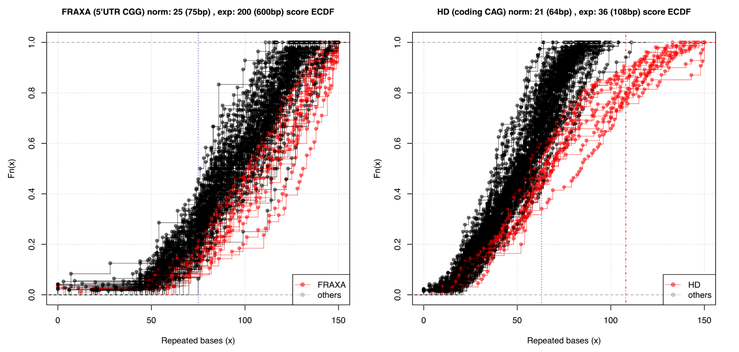
Whole exome sequencing (WES_PCR)



Whole genome sequencing with PCR library preparation (WGS_PCR_2)



Whole genome sequencing with PCR-free library preparation (WGS_PF)



We detect outliers by first creating a test statistic T, which is the sum of multiple t-statistics on quantiles q>0.5. We determine significance through simulation of a null background, based on robust estimates of the location and spread (median and MAD) at quantiles.

Performance

We have tested exSTRa[3], ExpansionHunter[4], STRetch[5] and TREDPARSE[6] on 4 cohorts with Illumina sequencing including WES (Agilent SureSelect V5+UTR), and WGS (HiSeq X) with and without a PCR-free library protocol.

Cohort	Cases	Controls	Method	TP	FN	TN	FP	Sens	Spec
WES_PCR (10 loci)	4	58	exSTRa	4	0	607	9	1	0.99
			ExpansionHunter	2	2	616	0	0.5	1
			STRetch	3	1	613	3	0.75	1
			TREDPARSE-T	4	0	585	31	1	0.95
			TREDPARSE-L	4	0	574	42	1	0.93
WGS_PCR_1	3	14	exSTRa	2	1	343	11	0.67	0.97
			ExpansionHunter	3	0	354	0	1	1
			STRetch	1	2	336	1	0.33	1
			TREDPARSE-T	3	0	354	0	1	1
			TREDPARSE-L	3	0	354	0	1	1
WGS_PCR_2	16	2	exSTRa	13	3	352	10	0.81	0.97
			ExpansionHunter	8	8	362	0	0.5	1
			STRetch	11	5	338	6	0.69	0.98
			TREDPARSE-T	12	4	362	0	0.75	1
			TREDPARSE-L	11	5	362	0	0.69	1
WGS_PCR_2 _30X_1	16	2	exSTRa	13	3	357	5	0.81	0.99
			ExpansionHunter	8	8	362	0	0.5	1
			STRetch	11	5	340	4	0.69	0.99
			TREDPARSE-T	13	3	362	0	0.81	1
			TREDPARSE-L	9	7	362	0	0.56	1
WGS_PCR_2 _30X_2	16	2	exSTRa	12	4	354	8	0.75	0.98
			ExpansionHunter	8	8	362	0	0.5	1
			STRetch	11	5	336	8	0.69	0.98
			TREDPARSE-T	13	3	362	0	0.81	1
			TREDPARSE-L	10	6	362	0	0.62	1
WGS_PF*	78	40	exSTRa	60	18	2329	71	0.77	0.97
			ExpansionHunter**	62	16	2394	6	0.79	1
			STRetch	62	16	2206	76	0.79	0.97
			TREDPARSE-T	52	26	2383	17	0.67	0.99
	66	52	TREDPARSE-L	34	32	2396	16	0.52	0.99
WGS_PF (FRAXA pre)*	96	22	exSTRa	63	33	2314	68	0.66	0.97
			ExpansionHunter**	95	1	2374	8	0.99	1
			STRetch	62	34	2188	76	0.65	0.97
			TREDPARSE-T	72	24	2364	18	0.75	0.99
	72	46	TREDPARSE-L	48	24	2383	23	0.67	0.99
WGS_PF (no FRAXA)*	62	56	exSTRa	52	10	2231	67	0.84	0.97
			ExpansionHunter**	61	1	2292	6	0.98	1
			STRetch	62	0	2104	76	1	0.97
			TREDPARSE-T	52	10	2281	17	0.84	0.99
	51	67	TREDPARSE-L	34	17	2293	16	0.67	0.99

*Controls have no known repeat expansion, that is, do not have symptoms for a repeat expansion disorder. Non-expanded loci in cases are negative conditions. Tests Bonferroni corrected by number of tests at each locus (including STRetch results). TREDPARSE-T is a repeat expansion size threshold method. TREDPARSE-L uses a likelihood ratio test based method and uses the inheritance model (recessive loci require both alleles expanded). *The WGS_PF cohort is from Dolzhenko et al[4]. These were also tested using a lower premuation threshold for FRAXA and without FRAXA. **ExpansionHunter results of WGS_PF cohort used original Dolzhenko et al results that were aligned with a different aligner (Issac).*

Overall, we found that no particular method outperformed the others. The correctly identified expansions differed between methods, suggesting the use of multiple methods to obtain greater sensitivity. Findings from these methods will require validation, but the use of these methods can significantly reduce loci requiring expensive testing.

We have shown that repeat expansions can be detected in WGS data with or without PCR-free library preparation, as well WES data. These methods may be used on preexisting NGS data to find repeat expansions that were previously difficult to find.

The exSTRa package is available from:
https://github.com/bahlolab/exSTRa

Acknowledgments

We thank Egor Dolzhenko and Michael Eberle (Illumina) for access to the WGS_PF cohort dataset EGA00001003562 from the European Genome-Phenome Archive, along with specification files for ExpansionHunter. Additional control datasets provided by Leslie Burnett, Ben Lundie, Katie Ayres, and Andrew Sinclair. Kate Pope and Greta Gillies assisted with recruitment and sample preparation.

References

1. Strachan, T. & Read, A. P. Human Molecular Genetics. (Garland Science: 2011).

2. Fondon, J. W., Hammock, E. A. D., Hannan, A. J. & King, D. G. Simple sequence repeats: genetic modulators of brain function and behavior. Trends Neurosci. 31, 328–334 (2008).

3. Tankard, R. M. et al., Detecting expansions of tandem repeats in Cohorts sequenced with Short-Read Sequencing Data, The American Journal of Human Genetics (2018), https://doi.org/10.1016/j.ajhg.2018.10.015 (in press)

4. Dolzhenko, E., et al.; US–Venezuela Collaborative Research Group. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. 27, 1895–1903 (2017).

5. Dashnow, H., et al. STRetch: Detecting and discovering pathogenic short tandem repeat expansions. Genome Biol. 19, 121 (2018).

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