Sequence analysis

REscan: inferring repeat expansions and structural variation in paired-end short read sequencing data

Russell Lewis McLaughlin 1,∗

1 Smurfit Institute of Genetics, Trinity College Dublin, Dublin, D02 PN40, Ireland.

∗To whom correspondence should be addressed.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Repeat expansions are an important class of genetic variation in neurological diseases. However, the identification of novel repeat expansions using conventional sequencing methods is a challenge due to their typical lengths relative to short sequence reads and difficulty in producing accurate and unique alignments for repetitive sequence. However, this latter property can be harnessed in paired-end sequencing data to infer the possible locations of repeat expansions and other structural variation.

Results: This paper presents REscan, a command line utility that infers repeat expansion loci from paired-end short read sequencing data by reporting the proportion of reads orientated towards a locus that do not have an adequately mapped mate. A high REscan statistic relative to a population of data suggests a repeat expansion locus for experimental follow-up. This approach is validated using genome sequence data for 259 cases of amyotrophic lateral sclerosis, of which 24 are positive for a large repeat expansion in C9orf72, showing that REscan statistics readily discriminate repeat expansion carriers from non-carriers.

Availability: C source code at https://github.com/rlmcl/rescan (GNU General Public Licence v3).

Contact: mclaugr@tcd.ie

1 Introduction

Repeat expansions (REs) are genetic variants characterised by an increase in the number of units in a tandem repeat sequence (figure 1A). REs are causative in many human diseases, especially nervous system disorders (van Blitterswijk et al., 2012), suggesting convergent pathological mechanisms driven by long stretches of repetitive DNA in certain transcripts. The discovery of novel REs is thus important for understanding the biology of disease but can be challenging using conventional methods. For example, expanded alleles may not amplify by polymerase chain reaction (PCR) due to their length and, as often seen in neurological disorders, high GC content. With short read sequencing, expanded alleles longer than the read length lead to poor alignment and false negative variant calls. However, with paired-end data, poorly-mapped reads from the edges of expanded alleles often have mates with unique sequences that adequately align to the locus (‘anchoring’ reads). This is harnessed by algorithms such as ExpansionHunter (Dolzhenko et al., 2017) and STRetch (Dashnow et al., 2018); however, ExpansionHunter requires prior knowledge of the nature of the locus and the RE (which is not necessarily available for novel REs), and STRetch relies on computationally-intensive realignment of reads to a reference genome supplemented with RE-containing decoy chromosomes. For pipelines aimed at efficient identification of novel REs from genome sequence data, a tool that reports observational data on unmapped reads without the need for a priori knowledge or preprocessing is useful. These data can then be used to statistically infer RE loci for technical and experimental validation. This paper presents REscan, a command-line tool for exploratory analysis of paired-end sequence data to infer the locations of long REs and other structural variation.

2 Methods

2.1 The REscan algorithm

REscan reads a position-sorted SAM-format stream from standard input and reports a straightforward, locus-wise statistic \( r_x / r_t \), where \( r_t \) is the total number of nearby reads orientated towards the locus and \( r_x \) is the number with a poorly mapped mate (figure 1B). The window within which a mate is considered nearby is modifiable (default 200 bases) and a mate is considered poorly mapped if any of the following conditions are met:

- The next segment in template is unmapped (bit 0x8 in FLAG is set);
- The mate’s mapping quality is lower than 20 or user-specified value;

© The Author(s) 2020. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
3.1 Discovery pipeline for novel repeat expansions

REscan provides a flexible method for generating exploratory data to enable a discovery pipeline for novel REs, incorporating the following steps: 1) REscan is run individually per-sample for (e.g.) all transcribed regions; 2) Resulting data are appropriately processed (e.g. normalised/standardised) and statistically analysed in a regression framework or similar to identify candidate RE loci associated with the trait of interest; 3) Candidate loci are cross-referenced against structural variants called by alternative algorithms to filter false positives; 4) The likely repeat motif is obtained from reference genome data, and common off-target alignment regions are identified from sequence data; 5) Using candidates retained in step 3 along with information obtained from step 4, RE lengths are estimated in a more comprehensive framework such as ExpansionHunter; 6) Expanded alleles are validated experimentally using Southern blot, repeat-primed PCR or targeted long-read sequencing.

Acknowledgements

Thanks to ALS patients, Laura O’ Briain and Orla Hardiman (TCD).

Funding

Supported by Science Foundation Ireland (17/CDA/4737).

References

Byrne S. et al. (2012) Cognitively and clinically characteristic patients with amyotrophic lateral sclerosis carrying a C9orf72 repeat expansion: a population-based cohort study. *Lancet Neurol.* 11(3), 232–240.

Dushnow JR et al. (2018) STRech: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.* 19(1), 121.

Dolsken E et al. (2017) Detection of long repeat expansions from PCR-free whole-genome sequence data. *Bioinformatics.* 29(16), 2041–2043.

van Blitterswijk M et al. (2012) How do C9ORF72 repeat expansions cause amyotrophic lateral sclerosis and frontotemporal dementia: can we learn from other noncoding repeat expansion disorders? *Cure Open Neurol.* 2(6), 689–700.

van Rheenen W et al. (2018) Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *Eur. J. Hum. Genet.* 26(10), 1537–1546.