Bioinformatics, 35(22), 2019, 4754–4756
doi: 10.1093/bioinformatics/btz431
Advance Access Publication Date: 27 May 2019
Applications Note

Genome analysis

ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions

Egor Dolzhenko [1], Viraj Deshpande[1], Felix Schlesinger[1], Peter Krusche[2], Roman Petrovski[2], Sai Chen[1], Dorothea Emig-Agius[1], Andrew Gross[1], Giuseppe Narzisi[3], Brett Bowman[1], Konrad Scheffler[1], Joke J. F. A. van Vugt[4], Courtney French[5], Alba Sanchis-Juan[6,7], Kristina Ibáñez[8], Arianna Tucci[8], Bryan R. Lajoie[1], Jan H. Veldink[4], F. Lucy Raymond[5], Ryan J. Taft[1], David R. Bentley[2] and Michael A. Eberle[1,*]

[1]Illumina Inc., San Diego, CA 92122, USA, [2]Illumina Cambridge Ltd, Illumina Centre, 19 Granta Park, Great Abington, Cambridge CB21 6DF, UK, [3]Computational Biology, New York Genome Center, New York, NY 10013, USA, [4]UMC Utrecht Brain Center, Utrecht University, 3508 AB Utrecht, The Netherlands, [5]Department of Medical Genetics, [6]Department of Haematology, University of Cambridge, NHS Blood and Transplant Centre, Cambridge, CB2 0PT, UK, [7]NIHR BioResource, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK and [8]Genomics England, Queen Mary University London, London EC1M 6BQ, UK

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on March 10, 2019; revised on April 26, 2019; editorial decision on May 15, 2019; accepted on May 23, 2019

Abstract

Summary: We describe a novel computational method for genotyping repeats using sequence graphs. This method addresses the long-standing need to accurately genotype medically important loci containing repeats adjacent to other variants or imperfect DNA repeats such as polyalanine repeats. Here we introduce a new version of our repeat genotyping software, ExpansionHunter, that uses this method to perform targeted genotyping of a broad class of such loci.

Availability and implementation: ExpansionHunter is implemented in C++ and is available under the Apache License Version 2.0. The source code, documentation, and Linux/macOS binaries are available at https://github.com/Illumina/ExpansionHunter/.

Contact: meberle@illumina.com

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

1 Introduction

Short tandem repeats (STRs) are ubiquitous throughout the human genome. Although our understanding of STR biology is far from complete, emerging evidence suggests that STRs play an important role in basic cellular processes (Gymrek *et al.*, 2016; Hannan, 2018). In addition, STR expansions are a major cause of over 20 severe neurological disorders including amyotrophic lateral sclerosis, Friedreich ataxia (FRDA) and Huntington's disease (HD).

ExpansionHunter was the first computational method for genotyping STRs from short-read sequencing data capable of consistently genotyping repeats longer than the read length and, hence, detecting pathogenic repeat expansions (Dolzhenko *et al.*, 2017). Since the initial release of ExpansionHunter, several other methods have been developed and were shown to accurately identify long (greater than read length) repeat expansions (Dashnow *et al.*, 2018; Mousavi *et al.*, 2019; Tang *et al.*, 2017; Tankard *et al.*, 2018).

Current methods are not designed to handle complex loci that harbor multiple repeats. Important examples of such loci include the CAG repeat in the *HTT* gene that causes HD flanked by a CCG repeat, the GAA repeat in *FXN* that causes FRDA flanked by an adenine homopolymer and the CAG repeat in *ATXN8* that causes Spinocerebellar ataxia type 8 (SCA8) flanked by an ACT repeat. An even more extreme example is the CAGG repeat in the *CNBP* gene whose expansions cause Myotonic Dystrophy type 2(DM2). This repeat is adjacent to polymorphic CA and CAGA repeats (Liquori *et al.*, 2001) making it particularly difficult to accurately align reads to this locus. Another type of complex repeat is the polyalanine repeat which has been associated with at least nine disorders to date (Shoubridge and Gecz, 2012). Polyalanine repeats consist of repetitions of α-amino acid codons GCA, GCC, GCG or GCT (i.e. GCN).

Clusters of variants can affect alignment and genotyping accuracy (Lincoln *et al.*, 2019). Variants adjacent to low complexity polymorphic sequences can be additionally problematic because methods for variant discovery can output clusters of inconsistently represented or spurious variant calls in such genomic regions. This, in part, is due to the elevated error rates of such regions in sequencing data (Benjamini and Speed, 2012; Dolzhenko *et al.*, 2017). One example is a single-nucleotide variant (SNV) adjacent to an adenine homopolymer in *MSH2* that causes Lynch syndrome I (Froggatt *et al.*, 1999).

Here we present a new version (v3.0.0) of ExpansionHunter that was reimplemented to handle complex loci such as those described above. The implementation uses sequence graphs (Dilthey *et al.*, 2015; Garrison *et al.*, 2018; Paten *et al.*, 2017) as a general and flexible model of each target locus.

2 Implementation

ExpansionHunter works on a predefined variant catalog containing genomic locations and the structure of a series of targeted loci. For each locus, the program extracts relevant reads (Dolzhenko *et al.*, 2017) from a binary alignment/map file (Li *et al.*, 2009) and realigns them using a graph-based model representing the locus structure. The realigned reads are then used to genotype each variant at the locus (Fig. 1).

The locus structure is specified using a restricted subset of the regular expression syntax. For example, the *HTT* repeat region linked to HD can be defined by expression (CAG)*CAACAG(CCG)* that signifies that it harbors variable numbers of the CAG and CCG repeats separated by a CAACAG interruption (see Supplementary Materials); the *FXN* repeat region linked to the FRDA corresponds to expression (A)*(GAA)*; the *ATXN8* repeat region linked to SCA8 corresponds to (CTA)*(CTG)*; the *CNBP* repeat region linked to DM2 consists of three adjacent repeats defined by (CAGG)*(CAGA)*(CA)*; the *MSH2* SNV adjacent to an adenine homopolymer that causes Lynch syndrome I corresponds to (A|T)(A)*.

Additionally, the regular expressions are allowed to contain multi-allelic or 'degenerate' base symbols that can be specified using the International Union of Pure and Applied Chemistry notation (Cornish-Bowden, 1985). Degenerate bases make it possible to represent certain classes of imperfect DNA repeats where, e.g. different bases may occur at the same position. Using this notation, polyalanine repeats can be encoded by the expression (GCN)* and polyglutamine repeats can be encoded by the expression (CAR)*.

ExpansionHunter translates each regular expression into a sequence graph. Informally, a sequence graph consists of nodes that correspond to sequences and directed edges that define how these sequences can be connected together to assemble different alleles.

We implemented the basic sequence graph functionality used by ExpansionHunter in the GraphTools C++ library (Supplementary Materials). One of the key features of the library is its support for single-node loops in contrast to the traditional approaches that use fully acyclic graphs (Lee *et al.*, 2002). Single-node loops are the key to representing STRs and other sequences that can appear in any number of copies.

Genotyping is performed by analyzing the alignment paths associated with the presence or absence of each constituent allele. The repeats are genotyped as before (Dolzhenko *et al.*, 2017) and SNVs/indels are genotyped using a straightforward Poisson-based model (Supplementary Materials).

3 Results and discussion

To demonstrate the performance of ExpansionHunter we analyzed multiple complex STR regions. First, we analyzed a simulated dataset containing a wide range of CAG and CCG repeat sizes at the *HTT* locus. As expected, the accuracy of ExpansionHunter was substantially higher when the reads were aligned to a sequence graph that included both repeats compared to when the repeats were analyzed independently (Supplementary Fig. S2). ExpansionHunter also produced more accurate genotypes compared to other tools that were not designed to

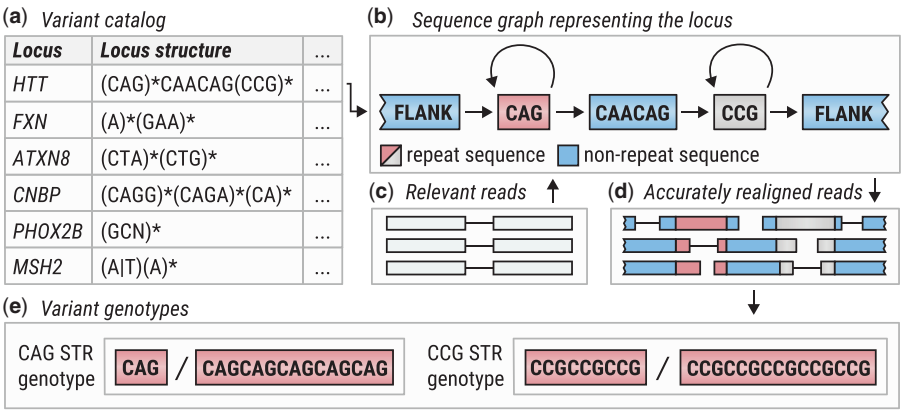

Fig. 1. Overview of ExpansionHunter. (**a**) A locus definition is read from the variant catalog file. (**b**) Sequence graph is constructed according to its specification in the variant catalog. (**c**) Relevant reads are extracted from the input binary alignment/map file. (**d**) Reads are aligned to the graph. (**e**) Alignments are pieced together to genotype each variant

handle loci harboring multiple nearby STRs, GangSTR and TREDPARSE (Supplementary Fig. S2). A recent study used ExpansionHunter to investigate mutations in the short sequence interrupting two repeats in the *HTT* locus across 1600 samples (Wright *et al.*, 2019) demonstrating usefulness of the program for analysis of complex loci in real data. ExpansionHunter also correctly detected the pathogenic SNV adjacent to an adenine homopolymer in the *MSH2* gene in three WGS replicates of a sample obtained from SeraCare Life Sciences (Supplementary Materials).

To demonstrate the utility of ExpansionHunter across both short and long repeats, we compared calls from ExpansionHunter, GangSTR and TREDPARSE on sequence data from samples with experimentally confirmed repeat expansions (Supplementary Materials and Fig. S3). ExpansionHunter had better accuracy (precision =0.91, recall =0.99) in detecting the expanded repeats in this dataset compared to GangSTR (precision =0.88, recall =0.83) and TREDPARSE (precision =0.84, recall =0.46).

Finally, we used ExpansionHunter to genotype degenerate DNA repeats by analyzing a polyalanine repeat in *PHOX2B* gene in 150 healthy controls and one sample harboring a known pathogenic expansion. *PHOX2B* contains a polyalanine repeat of 20 codons that can expand to cause congenital central hypoventilation syndrome. Consistent with what is known about this repeat (Amiel *et al.*, 2003), all but a few controls were genotyped 20/20. ExpansionHunter accurately genotyped the sole sample with the expansion as 20/27; the correctness of this genotype was confirmed by Sanger sequencing.

In summary, we have developed a novel method that addresses the need for more accurate genotyping of complex loci. This method can genotype polyalanine repeats and resolve difficult regions containing repeats in close proximity to small variants and other repeats. A catalog of difficult regions is supplied with the software and can be extended by the user. We expect that the flexibility of the sequence graph framework now adopted in ExpansionHunter will enable a variety of novel variant calling applications.

Conflict of Interest: none declared.

Acknowledgements

We would like to thank Dr Erik Garrison and two anonymous reviewers for the insightful comments and for suggesting great future directions for this project. We are grateful to Dr Stacie Taylor and Dr Kirsten Curnow for the help with preparing the manuscript. We would also like to acknowledge Dr Christopher Schröder and Prof. Christel Depienne from Essen University Hospital for helping us to improve the program.

References

Amiel,J. *et al.* (2003) Polyalanine expansion and frameshift mutations of the paired-like homeobox gene PHOX2B in congenital central hypoventilation syndrome. *Nat. Genet.*, **33**, 459.

Benjamini,Y. and Speed,T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*, **40**, e72.

Cornish-Bowden,A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.*, **13**, 3021.

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

Dilthey,A. *et al.* (2015) Improved genome inference in the MHC using a population reference graph. *Nat. Genet.*, **47**, 682.

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

Froggatt,N.J. *et al.* (1999) A common MSH2 mutation in English and North American HNPCC families: origin, phenotypic expression, and sex specific differences in colorectal cancer. *J. Med. Genet.*, **36**, 97–102.

Garrison,E. *et al.* (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.*, **36**, 875–879.

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

Hannan,A.J. (2018) Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.*, **19**, 286.

Lee,C. *et al.* (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–464.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lincoln,S.E. *et al.* (2019) A rigorous interlaboratory examination of the need to confirm next-generation sequencing–detected variants with an orthogonal method in clinical genetic testing. *J. Mol. Diagn.*, **21**, 318–329.

Liquori,C.L. *et al.* (2001) Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. *Science*, **293**, 864–867.

Mousavi,N. *et al.* (2019) Profiling the genome-wide landscape of tandem repeat expansions. *bioRxiv.*, doi: https://doi.org/10.1101/361162.

Paten,B. *et al.* (2017) Genome graphs and the evolution of genome inference. *Genome Res.*, **27**, 665–676.

Shoubridge,C. and Gecz,J. (2012). *Polyalanine Tract Disorders and Neurocognitive Phenotypes*. Springer, New York, NY, pp. 185–203.

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.

Wright,G.E.B. *et al.* (2019) Length of Uninterrupted CAG, Independent of Polyglutamine Size, Results in Increased Somatic Instability, Hastening Onset of Huntington Disease. *Am. J. Hum. Genet.*, **104**, 1116–1126.