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

Egor Dolzhenko[1] (edolzhenko@illumina.com)
Viraj Deshpande[1] (vdeshpande@illumina.com)
Felix Schlesinger[1] (fschlesinger@illumina.com)
Peter Krusche[2] (pkrusche@gmail.com)
Roman Petrovski[2] (RPetrovski@illumina.com)
Sai Chen[1] (schen6@illumina.com)
Dorothea Emig-Agius[1] (dagius@illumina.com)
Andrew Gross[1] (agross@illumina.com)
Giuseppe Narzisi[3] (gnarzisi@nygenome.org)
Brett Bowman[1] (bbowman@illumina.com)
Konrad Scheffler[1] (kscheffler@illumina.com)
Joke J.F.A. van Vugt[4] (J.F.A.vanVugt-2@umcutrecht.nl)
Courtney French[5] (cf458@cam.ac.uk)
Alba Sanchis-Juan[6,7] (as2635@cam.ac.uk)
Kristina Ibáñez[8] (kristina.ibanez-garikano@genomicsengland.co.uk)
Arianna Tucci[8] (arianna.tucci@genomicsengland.co.uk)
Bryan Lajoie[1] (blajoie@illumina.com)
Jan H. Veldink[4] (J.H.Veldink@umcutrecht.nl)
Lucy Raymond[5] (flr24@cam.ac.uk)
Ryan J. Taft[1] (rtaft@illumina.com)
David R. Bentley[2] (DBentley@illumina.com)
Michael A. Eberle[1] (meberle@illumina.com)

[1]Illumina Inc., 5200 Illumina Way, San Diego, CA, USA
[2]Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, UK
[3]New York Genome Center, 101 Avenue of the Americas, New York, NY, USA
[4]Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands
[5]Department of Medical Genetics, University of Cambridge, Cambridge, UK
[6]Department of Haematology, University of Cambridge, Cambridge, CB2 0PT, UK
[7]NIHR BioResource, Cambridge University Hospitals NHS Foundation Trust, Cambridge, CB2 0QQ, UK
[8]Genomics England, Queen Mary University London, Dawson Hall, London, EC1M 6BQ

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

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 (Hannan 2018; Gymrek et al. 2016). 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; Tang et al. 2017; Tankard et al. 2018; Mousavi et al. 2019).

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 re-implemented to handle complex loci such as those described above. The implementation uses sequence graphs (Garrison et al. 2018; Paten et al. 2017; Dilthey et al. 2015) as a general and flexible model of each target locus.

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 (BAM) 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 (Figure 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 is 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 (IUPAC) notation (“Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences. Recommendations 1984. Nomenclature Committee of the International Union of Biochemistry (NC-IUB)” 1986). Degenerate bases make it possible to represent certain classes of imperfect DNA repeats where, for example, 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, Grasso, and Sharlow 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).

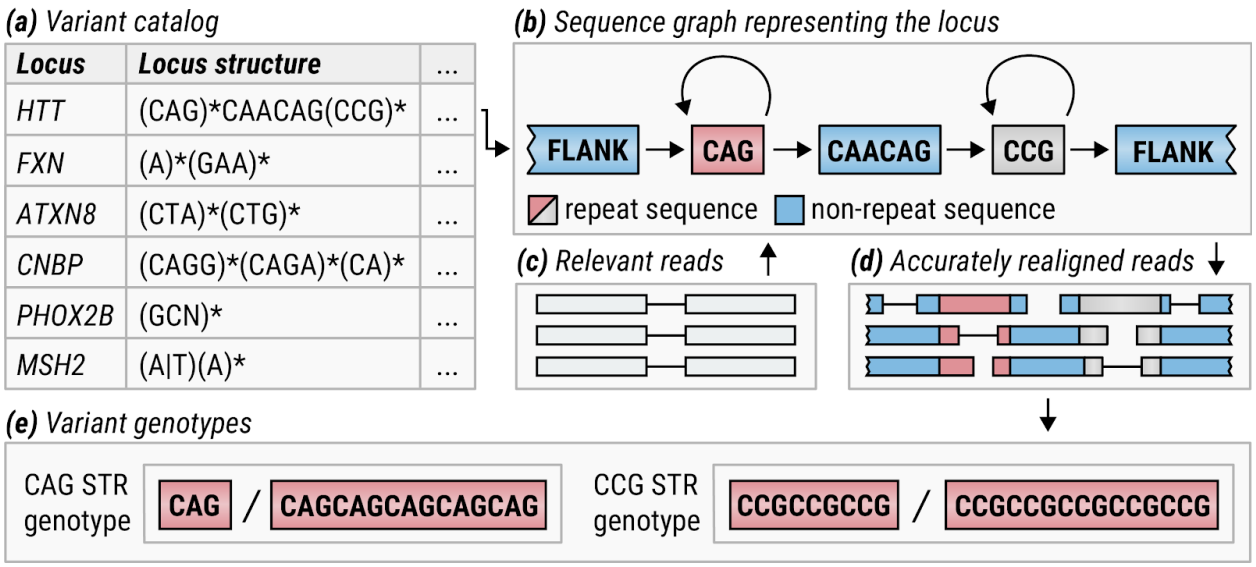

Figure 1: Overview of ExpansionHunter. (a) A locus definition is read from the variant catalog file. (b) Sequence graph is constructed according to its specification in the variant catalog. (c) Relevant reads are extracted from the input BAM file. (d) Reads are aligned to the graph. (e) Alignments are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

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 Figure S2). ExpansionHunter also produced more accurate genotypes compared to other tools that were not designed to handle loci harboring multiple nearby STRs, GangSTR and TREDPARSE (Supplementary Figure S2). A recent study used ExpansionHunter to investigate mutations in the short sequence interrupting two repeats in the *HTT* locus across 1,600 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 Figure 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.

References

Amiel, Jeanne, Béatrice Laudier, Tania Attié-Bitach, Ha Trang, Loïc de Pontual, Blanca Gener, Delphine Trochet, et al. 2003. “Polyalanine Expansion and Frameshift Mutations of the Paired-like Homeobox Gene PHOX2B in Congenital Central Hypoventilation Syndrome.” *Nature Genetics* 33 (4): 459–61.

Benjamini, Yuval, and Terence P. Speed. 2012. “Summarizing and Correcting the GC Content Bias in High-Throughput Sequencing.” *Nucleic Acids Research* 40 (10): e72.

Dashnow, Harriet, Monkol Lek, Belinda Phipson, Andreas Halman, Simon Sadedin, Andrew Lonsdale, Mark Davis, et al. 2018. “STRetch: Detecting and Discovering Pathogenic Short Tandem Repeat Expansions.” *Genome Biology* 19 (1): 121.

Dilthey, Alexander, Charles Cox, Zamin Iqbal, Matthew R. Nelson, and Gil McVean. 2015. “Improved Genome Inference in the MHC Using a Population Reference Graph.” *Nature Genetics* 47 (6): 682–88.

Dolzhenko, Egor, Joke J. F. A. van Vugt, Richard J. Shaw, Mitchell A. Bekritsky, Marka van Blitterswijk, Giuseppe Narzisi, Subramanian S. Ajay, et al. 2017. “Detection of Long Repeat

Expansions from PCR-Free Whole-Genome Sequence Data." *Genome Research* 27 (11): 1895–1903.

Froggatt, N. J., J. Green, C. Brassett, D. G. Evans, D. T. Bishop, R. Kolodner, and E. R. Maher. 1999. "A Common MSH2 Mutation in English and North American HNPCC Families: Origin, Phenotypic Expression, and Sex Specific Differences in Colorectal Cancer." *Journal of Medical Genetics* 36 (2): 97–102.

Garrison, Erik, Jouni Sirén, Adam M. Novak, Glenn Hickey, Jordan M. Eizenga, Eric T. Dawson, William Jones, et al. 2018. "Variation Graph Toolkit Improves Read Mapping by Representing Genetic Variation in the Reference." *Nature Biotechnology* 36 (9): 875–79.

Gymrek, Melissa, Thomas Willems, Audrey Guilmatre, Haoyang Zeng, Barak Markus, Stoyan Georgiev, Mark J. Daly, et al. 2016. "Abundant Contribution of Short Tandem Repeats to Gene Expression Variation in Humans." *Nature Genetics* 48 (1): 22–29.

Hannan, Anthony J. 2018. "Tandem Repeats Mediating Genetic Plasticity in Health and Disease." *Nature Reviews. Genetics* 19 (5): 286–98.

Lee, Christopher, Catherine Grasso, and Mark F. Sharlow. 2002. "Multiple Sequence Alignment Using Partial Order Graphs." *Bioinformatics* 18 (3): 452–64.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.

Lincoln, Stephen E., Rebecca Truty, Chiao-Feng Lin, Justin M. Zook, Joshua Paul, Vincent H. Ramey, Marc Salit, Heidi L. Rehm, Robert L. Nussbaum, and Matthew S. Lebo. 2019. "A Rigorous Interlaboratory Examination of the Need to Confirm Next-Generation Sequencing-Detected Variants with an Orthogonal Method in Clinical Genetic Testing." *The Journal of Molecular Diagnostics: JMD*, January. https://doi.org/10.1016/j.jmoldx.2018.10.009.

Liquori, C. L., K. Ricker, M. L. Moseley, J. F. Jacobsen, W. Kress, S. L. Naylor, J. W. Day, and L. P. Ranum. 2001. "Myotonic Dystrophy Type 2 Caused by a CCTG Expansion in Intron 1 of ZNF9." *Science* 293 (5531): 864–67.

Mousavi, Nima, Sharona Shleizer-Burko, Richard Yanicky, and Melissa Gymrek. 2019. "Profiling the Genome-Wide Landscape of Tandem Repeat Expansions." *bioRxiv*. https://doi.org/10.1101/361162.

"Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences. Recommendations 1984. Nomenclature Committee of the International Union of Biochemistry (NC-IUB)." 1986. *Proceedings of the National Academy of Sciences of the United States of America* 83 (1): 4–8.

Paten, Benedict, Adam M. Novak, Jordan M. Eizenga, and Erik Garrison. 2017. "Genome Graphs and the Evolution of Genome Inference." *Genome Research* 27 (5): 665–76.

Shoubridge, Cheryl, and Jozef Gecz. 2012. "Polyalanine Tract Disorders and Neurocognitive Phenotypes." *Advances in Experimental Medicine and Biology* 769: 185–203.

Tang, Haibao, Ewen F. Kirkness, Christoph Lippert, William H. Biggs, Martin Fabani, Ernesto Guzman, Smriti Ramakrishnan, et al. 2017. "Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes." *American Journal of Human Genetics* 101 (5): 700–715.

Tankard, Rick M., Mark F. Bennett, Peter Degorski, Martin B. Delatycki, Paul J. Lockhart, and Melanie Bahlo. 2018. "Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data." *American Journal of Human Genetics* 103 (6): 858–73.

Wright, Galen E. B., Jennifer A. Collins, Chris Kay, Cassandra McDonald, Egor Dolzhenko,

Qingwen Xia, Kristina Bečanović, et al. 2019. “Length of Uninterrupted CAG Repeats, Independent of Polyglutamine Size, Results in Increased Somatic Instability and Hastened Age of Onset in Huntington Disease.” *bioRxiv*. https://doi.org/10.1101/533414.