Targeted Genotyping of Variable Number Tandem Repeats with adVNTR

Mehrdad Bakhtiari[1], Sharona Shleizer-Burko[2], Melissa Gymrek[1,2], Vikas Bansal[3], and Vineet Bafna[*1]

[1]Department of Computer Science & Engineering, University of California, San Diego, La Jolla, CA 92093, USA
[2]Department of Medicine, University of California, San Diego, La Jolla, CA 92093, USA
[3]Department of Pediatrics, University of California, San Diego, La Jolla, CA 92093, USA

November 29, 2017

Abstract

Abstract.

Keywords. key1, key2

[*]Correspondence: vbafna@eng.ucsd.edu

1 Introduction

Introduction.

2 Method

Methods.

3 Results

Results.

4 Discussion

Discussion.

Acknowledgements. The analyses presented in this paper are based on the use of study data downloaded from the dbGaP web site, under phs001095.v1.p1, phs001096.v1.p1 and phs001097.v1.p1.

References

[1] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[2] Kin Fai Au, Jason G Underwood, Lawrence Lee, and Wing Hung Wong. Improving PacBio long read accuracy by short read alignment. *PloS one*, 7(10):e46679, 2012.

[3] Francesco Benedetti, Sara Dallaspezia, Cristina Colombo, Adele Pirovano, Elena Marino, and Enrico Smeraldi. A length polymorphism in the circadian clock gene Per3 influences age at onset of bipolar disorder. *Neuroscience letters*, 445(2):184–187, 2008.

[4] Gary Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*, 27(2):573, 1999.

[5] Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P Drake, Jane M Landolin, and Adam M Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology*, 33(6):623–630, 2015.

[6] KJ Brookes. The VNTR in complex disorders: The forgotten polymorphisms? A functional way forward? *Genomics*, 101(5):273–281, 2013.

[7] Amy L Byrd and Stephen B Manuck. MAOA, childhood maltreatment, and antisocial behavior: meta-analysis of a gene-environment interaction. *Biological psychiatry*, 75(1):9–17, 2014.

[8] A Cervera, D Tassies, V Obach, S Amaro, JC Reverter, and A Chamorro. The BC genotype of the VNTR polymorphism of platelet glycoprotein Ibα is overrepresented in patients with recurrent stroke regardless of aspirin therapy. *Cerebrovascular Diseases*, 24(2-3):242–246, 2007.

[9] Mark J Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC bioinformatics*, 13(1):238, 2012.

[10] James Clarke, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, and Hagan Bayley. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature nanotechnology*, 4(4):265–270, 2009.

[11] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[12] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498, 2011.

[13] Egor Dolzhenko, Joke JFA van Vugt, Richard J Shaw, Mitchell A Bekritsky, Marka van Blitterswijk, Giuseppe Narzisi, Subramanian S Ajay, Vani Rajan, Bryan Lajoie, Nathan H Johnson, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Research*, pages gr–225672, 2017.

[14] Ivana Durinovic-Belló, RP Wu, VH Gersuk, S Sanda, HG Shilling, and GT Nepom. Insulin gene VNTR genotype associates with frequency and phenotype of the autoimmune response to proinsulin. *Genes and immunity*, 11(2):188–193, 2010.

[15] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, et al. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.

[16] O Eser, B Eser, M Cosar, MO Erdogan, A Aslan, H Yildiz, M Solak, and A Haktanir. Short aggrecan gene repetitive alleles associated with lumbar degenerative disc disease in Turkish patients. *Genet Mol Res*, 10(3):1923–1930, 2011.

[17] Barbara Franke, Alejandro Arias Vasquez, Stefan Johansson, Martine Hoogman, Jasmin Romanos, Andrea Boreatti-Hümmer, Monika Heine, Christian P Jacob, Klaus-Peter Lesch, Miguel Casas, et al. Multicenter analysis of the SLC6A3/DAT1 VNTR haplotype in persistent ADHD suggests differential involvement of the gene in childhood and persistent ADHD. *Neuropsychopharmacology*, 35(3):656, 2010.

[18] Christian Fuchsberger, Jason Flannick, Tanya M Teslovich, Anubha Mahajan, Vineeta Agarwala, Kyle J Gaulton, Clement Ma, Pierre Fontanillas, Loukas Moutsianas, Davis J McCarthy, et al. The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41–47, 2016.

[19] Daniela Galimberti, Elio Scarpini, Eliana Venturelli, Alexander Strobel, Sabine Herterich, Chiara Fenoglio, Ilaria Guidi, Diego Scalabrini, Francesca Cortini, Nereo Bresolin, et al. Association of a NOS1 promoter repeat with Alzheimer's disease. *Neurobiology of aging*, 29(9):1359–1365, 2008.

[20] Yevgeniy Gelfand, Yozen Hernandez, Joshua Loving, and Gary Benson. VNTRseek-a computational tool to detect tandem repeat variants in high-throughput sequencing data. *Nucleic acids research*, 42(14):8884–8894, 2014.

[21] Melissa Gymrek, Thomas Willems, Audrey Guilmatre, Haoyang Zeng, Barak Markus, Stoyan Georgiev, Mark J Daly, Alkes L Price, Jonathan K Pritchard, Andrew J Sharp, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature genetics*, 48(1):22–29, 2016.

[22] Thomas Hackl, Rainer Hedrich, Jörg Schultz, and Frank Förster. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics*, 30(21):3004–3011, 2014.

[23] K Haddley, VJ Bubb, G Breen, UM Parades-Esquivel, and JP Quinn. Behavioural genetics of the serotonin transporter. In *Behavioral Neurogenetics*, pages 503–535. Springer, 2011.

[24] Minako Hijikata, Ikumi Matsushita, Goh Tanaka, Tomoko Tsuchiya, Hideyuki Ito, Katsushi Tokunaga, Jun Ohashi, Sakae Homma, Yoichiro Kobashi, Yoshio Taguchi, et al. Molecular cloning of two novel mucin-like genes in the disease-susceptibility locus for diffuse panbronchiolitis. *Human genetics*, 129(2):117–128, 2011.

[25] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, 2011.

[26] Andrew Kirby, Andreas Gnirke, David B Jaffe, Veronika Barešová, Nathalie Pochet, Brendan Blumenstiel, Chun Ye, Daniel Aird, Christine Stevens, James T Robinson, et al. Mutations causing medullary

cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. *Nature genetics*, 45(3):299–303, 2013.

[27] J Kirchheiner, K Nickchen, J Sasse, M Bauer, I Roots, and J Brockmöller. A 40-basepair VNTR polymorphism in the dopamine transporter (DAT1) gene and the rapid response to antidepressant treatment. *The pharmacogenomics journal*, 7(1):48, 2007.

[28] GJ LaHoste, JMet Swanson, SB Wigal, C Glabe, T Wigal, N King, and JL Kennedy. Dopamine D4 receptor gene polymorphism is associated with attention deficit hyperactivity disorder. *Mol Psychiatry*, 1(2):121–124, 1996.

[29] Maria D Lalioti, Hamish S Scott, Catherine Buresi, Colette Rossier, Armand Bottani, Michael A Morris, Alain Malafosse, and Stylianos E Antonarakis. Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature*, 386(6627):847, 1997.

[30] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[31] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[32] Hayan Lee, James Gurtowski, Shinjae Yoo, Shoshana Marcus, W Richard McCombie, and Michael Schatz. Error correction and assembly complexity of single molecule sequencing reads. *BioRxiv*, page 006395, 2014.

[33] Richard JLF Lemmers, Peggy de Kievit, Lodewijk Sandkuijl, George W Padberg, Gert-Jan B van Ommen, Rune R Frants, and Silvère M van der Maarel. Facioscapulohumeral muscular dystrophy is uniquely associated with one of the two variants of the 4q subtelomere. *Nature genetics*, 32(2):235, 2002.

[34] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[35] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[36] Qian Liu, Peng Zhang, Depeng Wang, Weihong Gu, and Kai Wang. Interrogating the unsequenceable genomic trinucleotide repeat disorders by long-read sequencing. *Genome medicine*, 9(1):65, 2017.

[37] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[38] Giles Miclotte, Mahdi Heydari, Piet Demeester, Stephane Rombauts, Yves Van de Peer, Pieter Audenaert, and Jan Fostier. Jabba: hybrid error correction for long sequencing reads. *Algorithms for Molecular Biology*, 11(1):10, 2016.

[39] Satoshi Okazaki, Marta Schirripa, Fotios Loupakis, Shu Cao, Wu Zhang, Dongyun Yang, Yan Ning, Martin D Berger, Yuji Miyamoto, Mitsukuni Suenaga, et al. Tandem repeat variation near the HIC1 (hypermethylated in cancer 1) promoter predicts outcome of oxaliplatin-based chemotherapy in patients with metastatic colorectal cancer. *Cancer*, 2017.

[40] Antonia L Pritchard, Colin W Pritchard, Peter Bentham, and Corinne L Lendon. Role of serotonin transporter polymorphisms in the behavioural and psychological symptoms in probable Alzheimer disease patients. *Dementia and geriatric cognitive disorders*, 24(3):201–206, 2007.

[41] Alberto Pugliese, Markus Zeller, Alarico Fernandez, Laura J Zalcberg, Richard J Bartlett, Camillo Ricordi, Massimo Pietropaolo, George S Eisenbarth, Simon T Bennett, and Dhavalkumar D Patel. The insulin gene is transcribed in the human thymus and transcription levels correlate with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes. *Nature genetics*, 15(3):293–297, 1997.

[42] Helge Ræder, Stefan Johansson, Pål I Holm, Ingfrid S Haldorsen, Eric Mas, Véronique Sbarra, Ingrid Nermoen, Stig Å Eide, Louise Grevle, Lise Bjørkhaug, et al. Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction. *Nature genetics*, 38(1):54, 2006.

[43] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nature biotechnology*, 29(1):24–26, 2011.

[44] Leena Salmela and Eric Rivals. LoRDEC: accurate and efficient long read error correction. *Bioinformatics*, 30(24):3506–3514, 2014.

[45] Leena Salmela, Riku Walve, Eric Rivals, and Esko Ukkonen. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics*, 33(6):799–806, 2016.

[46] Mark D Shriver, Li Jin, Ranajit Chakraborty, and Eric Boerwinkle. VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. *Genetics*, 134(3):983–993, 1993.

[47] Bianca K Stöcker, Johannes Köster, and Sven Rahmann. SimLoRD: Simulation of Long Read Data. *Bioinformatics*, 32(17):2704–2706, 2016.

[48] Cath Tyner, Galt P Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Christopher Eisenhart, Clayton M Fischer, David Gibson, Jairo Navarro Gonzalez, Luvina Guruvadoo, et al. The UCSC Genome Browser database: 2017 update. *Nucleic acids research*, 45(D1):D626–D634, 2016.

[49] Ajay Ummat and Ali Bashir. Resolving complex tandem repeats with long reads. *Bioinformatics*, 30(24):3491–3498, 2014.

[50] Biju Viswanath, Meera Purushottam, Thennarasu Kandavel, YC Janardhan Reddy, Sanjeev Jain, et al. DRD4 gene and obsessive compulsive disorder: do symptom dimensions have specific genetic correlates? *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 41:18–23, 2013.

[51] David A Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, et al. The complete genome of an individual by massively parallel DNA sequencing. *nature*, 452(7189):872–876, 2008.

[52] Thomas Willems, Dina Zielinski, Jie Yuan, Assaf Gordon, Melissa Gymrek, and Yaniv Erlich. Genome-wide profiling of heritable and de novo STR variations. *Nature Methods*, 2017.

[53] Bradford B Worrall, Thomas G Brott, Robert D Brown, W Mark Brown, Stephen S Rich, Sampath Arepalli, Fabienne Wavrant-De Vrièze, Jaime Duckworth, Andrew B Singleton, John Hardy, et al. IL1RN VNTR polymorphism in ischemic stroke. *Stroke*, 38(4):1189–1196, 2007.

[54] Jonathan M Wright. Mutation at vntrs: Are minisatellites the evolutionary progeny of microsatellites? *Genome*, 37(2):345–347, 1994.

[55] Justin M Zook, David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng, Yuling Liu, Christopher E Mason, Noah Alexander, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data*, 3, 2016.

Supplementary Material

A. Appendix