Haplotype resolved genomes

Porubský, David

Bibliography

Aguiar, Derek, and Sorin Istrail. 2012. HapCompass: A Fast Cycle Basis Algorithm for Accurate Haplotype Assembly of Sequence Data. *Journal of Computational Biology* 19(6): 577–90.

Aguiar, Derek, and Sorin Istrail. 2013 Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics*, 29: 352–360.

Amini, Sasan, Dmitry Pushkarev, Lena Christiansen, Emrah Kostem, Tom Royce, Casey Turk, Natasha Pignatelli, et al. 2014. Haplotype-Resolved Whole-Genome Sequencing by Contiguity-Preserving Transposition and Combinatorial Indexing. *Nature Genetics* 46(12).

Ammar, Ron, Tara A Paton, Dax Torti, Adam Shlien, and Gary D Bader. 2015. Long Read Nanopore Sequencing for Detection of HLA and CYP2D6 Variants and Haplotypes. *F1000Research* 4(0).

Bansal, Vikas, Ryan Tewhey, Eric J Topol, Nicholas J Schork, Meng Amy Li, and Allan Bradley. 2011. The next Phase in Human Genetics. *Nature Biotechnology* 29(1): 38–39.

Bakker, Paul I W De, Gil Mcvean, Pardis C Sabeti, Marcos M Miretti, Todd Green, Jonathan Marchini, Xiayi Ke, et al. 2006. A High-Resolution HLA and SNP Haplotype Map for Disease Association Studies in the Extended Human MHC. *Nature Genetics* 38(10): 1166–72.

Bakker, Bjorn, Aaron Taudt, Mirjam E Belderbos, David Porubsky, Diana C J Spierings, Tristan V De Jong, Nancy Halsema, et al. 2016. Single-Cell Sequencing Reveals Karyotype Heterogeneity in Murine and Human Malignancies. *Genome Biology*. 1–15.

Baslan, Timour, Jude Kendall, Linda Rodgers, Hilary Cox, Mike Riggs, Asya Stepansky, Jennifer Troge, et al. 2016. Genome-Wide Copy Number Analysis of Single Cells. *Nature Protocols* **7**(6): 1024–1041.

Broman, K W, J C Murray, V C Sheffield, R L White, and J L Weber. 1998. Comprehensive Human Genetic Maps: Individual and Sex-Specific Variation in Recombination. *American Journal of Human Genetics* 63: 861–69.

Brown, Pamela J B, Miguel A De Pedro, T Kysela, Charles Van Der Henst, Jinwoo Kim, Xavier De Bolle, Clay Fuqua, and Yves V Brun. 2012. Correction for Yang et Al., Completely Phased Genome Sequencing through Chromosome Sorting. *Proceedings of the National Academy of Sciences* 109(8): 3190–3190.

Browning, Sharon R, and Brian L Browning. 2007. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *American Journal of Human Genetics* 81(11): 1084–97.

Browning, Sharon R., and Brian L. Browning. 2011. Haplotype Phasing: Existing Methods and New Developments. *Nature Reviews Genetics* 12(10): 703–14.

Burton, Joshua N, Andrew Adey, Rupali P Patwardhan, Ruolan Qiu, Jacob O Kitzman, and Jay Shendure. 2013. Chromosome-Scale Scaffolding of de novo Genome Assemblies Based on Chromatin Interactions. *Nature Biotechnology* 31(12): 1119–25.

Cao, Hongzhi, Honglong Wu, Ruibang Luo, Shujia Huang, Yuhui Sun, Xin Tong, Yinlong Xie, et al. 2015. De Novo Assembly of a Haplotype-Resolved Human Genome. *Nature Biotechnology* 33(6): 617-622.

Carvalho, Antonio Bernardo, Eduardo G Dupim, and Gabriel Goldstein. 2016. Improved Assembly of Noisy Long Reads by K- Mer Validation. *Genome Research* 26:1–11.

Chaisson, Mark J P, John Huddleston, Megan Y Dennis, Peter H Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, et al. 2015. Resolving the Complexity of the Human Genome Using Single-Molecule Sequencing. *Nature* 517(7536): 608–11.

Chaisson, Mark J P, Richard K Wilson, and Evan E Eichler. 2015. Genetic Variation and the de Novo Assembly of Human Genomes. *Nature Reviews* 16(11): 627–40.

Chen, Zhi-zhong, Fei Deng, and Lusheng Wang. 2013. Exact Algorithms for Haplotype Assembly from Whole-Genome Sequence Data. *Bioinformatics* 29(16): 1938–45.

Church, Deanna M, Valerie A Schneider, Karyn Meltz Steinberg, Michael C Schatz, Aaron R Quinlan, Chen-shan Chin, Paul A Kitts, et al. 2015. Extending Reference Assembly Models. *Genome Biology* 16(13): 2–6.

Cilibrasi, Rudi, Leo Van Iersel, Steven Kelk, John Tromp 2005. On the Complexity of the Single Individual SNP Haplotyping Problem. *ArXiv* 1–20.

Conrad, Donald F, Jonathan E M Keebler, Mark A DePristo, Sarah J Lindsay, Yujun Zhang, Ferran Casals, Youssef Idaghdour, et al. 2011. Variation in Genome-Wide Mutation Rates within and between Human Families. *Nature Genetics* 43(7): 712–14.

De Bourcy, Charles F A, Iwijn De Vlaminck, Jad N. Kanbar, Jianbin Wang, Charles Gawad, and Stephen R. Quake. 2014. A Quantitative Comparison of Single-Cell Whole Genome Amplification Methods. *PLoS ONE* 9(8).

Delaneau, Olivier, Bryan Howie, Anthony J. Cox, Jean François Zagury, and Jonathan Marchini. 2013. Haplotype Estimation Using Sequencing Reads. *American Journal of Human Genetics* 93(4): 687–96.

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

Duitama, Jorge, Gayle K. McEwen, Thomas Huebsch, Stefanie Palczewski, Sabrina Schulz, Kevin Verstrepen, Eun Kyung Suk, and Margret R. Hoehe. 2012. Fosmid-Based Whole Genome Haplotyping of a HapMap Trio Child: Evaluation of Single Individual Haplotyping Techniques. *Nucleic Acids Research* 40(5): 2041–53.

Falconer, Ester, Mark Hills, Ulrike Naumann, Steven S S Poon, Elizabeth a Chavez, Ashley D Sanders, Yongjun Zhao, Martin Hirst, and Peter M Lansdorp. 2012. DNA Template Strand Sequencing of Single-Cells Maps Genomic Rearrangements at High Resolution. *Nature Methods* 9(11): 1107–12.

Fan, H Christina, Wei Gu, Jianbin Wang, Yair J Blumenfeld, Yasser Y El-sayed, and Stephen R Quake. 2012. Non-Invasive Prenatal Measurement of the Fetal Genome. *Nature* 487(19): 320–324.

Fan, H Christina, Jianbin Wang, Anastasia Potanina, and Stephen R Quake. 2011. Whole-Genome Molecular Haplotyping of Single Cells. *Nature Biotechnology* 29(1): 51–57.

Francioli, L C, A Menelaou, S L Pulit, F van Dijk, P F Palamara, C C Elbers, P B T Neerincx, et al. 2014. Whole-Genome Sequence Variation, Population Structure and Demographic History of the Dutch Population. *Nature Genetics* 46(8): 818–25.

Geraci, Filippo. 2010. A Comparison of Several Algorithms for the Single Individual SNP Haplotyping Reconstruction Problem. *Bioinformatics* 26(18): 2217–25.

Glusman, Gustavo, Hannah C Cox, and Jared C Roach. 2014. Whole-Genome Haplotyping Approaches and Genomic Medicine. *Genome Medicine* 6(73): 1–16.

He, D., A. Choi, K. Pipatsrisawat, A. Darwiche, and E. Eskin. 2010. Optimal Algorithms for Haplotype Assembly from Whole-Genome Sequence Data. *Bioinformatics* 26(12): 183–90.

Hoehe, Margret R, George M Church, Hans Lehrach, Thomas Kroslak, Stefanie Palczewski, Katja Nowick, Sabrina Schulz, Eun-kyung Suk, and Thomas Huebsch. 2014. Population Patterns of Gene and Protein Diplotypes. *Nature Communications* 5(5): 1–12.

Hills, Mark, Kieran O'Neill, Ester Falconer, Ryan Brinkman, and Peter M Lansdorp. 2013. BAIT: Organizing Genomes and Mapping Rearrangements in Single Cells. *Genome Medicine* 5(9): 82.

Hou, Yu, Wei Fan, Liying Yan, Rong Li, Ying Lian, Jin Huang, Jinsen Li, et al. 2013. Genome Analyses of Single Human Oocytes. *Cell* 155(7): 1492–1506.

Huang, Yu-chuen, Cheng-ming Lee, Marcelo Chen, Ming-yi Chung, Yen-hwa Chang, William Ji-shian Huang, Donald Ming-tak Ho, Chin-chen Pan, Tony T Wu, and Stone Yang. 2007. Haplotypes , Loss of Heterozygosity , and Expression Levels of Glycine N -Methyltransferase in Prostate Cancer. *American Association for Cancer Research* 13(5): 1412–20.

International Human Genome Sequencing Consortium. 2001. Initial Sequencing and Analysis of the Human Genome. *Nature* 409(2): 860–921.

James, T, and P Jill. 2012. Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration. *Briefings in Bioinformatics* 14(2): 178–92.

Kent, W James, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. 2002. The Human Genome Browser at UCSC. *Genome Research* 12: 996–1006.

Kirkness, Ewen F., Rashel V. Grindberg, Joyclyn Yee-Greenbaum, Christian R. Marshall, Stephen W. Scherer, Roger S. Lasken, and J. Craig Venter. 2013. Sequencing of Isolated Sperm Cells for Direct Haplotyping of a Human Genome. *Genome Research* 23(5): 826–32.

Kitzman, Jacob O, Alexandra P MacKenzie, Andrew Adey, Joseph B Hiatt, Rupali P Patwardhan, Peter H Sudmant, Sarah B Ng, et al. 2011. Haplotype-Resolved Genome Sequencing of a Gujarati Indian Individual. *Nature Biotechnology* 29(1): 59–63.

Kitzman, Jacob O, Matthew W Snyder, Mario Ventura, Alexandra P Lewis, Ruolan Qiu, Lavone E Simmons, Hilary S Gammill, et al. 2012. Noninvasive Whole-Genome Sequencing of a Human Fetus. *Genomics* 4(137): 1–8.

Klambauer, Günter, Karin Schwarzbauer, Andreas Mayr, Djork Arné Clevert, Andreas Mitterecker, Ulrich Bodenhofer, and Sepp Hochreiter. 2012. Cn.MOPS: Mixture of Poissons for Discovering Copy Number Variations in next-Generation Sequencing Data with a Low False Discovery Rate. *Nucleic Acids Research* 40(9): 1–14.

Kloosterman, Wigard P, Laurent C Francioli, Fereydoun Hormozdiari, Tobias Marschall, Jayne Y Hehir-kwa, Abdel Abdellaoui, Eric-wubbo Lameijer, et al. 2015. Characteristics of de novo Structural Changes in the Human Genome. *Genome Research* 25:792–801.

Kong, Augustine, Gudmar Thorleifsson, Daniel F. Gudbjartsson, Gisli Masson, Asgeir Sigurdsson, Aslaug Jonasdottir, G. Bragi Walters, et al. 2010. Fine-Scale Recombination Rate Differences between Sexes, Populations and Individuals. *Nature* 467(7319): 1099–1103.

Kuleshov, Volodymyr, Dan Xie, Rui Chen, Dmitry Pushkarev, Zhihai Ma, Tim Blauwkamp, Michael Kertesz, and Michael Snyder. 2014. Whole-Genome Haplotyping Using Long Reads and Statistical Methods. *Nature Biotechnology* 32(3): 261–66.

Langmead, Ben, and Steven L Salzberg. 2012. Fast Gapped-Read Alignment with Bowtie 2. *Nature Methods* 9(4): 357–59.

Lander, Eric S. 2011. Initial Impact of the Sequencing of the Human Genome. *Nature* 470(7333): 187–97.

Leung, Danny, Inkyung Jung, Nisha Rajagopal, Anthony Schmitt, Siddarth Selvaraj, Ah Young Lee, Chia-An Yen, et al. 2015. Integrative Analysis of Haplotype-Resolved Epigenomes across Human Tissues. *Nature* 518(7539): 350–54.

Li, Heng, and Richard Durbin. 2010. Fast and Accurate Long-Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 26(5): 589–95.

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.

Lippert, Ross, Russell Schwartz, Giuseppe Lancia, and Sorin Istrail. 2002. Algorithmic Strategies for the Single Nucleotide Polymorphism Haplotype Assembly Problem. *Briefings in Bioinformatics* 3(1): 23–31.

Liu, Nianjun, Kui Zhang, and Hongyu Zhao. 2008. Haplotype-Association Analysis. Advances in Genetics 60(7): 335–405.

Lo, C., R. Liu, J. Lee, K. Robasky, S. Byrne, C. Lucchesi, J. Aach, G. Church, V. Bafna, and K. Zhang. 2013. On the Design of Clone-Based Haplotyping. *Genome Biology* 14(9): 1–12

Lo, Y M Dennis, K C Allen Chan, Hao Sun, Eric Z Chen, Peiyong Jiang, Fiona M F Lun, Yama W Zheng, et al. 2010. Maternal Plasma DNA Sequencing Reveals the Genome-Wide Genetic and Mutational Profile of the Fetus. *Prenatal Diagnosis* 2(61): 1–13.

Lu, Sijia, Chenghang Zong, Wei Fan, Mingyu Yang, Jinsen Li, Alec R Chapman, Ping Zhu, et al. 2012. Probing Meiotic Recombination and Aneuploidy of Single Sperm Cells by Whole-Genome Sequencing. *Science* 338(6114): 1627–30.

Ma, Li, Yan Xiao, Hui Huang, Qingwei Wang, Weinian Rao, Yue Feng, Kui Zhang, and Qing Song. 2010. Direct Determination of Molecular Haplotypes by Chromosome Microdissection. *Nature Methods* 7(4): 299–301.

Marcus, Shoshana, Hayan Lee, and Michael C Schatz. 2014. Genome Analysis SplitMEM: A Graphical Algorithm for Pan-Genome Analysis with Suffix Skips. *Bioinformatics* 30(24): 3476–83.

Mardis, Elaine R. 2008. The Impact of next-Generation Sequencing Technology on Genetics. *Trends in Genetics* 24(3): 133–41.

The Computational Pan-Genomics Consortium. 2016. Computational pan-genomics: status, promises and challenges. *Bioinformatics* 1–18.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA, 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20: 1297–303

Mostovoy, Yulia, Michal Levy-Sakin, Jessica Lam, Ernest T Lam, Alex R Hastie, Patrick Marks, Joyce Lee, et al. 2016. A Hybrid Approach for de Novo Human Genome Sequence Assembly and Phasing. *Nature Methods* 13: 12–17.

Moynahan, Mary Ellen, and Maria Jasin. 2010. Mitotic Homologous Recombination Maintains Genomic Stability and Suppresses Tumorigenesis. *Nature Reviews Molecular Cell Biology* 11(3): 196–207.

Navin, Nicholas, and James Hicks. 2011. Future Medical Applications of Single-Cell Sequencing in Cancer. *Genome Medicine* 3(31): 1–12.

Pasaniuc, Bogdan, Nadin Rohland, Paul J Mclaren, Kiran Garimella, Noah Zaitlen, Heng Li, Namrata Gupta, et al. 2012. Extremely Low-Coverage Sequencing and Imputation Increases Power for Genome-Wide Association Studies. *Nature Genetics* 44(6): 631–35.

Paten, Benedict, Adam Novak, and David Haussler. Mapping to a Reference Genome Structure. *BioArx* 1–26.

Patterson, M., Marschall, T., Pisanti, N., van Iersel, L., Stougie, L., Klau, G.W., Schonhuth, A. 2015 WhatsHap: Weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology* 22(6), 498–509

Pendleton, Matthew, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch, Adrian M Stütz, et al. 2015. Assembly and Diploid Architecture of an Individual Human Genome via Single-Molecule Technologies. *Nature Methods* 12(8): 780–86.

Peters, Brock A., Bahram G. Kermani, Andrew B. Sparks, Oleg Alferov, Peter Hong, Andrei Alexeev, Yuan Jiang, et al. 2012. Accurate Whole-Genome Sequencing and Haplotyping from 10 to 20 Human Cells. *Nature* 487(7406): 190–95.

Petersdorf, Effie W, Mari Malkki, Ted A Gooley, Paul J Martin, and Zhen Guo. 2007. MHC Haplotype Matching for Unrelated Hematopoietic Cell Transplantation. *PLOS Medicine* 4(1): 60–68.

Pirola, Yuri, Simone Zaccaria, Riccardo Dondi, Gunnar W. Klau, Nadia Pisanti, and Paola Bonizzoni. 2015. HapCol: Accurate and Memory-Efficient Haplotype Assembly from Long Reads. *Bioinformatics* 1–8

Porubský, David, Ashley D Sanders, Niek Van Wietmarschen, Ester Falconer, Mark Hills, Diana C J Spierings, Marianna R Bevova, Victor Guryev, and Peter M Lansdorp. 2016. Direct Chromosome-Length Haplotyping by Single-Cell Sequencing. *Genome Research* 26: 1565–74.

Putnam, Nicholas H, Brendan L O Connell, Jonathan C Stites, Brandon J Rice, Marco Blanchette, Robert Calef, Christopher J Troll, et al. 2016. Chromosome-Scale Shotgun Assembly Using an in Vitro Method for Long-Range Linkage. *Genome Research* 26: 1–9.

Raymond, Christopher K, Sandhya Subramanian, Marcia Paddock, Ruolan Qiu, Chloe Deodato, Anthony Palmieri, Jean Chang, et al. 2005. Targeted, Haplotype-Resolved Resequencing of Long Segments of the Human Genome. *Genomics* 86: 759–66.

Roach, Jared C, Gustavo Glusman, Arian F A Smit, Chad D Huff, Robert Hubley, Jay Shendure, Radoje Drmanac, Lynn B Jorde, Leroy Hood, and David J Galas. 2010. Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science* 328: 636–39.

Roberts, Richard J, Mauricio O Carneiro, and Michael C Schatz. 2013. The Advantages of SMRT Sequencing. *Genome Biology* 14(405): 2–5.

Ross, Michael G, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J Lennon, Ryan Hegarty, Chad Nusbaum, and David B Jaffe. 2013. Characterizing and Measuring Bias in Sequence Data. *Genome Biology* 14(51): 2–20.

Sanders, Ashley D, Mark Hills, David Porubský, Victor Guryev, Ester Falconer, Peter M Lansdorp, British Columbia, and Cancer Agency. 2016. Characterizing Polymorphic Inversions in Human Genomes by Single Cell Sequencing. *Genome Researc*h 26:1575–1587

Schatz, Michael C, Arthur L Delcher, and Steven L Salzberg. 2010. Assembly of Large Genomes Using Second-Generation Sequencing. *Genome Research* 20:1165–1173

Selvaraj, Siddarth, Jesse R Dixon, Vikas Bansal, and Bing Ren. 2013. Whole-Genome Haplotype Reconstruction Using Proximity-Ligation and Shotgun Sequencing. *Nature Biotechnology* 31(12): 1111–18.

Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, Hastie A, Cao H, Yun JY, Kim J, Kuk J, Park GH, Kim J, Ryu H, Kim J, Roh M, Baek J, Hunkapiller MW, Korlach J, Shin JY, Kim C. De novo assembly and phasing of a Korean human genome. *Nature* 538(7624):243–247

Snyder, Matthew W, Andrew Adey, Jacob O Kitzman, and Jay Shendure. 2015. Haplotype-Resolved Genome Sequencing: Experimental Methods and Applications. *Nature Reviews* 16(6): 344–58.

Steinberg, Karyn Meltz, Valerie A Schneider, Tina A Graves-lindsay, Robert S Fulton, Richa Agarwala, John Huddleston, Sergey A Shiryev, et al. 2014. Single Haplotype Assembly of the Human Genome from a Hydatidiform Mole. *Genome Researc*h 24: 1–12.

Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. An Integrated Map of Structural Variation in 2,504 Human Genomes. *Nature* 526(7571): 75–81.

Suk, Eun Kyung, Gayle K. McEwen, Jorge Duitama, Katja Nowick, Sabrina Schulz, Stefanie Palczewski, Stefan Schreiber, et al. 2011. A Comprehensively Molecular Haplotype-Resolved Genome of a European Individual. *Genome Research* 21(10): 1672–85.

Tewhey, Ryan, Vikas Bansal, Ali Torkamani, Eric J. Topol, and Nicholas J. Schork. 2011. The Importance of Phase Information for Human Genomics. *Nature Reviews Genetics* 12(3): 215–23.

Tilgner, H., F. Grubert, D. Sharon, and M. P. Snyder. 2014. Defining a Personal, Allele-Specific, and Single-Molecule Long-Read Transcriptome. *Proceedings of the National Academy of Sciences* 111(27): 9869–74.

The International HapMap Consortium 2007. A Second Generation Human Haplotype Map of over 3.1 Million SNPs. *Nature* 449(7164): 851–61.

The International HapMap 3 Consortium 2010. Integrating Common and Rare Genetic Variation in Diverse Human Populations. *Nature* 467(7311): 52–58.

Venter, J Craig, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, et al. 2001. The Sequence of the Human Genome. *Science* 291: 1304–1351

Walter, Klaudia, Josine L. Min, Jie Huang, Lucy Crooks, Yasin Memari, Shane McCarthy, John R. B. Perry, et al. 2015. The UK10K Project Identifies Rare Variants in Health and Disease. *Nature* 526(7571): 82–90.

Wang, Jianbin, H. Christina Fan, Barry Behr, and Stephen R. Quake. 2012. Genome-Wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. *Cell* 150(2): 402–12.

Wang, Lu, Jun Zhang, Jialei Duan, Xinxing Gao, Wei Zhu, Xingyu Lu, Lu Yang, et al. 2014. Resource Programming and Inheritance of Parental DNA Methylomes in Mammals. *Cell* 157(4): 979–91.

Weisenfeld, Neil I, Vijay Kumar, Preyas Shah, Deanna M Church, David B Jaffe 2016. Direct Determination of Diploid Genome Sequences. *BioRxiv* 1–21.

Xie, Wei, Cathy L Barr, Audrey Kim, Feng Yue, Ah Young Lee, James Eubanks, and Emma L Dempster. 2011. Resource Base-Resolution Analyses of Sequence and Parent-of-Origin Dependent DNA Methylation in the Mouse Genome. *Cell* 148(4): 816–31.

Yang, Wen-yun, Farhad Hormozdiari, Zhanyong Wang, Dan He, Bogdan Pasaniuc, and Eleazar Eskin. 2013. Sequence Analysis Leveraging Reads That Span Multiple Single Nucleotide Polymorphisms for Haplotype Inference from Sequencing Data. *Bioinformatics* 29(18): 2245–52.

Yin, Tengfei, Dianne Cook, and Michael Lawrence. 2012. Ggbio: An R Package for Extending the Grammar of Graphics for Genomic Data. Genome Biology 13(8).

Zerbino, Daniel R, and Ewan Birney. 2008. Velvet: Algorithms for de novo Short Read Assembly Using de Bruijn Graphs. *Genome Research* 18:821–829.

Zhang, Kui, and Degui Zhi. 2013. Joint Haplotype Phasing and Genotype Calling of Multiple Individuals Using Haplotype Informative Reads. *Bioinformatics* 29(19): 2427–34.

Zheng, Grace X Y, Billy T Lau, Michael Schnall-Levin, Mirna Jarosz, John M Bell, Christopher M Hindson, Sofia Kyriazopoulou-Panagiotopoulou, et al. 2016. Haplotyping Germline and Cancer Genomes with High-Throughput Linked-Read Sequencing. *Nature Biotechnology* 34: 303–311.

Zook, Justin M, Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide, and Marc Salit. 2014. Integrating Human Sequence Data Sets Provides a Resource of Benchmark SNP and Indel Genotype Calls. *Nature Biotechnology* 32(3): 246–51.

Dutch summary (Nederlandse samenvatting)

Het in dit proefschrift beschreven werk is gericht op de uitdagingen en mogelijkheden die de analyse van compleet gephaseerde diploide genomen middels single-cell strand sequencing (Strand-seq) ons bieden. Ons initiële doel was het ontwikkelen van een robuuste, gecomputeriseerde methode voor het samenstellen van accurate haplotypes uit single cell Strands-seq libraries. Het volgende doel was het valideren van de resultaten van de phasing methode tegenover de "gold standard" haplotypes uit het HapMap project, alsmede het demonstreren van de toepasbaarheid van mijn Strand-seq phasing methode. Tot slot werden de mogelijkheden onderzocht om Strand-seq met andere sequencing methoden te combineren om de kosten te verminderen en de compleetheid van de geanalyseerde haplotypes te vergroten.

Hoofdstuk 1 bevat een algehele introductie bestaande uit een samenvatting van huidige en voorgaande haplotyping technologieën. Hierin wijs ik op het belang van haplotype informatie in zowel fundamentele als klinische onderzoeken. In dit hoofdstuk leg ik een sterke nadruk op opkomende technologieën zoals single-cell sequencing, linked-read sequencing en long-read technologieën en analyseer ik de capaciteit van deze technologieën op genomen te genereren waarbij de haplotype informatie correct onderscheiden wordt. Ondanks de kracht en het potentieel van Strand-seq zijn er op het moment weinig rekenkundige en gecomputeriseerde methoden beschikbaar voor de analyse van deze data. In hoofdstuk 2 omschrijf ik de ontwikkeling van een nieuwe bio-informatica pipeline die in staat is om kleine nuances in Strand-seq data te detecteren. Met name beschrijf ik een methode om breekpunten en haplotype phasing te extraheren uit Strand-seq data.

De validatie van deze tools wordt gepresenteerd in hoofdstuk 3. Voor de validatie hebben wij gekozen voor een bekend familie trio van het HapMap project te gebruiken, alsmede onafhankelijke bronnen als PacBio RNA-seq en andere single cell phasing methoden. Dit onderzoek bevestigde de hoge accuraatheid van Strand-seq phasing. Deze werd verder geverifieerd in een vergelijking met *de novo* assembly gebasseerde phasing. In hoofdstuk 4 bijschrijf ik de toepassing van genoom-wijde haplotypering met Strand-seq om zowel meiotische recombinaties alsmede haplotype verschillen in kaart te brengen in een familie trio. Verder demonstreer ik de phasing van grotere genetische varianten als deleties, duplicaties en inversies.

Om de kosten en de vereiste werklast die nodig is voor het phaseren van het genoom van een enkel individu te verlagen heb ik de mogelijkheden verkent om de globale phasering van Strand-seq te integreren met andere sequencing

technologieën zoals PacBio of Illumina. In hoofdstuk 5 stel ik een geïntegreerde phasing methode voor. Hierbij worden globale Strand-seq haplotypes gecombineerd met DNA fragmenten waarvan de sequentie bepaald is. Een combinatie van Strand-seq phasing met andere technologieën kunnen complete genomen voor lagere kosten phaseren, met een grotere compleetheid en een hogere accuraatheid. Op basis van onze resultaten verwachten wij dat toekomstige studies naar haplotypes een combinatie van Strand-seq en long-read technologieën zullen gaan gebruiken om complete haplotypes over hele chromosomen vast te stellen. Single-cell sequencing technieken worden steeds toegankelijker. Wij verwachten dat het haplotyperen middels Strand-seq een belangrijke rol zal gaan spelen in de *de novo* constructie van haplotype-bepaalde persoonlijke genomen. Naar verwachting zal dit belangrijke gevolgen hebben voor studies naar genetische variatie in relatie met gezondheid en ziekte.

Glossary

Allele – two or more alternative forms of a piece of DNA that resides on the same locus in the genome.

BreakPointR – software package specifically tailored to search for change-points in template strand inheritance using Strand-seq data.

Compound heterozygosity – the presence of two deleterious variants located in the same gene either in cis (on the same homologue) or in trans (on different homologues) conformation.

Crick – a positive, plus ('+') strand of the reference genome and also a read that aligns in this direction.

Direct (experimental) haplotyping – direct observation of alleles on a single molecule of DNA which represent haploid part of the genome.

Fosmid – an artificial construct consisting of bacterial DNA that includes section of cloned genomic DNA of ~ 40kb in length.

Genetic haplotyping – the process of assignment the phase to the observed alleles in a form of genotypes according to the principles of Mendelian segregation of alleles in pedigrees.

Genotype – represent particular combination of alleles of a given organism, however, relative position of alleles along the single homologue is unknown.

Haplotype – contiguous set of genetic variants that are co-located on the same homologous chromosome and are inherted from the same parent.

Linkage disequilibrium – represents nonrandom segregation of alleles at different loci in a population. Linkage disequilibrium decreases with genomic distance and is not present between alleles residing on different chromosomes.

Long-read sequencing – sequencing technology with raw read average size longer than 1kb. Technologies that rely on assembly of short reads into a long ones belongs here as well.

N50 value – standardized value used to evaluate achieved contiguity of assembled haplotypes and represents the smallest haplotype block in which the sum of that block and all larger blocks total to 50% of the complete haplotype assembly.

Population-based haplotyping – the process of inferring the most likely phasing of common alleles from unordered genotype information based on the frequency of shared haplotype block in a large populations

Homologous chromosomes – pair of chromosomes each inherited from one parent.

Homozygous allele – a locus in the genome where both homologues share the same allele.

Heterozygous allele – a locus in the genome where both homologues chromosomes differ and carry different alleles.

Homozygosity regions – localized regions of the genome at which both homologues chromosomes are identical.

Indel – genetic variant that includes insertions and deletions of relatively short size (< 50bp).

Loss of heterozygosity – loss of the normal, functional allele at a heterozygous locus resulting in a homozygous conformation of alleles.

Paired-end reads – two reads sequenced from the opposite ends of the same DNA fragment, further defined by a specific insert size

Phasing – process of assignment of genetic variants (alleles) to one of two homologous chromosomes. Phase then denotes the relative position of alleles at a given locus.

Recombination event – position where two homologous chromosomes cross-over and exchange pieces of DNA information during meiosis.

Strand-seq – single cell sequencing technique able to distinguish inherited parental template strands based on the directionality they map to the reference genome.

StrandPhase – software package that specifically interrogates haplotype informative (WC regions) in every Strand-seq library and uses greedy algorithm to obtain consensus haplotypes.

StrandPhaseR – software package that specifically interrogates haplotype informative (WC regions) in every Strand-seq library and uses binary sorting of Watson and Crick strands to obtain consensus haplotypes.

Structural variation – copy number variants (insertion, deletions) or copy number neutral differences between two homologous chromosomes.

Template state – the relative proportion of Watson and Crick reads in a chromosome (or shorter chromosomal region) in a Strand-seq library. We distinguish WW, WC or CC template state in a Strand-seq library.

Watson – a negative , minus (‘-’) strand of the reference genome and also a read that aligns in this direction.

WC region – a template strand state that consists of approximately equal proportions of reads aligned to the minus (‘-’) and plus (‘+’) strand of the reference genome. Such regions are haplotype informative.

Minimal error correction problem – focuses on correction of a minimal number of bases in the error-prone sequencing reads such that they can be partitioned into two conflict free sets representing the two haplotypes. Weighted version of this problem is abbreviated as wMEC.

List of abbreviations

BAIT	Bioinformatic analysis of inherited templates
BAM	Binary alignment map
bp	base pair
BrdU	5-Bromo-2'-deoxyuridine
C	Crick (template strand orientation) or Child's homologue
CC	Crick-Crick (template strand inheritance)
CNV	Copy number variation
Chr	Chromosome
CPT	Contiguity-preserving transposition
DNA	Deoxyribonucleic acid
F	Father's homologue
FACS	Fluorescent activated cell sorting
GC	Guanin-Cytosine pair
H1,2	Homologous one or two
HLA	Human leukocyte antigen
HMW	High molecular weight
kb	kilobase(s)
LD	Linkage disequilibrium
LOH	Loss of heterozygosity
M	Mother's homologue
Mb	Megabase(s)
MDA	Multiple displacement amplification
MEC	Minimal error correction
NGS	Next-generation sequencing
NCBI	National Center for Biotechnology Information
IGV	Integrated genome viewer
PCR	Polymerase Chain Reaction
PE	Paired-end (read)
SCE	Sister chromatid exchange
SCS	Single-cell sequencing
SD	Standard deviation
SE	Single-end (read)
SNV	Single nucleotide variant
SV	Structural variant
VCF	Variant calling format
W	Watson (template strand orientation)
WC	Watson-Crick (template strand inheritance)
WGA	Whole genome amplification
WGS	Whole genome sequencing
WW	Watson-Watson (template strand inheritance)
UCSC	University of California, Santa Cruz

Acknowledgements

My PhD journey started more than 4 years ago. The last 4 years can be summarized as a time of tremendous personal and scientific growth, which would not have been possible without all the people I would like to thank below.

First of all, I would like to thank my supervisor **Peter Lansdorp**. Thank you **Peter** for giving me the opportunity to be the part of your lab in Groningen. I really admire your passion for science and your desire to always push the limitation of current knowledge one step further. I appreciate the freedom you gave me to pursue my own ideas and to drive my PhD project forward. I value your openness and your constructive feedback on my work, especially stressing the importance to finish every task to the very end. I enjoyed working in your lab, and I believe that the future will bring more collaborative work for us.

Next I would like to thank my second supervisor **Victor Guryev**. I see **Victor** as my mentor who was always willing to share his vast knowledge of bioinformatics analysis. Thank you for always having your door open whenever I needed to discuss any data analysis issues, regardless how trivial or difficult they might have been. Doubtlessly, your experience and expertise in bioinformatics has been crucial for my scientific and professional development. I will always value your opinion and I'm grateful for all the time you have invested in me.

I would also like to thank my third supervisor **Marianna Bevova**. Thank you **Marianna** for all the comments and suggestions regarding my main PhD project, as well as for your careful scrutiny of all my writing attempts, thesis included. I highly appreciate all your critical comments, suggestions and advice during our numerous meetings. I value you caring and supportive attitude, especially towards the end of my thesis writing.

I would also like to send thanks to the members of my assessment committee **Prof. Jan Korbel**, **Prof. Edwin Cuppen** and **Prof. Marcel van Vugt** for being so kind to provide comments on this thesis in such a short time. I feel honoured that my thesis could be evaluated by the leading experts in human genome biology.

I can't forget to thank both **Lansdorp labs** situated at opposite sites of the Atlantic ocean. From the Groningen lab I would like to thank **Diana Spierings** for leading the efforts to drive Strand-seq technology further, and for all her help to produce raw data for my research. I thank **Niek van Wietmarschen** for generating all the Strand-seq libraries for my main project. I highly appreciate all the technical support in the lab from **Nancy Halsema**, **Inge Kazemier** and **Karina Hoekstra-Wakker**. Also I want to thank all the current (**Hilda**, **Anne Margriet** and **Jorn**) and past members (**Carles**, **Sandra** and **Evert-Jan**) of the Lansdorp lab in Groningen who have shaped a productive and collaborative environment.

The Lansdorp lab on the other side of the ocean was at least as cool as the one in Groningen. I'm thankful to have had the opportunity to spend two summers (and more) with them. In Vancouver I met very talented and inspiring scientists: **Ester Falconer**, **Mark Hills, Uli Naumann**, **Geraldine Aubert**, **Ashley Sanders** and **Michael Yuen**. Some people even know them as 'Team Success'. I'm happy to have a chance to work with **Ester** and learn from her. I see **Ester** as a very important driver of scientific progress in our lab. Also I would like to thank **Mark** for all his valuable comments and suggestions on my work, as well as for reading through my very first draft of my paper, which I would describe as "NOT GOOD". Big thank you goes also to the "inversion guru" **Ashley Sanders**. **Ashley** thank for all your support and valuable advice in becoming a better scientist. It has always been great fun and a pleasure to work on our joined Strand-seq projects. I very much appreciate all your effort to make my Vancouver internship a unique experience by organizing all fun lab events.

My gratitude goes also to Germany where I have spent very productive months in the lab of **Tobias Marschall**. I'm happy that my next move in science will go to your lab. Looking forward to work with you again.

Next, I thank to **Magda Grudniewska** and **Seka Lazare** who accepted such a demanding task to be my paranymhps. Dear **Magda,** it was a long journey we have went through together and I'm happy to call you my friend and appreciate that you have never gave up on me. Thank you for all your encouragement and the support you gave me. I'm very fortunate to have such a friend like you and I hope our friendship will continue even after we move on to the next phase of our professional and personal lives. Dear **Seka** thank you for being such a positively tempered person.

I always enjoyed our competition of who will be the first at work, and vice versa, whom will be the last man standing. I'm sorry that I have almost always won. Seka I wish you a lot of success in your future career, and hope to meet you in Dominica (not Dominican Republic) one day.

Finally, I would like to thank all my friends and people close to me. I would like to start with very special person to me, also known as **Sla**. **Sla** I have always admired your infectious positivity and your positive attitude to life. Your 'never give up' attitude and your way to overcome obstacles of life was an inspiration for me. You have been a great support to me, and you always knew how to push me to overcome my doubts and fears. However, I still think that there should never be more than one person on a waterslide at a time. I'm looking forward to the next chapter of our lives, and I want you to be a part of it.

Next to receive my gratitude are my french friends: **Clemence** and **Celine**. **Clemence** I'm lucky to share my office with you from my early times in ERIBA. I can imagine I could be quite annoying while I was trying to lead a conversation with you while you were trying to focus on your work. This way I would like to thank you for your patience with me, and I wish you all the best in your future career. I'm convinced you will do great with your determination and passion for science. **Celine** thank you for being not only my colleague, but also a very good friend. Also thank you for letting me beat you at table tennis so many times.

... And many others I'm fortunate to call my friends:
THANK YOU: **Floor** for being my only 'outside work' friend and for being there when I occasionally fell. **Aaron** for all amazing collaborative work we did together, as well as the fun stuff we did outside of the work. **Stein** for being the best wingman one can imagine. **Alejandra** for checking my back on 10km run Nacht van Groningen - I was feeling much safer knowing you are there. **Tristan** for suppling me with all the dutch sweets and also for translating my summary into Dutch. **Katya** for knowing the exact location of my origin (Balkan peninsula?). **Kirill** for all the table football practice over the past moths. **Jakub** for just being my Polish friend with whom I could relate in many Slovak-Polish stereotypes.

Also I would like to thank the honorable members of the Gentelmen's club back in Slovakia: **Miroslav**, **Tinec**. **Bobulus**, **Janči**, **Pánko C**, **Škvíďo** and **Martin**.

Na záver by som sa chcel poďakovať svojej rodine za všetkú podporu, ktorú do mňa vložili. Všetko čo som doteraz dokázal je aj výsledkom úžasného rodinného zázemia. Ďakujem svojim **rodičom** za to, že ma vždy podporovali vo všetkom čo som si zaumienil, aj keď to pre nich znamenalo, že ich syn opustil rodnú krajinu a už ho nemohli vídať tak často ako by chceli. Taktiež ďakujem svojmu bratovi **Michalovi** za podporu a naše takmer každodenné telefonáty. Môj brat bol a bude pre mňa vždy vzorom hodný nasledovania v tom ako si treba ísť tvrdohlavo za svojím cieľom. Rovnako ďakujem aj dalším členom rodiny: **babičke**, **Ať'ke** a deťom (**Miškovi** a **Lenke**) a ostatným.

Curriculum Vitae

PERSONAL DETAILS

Name: **David Porubsky**
Date and place of birth: 02/01/1985, Trencin, Slovakia

EDUCATION

2007 – 2009 M.Sc. in Natural Sciences (molecular biology)
Comenius University in Bratislava, Faculty of Natural Sciences, Bratislava, Slovakia
Title: "*Approaches and methods of information integration in biological sciences"*
(graduation with excellent study results)

2004 – 2007 B.Sc. in Natural Sciences (molecular biology)
Comenius University in Bratislava, Faculty of Natural Sciences, Bratislava, Slovakia

RESEARCH POSITIONS

09/2012 – 03/2017 **Doctoral candidate**
Laboratory of Prof. Peter M. Lansdorp, Department of European Research Institute for the Biology of Ageing, University Medical Centre Groningen, Groningen, The Netherlands

03/2016 – 05/2016 **Short-term EMBO fellowship**
Laboratory of assist. Prof. Tobias Marschall, Department of Algorithms for Computational Genomics, Max-Planck-Institut für Informatik, Saarbrücken, Germany.
Project: Integrative experimental and read-based phasing of a single individual.
Supervisor: assist. Prof. Tobias Marschall

07/2014 – 08/2014 **Internship student**
Laboratory of Prof. Peter M. Lansdorp, Department of Genome Instability, Terry Fox Laboratory, BC Cancer Agency, Vancouver, Canada
Project: Development of novel tools for bioinformatic analysis of single cell sequencing data
Supervisor: Ester Falconer, PhD; Mark Hills, PhD

01/2012 – 06/2012 **Internship student**
Laboratory of Prof. Edwin Cuppen, Department of Medical Genetics in collaboration with the Hubrecht Institute, University Medical Centre Utrecht, Utrecht, The Netherlands
Project: Non-invasive prenatal diagnosis of fetal chromosomal aneuploidies using next-generation sequencing.
Supervisor: Gijs van Haaften, PhD

09/2010 – 06/2012 **Research staff**
Laboratory of Genomics and Bioinformatics, Department of Molecular biology, Comenius University in Bratislava, Bratislava, Slovakia
Responsibilities: Sanger sequencing service, Sequencing data analysis, Teaching undergraduate students

HONORS AND AWARDS

2012 National Scholarship Program of the Slovak Republic
2014 Best poster presentation award, 2nd Annual PhD student meeting of the Cancer Research Center Groningen
2015 Short-term EMBO fellowship
2016 Selected poster walk at ASHG meeting in Vancouver

List of publications

David Porubsky, Shilpa Garg, Ashley D. Sanders, Victor Guryev, Peter M. Lansdorp, Tobias Marschall. Integrative experimental and read-based phasing of a single individual. [manuscript in preparation]

Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. **2016** *Brief Bioinform*., 1–18

David Porubský, Ashley D Sanders, Niek Van Wietmarschen, Ester Falconer, Mark Hills, Diana C J Spierings, Marianna R Bevova, Victor Guryev, and Peter M Lansdorp. **2016**. Direct Chromosome-Length Haplotyping by Single-Cell Sequencing. *Genome Research* 26: 1565–1574

Ashley D Sanders, Mark Hills, **David Porubský**, Victor Guryev, Ester Falconer, Peter M Lansdorp, British Columbia, and Cancer Agency. **2016**. Characterizing Polymorphic Inversions in Human Genomes by Single Cell Sequencing. *Genome Researc*h 26:1575–1587

Bjorn Bakker, Aaron Taudt, Mirjam Belderbos, **David Porubsky**, Diana C.J. Spierings, Tristan de Jong, Nancy Halsema, et al. **2016.** Single Cell Sequencing Reveals Karyotype Heterogeneity in Murine and Human Tumours. *Genome Biology* 1–15.

Hilda van den Bos, Diana C. J. Spierings, Aaron S. Taudt, Bjorn Bakker, **David Porubský**, Ester Falconer, Carolina Novoa, et al. **2016**. Single-Cell Whole Genome Sequencing Reveals No Evidence for Common Aneuploidy in Normal and Alzheimer's Disease Neurons. *Genome Biology* 17 (1).

Yosef Buganim, Styliani Markoulaki, Niek van Wietmarschen, Heather Hoke, Tao Wu, Kibibi Ganz, Batool Akhtar-Zaidi, Yupeng He, Brian J. Abraham, **David Porubsky** et al. **2014**. The Developmental Potential of iPSCs Is Greatly Influenced by Reprogramming Factor Selection. *Cell Stem Cell* 15 (3).

Gabriel Minárik, Lukáš Plank, Zora Lasabová, Tomáš Szemes, Tatiana Burjanivová, Peter Szépe, Veronika Buzalková, **David Porubský**, and Jozef Sufliarsky. **2013**. Spectrum of Mutations in Gastrointestinal Stromal Tumor Patients - a Population-Based Study from Slovakia. *APMIS : Acta Pathologica, Microbiologica, et Immunologica Scandinavica* 121 (6): 539–48.

THE END