Algorithms for Computational Genomics

Tobias Marschall
Center for Bioinformatics, Saarland University, Saarbrücken, Germany
Max-Planck-Institute for Informatics, Saarbrücken, Germany
t.marschall@mpi-inf.mpg.de

Abstract: The topics studied in the Algorithms for Computational Genomics group range from theoretical foundations in algorithmic statistics, combinatorial optimization, and sequence algorithms to applied studies on population genetics, structural variation in human, and horizontal gene transfer in bacteria. We aim to develop algorithmic concepts as well as to provide production quality software tools. Current topics addressed in the group include structural variation calling and genotyping, read-based phasing of diploid individuals and viral quasispecies, methods for detecting horizontal gene transfer, as well as computational pan-genomics.

1 Group Development

The *Algorithms for Computational Genomics* group was established in April 2014, when Tobias Marschall was appointed assistant professor ("Juniorprofessor") at the Center for Bioinformatics at Saarland University. Since then, the group is also affiliated with the Max-Planck-Institute for Informatics where Tobias has been appointed Senior Researcher. Two PhD students (Shilpa Garg and Ali Ghaffaari) joined the group in November 2014 and April 2015, respectively. Furthermore, five Master and six Bachelor students are members of the group, working on their respective thesis projects.

2 Research Strategy

The group develops algorithms and statistical methods for computational genomics. In particular, we work on methods to analyze high-throughput sequencing data to study genetic diversity in human populations, bacterial adaptation, and cancer. On the one hand, we develop the required theoretical foundations in algorithmic statistics, combinatorial optimization, and sequence algorithms and, on the other hand, we apply the resulting methods in collaboration with biomedical researchers to gain biological insights in the aforementioned domains.

3 Research Areas

Topics addressed in the group range from algorithms for low level data processing to questions of population genetics. At present, we are particularly focusing on the following projects.

3.1 Structural Variation Calling and Genotyping

Beyond SNPs and short indels, larger genetic differences between individuals make an important contribution to genetic diversity in human populations [KUA+07, MWS+11]. Such larger events, called *structural variants (SVs)*, come in the form of deletions, insertions, duplications, translocations, inversions, and also more complex events. Detecting SVs from next-generation sequencing (NGS) data has been subject to active research, as reviewed in [MSB09] and [ACE11]. While still a postdoc at CWI Amsterdam, Tobias Marschall developed CLEVER [MCC+12] and MATE-CLEVER [MHS13], two ap-

proaches to detect deletions and insertions. The main contribution of these methods was to achieve good performance also for the particularly difficult mid-size deletions between 30–250bp (called deletion twilight zone by some). MATE-CLEVER also introduced a novel Bayesian approach for Mendelian-inheritance-aware genotyping of insertions and deletions. In a current project, we show that extending these approaches to inversions and duplications yields performance clearly superior to existing genotyping methods (yet unpublished).

In a recent publication [LMP+15], we furthermore contributed to establishing a virtual-machine based platform for benchmarking and running a multitude of SV calling algorithms. This helps to alleviate practical problems like (missing) software dependencies or incompatible data formats and, more importantly, facilitates reliable and reproducible research.

Beyond such practical problems, more fundamental issues exist regarding the seemingly simple task of comparing or merging multiple sets of SV calls. The interplay of two effects renders this a non-trivial task: on the one hand, SV callers in general do not deliver single-base-pair resolution and, on the other hand, two SVs with different breakpoint coordinates can be equivalent in the presence of repeats (in the sense that the resulting donor sequences are identical). We recently introduced a framework addressing both aspects simultaneously and provided an efficient implementation [WMSM15].

3.2 Structural Variations in the Genome of the Netherlands.

The Genome of the Netherlands (GoNL) project has sequenced the whole genomes of 750 Dutch individuals from 250 families. Applications of these data include building high-quality reference panels for imputation, studying *de novo* mutations and the corresponding mechanisms, estimating the rate of such events, and analyzing population structure, among many others. We contributed to this project [The14] as part of the Structural Variations subgroup and provided algorithms for the discovery and genotyping of structural variations, especially for "difficult" types like mid-size deletions and insertions. Furthermore, we addressed the particularly challenging task of detecting *de novo* SVs, i.e. structural variants present in a child and *not* inherited from any parent, published as [KFH+15]. Presently, we work on phasing and imputation of structural variations found in the GoNL.

3.3 Haplotype Reconstruction—Diploid Case.

Reconstructing the two haplotypes of a diploid organism (also known as phasing) is an important problem with applications in fundamental research but also in clinical settings, as discussed in [GCR14]. Emerging sequencing technologies hold the promise of allowing for read-based phasing through longer reads. On the computational side, most formalizations of the corresponding optimization problem are NP-hard. In an approach called WhatsHap [PMP+15], we demonstrated that (i) the problem instances encountered in practice can be solved using a fixed parameter tractable (FPT) algorithm and (ii) that read-based phasing indeed delivers excellent performance for long reads. In follow-up work, we contributed to an optimized parallel implementation [ABM+ar]. At present, we are extending and improving these approaches with respect to both basic methodology and algorithm engineering and work towards a production-quality software implementation (see `https://bitbucket.org/whatshap/whatshap`).

3.4 Haplotype Reconstruction—Viral Quasispecies.

Viruses like HIV exhibit a fast mutation rate and hence evolve within a host. As a result, the host is not infected by a single virus type, but by a population of genetically diverse viruses, called a *viral quasispecies* [VSA+06]. Knowledge of the spectrum of present virus haplotypes and their relative

abundances can be important for the choice of treatment. On current second-generation sequencing machines, such a virus population can be sequenced to very deep coverage at moderate cost. Reconstructing haplotypes from the resulting sequencing reads is computationally challenging, see [BGRM12]. In prior work, we met these challenges and introduced a haplotype reconstruction algorithm that is able to reconstruct full-length haplotypes and to deliver error rates that are are lower by about two orders of magnitude compared to previous approaches on simulated data [TMB+14]. In a current project, we apply algorithm engineering techniques to speed-up the enumeration of maximal cliques, which is the core algorithmic component of this method. Moreover, we study and address artifacts present in real sequencing data and reconstruct the quasispecies of a large cohort of patient plasma samples provided by our collaborators.

3.5 Computational Pan-Genomics.

Many bioinformatics methods use the reference genome of a species under study. The used reference genomes are linear, i.e. they consist of one DNA sequence per chromosome. For instance, programs to align next-generation sequencing reads will map the reads to such a linear reference genome. Likewise, tools to call variants like SNPs and structural variations do that with respect to this reference genome. Today, however, information on common and rare variants is available for many species (and, most prominently, for *Homo sapiens*). To leverage this additional information, linear reference genomes should be replaced by variant-aware reference genomes, which comes with considerable computational challenges. We develop data structures and algorithms to overcome these challenges.

Together with four co-applicants (Victor Guryev, Alexander Schönhuth, Fabio Vandin, and Kai Ye), we successfully applied at the Lorentz Center (Leiden, Netherlands) to host a workshop on this topic. The workshop was held in June 2015 and enjoyed the participation of many internationally renowned scientists[1]. At this very productive meeting, the participants drafted a white paper summarizing the state-of-the-art and pointing out future challenges in computation pan-genomics, to be submitted soon.

References

[ABM+ar] Marco Aldinucci, Andrea Bracciali, Tobias Marschall, Murray Patterson, Nadia Pisanti, and Massimo Torquati. High-Performance Haplotype Assembly. In *Proceedings of the 11th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB)*, LNBI, Cambridge, UK, June to appear. Springer.

[ACE11] Can Alkan, Bradley P. Coe, and Evan E. Eichler. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12(5):363–376, May 2011.

[BGRM12] Niko Beerenwinkel, Huldrych F. Gnthard, Volker Roth, and Karin J. Metzner. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in Microbiology*, 3, September 2012.

[GCR14] Gustavo Glusman, Hannah C. Cox, and Jared C. Roach. Whole-genome haplotyping approaches and genomic medicine. *Genome Medicine*, 6(9):73, September 2014.

[KFH+15] Wigard P. Kloosterman, Laurent C. Francioli, Fereydoun Hormozdiari, Tobias Marschall, Jayne Y. Hehir-Kwa, Abdel Abdellaoui, Eric-Wubbo Lameijer, Matthijs H. Moed, Vyacheslav Koval, Ivo Renkens, Markus J. van Roosmalen, Pascal Arp, Lennart C. Karssen, Bradley P. Coe, Robert E. Handsaker, Eka D. Suchiman, Edwin Cuppen, Djie T. Thung, Mitch McVey, Michael C. Wendl, Genome of the Netherlands Consortium, Andre Uitterlinden, Cornelia M. van Duijn, Morris Swertz, Cisca Wijmenga, Gertjan van Ommen, Eline Slagboom, Dorret I. Boomsma, Alexander Schönhuth, Evan E. Eichler, Paul I. W. de Bakker, Kai Ye, and Victor Guryev. Origin, frequency and functional impact of de novo structural changes in the human genome. *Genome Research*, 25:792–801, mar 2015.

[1] See `http://www.lorentzcenter.nl/lc/web/2015/698/participants.php3?wsid=698&venue=Oort`

[KUA+07] Jan O. Korbel, Alexander Eckehart Urban, Jason P. Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M. Kim, Dean Palejev, Nicholas J. Carriero, Lei Du, Bruce E. Taillon, Zhoutao Chen, Andrea Tanzer, A. C. Eugenia Saunders, Jianxiang Chi, Fengtang Yang, Nigel P. Carter, Matthew E. Hurles, Sherman M. Weissman, Timothy T. Harkins, Mark B. Gerstein, Michael Egholm, and Michael Snyder. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science*, 318(5849):420–426, October 2007.

[LMP+15] Wai Y. Leung, Tobias Marschall, Yogesh Paudel, Laurent Falquet, Hailiang Mei, Alexander Schönhuth, and Tiffanie Y. Maoz. SV-AUTOPILOT: optimized, automated construction of structural variation discovery and benchmarking pipelines. *BMC Genomics*, 16(1):238, March 2015.

[MCC+12] Tobias Marschall, Ivan G. Costa, Stefan Canzar, Markus Bauer, Gunnar W. Klau, Alexander Schliep, and Alexander Schönhuth. CLEVER: clique-enumerating variant finder. *Bioinformatics*, 28(22):2875–2882, November 2012.

[MHS13] Tobias Marschall, Iman Hajirasouliha, and Alexander Schönhuth. MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. *Bioinformatics*, 29(24):3143–3150, dec 2013.

[MSB09] Paul Medvedev, Monica Stanciu, and Michael Brudno. Computational methods for discovering structural variation with next-generation sequencing. *Nat Meth*, 6(11s):S13–S20, November 2009.

[MWS+11] Ryan E. Mills, Klaudia Walter, Chip Stewart, Robert E. Handsaker, Ken Chen, Can Alkan, Alexej Abyzov, Seungtai Chris Yoon, Kai Ye, R. Keira Cheetham, Asif Chinwalla, Donald F. Conrad, Yutao Fu, Fabian Grubert, Iman Hajirasouliha, Fereydoun Hormozdiari, Lilia M. Iakoucheva, Zamin Iqbal, Shuli Kang, Jeffrey M. Kidd, Miriam K. Konkel, Joshua Korn, Ekta Khurana, Deniz Kural, Hugo Y. K. Lam, Jing Leng, Ruiqiang Li, Yingrui Li, Chang-Yun Lin, Ruibang Luo, Xinmeng Jasmine Mu, James Nemesh, Heather E. Peckham, Tobias Rausch, Aylwyn Scally, Xinghua Shi, Michael P. Stromberg, Adrian M. Stutz, Alexander Eckehart Urban, Jerilyn A. Walker, Jiantao Wu, Yujun Zhang, Zhengdong D. Zhang, Mark A. Batzer, Li Ding, Gabor T. Marth, Gil McVean, Jonathan Sebat, Michael Snyder, Jun Wang, Kenny Ye, Evan E. Eichler, Mark B. Gerstein, Matthew E. Hurles, Charles Lee, Steven A. McCarroll, and Jan O. Korbel. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, February 2011.

[PMP+15] Murray Patterson*, Tobias Marschall*, Nadia Pisanti, Leo van Iersel, Leen Stougie, Gunnar W. Klau, and Alexander Schönhuth. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *Proceedings of RECOMB / Journal of Computational Biology*, 22(6):498–509, feb 2015.

[The14] The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, 46:818–825, June 2014.

[TMB+14] Armin Töpfer, Tobias Marschall, Rowena A. Bull, Fabio Luciani, Alexander Schönhuth, and Niko Beerenwinkel. Viral Quasispecies Assembly via Maximal Clique Enumeration. *Proceedings of RECOMB / PLoS Computational Biology*, 10(3):e1003515, mar 2014.

[VSA+06] Marco Vignuzzi, Jeffrey K. Stone, Jamie J. Arnold, Craig E. Cameron, and Raul Andino. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature*, 439(7074):344–348, January 2006.

[WMSM15] Roland Wittler*, Tobias Marschall*, Alex Schönhuth, and Veli Mäkinen. Repeat- and Error-Aware Comparison of Deletions. *Bioinformatics*, advance online, 2015.