
Computational challenges in disease mapping

Technological improvements over the last decade have greatly increased our ability to study biological systems: We can track the regulation of genes over time, in different cell types and under different conditions; we can measure the vast number of proteins present in different cells under different conditions; and we can measure the genome wide genetic variation between individuals. The dramatic increase in data collection technology, however, has not been followed by an equal development in the computational methods for analysing the data. An important challenge in the coming years is the development of analysis methods, both efficient at extracting relevant information from the data and computationally feasible when applied to massive datasets.

The aim of this project is to develop analysis methods for genetic disease mapping – the search for the genes behind genetic diseases – scalable to large genome wide studies. The main focus will be on *i)* gene-gene and gene-environment interaction; *ii)* integration of various sources of genomics, population genetics, and systems biology data; *iii)* dealing with complex phenotypes; and *iv)* coping with new types of data sources emerging over the next decade, such as structural variation data and complete genomic re-sequencing.

By improving the methods we use to analyse the genetic component of diseases we will hopefully discover more of the genes affecting disease risk, learn more about the biological processes underlying the diseases, and eventually be able to develop improved medicine for curing or preventing diseases. The disease mapping step is only a small part of this process, but an important first step.

Background, state-of-the-art, and future challenges

Sequencing of the human genome has been followed up by a survey of genetic variation in humans. Today, more than 10 million positions in the genome are known to show variation and have been surveyed in the human HapMap project.¹ This has allowed the design of cost effective technology for measuring the genetic differences between individuals on a genome wide scale, and opened up for unprecedented large-scale studies of genetic causes of common diseases, e.g. prostate and breast cancer,²⁻⁴ type 1 and type 2 diabetes,^{3,5-10} and Crohn's disease.^{11,12}

Disease mapping approaches: Methods for efficient analysis of such genetic data are lagging behind the data acquisition technology, however. With the current technology, the full genetic variation is not explicitly measured, and instead it must be inferred through statistical methods. Analysis is typically carried out by testing each measured genetic variant independently, but such tests are not efficient unless the disease causing variant is included among the variants explicitly measured,¹³ which is usually not the case. Indeed, it has been calculated that for several of the confirmed findings in a recent very large study⁵ the chance of recognising the disease association was less than 1%.¹⁴

Methods analysing several genetic variants at the same time can improve on the statistical power to detect disease causing genes, and several methods have been developed.¹⁵⁻²³ These methods, however, are very computational intensive, and can presently not handle large datasets. New approaches must be developed, balancing method sophistication with pragmatic resource constraints; approaches combining algorithmic techniques with biological knowledge to extract a maximum of information from the data, at a minimal computational cost.

Gene-gene and gene-environment interaction: There is growing evidence that many diseases are affected by complex interactions of multiple genetic and environmental effects, and constructing analysis methods for dealing with this is an area of active research.²⁴⁻³⁰ Some genes affect the disease risk only when certain other genes are present and not otherwise – or even reduce disease risk when some genes are present and increase the risk when other genes are present.^{31,32} In such case, the gene may show no disease association when considered alone; only when other genes are taken into account does the disease signal manifest. Ideally, a study should consider e.g. all pairs of gene variants^{33,34} – and for simple methods, this is a feasible approach. For more sophisticated methods, however, the computing resources needed make this prohibitive.

Integrating several sources of data: Using sequence information from the human genome project in a mapping context has proven successful and improves efficiency e.g. by guiding the search for disease genes to areas known to be functionally important,²⁰ or by using so-called recombination hotspots^{35,36} to select multiple variants for simultaneous analysis.² Adding more data, and more types of data, holds the promise of improving our mapping efficiency even further, but integrating such varied types of data will require more sophisticated methods with additional computation challenges. Furthermore, the integration of various data sources and the structuring of data in forms that can efficiently be analysed is an informatics challenge in itself.

Types of genotype data: Not only can we expect the amount of data to grow in the coming years, we will also see a change in the type of (primary) data. Most studies to date, have used so-called SNP (*single nucleotide polymorphism*) data. Recently, however, it has become clear that another kind of variation, *copy-number variants* or CNV data, contributes significantly more to the variation between individuals (estimates are 0.12%–0.2% differences in CNV between two random individuals compared to 0.08% for SNP data).³⁷ Early studies have found association between copy number variants and disease both for sporadic diseases,^{38,39} mendelian diseases,^{40,41} and complex diseases/phenotypes.^{42,43} Technology is now being developed to enable genome wide measurements of CNV data, and in the next year or two, we can expect the first CNV based studies. Disease mapping based on CNV data introduces a set of new challenges, including: high measurement noise in measuring genotypes⁴⁴ and lack of population genetic models (although some work is being done in this direction⁴⁵). The latter means that techniques based on inference of genealogies^{19,22,23,46,47} cannot immediately be applied; the former that the genotyping uncertainty needs to be explicitly modelled in the mapping method.

The ultimate data for disease studies is of course the full genomic sequences of each individual, i.e. the complete DNA sequence for each individual in the study. Obtaining this is currently technically as well as economically prohibitive, but sequencing technology is improving rapidly, and within a few years it will be a cost-effective alternative. Methods dealing with full sequences will be very different from the current methods: for complete re-sequencing, inference of unobserved variation based on the measured variation will no longer be necessary: The disease variants will always be directly observed. Instead, the size of the data and the increased risk of false positives will be a major problem.

My own background

As a computer scientist, I am experienced in algorithm development⁴⁸⁻⁶⁰ and computer-tool development.^{23,61-64} This background gives me a different approach to the field of disease association mapping that is otherwise generally approached statistically. A main approach in the field so far has been accurate and sophisticated modelling in

the methods development to achieve maximal statistical power – with the result that the methods do not scale to realistic datasets, thus actual data analysis has been done with very unsophisticated methods.

The last three years, association mapping has been my main research interest.^{23,46,61,65} A key focus has been the development of computationally efficient methods and computer tools – a list of developed software is available at <http://www.birc.au.dk/~mailund/association-mapping/> – and the trade-off between computational efficiency and method sophistication. Of special interest in this regard is the *Blossoc* method.⁴⁶ This method combines a population genetics model with very fast algorithms leading to a mapping approach scalable to large data sets, with a mapping accuracy comparable or superior to methods orders of magnitude slower.

Project objectives

The main objectives of the first part of the proposed project are to extend the methods I have already developed – and to develop new methods as needed – to move beyond a simple genotype data mining approach and exploit the wealth of information available about the human genome and human population from various other sources. The methods I have developed so far attempt to locate disease risk variations based only on the data collected for the study in question – as do most methods in the field. Integrating disease mapping data with population genetics data from e.g. the HapMap project, or with genomics information such as gene interaction databases, can potentially greatly improve these methods and help guide the search for associated markers.

The objective of the second part of the project is development of methods for new kinds of primary data: structural variation (CNV data) and complete genomic re-sequencing. These objectives require major changes to existing methods, or completely new methods: The existing population genetics and molecular evolution theory will be used, but the new data will change how we search for disease-related genotypic variants. To avoid a lag between genotyping technology and analysis methods – similar to the lag existing today – method development of the new types of data should start as soon as possible and not wait for the availability of the data.

The areas I wish to focus on are the following:

Gene-gene and gene-environment interaction: If interaction between genes, and between genes and environment, need to be taken into account, the search space to explore when searching for disease genes explodes in size. For simple statistical tests each pair of genes can feasibly be explored, but for more sophisticated tests – necessary to avoid too many false positives in the search – the search time makes an exhaustive exploration impractical. Considering three or more genes, and all combinations of these, will be impossible even for simple and fast methods.

One way to alleviate this and reduce the state space to something more manageable, is to exploit the existing knowledge about gene-gene interactions. The knowledge about which genes interact – at the protein level or through regulatory mechanisms – is growing fast, and it is conceivable that we will know most gene interactions within a few years. By only considering genes known to interact, we can potentially reduce the number of tests by orders of magnitude, improving on both CPU usage and statistical power. For gene-environment interaction, integrating environment as co-variables in the analysis³⁰ should be attempted. This will, however, increase the search space, thus reduction techniques will be needed here as well.

Considering interactions between more than just pairs of genes, even with a reduced search space, will probably be infeasible through exhaustive exploration, and instead heuristic search methods must be developed. To improve the search time further, ap-

proaches to parallelise computations should be considered. Parallelisation needs to be considered early in the method development, since some heuristic search strategies tend to be easier to parallelise than others (e.g. genetic algorithms vs. Markov-Chain Monte Carlo).

Incorporating genomics and population genetics data: All my methods analyse primary genotype data only, and they are oblivious to knowledge gained from e.g. the human genome project or the HapMap project. Incorporating such knowledge should improve the methods. A starting point here would be to extend my methods – especially the *Blossoc* method – to make use of the recombination rate map of the human genome,^{35,36,66} and to make use of known so-called *haplotype blocks* and known *haplotype phase*. With such information available, the statistical inference in the *Blossoc* method will be both faster and more accurate. Achieving this improvement will require both informatics development – dealing with the different kinds of data and combining them in an optimal way – and algorithmic development – improving the algorithms underlying *Blossoc* to exploit the additional information.

Complex phenotypes: With improved technology for measuring biological systems, we can expect more complex phenotypes (observable characteristics) to appear in future studies, e.g. gene expression levels for large sets of genes, or time series of expression levels of selected genes. Some dimensionality reduction will be necessary to deal with such phenotypes, but reducing these high-dimensional data points to one-dimensional values, such that current mapping methods can cope with them, will be too drastic a reduction. Instead I propose to use a clustering based association scoring, combining genotype clustering^{18,21,46} with phenotype clustering (similar to how the *Blossoc* method deals with association mapping of simple phenotypes): when genotype clustering matches phenotype clustering, that would be taken as a signal of association. Essential for this approach is computationally efficient clustering methods and, dependent on the type of complex phenotypes, memory and IO efficient management of the data.

New genotyping and sequencing technologies: My approach to dealing with CNV data will depend on the development of population genetic theory for such data. If promising models appear, I will attempt to modify the *Blossoc* method to deal with CNV data. Alternatively, I will take a clustering approach: Local similarity measures^{18,21} will be used to cluster genotypes, and association will be measured by the corresponding clustering of phenotypes.⁴⁶

For full genomic re-sequencing, my approaches will also be based on co-clustering of genotypes and phenotypes. The major challenge here, however, will most likely be efficient management of the very large dataset, when the full genome of hundreds or thousands of individuals must be analysed.

Time schedule

1st year: Explore the benefit of using known gene-gene interactions for detecting disease association. Extend my methods (at least the *Blossoc* method) to test for association in pairs of genes and to use gene-gene interaction networks to select the pairs to examine. Extend methods to consider co-variates such as environmental factors. Develop software framework for parallel execution of methods (together with B. Vinter's group; see collaborators below). Promising methods will be applied to the POLYGENE dataset (see collaborators below).

2nd year: Extend methods to exploit recombination rate maps and known haplotype blocks from e.g. HapMap. Extend methods to consider complex phenotypes. Again,

promising methods will be used on the POLYGENE datasets. Explore and develop models for CNV data and models for association mapping based on CNV data.

3rd year: Develop methods for analysing full re-sequencing data. This will be as much an informatics challenge as a computational challenge. None of my existing methods will be able to scale to this challenge, so it is necessary to develop completely new methods. The parallel computation framework developed for the existing methods should be extendable to this, however.

Research environment

The project will be carried out at the Bioinformatics Research Center (BiRC), University of Aarhus. BiRC is a collaborative effort between The Faculty of Science and The Faculty of Health Sciences, and hosts an equal mix of statisticians, computer scientists and biologists. At BiRC, one professor, two associate professors, two postdocs (myself included) and three Ph.D. students work on related problems and will be close discussion partners and collaborators in this project.

Collaborators: A close collaboration is already established with Prof. Jotun Hein's group at University of Oxford, UK; Prof. Bart Kiemeneij, Radboud University Nijmegen Medical Centre, Netherlands, and the statistical department at DeCODE Genetics, Iceland, within the EU funded POLYGENE project. The POLYGENE project – www.polygene.eu – gives me a unique opportunity to apply my methods on real data. At DeCODE Genetics, one of the world's largest genotype datasets is available and the statisticians at DeCODE Genetics are among the world leaders in genome wide disease mapping studies.

During my postdoctoral stay at University of Oxford I established collaborations with Prof. David Balding, Imperial College, London, UK, and Dr. Yun S. Song, UC Davis and Berkley, US. Both collaborations concern methods and algorithmical development and will continue in this proposed project.

For CPU intensive computations, I am collaborating with Prof. Brian Vinter's group at University of Copenhagen.^{67,68} Vinter's group develop grid computing technology, and for the last three years I have had access to their resources as a test user. I will have access to the computer resources on their grid system and collaborate in the further development of their system.

At BiRC we also have close collaborations with Prof. Lars Bolund, Institute of Human Genetics, and through him, Prof. Wang Jun, Beijing Genomics Institute, China. In the proposed project we will use re-sequencing technology to analyse a candidate gene currently being investigated at the Institute of Human Genetics.

Publication outcomes and career plan

Developed methods will be published in relevant journals (e.g. *Genetics*, *Genetic Epidemiology*, *Bioinformatics*) and implemented in computer tools that will be made available on the Internet as open source software for other research groups. Through my collaborations, the computer tools will be applied in real disease mining studies. Any findings in such studies will be published together with my collaborators.

The project will be used to develop my profile at BiRC, hopefully resulting in an associate professorship in this research centre. BiRC has a strong profile in statistical modelling and population genetics, and has several people working on problems related to the proposed project. My computer science background and focus on computational/algorithmic challenges complement the more statistical/modelling focus of my colleagues at BiRC and have, so far, lead to fruitful collaborations.

References

1. International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437 (7063):1299–1320, 2005.
2. J. Gudmundsson *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet*, 39(5):631–7, 2007.
3. J. Gudmundsson *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet*, 39(8):977–83, 2007.
4. D.F. Easton *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148):1087–93, 2007.
5. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, 2007.
6. R. Saxena *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829):1331–6, 2007.
7. E. Zeggini *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, 316(5829):1336–41, 2007.
8. J.A. Todd *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet*, 39(7):857–64, 2007.
9. L.J. Scott *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, 316(5829):1341–5, 2007.
10. D.J. Smyth *et al.* A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nat Genet*, 38(6):617–619, 2006.
11. R.H. Duerr *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*, 314(5804):1461–3, 2006.
12. M. Cargill *et al.* A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes. *Am J Hum Genet*, 80(2):273–90, 2007.
13. I. Pe'er *et al.* Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet*, 38(6):663–667, 2006.
14. D. Altshuler and M. Daly. Guilt beyond a reasonable doubt. *Nat Genet*, 39(7):813–5, 2007.
15. F. Larribe *et al.* Gene mapping via the ancestral recombination graph. *Theor Popul Biol*, 62 (2):215–229, 2002.
16. J. Li *et al.* Haplotype-based quantitative trait mapping using a clustering algorithm. *BMC Bioinformatics*, 7(1):258, May 2006. doi: 10.1186/1471-2105-7-258.
17. J.S. Liu *et al.* Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res*, 11(10):1716–1724, 2001. doi: 10.1101/gr.194801.
18. J. Molitor *et al.* Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *Am J Hum Genet*, 73(6):1368–1384, 2003.
19. A.P. Morris *et al.* Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am J Hum Genet*, 70(3):686–707, 2002.
20. B. Rannala and J.P. Reeve. High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *Am J Hum Genet*, 69(1):159–178, 2001.
21. E.R.B. Waldron *et al.* Fine mapping of disease genes via haplotype clustering. *Genet Epidemiol*, 30(2):170–179, 2006.

22. S. Zöllner and J.K. Pritchard. Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, 169(2):1071–1092, 2005. doi: 10.1534/genetics.104.031799.
23. T. Mailund *et al.* GeneRecon—a coalescent based tool for fine-scale association mapping. *Bioinformatics*, 22(18):2317–8, 2006.
24. H. Zhao *et al.* Test for interaction between two unlinked loci. *Am J Hum Genet*, 79(5):831–45, 2006.
25. J. Millstein *et al.* A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Genet*, 78(1):15–27, 2006.
26. M.D. Ritchie *et al.* Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol*, 24(2):150–7, 2003.
27. L.W. Hahn *et al.* Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, 19(3):376–82, 2003.
28. J.H. Moore and L.W. Hahn. A cellular automata approach to detecting interactions among single-nucleotide polymorphisms in complex multifactorial diseases. *Pac Symp Biocomput*, pages 53–64, 2002.
29. M.D. Ritchie *et al.* Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, 69(1):138–47, 2001.
30. A. Albrechtsen *et al.* A Bayesian multilocus association method: allowing for higher-order interaction in association studies. *Genetics*, 176(2):1197–208, 2007.
31. J.A. Staessen *et al.* Effects of three candidate genes on prevalence and incidence of hypertension in a Caucasian population. *J Hypertens*, 19(8):1349–58, 2001.
32. A. Balmain and C.C. Harris. Carcinogenesis in mouse and human cells: parallels and paradoxes. *Carcinogenesis*, 21(3):371–7, 2000.
33. D.M. Evans *et al.* Two-stage two-locus models in genome-wide association. *PLoS Genet*, 2(9):e157, 2006.
34. J. Marchini *et al.* Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*, 37(4):413–417, 2005.
35. S. Myers *et al.* A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–4, 2005.
36. G.A. McVean *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–4, 2004.
37. J. Sebat. Major changes in our DNA lead to major changes in our thinking. *Nat Genet*, 39(7):S3–S5, 2007.
38. J.R. Lupski. Genomic rearrangements and sporadic disease. *Nat Genet*, 39(7):S43–47, 2007.
39. K. Inoue and J. R. Lupski. Molecular mechanisms for genomic disorders. *Annu Rev Genomics Hum Genet*, 3:199–242, 2002.
40. C. Le Marechal *et al.* Hereditary pancreatitis caused by triplication of the trypsinogen locus. *Nat Genet*, 38(12):1372–4, 2006.
41. Q.S. Padiath *et al.* Lamin B1 duplications cause autosomal dominant leukodystrophy. *Nat Genet*, 38(10):1114–23, 2006.
42. T.J. Aitman *et al.* Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature*, 439(7078):851–5, 2006.

43. E. Gonzalez *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, 307(5714):1434–40, 2005.
44. S.A. McCarroll and D.M. Altshuler. Copy-number variation and association studies of human disease. *Nat Genet*, 39(7):S37–42, 2007.
45. D.F. Conrad and M.E. Hurles. The population genetics of structural variation. *Nat Genet*, 39(7):S30–36, 2007.
46. T. Mailund *et al.* Whole genome association mapping by incompatibilities and local perfect phylogenies. *BMC Bioinformatics*, 7(454), 2006.
47. M. J. Minichiello and R. Durbin. Mapping trait loci by use of inferred ancestral recombination graphs. *Am J Hum Genet*, 79(5):910–22, 2006.
48. C. Christiansen *et al.* Fast calculation of the quartet distance between trees of arbitrary degrees. *Algorithms Mol Biol*, 1:16, 2006.
49. T. Mailund *et al.* Recrafting the neighbor-joining method. *BMC Bioinformatics*, 7:29, 2006.
50. C. Christiansen *et al.* Algorithms for Computing the Quartet Distance between Trees of Arbitrary Degree. In *Proceedings of Workshop on Algorithms in Bioinformatics (WABI)*, volume 3692 of LNBI, pages 77–88. Springer-Verlag, 2005.
51. C. Christiansen *et al.* Quartet Distance between General Trees. In *Proceedings of International Conference on Numerical Analysis and Applied Mathematics (ICNAAM)*, pages 796–799. Wiley-VCH Verlag GmbH & Co., 2005.
52. M. Stissing *et al.* Computing the All-Pairs Quartet Distance on a Set of Evolutionary Trees. In *Proceedings of the Asia-Pacific Bioinformatics Conference (APBC)*, volume 5 of *Series on Advances in Bioinformatics and Computational Biology*, pages 91–100. Imperial College Press, 2007.
53. M. Stissing *et al.* Computing the Quartet Distance between Evolutionary Trees of Bounded Degree. In *Proceedings of the Asia-Pacific Bioinformatics Conference (APBC)*, volume 5 of *Series on Advances in Bioinformatics and Computational Biology*, pages 101–110. Imperial College Press, 2007.
54. S. Christensen *et al.* A Sweep-Line Method for State Space Exploration. In *Proceedings of Tools and Algorithms for the Construction and Analysis of Systems (TACAS 2001)*, volume 2031 of LNCS, pages 450–464. Springer-Verlag, 2001.
55. T. Mailund. Analysing Infinite-State Systems by Combining Equivalence Reduction and the Sweep-Line Method. In *Proceedings of International Conference on Application and Theory of Petri Nets (ICATPN 2002)*, volume 2360 of LNCS, pages 314–333. Springer-Verlag, 2002.
56. L.M. Kristensen and T. Mailund. A Compositional Sweep-Line State Space Exploration Method. In *Proceedings of Formal Description Techniques for Distributed Systems and Communication Protocols (FORTE 2002)*, volume 2529 of LNCS, pages 327–343. Springer-Verlag, 2002.
57. L.M. Kristensen and T. Mailund. A Generalised Sweep-Line Method for Safety Properties. In *Proceedings of Formal Methods Europe (FME 2002)*, volume 2391 of LNCS, pages 549–567. Springer-Verlag, 2002.
58. L.M. Kristensen and T. Mailund. Efficient Path Finding with the Sweep-Line Method using External Storage. In *Proceedings of International Conference on Formal Engineering Methods (ICFEM 2003)*, volume 2885 of LNCS, pages 319–337. Springer-Verlag, 2003.
59. T. Mailund and M. Westergaard. Obtaining Memory-Efficient Reachability Graph Representations Using the Sweep-Line Method. In *Proceedings of Tools and Algorithms for the Construction and Analysis of Systems (TACAS 2004)*, volume 2988 of LNCS, pages 177–191. Springer-Verlag, 2004.

60. J. Billington *et al.* Exploiting Equivalence Reduction and the Sweep-Line Method for Detecting Terminal States. *IEEE Transactions on Systems, Man, and Cybernetics*, 34(1):23–37, 2004.
61. T. Mailund *et al.* CoaSim: a flexible environment for simulating genetic data under coalescent models. *BMC Bioinformatics*, 6:252, 2005.
62. S. Besenbacher *et al.* RBT—a tool for building refined Buneman trees. *Bioinformatics*, 21(8):1711–2, 2005.
63. T. Mailund and C.N.S. Pedersen. QDist—quartet distance between evolutionary trees. *Bioinformatics*, 20(10):1636–7, 2004.
64. T. Mailund and C.N.S. Pedersen. QuickJoin—fast neighbour-joining tree reconstruction. *Bioinformatics*, 20(17):3261–2, 2004.
65. T. Bataillon *et al.* The effective size of the Icelandic population and the prospects for LD mapping: inference from unphased microsatellite markers. *Eur J Hum Genet*, 14(9):1044–53, 2006.
66. G.A. McVean *et al.* Perspectives on human genetic variation from the HapMap Project. *PLoS Genet*, 1(4):e54, 2005.
67. T. Mailund *et al.* Experiences with GeneRecon on MiG. *Future Generation Computer Systems*, 23:580–586, 2007.
68. T. Mailund *et al.* Initial experiences with GeneRecon on MiG. In *Proceedings of The 2005 International Conference on Grid Computing and Applications (GCA'05)*, 2005.