Automatic fusion of genetic variant calling pipelines and the Bayesian relevance analysis in candidate gene association studies

PhD thesis

András Gézsi

Semmelweis University
Doctoral School of Molecular Medicine

Supervisor:
Csaba Szalai DSc

Opponents:
Zsolt Rónai PhD
Zoltán Maróti PhD

Chairman of the Final Exam Committee:
Elek Dinya CSc

Members of the Final Exam Committee:
András Kiss PhD
Béla Pataki PhD

Budapest
2016

Introduction

Genetic and genomic research are of growing importance in medical sciences. As a result of sequencing the entire human genome and of the rapid development of ever faster and cheaper sequencing technologies, personalized medicine has become part of the clinical routine in some areas. However, the unprecedented increase of sequencing data poses significant challenges for the physicians, biologists and bioinformatics who wish to interpret and analyze the data. The biological conclusions drawn from next-generation sequencing (NGS) studies are heavily based on the accuracy of the called variants and genotypes, which still does not always reach the level of clinical acceptance. Therefore, bioinformatic methods that can improve the accuracy of the variant calls, can greatly contribute to the broadest possible application of the technology. In my work, I developed a software, which combines different variant callers, and achieves improved performance over the individual methods.

Analysis of genetic variants has a central role in exploring the patomechanism of diseases, in detecting factors that predispose to or affect the recovery from a disease, and in developing more effective treatments and therapeutic protocols. Methods based on Bayesian statistics are of growing use in genetic data analysis. In the course of my work I was involved in the development of a Bayesian relevance analysis methodology that can be used to characterize the complex dependencies of genetic variants and phenotypic features, and by this it provides an effective alternative to frequentist statistical methods to analyze data in association studies. I present the usability and benefits of the Bayesian method through analyzing polymorphisms that influence the susceptibility and survival rate of childhood acute lymphoblastic leukemia.

Next-generation sequencing

The next-generation sequencing technology has revolutionized among others the human genetic and genomic research. By sequencing the complete human genome or the exome, the genetic background of rare and complex diseases can be discovered. Due to the continuous development and competition of the manufacturing companies, sequencing machines continuously appear with increasingly higher throughput,

and the unit cost for determining a base is getting cheaper.

The measurement process includes a complex computational variant calling pipeline, which contains many alternative elements with various parameters, heavily influencing the unique characteristics and performance of the whole procedure. Several studies showed that (1) currently there is no single best general individual variant calling method with both superior sensitivity and precision at all circumstances, and (2) there are significant discrepancies between commonly used variant calling pipelines, even when applied to the same set of sequence data.

Bayesian network-based relevance analysis

As a member of the bioinformatic working group of the Budapest University of Technology and Economics, with the leadership of Peter Antal PhD, I was involved in the development of a statistical methodology that can be used to analyze genetic association data among other things. The methodology uses so called Bayesian networks to model the variables of the subject area and Bayesian statistical methods to determine the probability of the complex dependencies between the variables.

In genetic association studies, our goal is to identify the genetic variants that influence a particular phenotype (e.g. the development of a disease). In other words, we actually want to know the complex interrelationship of genotype and phenotype. Each observation (sample) can be considered a particular state of the system that we want to learn, which state can be described by the genotype and phenotypic features of the sample. If these are considered random variables, then the goal can also be formulated as to describe the joint probability distribution or the structure of the subject area. For this task, we use Bayesian network-based relevance analysis.

Acute lymphoblastic leukemia

Leukemias are hematopoietic malignancies in which abnormal proliferation of white blood cells overgrow normal cells and infiltrates various organs (e.g. bone marrow, nervous system, eye etc.). Among the different types of leukemia, acute lymphoblastic leukemia (ALL) is the most common (about 80%) which is a multifactorial disease; both genetic and environmental factors influence the development of the disease.

The five-year survival rate of ALL today is about $80 - 90\%$. In the future, by identifying new drug targets, by personalized therapy, and by eliminating late side effects and drug toxicity hopefully this ratio can be further improved. To help to achieve these goals various pharmacogenetic studies can be performed that analyze the effects of variants in genes encoding drug metabolizing enzymes, drug targets and transporters.

In the human body the members of the cytochrome P450 enzyme family play a role in the oxidative metabolism of several different types of pharmaceutical and endogenous substances. CYP3A4 is the most abundant CYP450 enzyme in the liver and the gut, and the main drug-metabolizing protein in humans. It has an important role in the metabolisms of many drugs used in ALL therapy, for example, vincristine, cyclophosphamide, dexamethasone and doxorubicin. The substrate specificity of CYP3A5 overlaps with that of CYP3A4. Still, there are practically no publications about the role of *CYP3A4* or *CYP3A5* polymorphisms in ALL pharmacogenomics. One possible reason for this is that the frequency of the functionally relevant variations in the gene that affect their expression level are relatively low (maximum about $4 - 5\%$) and thus the studies were underpowered in the usually small ALL populations. However, the size of the biobank available at the Department of Genetics, Cell- and Immunobiology allows the analysis of these genes is much higher number of samples.

It is well known that individual alleles do not act alone, but in interaction with other genetic and environmental factors. Unfortunately, these interactions are very difficult to detect, especially, when larger numbers of variables are studied. Therefore, those methods play an important role in pharmacogenetic studies that are capable of discovering the combined effects of variables, as well as their interactions.

Objectives

In the course of my work I had the following objectives:

1. To assess and compare the performance and concordance rates of different variant calling pipelines, in particular to analyze the impact of the alignment software and the sequencing depth. To assess the impact of the current hard filtering recommendations on the sensitivity and precision of the variant calling pipelines.

2. To develop a software that combines the results of individual variant calling methods, over which it utilizes variant quality annotation information. The software should allow the precision based filtering of the variants by estimating their probability for smaller genomic regions or in case of a few samples, *without* using highly reliable variant call sets. To compare the performance of the newly developed method with the individual variant callers, and with an alternative combinational software (BAYSIC).

3. To analyze the effects of common *CYP3A4* and *CYP3A5* polymorphisms on the survival of childhood ALL, to search for interactions using Bayesian relevance analysis.

4. To test the Bayesian network-based relevance analysis methodology and to compare it with frequentist statistical methods in association studies. In doing so, test the Bayesian methodology, compare its results with that of the frequentist tests, analyze the results from a methodological point of view, and based on these, develop further the Bayesian methodology.

Methods

Analyzing the performance and concordance of variant calling pipelines, combining the variant calls with VariantMetaCaller

In order to compare the performance of previous variant calling pipelines to that of our newly developed method, VariantMetaCaller, we created synthetic sequencing data with known variations in the reference genome. The true variants of the artificially generated samples served as the reference call set during the comparisons, i.e. we measured sensitivity and precision of the methods with respect to these variants. Besides simulated data, we used real sequencing data downloaded from the Illumina base space website. For this genome, a high-confidence reference call set is also available which we used as "gold standard" during the evaluations.

We aligned the quality filtered sequencing reads to the hg19 reference genome with BWA–MEM and, alternatively, with Bowtie 2. Then, we used four variant callers to detect SNPs and short indels based on the two different alignments: GATK UnifiedGenotyper and HaplotypeCaller, FreeBayes and SAMtools combined with BCFtools. Hard filters were set according to the GATK Best Practices recommendations for GATK called variant sets. For FreeBayes and SAMtools we filtered the variants by setting a minimum variant quality threshold.

VariantMetaCaller combines the unfiltered results of multiple variant callers using support vector machines (SVM). After merging the unfiltered variant calls, the program creates a data set for each input method from annotations generated by the callers coupled with annotations computed by VariantMetaCaller. On these data sets, SVMs are trained separately for SNPs and indels using fully concordant and singly-called variants as positive and negative training examples, respectively. Given a variant caller, we compute the conditional probability of each variant being a "real" variant (i.e. belonging to the positive class), where callers have equal probabilities. Then, the final score is the probability of variants marginalized over all variant callers.

We compared the performance of the variant callers and VariantMetaCaller by plotting precision-recall curves.

Analyzing the potential effects of CYP3A4 and CYP3A5 on the pharmacogenetics of childhood acute lymphoblastic leukemia

During the analyses of genetic and environmental factors that affect the survival of childhood acute lymphoblastic leukemia, we used the biobank available at the Institute of Genetics, Cell-, and Immunobiology, Semmelweis University. In a retrospective manner, DNA was obtained from peripheral blood samples of 511 children who were diagnosed with ALL between 1990 and 2010.

We selected six SNPs in the *CYP3A4* gene and two SNPs in the *CYP3A5* gene based on their minor allele frequencies in Caucasian populations or possible relevancy in ALL chemotherapy.

Genomic DNA was isolated with QIAmp DNA Blood Maxi Kit (Qiagen, Hilden, Germany). SNPs were genotyped by Sequenom iPLEX Gold MassARRAY technology at the McGill University and Génome Québec Innovation Centre, Montréal, Canada.

For data analysis, R statistical software (R Foundation for Statistical Computing, Vienna, Austria; version 3.0.3) was used. Cox proportional hazards regression models were applied for uni- and multivariate analysis. Power analysis was conducted by bootstrapping.

To determine the discriminative power of different risk-group assessment variables, we calculated the C-index (condordance index). Confidence intervals were estimated by bootstrapping. The differences between the risk-group assessment variables were evaluated by a two-sided t-test at each time point. P-values were adjusted using the Benjamini and Hochberg method.

In addition, the Bayesian network-based relevance analysis was used to detect strongly relevant variables with respect to 5-year event-free survival and overall survival as a target variable. We used Metropolis-coupled Markov Chain Monte Carlo (MC^3) methods to draw samples from the possible Bayesian networks. We computed *a posteriori* probabilities for different dependency types between the variables, and for the strongly relevant sets with respect to the target variables. We also computed the interactions and redundancies between the variables.

Results

Precision and concordance of variant calling pipelines

We systematically assessed the performance of the individual variant callers using synthetic sequencing data with known variations in the reference genome. Our results showed, in agreement with previous findings, that there was not a single best general individual variant calling method with superior sensitivity and precision at all read depths, neither for SNPs nor indels, although HaplotypeCaller performed quite well in case of indels and was the most precise in case of SNPs. The sensitivity of the variant callers increased with increasing coverage depth in most cases. Interestingly however, the number of falsely called variants also increased above a certain read depth, i.e. the precision of the individual variant callers decreased with increasing coverage depth.

The sensitivity and the precision of the variant callers were markedly higher for SNPs than for indels at the same coverage depths. Our results show that a significant increase of coverage depth is needed to achieve the same sensitivity levels (for example $200\times$ coverage for indels and $16\times$ coverage for SNPs provides approximately equal sensitivity in case of HaplotypeCaller).

The choice of the aligner software heavily influenced the results of variant calling. The variant callers generally achieved significantly higher accuracy when BWA, as opposed to Bowtie 2 was used.

We also showed that there were significant discrepancies between commonly used variant calling pipelines. The choice of the aligner software also influenced the concordance of the variant calling methods. The concordance rates were generally lower for variant call sets based on the Bowtie 2 with respect to BWA, alignments.

Hard filters are extensively used for improving precision of variant calls. However, as there is no direct score or combination of scores that clearly separates true variants from erroneously called variants, the precision and sensitivity are inversely related to each other, and one can improve precision only at the price of losing some sensitivity. We applied hard filters to the individual variant callers based on the current recommendations, bearing in mind, that these are not directly optimal and would need experimenting. The coverage dependency of hard filtering was considerably

different between GATK and non-GATK variant callers. In case of SAMtools and FreeBayes, according to the recommendations, we used only the quality estimation of the variants with a predefined threshold for filtering. Evidently, the quality scores of the variants increase with increasing depth, and we filter less and less variants. In case of HaplotypeCaller and UnifiedGenotyper, the filtering criteria depended on multiple annotations, therefore, the coverage dependency of hard filtering was more complex.

In summary, our results show that the utility of hard filtering is limited in our case: (1) generally higher sensitivity loss was observed for increasing coverage depths while the precision increased only a little and (2) the same hard filter settings were not appropriate for all coverage depths.

Combining variant calls: VariantMetaCaller

VariantMetaCaller combines the results of individual variant callers, exploiting their strength and complementarity.

As in case of every variant caller generally there exist real variants not found by this caller but found by one or more other callers, the variant call sets combined by VariantMetaCaller achieved higher maximal sensitivity than all other individual methods. However, it is also important to maximize the precision of the variant calls. Therefore, we sould also investigate how can a given score computed by a given method differentiate between real and false variants. Based on our results, the probability score computed by VariantMetaCaller meets this requirement: by ordering the variants based on their probability, the precision decreases slowly with increasing sensitivity, and it drops sharply only at high sensitivity values.

In summary, based on the analyses of simulated and real sequencing data, VariantMetaCaller achieved higher precision at all sensitivity levels than any of the individual variant callers irrespective of the depth of coverage, the aligner and the type of the variants.

To measure the performance of different scores, we calculated the area under the precision–recall curves (AUPRC), which is a summary statistic that reflects the ability of a score to correctly identify true variants. Based on the analyses of simulated and real sequencing data, the AUPRC of VariantMetaCaller was the highest among

all methods independently of coverage depth, the aligner or the type of the variants. Using simulated sequencing data, we showed that VariantMetaCaller performed better also on smaller genomic scales, especially in case of target region sizes that are typical in targeted gene panels.

One of our objective was to show the advantage of annotation information fusion over combining variants without utilizing this type of information. Therefore, we compared VariantMetaCaller to the late information fusion method BAYSIC, which utilizes only variant calls during combination. We found that the difference between the AUPRC of VariantMetaCaller and BAYSIC was in the range of $1 - 4\%$. This is remarkable, because 1% difference means prioritizing approximately 473 SNPs, and 49 indels more accurately. Additionally, we also computed the AUPRC for VariantMetaCaller and BAYSIC for each chromosome, and found that the AUPRC for VariantMetaCaller was higher than that for BAYSIC in most cases regardless of the type of the variant or the aligner and the differences were strongly statistically significant.

The other objective of our work was to provide a flexible solution for quantitative support of variant filtering, similarly to the false discovery rate based paradigm. This can be achieved by the precise estimation of the probability of the variants. Specifically, probabilities can be used to order the variants, and for a given threshold, unlike scores, probabilities can be used to estimate precision. Precision then can be directly translated to the number of true called variants, or equivalently to the number of false calls. As VariantMetaCaller provided more accurate probabilistic scores for calls than the other methods, the newly developed software supports quantitative, application-specific filter adjustment.

Effect of the *CYP3A4* and *CYP3A5* SNPs on the survival rates of childhood acute lymphoblastic leukemia

In the course of our work, we studied the influence of selected polymorphisms of *CYP3A4* and *CYP3A5* genes on the survival of childhood acute lymphoblastic leukemia. We found that a common SNP (rs2246709) in the *CYP3A4* gene significantly influenced the survival rates of the ALL patients after chemotherapy, which was strongly affected by the gender of the patients. Difference between genders was

especially large in the AG heterozygotes where male gender was associated with a significantly higher survival rate compared with that of females. In contrast, the wild homozygous AA genotype was associated with worse survival rate in males than in females.

We drew relative simple rules from the results of the rs2246709 gender interactions. We calculated new risk assessments and showed that these, in respect of survival rates, significantly outperformed the earlier risk-group assessment in our population.

With the help of Bayesian network-based relevance analysis we searched for interactions that could influence the survival of the patients. Involving additional SNPs in 34 genes and applying the Bayesian methodology, we also detected that the effect of the *CYP3A4* rs2246709 on survival was significantly influenced by a SNP in the *MTHFD1* gene. This gene is part of the folate pathway, which is the target of methotrexate. We found that the GG genotype of this SNP (rs1076991) increased the risk of B-cell ALL, but did not influence the survival rate. CYP3A4 does not metabolize the methotrexate, and thus the effects of the two SNPs on the different pathways seem to be superimposing each other.

Investigation of possible applications of the Bayesian relevance analysis methodology in association studies

During our work, we applied the Bayesian relevance analysis methodology (outside of the analysis of *CYP3A4* gene) in two candidate gene association studies to detect predisposing factors for childhood ALL, and in a partial genome screening study to detect genetic predisposing factors for asthma.

During the analyses we managed to highlight a number of advantages of the Bayesian relevance analysis methodology which are as follows:

- Because of the Bayesian statistical approach, the individual results (hypotheses) can be formulated in the form of direct probabilistic statements which accurately reflect the amount of information in our data. This, as opposed to the frequentist approach, might as well make it possible to build probabilistic data- and knowledge bases that support complex probabilistic queries, easier meta-analysis, or the fusion of the results with background knowledge.

- Since the method is inherently multivariable, we can simultaneously analyze the dependence of the target variable on all polymorphisms, environmental factors and phenotypic descriptors, or the complex dependence relations of the variables. Thus, the direct and transitive effects of the variables can be distinguished from each other, and the method offers a much richer language to describe and interpret the impact of factors affecting the target variable. Relationships between variables at this level of refinement could only be inferred using traditional statistical methods in a limited way (e.g. using pairwise association tests), which is further exacerbated by the need for multiple hypothesis testing correction.

- Because of the multivariate modeling, possible redundancies and interactions between the variables can be explored. For example, when examining the genes that play a role in the folate metabolism, the Bayesian relevance analysis method was able to detect a complex interaction effect in the hyperdiploid ALL subgroup.

- Bayesian statistics offer an automated and normative solution for the multiple hypothesis testing problems, so the results need not to be corrected.

- It is also possible to treat more than one target variable by virtually calculating the *a posteriori* probability of any structural question. The importance of this was demonstrated during an asthma partial genome screening study.

Conclusions

Based on our work, we can conclude the following:

1. We have developed a novel method (VariantMetaCaller), which combines the results of multiple next-generation sequencing variant callers. The method (1) utilizes the low concordance and complementarity of the individual variant callers, (2) uses the high-dimensional annotation information produced by the callers and (3) provides accurate probabilistic scores for variant calls. Based on simulated and real sequencing data, we showed that VariantMetaCaller significantly outperformed individual variant callers under a wide range of conditions, i.e. ranging from a few hundred kilobases to whole exomes.

2. Using real sequencing data we investigated the accuracy of the methods that predict the probability of variants. Our results showed that VariantMetaCaller predicted the expected precision more accurately than the alternative methods. Thus, VariantMetaCaller provides a quantitative, *precision-based filtering* method, supports the optimal selection of variants depending on the preferences and cost functions of the researchers, and allows finding problem-specific balance between sensitivity and precision. Our study also shows that VariantMetaCaller can be used even in areas that have been inaccessible for existing solutions, such as in targeted gene panels or organisms without accurate call sets. Thus, VariantMetaCaller broadens the application scope of precision-based filtering, and it can be a viable alternative to hard filtering.

3. Analyzing the genetic polymorphisms of one of the most important drug-metabolizing enzyme, CYP3A4, we found that the rs2246709 SNP significantly influenced the overall and event-free survival rate of acute lymphoblastic leukemia. Using the Bayesian relevance analysis we also detected that the effect of the *CYP3A4* rs2246709 on survival was significantly influenced by a SNP in the *MTHFD1* gene that plays a role in the folate pathway.

4. Using the Bayesian relevance analysis we detected, and by frequentist statistical analysis we confirmed that the interaction of rs2246709 SNP (*CYP3A4* gene) and of the gender of the patient significantly influenced the survival

rates of the ALL patients after chemotherapy. With simple rules describing this interactional effect, we could create a new risk assessment variable which significantly outperformed the earlier risk group assessment.

5. We have shown that the Bayesian relevance analysis method, over conventional frequentist statistical methods, allows a much more detailed analysis in genetic association studies by (1) distinguishing direct and transitive effects of the variables, by (2) defining various types of relationships and by (3) automatic discovery of redundancies and interaction between the variables.

List of scientific publications

Publications directly related to the PhD thesis:

Gézsi A, Bolgár B, Marx P, Sarkozy P, Szalai C, Antal P. (2015) VariantMetaCaller: Automated fusion of variant calling pipelines for quantitative, precision-based filtering. BMC Genomics, 16 (1): 875. IF: 3,986

Gézsi A, Lautner-Csorba O, Erdélyi DJ, Hullám G, Antal P, Semsei ÁF, Kutszegi N, Hegyi M, Csordás K, Kovács G, Szalai C. (2015) In interaction with gender a common CYP3A4 polymorphism may influence the survival rate of chemotherapy for childhood acute lymphoblastic leukemia. Pharmacogenomics J, 15 (3): 241–247. IF: 4,229

Lautner-Csorba O, **Gézsi A**, Erdélyi DJ, Hullám G, Antal P, Semsei ÁF, Kutszegi N, Kovács G, Falus A, Szalai C. (2013) Roles of Genetic Polymorphisms in the Folate Pathway in Childhood Acute Lymphoblastic Leukemia Evaluated by Bayesian Relevance and Effect Size Analysis. PLoS One, 8 (8): e69843. IF: 3,534

Lautner-Csorba O, **Gézsi A**, Semsei AF, Antal P, Erdélyi DJ, Schermann G, Kutszegi N, Csordás K, Hegyi M, Kovács G, Falus A, Szalai C. (2012) Candidate gene association study in pediatric acute lymphoblastic leukemia evaluated by Bayesian network based Bayesian multilevel analysis of relevance. BMC Med Genomics, 5: 42. IF: 3,466

Ungvári I, Hullám G, Antal P, Kiszel PS, **Gézsi A**, Hadadi E, Virág V, Hajós G, Millinghoffer A, Nagy A, Kiss A, Semsei AF, Temesi G, Melegh B, Kisfali P, Széll M, Bikov A, Gálffy G, Tamási L, Falus A, Szalai C. (2012) Evaluation of a partial genome screening of two asthma susceptibility regions using bayesian network based bayesian multilevel analysis of relevance. PLoS One, 7 (3): e33573. IF: 3,730

Antal P, Hullam G, **Gézsi A**, Millinghoffer A. (2006) Learning complex bayesian network features for classification. Proc. of third European Workshop on Probabilistic Graphical Models. Prague, Czech Republic; 9–16.

Cumulative impact factor of publications directly related to the PhD thesis: 18,945

Book chapters directly related to the PhD thesis:

Hullám G, **Gézsi A**, Millinghoffer A, Sárközy P, Bolgár B, Srivastava SK, Pál Z, Buzás EI, Antal P. Bayesian systems-based genetic association analysis with effect strength estimation and omic wide interpretation: a case study in rheumatoid arthritis. Methods Mol Biol 2014, 1142: 143–176.

Antal P, Millinghoffer A, Hullam G, Hajos G, **Gezsi A**, Szalai C, Falus A. Bayesian, Systems-based, Multilevel Analysis of Associations for Complex Phenotypes: from Interpretation to Decision. In: Christine Sinoquet, Raphael Mourad (szerk.) Probabilistic graphical models for genetics. Oxford University Press, New York, 2014: 319-360.

Gézsi A. Génexpressziós adatok standard asszociációs elemzése. In: Antal P (szerk.), Bioinformatika: Molekuláris méréstechnikától az orvosi döntéstámogatásig. Typotex Kiadó, Budapest, 2014: 107-120.

Other publications:

Kutszegi N, Semsei AF, **Gézsi A**, Sági JC, Nagy V, Csordás K, Jakab Z, Lautner-Csorba O, Gábor KM, Kovács GT, Erdélyi DJ, Szalai C. (2015) Subgroups of Paediatric Acute Lymphoblastic Leukaemia Might Differ Significantly in Genetic Predisposition to Asparaginase Hypersensitivity. PLoS One, 10 (10): e0140136. IF: 3,234

Temesi G, Virág V, Hadadi E, Ungvári I, Fodor LE, Bikov A, Nagy A, Gálffy G, Tamási L, Horváth I, Kiss A, Hullám G, **Gézsi A**, Sárközy P, Antal P, Buzás E, Szalai C. (2014) Novel genes in Human Asthma Based on a Mouse Model of Allergic Airway Inflammation and Human Investigations. Allergy Asthma Immunol Res, 6 (6): 496–503. IF: 2,160

Béres A, Lelovics Z, Antal P, Hajós G, **Gézsi A**, Czéh A, Lantos E, Major T. (2011)

"Does happiness help healing?" Immune response of hospitalized children may change during visits of the Smiling Hospital Foundation's Artists. Orv Hetil, 152 (43): 1739–1744.

Gézsi A, Budde U, Deák I, Nagy E, Mohl A, Schlammadinger Á, Boda Z, Masszi T, Sadler JE, Bodó I. (2010) Accelerated clearance alone explains ultra-large multimers in von Willebrand disease Vicenza. J Thromb Haemost, 8 (6): 1273–1280. IF: 5,439

Other book chapters:

Gézsi A. Metagenomika. In: Antal P (szerk.), Bioinformatika: Molekuláris méréstechnikától az orvosi döntéstámogatásig. Typotex Kiadó, Budapest, 2014: 264-273.

Cumulative impact factor of all publications: 29,778