Guideline for Final Reports

NWO-DFG cooperation programme

The report should cover the whole period since the start of the project.

Total length: approximately 8 pages

Objectives of the report:

- Scientific: Demonstrate the scientific progress
- Administrative: Address administrative and /or management problems
- Financial: Show how the grant has been spent and justify the payments being made
- Other Write a one-page popular scientific report of your project

1 General Information

File number:	047.016.009
Starting date of the project:	01.06.2004

Dutch project leader:	Prof. Cornelia M. van Duijn
Russian co-leader:	Prof. Tatiana I. Axenovich

2 Scientific Part

Overview of Research activities: Please specify which activities have been carried out.

During the course of the project, we worked on (1) research training of young Russian scientists, (2) identification and analysis of the problems, which appear in analysis of complex traits and application of high performance solutions to these problems, (3) development of new methodology for genetic epidemiologic research of complex traits, (4) development of software for data management and analysis with emphasis on high-performance solutions for genetic epidemiology, (5) development of computational infrastructure, (6) analysis of empirical data, and (7) dissemination of the results obtained.

Scientific Results: What are the main results achieved and what is their scientific significance?

Include references to the list below.

1. Research training: In 2004, two young Russian researchers have visited Erasmus for one month (August-September). During this period they became acquainted with the Erasmus team and participated in several courses within Erasmus Summer Program (Genetic Epidemiology). In 2006, three young Russian researchers have visited Erasmus. Two of them (M. Struchalin and N. Belonogova) took courses in Genetic Epidemiology and discussed our research projects with Erasmus part of the team (August 2006). A. Kirichenko, who is expert in high-performance computing, worked on several projects together with local staff (July-August 2006). In 2007, one young Russian researcher (N. Belonogova) visited Erasmus for 2 months for joint work on methodology development.

2. Problem identification and analysis:

2.1 In 2004, a draft of list of methodological, algorithmic and software problems we face in our research in Erasmus and Novosibirsk was generated (available on the Web). In 2005, 2006 and 2007 the list has been circulated between researchers from different countries, who work on similar topics, and was updated with their suggestions.

2.2 In 2004, testing of free (MySQL) and commercial (MS SQL) DBMSs was performed on simulated data. We have accessed the speed of data import and export and the amount of disk space required for storage. It has been shown that these DBMSs are not suitable for managing large amounts of genetic-epidemiologic data. They need large disk space and a lot of time for importing information into database and handling queries. The report, which was generated in 2005, is now available on the Web. As a result of this work, we decided to keep the data in plain text format.

2.3 In 2004, standard freely distributed software packages were tested using Erasmus data. We have shown that genotypic quality control tool PEDCHECK works well when autosomal inheritance is tested. On the contrary, Mendelian errors were not correctly identified for sex-linked markers. Therefore in 2005, we developed our own software for this type of data. Also in 2004, we tested PEDIG package which computes kinship coefficients for large pedigrees and found it capable to accommodate our data. In 2005, we found that standard linkage analysis packages which use Elston-Stewart algorithm for model-based linkage analysis (MENDEL, LINKAGE/FastLINK) are not capable of analysing pedigrees of very large size, even in absence of loops, because of underflow problem.

2.4 In 2005-2006, we performed a review of existing and potential applications of high-performance parallel computing (HPPC) to genetic epidemiological data. As a result we have identified a number of problems (e.g. derivation of empirical significance in linkage an association, whole-genome association studies, etc.) where HPPC is highly effective, and also a set of problems (e.g. linkage analysis using Lander-Green algorithm) where HPPC is of marginal significance.

2.5 Starting from 2005, we used simulated and real data from the Erasmus ERF project to compare performance and power of different existing and newly developed methods for pedigree-based association analysis. Our results indicate that in terms of power, different variants of measured genotype approach should be preferred over others, e.g. transmission-disequilibrium tests approach. The measured genotype approach in likelihood formulation is, however, very time consuming. Within general measured genotype approach we worked on implementation of software which facilitates high-performance parallel association analysis and development of fast (approximate) algorithms for high-speed analysis.

2.6 Pedigree information is crucial in linkage, but also association analysis of related individuals. This information, however, may be incomplete and/or contain errors. Starting with 2006, we work on methods which will rely on genomic, and not pedigree data, when accessing relation between study participants.

3. Methodology development:

3.1 In 2005, we developed algorithmic solution to the problem of underflow in genetic analysis identified in **2.3** (see Publications). We have already developed a set of programs implementing this solution in the context of linkage analysis (see Software).

3.2 For pedigrees with multiple loops exact calculation of likelihood is impossible. The approximation based on the breaking loops is used. In 2005 we developed a software package for breaking loops in pedigrees of arbitrary structure with multiple loops. Three algorithms are realized in this software (see Software). Two of them use the cost of edges for optimisation. We proposed the new approach for the cost calculation based on the lost of relationship after the edge elimination and realized this approach using parallel computation. In 2006, we worked on description of statistical properties of this approach for linkage analysis of quantitative traits.

3.3 In 2005, to increase throughput of genome-wide pedigree-based association analysis using measured genotype approach (problem identified at **2.5**), we proposed to develop fast approximate method for pedigree-based association analysis, which is based on two-stage procedure. At first stage, the genetic variance-covariance structure of the is estimated, and the trait residuals are computed using best linear unbiased prediction for the breeding values. These residuals, which are independent from pedigree structure, enter association analysis at stage 2, when simple regression tests may be used to estimate effect of a polymorphism on the trait. In 2006 and 2007, we worked on improved variants of the method and its' statistical characterisation.

3.4 In 2005 we estimated the power of new method of *in silico* mapping which has been proposed for localization of disease genes using inbred strains of mice. This method was aimed to facilitate a search for candidate genes of human diseases, however its power and limitations have not been analysed yet.
4. Software development: In 2004, several programs for the initial data management were created. These include pedigree structure verification and recoding program RECODE_PED. This program was tested using simulated and real data and is now distributed as a release candidate. Other program distributed now as release candidate is POOL_STR (2004), which allows pooling the data from several stages of genome scans performed using Short Tandem Repeats. In 2005, a number programs were developed for data quality control and management (GENOT_QC, PHENO_QC, GENOT_QC_X) and descriptive analysis (PHENO_QC, FCN). A set of programs (LOOP_PED, LOOP_EDGE and LOOP_STAR) have been developed for breaking loops in pedigrees of arbitrary structure with multiple loops. These programs achieve high performance through parallel computations. In 2006, we developed 7 software products, including a package for genome-wide association analysis (GenABEL). More detailed description is available at the Publications and Software sections.
5. Infrastructure development: in 2005, we have build the cluster with the following configuration: Hardware: 1 Server (2 Xeon 2.8GHz, 8Gb, 480Gb), 4 nodes (2 Xeon 2.8GHz, 6Gb, 80Gb). Software: OS Linus Slackware 10.2, LAM-MPI 7.1.1. Availability of this cluster in early 2006 greatly facilitated analysis of the Erasmus and Novosibirsk data. In 2006 and 2007, we maintained and actively used this infrastructure.
6. Real data analysis: In 2005-2007, the statistical quality control of Erasmus ERF project genotypic data was performed by the Novosibirsk group, using special software developed within the framework of this project. In 2006, we worked jointly on analysis of several complex traits.
7. Results dissemination: In 2004, we created the project's web page at http://mga.bionet.nsc.ru/nlru/. Results achieved by the project were published at this web-site, through national and international scientific journals, and through reports at scientific conferences.

Publications: which scientific papers, presentations or patents have resulted directly from this project? Please note: papers which were published before the project started must not be included.

Journal publications:

1. Y.S. Aulchenko, A.M. Bertoli-Avella, C.M. van Duijn (2005) A method for pooling alleles from different genotyping experiments. Annals of Human Genetics, 69: 233-238.
2. F. Liu, S. Elefante, C. M. van Duijn, Y. S. Aulchenko (2006) Ignoring distant genealogic loops leads to false-positives in homozygosity mapping. Annals of Human Genetics, 70: 965-970.
3. T. I. Axenovich (2006). Invited Review: Genetic mapping of common human diseases [in Russian]. Russian Journal of Clinical Genetics, 2(44): 11-15.
4. Y. S. Aulchenko, T. I. Axenovich (2006) Invited Review: Mapping genes for complex human disease: problems and perspectives [in Russian]. Vestnik VOGiS, 10(1): 189-202.
5. T. I. Axenovich, A. S. Zykovich (2006) Power of in silico mapping [in Russian]. Russian Journal of Genetics, 42(6): 850-857.
6. T. I. Axenovich, Y. S. Aulchenko (2006) Solution for underflow problem in linkage and segregation analysis. Computational Biology & Chemistry, 30: 382-385.
7. F. Liu, A. Arias-Vasques, K. Sleegers, Y. S. Aulchenko, M. Kayser, P. Sanchez-Juan, B. J. Feng, A. M. Bertoli-Avella, J. van Swieten, T. I. Axenovich, P. Heutink, C. vanBroeckhoven, B. A. Oostra, C. M. van Duijn (2007) A genomewide screen for late-onset Alzheimer disease in a genetically isolated dutch population. Am J Hum Genet. 81(1):17-31
8. M. J. van Rijn, A. F. Schut, Y. S. Aulchenko, J. Deinum, F. A. Sayed-Tabatabaei, M. Yazdanpanah, A. Isaacs, T. I. Axenovich, I. V. Zorkoltseva, M. C. Zillikens, H. A. Pols, J. C. Witteman, B. A. Oostra, C. M. van Duijn (2007) Heritability of blood pressure traits and the

genetic contribution to blood pressure variance explained by four blood-pressure-related genes. J Hypertens. 25(3):565-570.

9. N. M. Belonogova, T. I. Axenovich (2007) Optimal peeling order for pedigrees with incomplete genotypic information. Computational Biology and Chemistry 31: 173–177
10. Y. S. Aulchenko, S. Ripke, A. Isaacs, C. M. van Duijn (2007) GenABEL: an R library for genome-wide association analysis. Bioinformatics 23: 1294-1296.
11. Y. S. Aulchenko, D. J. de Koning, C Haley (2007) Grammar: a fast and simple method for genome-wide pedigree-based quantitative trait loci association analysis. Genetics (in press).
12. T. I. Axenovich, I. V. Zorkoltseva, F. Liu, A. V. Kirichenko, Y. S. Aulchenko (2007) Breaking loops in large complex pedigrees. Human Heredity (in press)

Presentations:

1. T.I. Axenovich Breaking loops for linkage analysis in complex pedigrees. EHGC 2006, invited speaker
2. I. Zorkoltseva, T. Axenovich, Yu. Aulchenko, M. Struchalin A package of programs for quality control of genetic data. EHGC 2006, poster presentation
3. Yu. Aulchenko, D. de Konig, C. M. van Duijn A fast and powerful mehtod for pedigree-based quantitative trait loci association analysis. EHGC 2006, poster presentation
4. L. M. Pardo, I. de Konig, K. Sleegers, M. Yazdanpanah, P. Sanches, J. van Swieten, Y. S. Aulchenko, B. Oostra, C. M. van Duijn. The role of the M235T polymorphism of the angiotensinogen gene on cognitive functions EHGC 2006, poster presentation
5. T.I. Axenovich, I. V. Zorkoltseva Parametric linkage analysis of large complex pedigrees. EHGC 2007, poster presentation
6. I. V. Zorkoltseva, T.I. Axenovich Model free Lods for linkage analysis of large complex pedigrees. EHGC 2007, poster presentation
7. N. M. Belonogova, Y. S. Aulchenko A powerfull approach to detect parent-of-origin effect in whole-genome association scans of quantitative traits. EHGC 2007, oral presentation
8. Y. S. Aulchenko, N. Amin, N. M. Belonogova, D. de Koning, C. Haley, C. M. van Dujin Powerfull methods for whole genome association scans of quantitative traits in samples of related individuals. EHGC 2007, oral presentation

Web publications (http://mga.bionet.nsc.ru/nlru/):

1. (2004) RN-list, a list of design, methodological and computational questions appearing in genetic-epidemiologic research in genetically isolated populations.
2. (2005) Performance and efficiency of several DBMSs for storage and retrieval of genetic-epidemiologycal data.

Programs, published on the Web (http://mga.bionet.nsc.ru/nlru/):

1. RECODE_PED (2004) is a program for verification of pedigree data and recoding pedigrees from free to standard format.
2. POOL_STR (2004) is a program for pooling alleles from different genotyping experiments.
3. FCN (2005) is a program to describe complex pedigrees.
4. PRE_PEDCHECK is a program for preparing pedigree and genotypes data for program PEDCHECK.
5. GENOT_QC (2005) is an interface to standard genotypic quality control program PEDCHECK
6. GENOT_QC_X (2005) is a program for finding Mendelian errors in the data from sex-chromsomes.
7. PHENO_QC (2005) is a program for qualoity control of phenotypic data and generation of simple statistics. LOOP_EDGE (2005) is a program for cutting and extension of pedigrees with multiple loops (classical Kruskal algorithm).
8. LOOP_PED (2005) is a program for cutting and extension of pedigrees with multiple loops (step by step breaking loops).

9. LOOP_STAR (2005) is a program for cutting and extension of pedigrees with multiple loops (algorithm described by Vitezica et al, Human Heredity 2004,57:1-9)
10. AFFY2MEGA (2006) is a program for converting Affymetrix SNP output data files, as written in text format, to mega2 or merlin format.
11. RECODE_SNP (2006) is a program for recoding alphanumerically coded systems into numerical alleles for PedCheck.
12. DESC_STA (2006) is a program for basic descriptive statistics for samples from the normal distribution.
13. TASK_MANAGER (2006) is a program for optimal running of multiple tasks on multiprocessor platform, when number of tasks is much larger than the number of processors available. Developed for Linux system with OpenMosix.
14. PEDCUT (2006) is a program for cutting deep pedigrees where patients are distantly related into computable sub-pedigrees based on user-specified MaxBit.
15. PED_STR (2007) is a program for cutting complex pedigrees with large number of patients which are close related into computable sub-pedigrees based on user-specified MaxBit size.
16. GenABEL (2006) is an R library for genome-wide association analysis.
17. PEDPEEL (2006) is a program for preparing pedigree data for calculation of Elston-Stewarts' likelihood function. It finds an optimal way to peel.

Please Summarise the scientific output (in numbers) below:

		Number of … with co-authorship	Number of … without co-authorship	**Academic Publications**
1	A	5	5	Publications in (international) refereed journals
	B	5	5	Publications in other (national) journals and other scientific output (abstracts in proceedings)
	C			Contribution to (chapters in) books
	D			Monograph
	E	3 MSc.		Thesis (MSc, PhD)
				Professional Products
2	A			Patent
	B	17		Other professional products
				Other output
3	A	2		**Web publications**
				Conferences attended
4	A	2		

In the course of the project, which contacts or cooperation did you have with parties outside the Scientific Community, if so can you please describe the outcome,:

In a course of several conferences we made contact with a small European enterprise “BC Platforms”. Their products are aimed to the researchers in the area of genetic epidemiology. As a result, they have integrated our GenABEL package as a part of their analysis facilities (consisting of freely distributed software only) of their SNPmax platform.

During the visits of Erasmus faculty to Novosibirsk, several lectures on genetic epidemiology were taught; these were open to general public.

3 Administrative Part

Please address administrative and /or management problems, if any:

At the beginning of the project, we experienced some problems with money transfer to Novosibirsk (individual grants and equipment); the first part on money have been received only at the end of November. Otherwise we experienced little administrative and almost no management problems.

In your opinion, did the cooperation proceed as planned?

Yes

5 Popular Scientific Report

Please write a 1-page popular scientific report about the research project
This general summary should present a description of background, aim, methodology, results and main conclusions. Include keywords.

Please note this summary can be used in annual reports and for other ways of dissemination information on NWO funded research.

Title: Development of algorithms and software for high-performance computing in genetic analysis of complex human traits

Keywords: genetic epidemiology, high performance computing, complex traits, pedigree, linkage analysis, association analysis

Background: Common human diseases and genetic characteristics are determined by complex interplay between environmental and genetic factors. Typically, the genes' contribution to complex human traits varies between 30 to 80%, as measured by heritability. Identification of genes whose variation is involved into control of complex traits is therefore of major importance for the better understanding of the biological bases of human health and disease, prediction, and, ultimately, development of new treatments. Recent developments in molecular technologies open a wide range of possibilities for gene identification. At this stage, however, methodology and computational technology start being more and more of a bottleneck.
Aim: Development of new high throughput methods, algorithms and software for identification of genes involved into control of variation of complex human traits. In particular, we aimed to solve methodological and computational problems associated with linkage and association analysis of complex human pedigrees.
Methodology: We identified and addressed critical methodological, algorithmic and computational problems met in linkage and association analysis of complex human pedigrees.
Results: We have developed a number of methods and algorithms which make linkage analysis of large complex pedigrees possible. The general approach adopted is based on approximate (likelihood) solution. This involves simplification of pedigree structure, while retaining maximum information, and analysis of simplified pedigrees using already available and newly developed software. All of the methods scale linearly with the number of processors and benefit highly from parallel implementation. We have also developed new fast approximate and powerful methods and software for association analysis of quantitative traits in samples of related individuals. The computational time required grows approximately linearly with the number of subjects and markers studied. In large studies, empirical computations may be speeded up through parallel computations. The methods developed were applied to study the genetics of different complex traits using the data generated in Erasmus MC.
Conclusions: We have developed and implemented a number of new methods for linkage and association analysis of complex human traits. Most of these methods are almost linearly scalable with the number of processors and highly benefit from high performance parallel implementation we provide. Availability of these products makes linkage and association analysis of large complex human pedigrees possible. Application of these products to real data has already provided insights into genetics of blood pressure, cognitive function, stature and other traits.