Guideline for Periodic Reports

Dutch Russian cooperation programme

The periodic report should cover the whole period since the start of the project, and clearly point out the progress achieved since the last report.

Total length: approximately 6 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

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

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

Include references to the list be[1]low.

1. Results dissemination: In 2004, we created a preliminary version of the project's web page at http://mga.bionet.nsc.ru/nlru/. This page was maintained and updated with most recent results in 2005. Results achieved by the project were published at this web-site and also in national and international scientific journals.

2. Reasearch training: In 2004, two young Russian researchers have visited Erasmus for one month (August-September). During this period they took courses in Genetic Epidemiology, met the members of Erasmus Lab and became familiar with the Erasmus research projects.

3. Problem identification and analysis:

3.1 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 the list has been circulated between researchers from different countries, who work on similar topics, and was updated with their suggestions.

1

3.2 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. Development of our own binary data warehouse for genetic-epidemiologic data, which is taking into account specificity of data and configuration of queries, was started in 2004, and is currently in progress.
3.3 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.
3.4 In 2005, we performed a review of existing and potential applications of high-performance parallel computing (HPPC) to genetic epidemiological data. Manuscript is currently in preparation.
3.5 In 2005, using simulated and real data from the Erasmus ERF project, we compared performance and power of different existing methods for pedigree-based association analysis. Our results indicate that in terms of power, measured genotype approach should be prefered over transmission-disequilibrium tests approach. The former, however, is very time consuming if complete pedigree is used. Analysis of a single marker may take hours, when standard software packages (SOLAR, ASReml) rae used. This implies that a single genome-wide association scan with 100,000 markers will take about 11 years of computations to finish. We work on solution to this problem along two lines: first, implementation of software which facilitates high-performance parallel association analysis and, second, development of methods which allow faster analysis.

4. Methodology development:

4.1 In 2005, we developed algorithmic solution to the problem of underflow in genetic analysis identified in 2.3 (see publications). We plan to either modify existing or write our own programs implementing this solution in 2006.
4.2 For pedigrees with multiple loops exact calculation of likelihood is impossible. The approximation based on the breaking loops is used. In 2005 we developed the software for breaking loops in pedigrees of arbitrary structure with multiple loops. Three algorithms are realized in this software (see Software in http://mga.bionet.nsc.ru/nlru/). Two of them use the cost of edges for optimization. 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.
4.3 To increase throughput of genome-wide pedigree-based association analysis using measured genotype approach (problem identified at 3.5) we currently 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. The characteristics of the method are now being tested.
4.4 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 analyzed yet.

5. 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. More detailed description is available at the Publications section.

6. 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 will greatly facilitate analysis of the Erasmus and Novosibirsk data.

7. Real data analysis: In 2005, 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.

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:

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.

F. Liu, S. Elefante, C. M. van Duijn, Y. S. Aulchenko (in press) Ignoring distant genealogic loops leads to false-positives in homozygosity mapping. Annals of Human Genetics.

T. I. Axenovich (in press). Invited Review: Genetic mapping of common human diseases [in Russian]. Russian Journal of Clinical Genetics

Y. S. Aulchenko, T. I. Axenovich (in press) Invited Review: Mapping genes for complex human disease: problems and perspectives [in Russian]. Vestnik VOGiS

T. I. Axenovich, A. S. Zykovich (in press) Power of in silico mapping [in Russian]. Russian Journal of Genetics

T. I. Axenovich, Y. S. Aulchenko (submitted) Solution for underflow problem in linkage and segregation analysis. Submitted to Computational Biology & Chemistry.

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

(2004) RN-list, a list of design, methodological and computational questions appearing in genetic-epidemiologic research in genetically isolated populations.

(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/):

RECODE_PED (2004) A program for verification of pedigree data and recoding pedigrees from free to standard format.

POOL_STR (2004) A program for pooling alleles from different genotyping experiments.

FCN (2005) A program to describe complex pedigrees.

GENOT_QC (2005) An interface to An interface to standard genotypic quality control program PEDCHECK

GENOT_QC_X (2005) Software for finding Mendelian errors in the data from sex-chromsomes.

PHENO_QC (2005) Program for qualoity control of phenotypic data and generation of simple statistics.

LOOP_EDGE (2005) Software for cutting and extension of pedigrees with multiple loops (classical Kruskal algorithm).

LOOP_PED (2005) Software for cutting and extension of pedigrees with multiple loops (step by step breaking loops).

LOOP_STAR (2005) Software for cutting and extension of pedigrees with multiple loops (algorithm described by Vitezica et al, Human Heredity 2004,57:1-9)

Guideline for Periodic Reports

Dutch Russian cooperation programme

Please Summarise the scientific output (in numbers) below:

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