

Place: Biomedicum, Helsinki

Attendees:

SWE twin reg: Litton Jan-Eric, Carlsson Jenny, Björklund Ann, Lindqvist Pontus,
Petterson Gunnar, Söderberg Johan
FGC: Muilu Juha, Leinonen Anne, Lehto Anne
KTL-BBU: Saharinen Juha
FIN twin reg: Heikkilä Kauko, Siivola Matti
NO twin reg: Brandt Ingunn
DK twin reg: Hvidberg Lars
IT twin reg: Cotichini Rodolfo, Fagnani Corrado
Morgam: Kuulasmaa Kari, Cepaitis Zygimantas
Leiden: van der Wijk Henk Jan, Putter Hein

PP = presentation

Mon 15.3.

Objectives (PP, Jan-Eric Litton)

Jan-Eric Litton presented objectives and milestones of database core for year 2004. Four new working groups will be established; WG1: Phenotype, leader Ann Björklund, WG2: Data Modeling, leader Juha Muilu, WG3: Database Integration, leader Pontus Lindqvist and WG4: Output and Analysis, leader Hein Putter.

VPN (Virtual Private Network) issues was included into WG3, but might be separated as own working group later, if needed.

Deliverables, milestones and deadlines of GenomEUtwin (PP, Markus Perola)

Markus Perola presented the progress of GenomEUtwin project. Among six phenotypes first analysed phenotypes will be BMI and stature. More phenotypes will be included into this project in possible extension of 4 year. Kaare Christensen has almost finished checking up the status of DNA collection. GenomEUtwin project is also linked into other European Projects (e.g. P3G <http://www.p3gconsortium.org/>). Markus reminded that GenomEUtwin milestones are flexible.

How to document our work (PP, Kari Kuulasmaa)

Steering Group has asked DB core to create documentation guidelines. Documentation concerns all areas of GenomEUtwin. Kari Kuulasmaa presented the principals of documentation and shared the experience from MORGAM. Kari suggested that all centers would have their own web site where to store and share their own documentation. Web sites will be linked into GenomEUtwin homepage and Anne Leinonen will provide stylesheet to all centers. Draft versions should also be available.

Jan-Eric: what happens to documents after finishing this project?

How to use data encryption (PP, Zygimantas Cepaitis)

Zygis showed us how to use GnuPG software for encryption. It was agreed that the common private-public keypair will be created for GenomEUtwin. Encryption instructions and keys will be put into GenomEUtwin homepage.

Statistics in GenomEUtwin (PP, Hein Putter)

Hein Putter presented the statistical core and how the selection of most informative twins to be genotyped for linkage was chosen. He also shared some experience of combined data analysis. The data sets are planned to analyse separately and summarize the results of analysis. In file formats it is recommended that each individual is in its own row (no pair in same row).

Information Integrator (PP, Juha Muilu)**Working Groups** (2 hour meetings)

WG1: Ann Björklund, Ingunn Brandt, Lars Hvidberg, Anne Leinonen

WG2: Juha Muilu, Jan-Eric Litton, Matti Siivola, Zygis Cepaitis

WG3: Pontus Lindqvist, Jenny Carlsson, Rodolfo Cotichini, Kauko Heikkilä, Johan Söderberg, Timo Miettinen

WG4: Hein Putter, Kari Kuulasmaa, Markus Perola, Henk Jan van der Wijk, Corrado Fagnani, Gunnar Petterson

Discover Link (PP)

Juha Muilu talked on experiences on using DB2 and Information Integrator (Discovery Link) in genotype database work. In the database remote data sources are accessed using relational (Oracle, Finnish Genome Center) and non relational wrappers (Uppsala flat files). The genotype database system contains two databases. Submission database take care of data collection and analysis database has discrete data releases in optimal data structure. In addition to genotype data, the database also federates phenotype data from Karolinska. Secure database connections between the centers are implemented using encrypted ssh tunnels. The federated database system and developed reporting interface was demonstrated on separate hands on session.

SQL exercises were done using DbVisualizer (<http://www.minq.se/products/dbvis/>) and xReporter as web interface reporting tool.

Tue 16.3.

WG1: Phenotype Data (PP, Ann Björklund)

Migraine and Coronary heart disease and stroke (CHD) phenotype data should be in Discovery Link by the 1st of October. Each phenotype's (6) variable formats will be as own appendix to the basic phenotype document. Variables will be commented by phenotype group (e.g. migraine) and statistical group. Statistical group will be asked the routine checks for each phenotype variable so each center is able to limit the dataset ("clean") in same way.

Statistical group will give feedback about data straight to centers. Elles Mulder, Markus Perola and Hein Putter will set a working group for migraine.

WG2: Data Modeling (PP, Juha Muilu)

Juha suggested a forum for data modelling ideas. UML is recommended as modelling language. Participating OMG standardization would be one way of creating universal modelling structures in this project. WG1 and WG2 together should try to create universal way of modelling all phenotypes. It maybe possible to start with migraine, epi-core's impact will be asked. Should the analysis results be modelled and stored into database?

WG:3 Data Integration (PP, Pontus Lindqvist)

Challenges of data integration are data security and preferred integration method. Preferred integration methods are distributed SQL, Information Integrator and file transfers. To avoid mess with SSH connections, VPN is desired option despite that there is not much experience in building such a network. Timo Miettinen at FGC and Juri Ahokas at BBU will start to build VPN network. Document describing the different environments should be ready by the end of April. By the June VPN and Information Integrator should work between Helsinki and Stockholm and by December, VPN should be between HKI/STO/RME.

WG: 4 Output and Analysis (PP, Hein Putter)

DBcore's responsibility is to provide data that has passed routine checks, quality assessment reports and data books (summaries of variables). Data will be given to statistical core after approval of GSG. Analysis request will be handled by Publication Policies and GSG. Errors will be reported to responsible database persons. Responsible person concerning phenotype databases is Jenny Carlsson and concerning GtDB Juha Muilu. Statistical analysis will be documented and data used will be archived in Leiden. Databases should keep log on all transactions and logins.

VPN status and Citrix and SSH (PP, Juha Saharinen)

Juha Saharinen reported that evaluating VPN Open Source Software is ongoing and they have met many difficulties with setting up VPN connections. VPN will be based on IPSEC protocol to provide compatibility to as many platforms and software as possible. SSH connections are much easier to setup, but the network topology is too complicated to be handled by SSH connections. One possibility would be to automate SSH scripts so administration would be easier. Based on Juha's suggestion, it was decided to start with SSH connections, because it is much easier and faster than VPN. When changing into VPN later, it will be invisible for the users. There is no need to use same VPN software through all GenomEUtwin project networks; network administrators are familiar with the software that is already in their use.

Accounts for BBU Citrix servers can be applied from BBU (<http://www.bioinformatics.helsinki.fi>). Copy-pasting, file transfer and printing on local printers works in practise better in Citrix compared to X-terminal.

SWOT – Strengths Weaknesses Opportunities Threats (chairperson: Kari Kuulasmaa)

Italy (PP)

S: Flexible relational database, optical questionnaires and possibility to reconstruct family relationships within the database based on individual codes.

W: Lack of experience in biological sample data management.

Denmark

S: Oracle experience.

O: Learn about VPN and experience of Information Integrator.

Finland

S: Started at 1975 and have lots of data, Oracle since 1990, include family members

W: Some old data is not in database, some researchers have their data in files.

Norway (PP)

S: Experience of data conversions and transferring data from several systems into relational model

O: Database model for genotype data and experience in web questionnaires needed. Norway has heavy restrictions what data can be transferred via web into database.

Sweden

S: Data since the sixties, transferring data into relational model (flat files into Oracle), creating questionnaires (also web-based, digital paper)

W: No family members

O: Relational or object model for family structure

FGC (PP)

S: All data in same database, good documentation, familiar environment

W: Lack of resources (new LIMS development is ongoing)

T: Synchronization between phe dbs and GtDB concerning zygosity, sex and family structure, overall data management, release version handling with Ii and genotype data harmonisation

It was decided that genotype laboratories will inform gender information to DNA Unit who will check the discrepancies and report to laboratory and twin centre. Twin registries will take responsibility to check for zygosity discrepancies and will inform genotyping labs.

Leiden is responsible for archiving data files used in statistical analyses. There is no need to reselect the dataset from database. Archives stored in fileserver must be protected against non-authorized access.

Spotfire (Stefan Hansen)

Stefan Hansen presented Spotfire's Decision Site analysis tool. This tool will help to outline and check the data for the analysis. Browser interface must support dynamic html (e.g. IE v.5.58 ->). Anthony Metzidis will find out if it works with Mozilla and MAC OS. Jenny Carlsson has contact information of Spotfire for those who are interested. Speed is based on RAM memory, no other optimization is used.

Biobank visit (Outi Repola)

Ossi Bister presented sample management database and his Java application developed for the DNA sample data handling. Outi Repola gave guided tour in DNA extraction laboratory and showed pipeting robots.

Other Issues

In order to get advantage of IBM scholarship, each center should have own responsible person and each user has to apply its own user account.

<http://www-306.ibm.com/software/info/university/members/registration.html>

Markus Perola encouraged applying some money to visit other centers. Jaakko Kaprio will handle the applications.

Next Meeting

December 13th -14th 2004 in Rome.

Topics:

- genetic / living family structures
- talk about ethics and laws from ethical core
- other core issues