Groningen, 18 Juni 2010

Betreft: uitgewerkt plan regenboog 'data management en analyse'

Geachte BBMRI-NL stuurgroepleden,

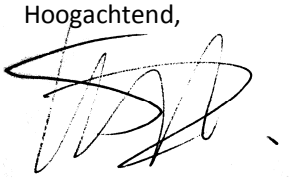
Op 15 Februari van dit jaar is door de BBMRI-NL stuurgroep besloten dat er een BBMRI-NL project gedefinieerd moet worden met als doel de harmonisatie van bioinformatica methoden, databases, software en formats voor data harmonisatie, verrijking, uitwisseling, management en analyse. Hierbij is aan ondergetekende verzocht dit project als 'regenboog' voorstel uit te werken in samenwerking met de participerende BBMRI-NL biobanken, nationale initiatieven zoals het Netherlands Bio Informatics Center (NBIC), Parelsnoer en Mondriaan en maximaal gebruik makend van ontwikkelingen in internationale context. Vind bijgesloten het resulterende project voorstel, conform de BBMRI-NL regenboog aanvraag procedure.

Tijdens de uitwerking van dit project plan is

- qua timing en werkzaamheden sterk rekening gehouden met de recente ontwikkeling van het Genoom van Nederland regenboog project om tijdig de benodigde bioinformatica infrastructuur te realiseren voor het managen, analyseren en delen van de te produceren sequentie data ter verrijking van alle BBMRI-NL GWAS biobanken via imputatie (Appendix 1).
- is een open netwerk van BBMRI-NL bioinformatica experts bijeen gebracht met sterke internationale wortels die allen aan het onderliggende voorstel hebben bijgedragen en waarmee het project bij toekenning inhoudelijk ingevuld kan worden (Appendix 2).
- is bij deelnemende BBMRI-NL partners de consensus onstaan dat het complexe vraagstuk van 'phenotype variabele harmonisatie' het best in een ander, dedicated project kan worden ontwikkeld in coordinatie met Parelsnoer, Mondriaan, P3G en CWA (Appendix 3).

Wij hopen dat dit regenboog plan op de ondersteuning van de BBMRI-NL stuurgroep kan rekenen.

Hoogachtend,



Dr. Morris Swertz	(UMCG)
Dr. Yurii Aulchenko	(EMC)
Dr. Paul de Bakker	(UMCU)
Prof. dr. Dorret Boomsma	(VU)
Prof. dr. Johan den Dunnen	(LUMC)
Dr. Barend Mons	(NBIC)

Bijlage: Grant request form BBMRI-NL Regenboog Project 'Dynamic bioinformatics infrastructures for biobank enrichment'

Grant request form BBMRI-NL Regenboog Project

1. Title of the project: Dynamic bioinformatics infrastructures for biobank enrichment

2. Project number (to be filled in by BBMRI-NL)**:**

3. Name of Principal Investigator (contact person)**:** dr. Morris Swertz
Institution: University Medical Center Groningen
Department: Genomics Coordination Center, Dept of Genetics & Groningen Bioinformatics Center
E-mail: m.a.swertz@rug.nl
Phone: +31 (0)50 361 7100 or +31 (0)6 52 606 501

Other participants
Dr. Yurii Aulchenko, EMC, Rotterdam (association analysis)
Dr. Paul de Bakker, UMCU, Utrecht (statistical genetics analysis)
Prof. dr. Dorret Boomsma, VUMC, Amsterdam (imputation)
Prof. dr. Johan den Dunnen, LUMC, Leiden (genetic variation analysis)
Dr. Barend Mons, Netherlands Bioinformatics Center (infrastructure and semantic integration)

4. Project aim
The next stage of epidemiological and genetic research will depend critically on large collections of high quality samples and data, also known as biobanks. This ambitious BBMRI-NL project aims to facilitate large scale data enrichment and data pooling between Dutch biobanks which will allow leading participation in the next generation of international etiological research. This will be achieved via harmonization and enrichment of bioinformatics databases, methods, models, software and tools, in particular focusing on high throughput analysis of sequencing and genome-wide association studies in the context of the 'Genoom van Nederland (GvNL)' rainbow project.

Background
In the past decades the Dutch university medical centers and research institutes have collected broad and deep collections of over 400.000 individuals, not including major new initiatives like the LifeLines project which will follow 165.000 individuals throughout the next 30 years. Currently ~200 organizations are in charge of the Dutch biobanks ranging from large and broad (inter-institute) cohorts to small and deep departmental boutique disease biobanks. Each of these biobanks has collected for each participating individual a unique set of materials, including tissue-derived samples (like blood plasma, serum, urine and/or DNA among others) and/or phenotypic data in the form of responses to questionnaires, results of measurements and information from hospital information systems.

Challenges
The science of biobanking involves major bioinformatics and IT challenges at multiple levels, and it is now clear that significant development is required to enable multiple biobanks to work together in a highly effective and integrated way. Novel high throughput measurement methods like SNP-chip based genome-wide association studies (GWAS) and next generation sequencing (NGS) enable massive genetic and molecular profiling of samples at an unprecedented rate. Suitable software infrastructures are needed to enable integrated analysis of all these data with sufficient statistical

power to unravel the complex interplay of genetic and environmental factors in determination of human health and disease.

Mission

This joint BBMRI-NL and NBIC project brings together a team of leading bioinformatics researchers on a mission to remove technical barriers to the integration and exploitation of the wealth of phenotype and genotype data available in the biobank community. This will involve the research & generation of suitable software protocols, models, formats, databases, hardware and tools building on, and in collaboration with, national NBIC/BioAssist biobanking, sequencing and interoperability task forces, Parelsnoer, Mondriaan, CTMM and LifeLines/Target and international efforts BBMRI-EU, ELIXIR, 1000 Genomes project, European Bioinformatics Institute, EU-GEN2PHEN (genotype to phenotype), P3G (Public Population Project in Genomics: Genome Canada and Genome Quebec), EU-GENECURE (GENomic StratEgies for Treatment and Prevention of Cardiovascular Death in Uraemia and End-stage REnal Disease: FP6), ENGAGE (European Network for Genetic and Genomic Epidemiology: FP7), EU-BioSHARE (Biobank Standardisation and Harmonisation for Research Excellence in the European Union), EU-NMD-chip (neuromuscular diseases chip, FP7), EU-TechGene (echnological innovation of high throughput molecular diagnostics of clinically and molecularly heterogeneous genetic disorders: FP7) and GEFOS (Genetic Factors of Osteoprosis:FP7). The aim is to harmonize data management, exchange and protocols for existing data within BBMRI-NL and to enrich Dutch biobanks with new models, software and tools for next generation data with scalable data archives, flexible and large scale processing pipelines and easy-to-connect systems for data exchange.

Expected output

This project aims to produce the bioinformatics resources needed by BBMRI-NL participating biobanks and rainbow and complementation projects, most notably in the context of Genoom van Nederland:

1. Sequence data management, QC and analysis pipelines to produce and share a Dutch catalog of variants.
2. GWAS data management, QC and imputation to produce a Dutch GWAS control cohort
3. Dutch (inter)national biobank catalog and data exchange formats
4. Scalable and easy to maintain software and web access tools underlying 1-3.

All these resources will be made publically available both as centralized, secured, web accessible national services, i.e. central hubs assembled in partnership to support the rainbow projects, as well as downloadable and customizable 'tools-in a-box' meant for local installation by biobanks and their local projects (local hubs). This project will develop in parallel the scientific, professional and physical infrastructures needed to effectively communicate expertise, procedures and tools between all Dutch biobanks as well as the provision of bioinformatics experts building on the infrastructure organized in the Netherlands Bioinformatics Center (NBIC) BioAssist program. This group will work in coordination with the BBMRI-NL ethical-legal working group to develop a code of practice and guidelines for large scale harmonized data pooling and for the use of data from multiple biobanks.

5. Approach

This project will combine a hub-and-spoke research & development organization to harmonize data between biobanks together with the provision of experts who will provide innovative model-driven

software methods to efficiently produce ready-to-use software infrastructures needed by biologists and researchers. This includes:

Agile hub-and-spoke organization

At the core of BBMRI there is the vision to develop all resources in a hub and spokes manner such that we maximize use of local expertise and innovation and minimize duplicated efforts and barriers to integration via centralized harmonization and enrichment. The smallest hubs within the Dutch biobank landscape are the individual biobanks, the larger hubs the participating institutes, and the largest hubs are central deployment of key data and analysis resources (which again can connect to pan-European hubs). This project will mirror this organization to bridge between biomedical researchers, bioinformaticians and hardcore software engineers to ensure the multi-disciplinary interplay needed:

- A central engineering team of hardcore programmers is responsible for the overarching infrastructure and will ensure harmonization of tools, pipelines and databases between working groups. This group will function as one of the eight NBIC task forces and will meet every week to ensure knowledge and method transfer.
- Participating experts will host programmers and scientific staff to pilot the planned tools and pipelines in close support to (their) BBMRI-NL complementation and rainbow projects. These bioinformaticians will be organized in themed working groups as described in appendix 1. Each working group will have a lead programmer that is part of the central engineering team. All members will meet monthly and will have weekly Skype meetings.
- This project is strongly linked with leading international sister projects to avoid duplicated efforts and efficiently achieve these aims by having project members participating in, or staying at, institutes like European Bioinformatics Institute (1KG, EGA, ArrayExpress), Netherlands Bioinformatics Center (NGS, eScience, CWA), projects like EU-GEN2PHEN, EU-BIOSHARE, OMII-UK, ESFRI/ELIXIR, Parelsnoer, Mondriaan, CTMM, TIFN, NPC, NMC, P3G, Human Variome Project and open source collaborations like ObiBa, MOLGENIS/XGAP, ABEL and Concept Web Alliance.

Model driven software

Flexible model driven software development as described in Swertz & Jansen (2007) has proven to be an efficient method to rapidly produce harmonized software infrastructures for life scientists while sharing the best models, software and tools notwithstanding large variation in research aims. This project will build and extend upon open source implementations of these methods such as MOLGENIS and Galaxy focusing on:

- Implementing extensible standard data models and software components developed internationally (we co-piloted data models for microarrays, QTLs, GWAS studies [Swertz 2010], and phenotypes in EU consortia like GEN2PHEN and EBI and participated in international GWAS and sequencing initiatives like the 1KG project).
- Making tools and protocols reusable in a user-friendly catalog of bioinformatics tools and workflows that captures all necessary inputs, outputs, optimization properties and user interactions in models to automatically incorporate existing tools (building or inspired on Taverna and Galaxy).
- Generating automatically from these data and tool models the scalable back-ends and front-ends needed. This automatic procedure ensures harmonized software results building on industry

standard databases for metadata and innovative approaches like cloud computing activities at SARA/Amsterdam, CIT/Groningen and BigGRID/Rotterdam to connect to the scalable compute power and storage needed.

- Ease finding and integration of resources using semantic and ontology technologies such as developed at EBI and NBIC/Concept Web Alliance to build bridges between data and tools, tapping into existing ontologies for data (e.g. HPO for human phenotype ontology) and for analysis protocols to help user and systems developers to bring tools together.

Ready-to-use databases and tools 'in-a-box' that can federate into national resources

As detailed below in the description of work section, this project aims to develop novel or incorporate internationally proven key bioinformatics tools, databases, models and software such can be re-used by the smallest hubs (to accommodate and improve local research and complementation projects) up to the larges hubs (supporting rainbow projects, starting with Genoom van Nederland). By sharing the same components between all hubs we provide an effective path to

- harmonize and enrich available data management, exchange and analysis protocols
- avoid duplicated efforts between local hubs
- make it more likely that everyone's needs are supported
- improve quality because more users test the available bioinformatics infrastructure
- preserving flexibility to go beyond standardization and accommodate specific local needs.

Harmonization / Enrichment (please tick)

6. Biobanks involved in this grant request

Name of biobank (1): All biobanks having GWAS data **Principal Investigator:** nvt
Current number of samples: 400.000 **Started in:** 2007 with BBMRI
Short description:
In the Netherlands there are current GWA data available from >85.000 individuals, geographically distributed throughout the country. This material is an excellent foundation to study whether regional differences have any relationship with genetic differences. In the Genoom van Nederland project this has been taken as starting point to select individuals for sequencing. This project will take the result of this sequencing project to elucidate Dutch subpopulations and genetic variation and use this information to enrich available GWAS data.
Content (please tick)**:**
Phenotypic data: clinical / anthropomorphic/ lifestyle / environment / biomarkers / medication / family / …
Biomaterials: DNA / RNA / Plasma / Serum / Urine /
'Omics' data: transcriptomics / proteomics / DNA sequence / GWA / metabolomics /

7. Added value of the project for BBMRI-NL. Please explain and indicate which biobanks, biobank researchers or other stakeholders will profit.
The bioinformatics rainbow connects to 6 of the goals as formulated in the so-called "meerjarenplan":

1. Het opzetten van een efficiënte en geïntegreerde nationale infrastructuur die bestaande biobanken in Nederland verbindt en verrijkt

2. De koepel en het 'gezicht' worden van de Nederlandse biobanken
3. Het vormen van een sterke nationale hub voor BBMRI-EU, zowel voor populatiebiobanken als voor klinische biobanken.
4. Zorgen voor optimale toegang tot materiaal en gegevens in bestaande biobanken voor wetenschappelijk onderzoek, die optimaal recht doet aan privacy en autonomie van donoren/participanten
5. Optimale aansluiting bij en benutting van resultaten uit bestaande initiatieven.
6. Het faciliteren van toekomstig multidisciplinair wetenschappelijk onderzoek naar het ontstaan en beloop van multifactoriële aandoeningen, ten bate van nieuwe concepten voor preventie, diagnostiek en behandeling

8. Duration of project:

3 years

Planning (matching GvNL planning where appropriate)	
Short read archive	Month 0 – 8
Biobank catalog pilot	Month 0 – 6
Sequence analysis Phase 1 (GvNL)	Month 4 – 16
Harmonized exchange formats	Month 6 - 24
Establish variation QC and analysis pipeline	Month 8 – 20
Sequence analysis Phase 2 (GvNL)	Month 8 – 20
Variation catalog/Dutch HapMap	Month 20
GWAS data release server	Month 0 – 12
GWAS QC and imputation protocols	Month 6 – 20
Dutch GWAS Control Cohort (DGCC)	Month 12 –24
Imputation of available GWA data (GvNL)	Month 20 – 30
Make sequence data available (GvNL)	Month 12 – 30
GWAS analysis tools catalog	Month 12 – 36
Web access tools	Month 22 – 30
Integrated DCGG and Variation catalog web access tools	Month 24 – 36

9. Deliverables

D1 Sequencing

- Short Read Archive (GvNL) – a database and user interface to manage and trace next generation sequencing data, associated sample annotations (metadata) and intermediate- and end-results.
- Variation analysis and QC pipelines (GvNL) – harmonization and enrichment of available processing pipelines for quality control and variation analysis for (exome) re-sequencing projects.
- Variation catalog/Dutch HapMap (GvNL) – release of the enrichment results of variation analysis of the GvNL 1000 genomes as produced using above tools as imputation data source.

D2 Genome-wide association analysis

- GWAS data release server– database and user interfaces to manage and query GWAS data, in particular to create GWAS (control cohort imputation) data releases.

- GWAS CQ and imputation protocols – harmonization and enrichment of tools and pipelines to verify and clean GWAS data sets and produce data sets ready for analysis by the researcher.
- GWAS data analysis – a catalog of established protocols and bioinformatic pipeline implementations thereof for GWAS analysis.
- GWAS control cohort and DCGG (GvNL) – collection of BBMRI-NL GWAS data into the DCGG database and release of imputed datasets using variation catalog produced by GvNL

D3 Biobank (meta)data finding and exchange

- Biobank and biobankers catalogue – central index of biobanks with aggregate metadata on biobank contents (protocols, features observed, optionally (aggregate) data) and semantic search functionality to enable researchers to find biobank(er)s and samples.
- Harmonized data exchange formats – harmonization of syntaxes / file formats to transfer sample annotations, phenotypic data and molecular data between biobanks and/or central hubs.
- Pseudonimization system – to ensure privacy of participants is protected and legal/ethical requirements are addressed (in collaboration with Parelsnoer).

D4 Core software platform (support of above to prevent reinvented wheels)

- Flexible 'model-driven' software platform – which allows to efficiently produce, configure and maintain all data models, databases, compute services and pipelines needed.
- Large data platform – to harmonize how to deal with the GWAS and NGS data within data archives (storage), algorithms (runtime) and data exchange (network)
- Flexible compute pipeline platform - to harmonize how to run large scale analyses without each pipeline having to bother about how difficult it is to run your algorithms on clusters, grids or clouds with suitable user interfaces
- Web access tools – harmonized user interfaces and programmers interfaces to provide a single point of access to all the resources developed in this project.

10. Required budget

Personnel / material	**Number / time period / ..**	**kEuro**
BBMRI-NL	8fte/3yr bioinformaticians*	1,440
NBIC	3fte/3yr software engineers*	540
BBMRI-NL	0.5fte/3yr fund for a NBIC technical coordinator**	135
BBMRI-NL	0.5fte/3yr system administrator***	90
BBMRI-NL	Monthly meetings and visits to international institutes	35
BBMRI-NL	250B/y1 + 500TB/y2 storage + backup + connects***	150
BBMRI-NL	300cpu compute service (dedicated) + 3500cpu (shared)***	100
BBMRI-NL	Running costs (power & cooling)/3yr***	50
Total:	**12fte/3 year + materials**	**2,540**
	Total BBMRI-NL	2,000
	Total NBIC	540

*A total of 11 fte/3yr will be allocated to produce Deliverable D1-D4 (3,3,2,3 fte, respectively), including 0.5fte project leader. **We will outsource software engineering coordination to NBIC. ***Based on confidential quotes from CIT/Groningen in collaboration with SARA and BigGrid.

11. Literature references (maximum 10)

1. Koboldt, DC (2010) Challenges of sequencing of human genomes. Briefings in Bioinformatics. Advance Access published June 2, 2010.
2. International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature. 18;449(7164):851-61.
3. The 1000 Genomes project. http://www.1000genomes.org
4. Thorisson GA, Muilu J, Brookes AJ (2009) Genotype-phenotype databases: challenges and solutions for the post-genomic era. Nature Reviews Genetics. 10(1):9-18.
5. Stein L (2008) Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. Nat Rev Genet 2008, 9:11.
6. Swertz MA, Jansen RC. (2007) Beyond standardization: dynamic software infrastructures for systems biology. Nature Reviews Genetics 8(3):235-43.
7. Rios D, McLaren WM, Chen Y, Birney E, Stabenau A, Fliceck P, Cunningham F (2010) A database and API for variation, dense genotyping and resequencing data. BMC Bioinformatics. 11:238.
8. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. (2007) GenABEL: an R library for genome-wide association analysis. Bioinformatics. 15;23(10):1294-6.
9. Swertz MA, Velde KJ, Tesson BM, Scheltema RA, Arends D, Vera G, Alberts R, Dijkstra M, Schofield P, Schughart K, Hancock JM, Smedley D, Wolstencroft K, Goble C, de Brock EO, Jones AR, Parkinson HE; Coordination of Mouse Informatics Resources (CASIMIR); Genotype-To-Phenotype (GEN2PHEN) Consortiums, Jansen RC. XGAP: a uniform and extensible data model and software platform for genotype and phenotype experiments. Genome Biology 2010;11(3):R27.
10. Estrada K, Abuseiris A, Grosveld FG, Uitterlinden AG, Knoch TA, Rivadeneira F (2009) GRIMP: a web- and grid-based tool for high-speed analysis of large-scale genome-wide association using imputed data. Bioinformatics, 25(20):2750-2.

Appendix 1 – Description of work

To ensure maximum utility and dissemination of this rainbow project we propose to develop a series of essential components for BBMRI-NL rainbow and complementation projects, most notable Genoom van Nederland (GvNL). All these activities have in common that they pose informatics infrastructure challenges that cannot be addressed by the participating laboratories individually. Figure 1 provides a vision of a suitable 'e-Science' infrastructure to address all these needs:

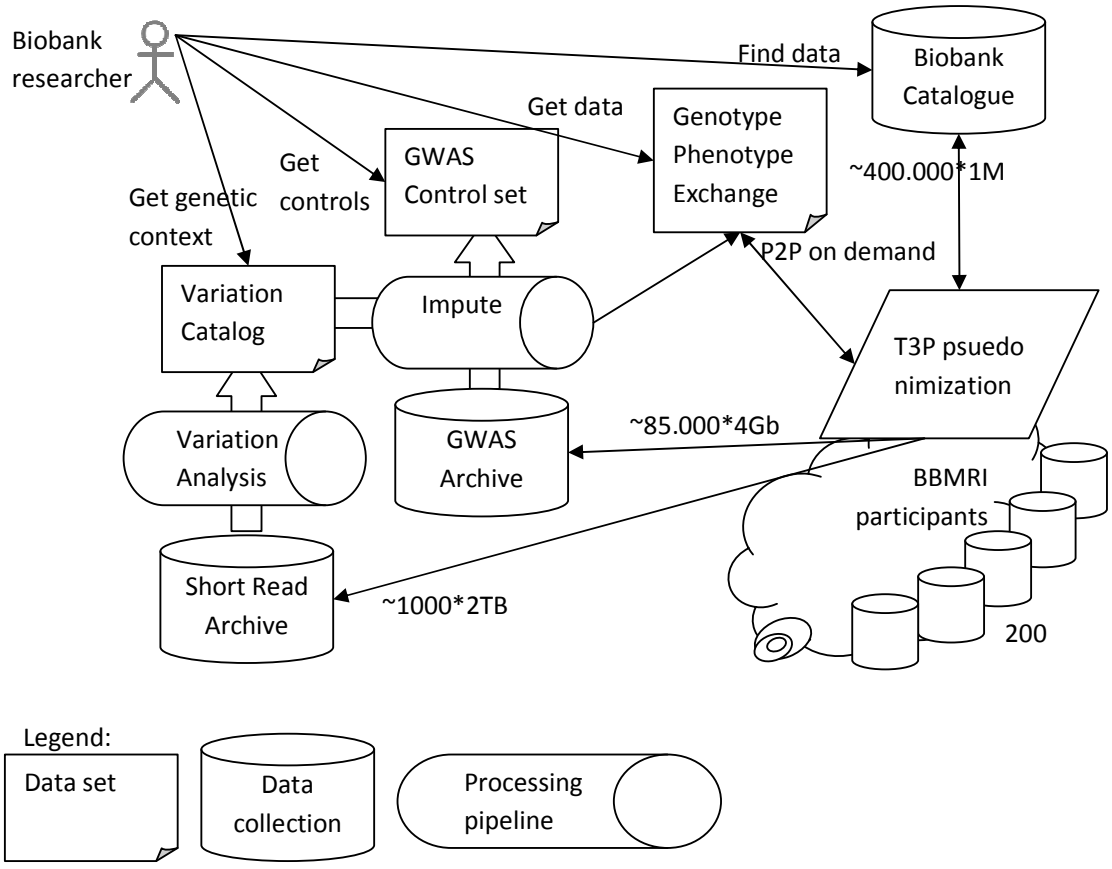


Figure 1: Conceptual overview of the e-Science infrastructure for biobanking

We organize the work in the following workpackages:

1. Variation analysis, addressing short read archive and variation analysis
2. GWAS archive, imputation and GWAS control set
3. Data search and exchange addressing biobank catalogue, formats and pseudonymization,
4. Flexible and scalable data and compute platform (cross cutting to support work of 1-3)

For each of the work packages we describe the purpose, tasks and organization below. Organizational benchmarking with for example the 1KG learned that one needs at least 8 bioinformatics analysis and 8 software engineers for workpackage 1 only. Because we expect to learn from such existing projects we aim for a much tighter budget of 12 fte for all four work packages (fte is shown per package in brackets).

Work package 1: Short read archive & variation analysis (3 fte)

Next generation sequencing technologies have greatly lowered the cost of (re)sequencing of individuals currently on the range of $10k. This technology push potentially enables the Dutch biobanks to enrich their collections with detailed DNA information from large numbers of individuals not only limited to the study of common genetic variation (like done in GWAS) but also less frequent genetic variations. This information will greatly help to pinpoint variation-disease relationships by providing much clearer hypotheses on the causative gen-transcript-protein-metabolomics mechanisms. The sequencing of up to 1000 individuals as part of BBMRI-NL Genoom van Nederland 'regenboog' project is already underway and will result in 500 – 2000 Terabytes of raw data files. Making possible the data management, processing and analysis of these large-scale data sets is the essence of the project described here. Therefore, this work package addresses the challenges in the management, QC, analysis and (predictive) annotation of next generation sequencing data in particular to support the Genoom van Nederland regenboog project to build a genetic variation catalog of the Dutch population:

Task 1.1: Sequencing Short Read archive

The first challenge in this work package is how to accommodate the large data produced by next generation sequencers as well as sample annotations ('study capture') and intermediate and resulting data from sequencing analyses . Already planned is the sequencing of up to 1000 individuals as part of BBMRI-NL Genoom van Nederland 'Rainbow' project which will result in 500 – 2000 Terabyte of raw data files. This obviously surpasses the capabilities of local labs to manage and process of all these new data and to suitably control the quality of data sets which is why they depend on the project described here to manage and process these data. Therefore the first task of this workpackage is to develop suitable 'database-in-a-box' storage infrastructure to manage and exchange NGS data and associated sample metadata and to track-and-trace downstream analysis results (see tasks below).

Task 1.2: Variation QC and analysis pipeline

The collection of short read data into suitable archives such as described above are just the first step to uncover genomic variation. The development of suitable analysis protocols is still undergoing major development: there are many questions still on what are the best practices to apply quality control procedures (e.g. verify against SNP arrays, see task 2.1), analyze extent of coverage, align against reference genomes, discover real DNA variation, detect and exclude sequencing errors, and functionally annotate with known or predicted consequences of discovered variation (modifier genes, protective other variants, influence on splicing, methylation). Many alternative and heterogeneous analytical components are available as open source (bowtie, BWA, SAMtools, Qcall, Dindel, Pindel and BreakDancer). Yet, most of these suits are very laborious to implement into suitable protocols. In addition, their execution is very computationally demanding both in compute power and runtime Random Access Memory. Finally, storage demands for intermediate and end results and the integration with available annotations in public repositories is very challenging. Therefore the second task in this work package is to assemble a scalable and configurable set of analysis pipeline 'tools-in-a-box' for variation analysis and its application to produce genomic variation reference sets on the data produced in Task 1.

Task 1.3: Dutch subpopulation structure and variation catalog

The purpose of the Genoom van Nederland project is to gain insight about variation in the genome of individuals from the Netherlands, including low frequency variants. The great benefit is that these low frequency variants can then be used to enrich the assessment of existing GWA data sets via imputation. It is estimated that there around 12-14 million SNPs in the Genome of which we expect to asses at least 5 million beyond statistical doubt. Hence, the third task in this workpackage which will compile a map of Dutch subpopulation structure and a catalog of rare variants/SNPs/CNVs and their frequency.

Milestones

- Short Read Archive
- Short Read QC protocols
- Availability of whole genome sequence of up to 1000 individuals (with GvNL)
- Genome variation analysis protocol
- GvNL catalog of structure and genetic variance in Dutch population (Dutch HapMap)

Work package 2: GWAS archive, imputation & data analysis (3fte)

Between all Dutch biobanks there is genotype data accumulated for >85.000 individuals, that is, each of these individuals has been genetically profiled on 300k to 1M genomic variants using genotyping microarrays enabling Genome Wide Association Studies (GWAS) that analyzes associations between these genomic variants and complex phenotypes. However, in order to find these associations vast numbers of individuals need to be examined – typically in the range at least 30.000 of subjects which is much greater than most biobanks can provide and pay for. Due to the nature of the data and the type of statistical analysis used in GWAS it is possible to pool the data from the various research groups into a control dataset collection which may greatly reduce costs while potentially increasing statistical power. This work package harmonizes and enriches available tools for the management, exchange, QC, imputation and analysis of GWAS data focused on building a national GWAS control cohort using data produced by the Genoom van Nederland rainbow project.

Task 2.1: GWAS data archive

The first challenge in this work package is how to securely track and trace all GWAS as well as intermediate and resulting (imputation) data, including software tools and methods for data input, curation, search and presentation. Moreover, because of the sensitivity of this data it is imperative that the data can be shared efficiently between labs for which the XGAP protocol will help (Swertz et al, 2010), for example to exchange data and tools with public repositories like EGA. Therefore the first task in this workpackage is to establish a GWAS cohort portal 'database-in-a-box' to address the challenge of storing and combining the large genotype raw data files (several gigabytes per measurement, thousands of measurements), minimal information on the samples, computationally intensive QC protocols, and derived information such as imputation data (see below).

Task 2.2: GWAS QC and imputation service

Quality control and imputation of GWAS data is already challenging on smaller populations of several thousand individuals and 2.5 million known variations requiring several months of processor-time. Now these procedures need to be scaled up 10x to 85.000 individuals and potentially much more than 2.5 million known variations as produced by the genomic variation analysis and major sequencing projects described above (estimations are we will get at least 5 million). Therefore the

second task is to provide the biobank community with a reusable QC and imputation services to enrich the large GWAS data collected into Task 1.3 with cleaned variant imputations. This task will build on WP4 to deal with the challenging storage and computational demands.

Task 2.3: Dutch GWAS Control Cohort

Tasks 2.1 and 2.2 work up towards the establishment of a national control cohort of control samples fo GWAS studies based on the HapMap-like catalogue of results derived from GvNL (task 1.3). There is a pressing need for implementing such a control cohort pooling this information which will increase the statistical efficiency of research efforts at a lower cost than having individual studies create control cohorts themselves. In collaboration with NBIC/BioAssist the VU (Dorret Boomsma) and UMCG (Cisca Wijmenga) have outlined what is needed to setup such a central resource, see (Kattenberg & Swertz, 2009[1]), including security constraints, concerns with ownership of the data, need for trusted third parties to host the data, and an application procedure and steering committee to govern data releases (related to PALGA). Also in WTCC, EGA and dbGaP there are experiences with similar systems. Based on this, this main task of this effort will be collecting available BBMRI-NL GWAS data into an instance (2.1) named the Dutch GWAS Control Cohort (DGCC) repository which will provide the community with secure web access to suitable (imputed) data releases and primary data exploration tools.

Task 2.4: Catalog of GWAS statistical analysis tools and pipelines

Harmonisation and integration of data and results can be achieved on two levels, mainly: genotype and phenotype data. Such data can be made available for pooled analysis or it can be provided on a higher level of aggregation, for example through reporting the sufficient statics, such as effect estimates, standard errors and their variance-covariance matrix. The latter approach has been very successful but depends on uniformity in the local analysis steps from QC all the way up to the GWAS and xQTL analysis itself. Therefore this tasks aims to harmonize and enrich statistical tools for GWAS (meta)analysis and to make them available via central catalogs integrated with the results of task 2.1 and task 2.2.

Milestones

- GWAS infrastructure built on XGAP and EGA projects
- Central database with all GWAS data
- Established GWAS QC procedures and protocols
- Scalable and reusable GWAS imputation pipeline
- Online catalog of GWAS data analysis protocols and tools
- Imputation of all Dutch GWA data sets on low frequency variants
- Make data available in Dutch GWAS Control Cohort database

Workpackage 3: biobank data search, exchange & integration (2fte)

At the heart of BBMRI-NL the mission is to make existing biobanks optimally accessible by providing detailed descriptions, easy interfaces and centralized, controlled data access. In sum, this workpackage will provide a catalog of information on the variables, high-throughput data and materials available in each biobank, and harmonize data formats for the exchange between biobanks (methodologies needed to enable valid phenotype variable pooling are not within the scope of this

[1] Kattenberg & Swertz (2009) Design specification of the DGCC Dutch Genotype Archive for GWA Control Cohorts. Technical report VU, UMCG, NBIC.

activity, i.e. we don't address the question: if one study measures 'number of beers drank/day', and the other 'glasses of wine/week' if these studies can be integrated on a common 'alcohol intake' variable. We recommend another, research-question driven rainbow project to answer these context specific questions, see Appendix 3).

Task 3.1: Central biobank catalogue which can be also used as local biobank archive

The first step in data harmonization is making accessible a general inventory on the available materials, data and contributing biobanks through a centralized database, linked to other BBMRI-EU catalogues. This catalogue will consist of aggregated information, such as the numbers and kinds of materials and the available variables for each biobank. The search of biobanks and biobankers could greatly benefit from the use of data models developed in initial BBMRI-EU catalogues, GEN2PHEN, P3G, BBMRI-EU, Parelsnoer and Mondriaan. This will also include models and semantic search technologies being developed in the ontology projects like NBIC/Concept Web Alliance to facilitate elucidating the overlaps between available biobank resources, particularly cross-language searches and linking to known semantic relationships between available information (under the common search Booleans 'also refered to as', 'is_a' and 'part_of'). Therefore, the first deliverable of this workpackage is a national biobank data portal that brings together all available knowledge on which data and materials are available across biobanks, which questionnaires, protocols and SOPs were used and finally which experts are available to help with the exploitation of available biobank data. The result compilation will be developed such that local groups can also install a local version of the catalogue for local data management (including individual level data) with additional benefits easing future integration.

Task 3.2: Data exchange formats and systems

In this task, we will intensively collaborate with other national hubs and initiatives like P3G, EBI, GEN2PHEN and Parelsnoer, in order to harmonize models, formats and software tools for data exchange and searchable indices. For the exchange of "omics" data, we will build on top of some parallel, recently successful harmonization initiatives like the international healthcare exchange format HL7, PRIM (parelsnoer Reference Information Model), FuGE (a general model for functional genomics experiments), XGAP (a flexible system for QTL, GWL and GWA studies), MAGE-TAB (MicroArray Gene Expression tabular format), the Investigation, Study and Assay (ISA-)TAB formats and the Minimum Information for Biological and Biomedical Investigations (MIBBI) guidelines, and established sequencing formats. Access to the individual level data should be well-regulated meeting legal requirements regarding privacy and informed consent.

Task 3.3: Integration with a pseudonymization system

Because of legal, ethical and privacy-related aspects, it is necessary that the delivered data cannot be related to individuals. For that reason, personal data are decoupled from medical investigation data by using separate identification numbers. This process is called pseudonymization. The relation between the pseudonymous numbers can be safely stored at a Trusted Third Part (TTP), thereby warranting the anonymity of the individual. Regarding pseudonymization, we expect to adopt the results of efforts at BBMRI-EU and Parelsnoer.

Milestones

- Creation of pilot/proof of concept biobank catalog
- Provide updated overview with metadata on all participating cohorts

- Release of extended biobank catalog including the data that biobanks can use for local data management/submission to exchange model
- Run evaluations and prototyping testing of standards for format harmonization, for example matching the PSI and GEN2PHEN initiatives for phenotypes and FUGE/XGAP/MAGE-TAB for 'omics' data
- Achieve toolkit development for harmonized genomics exchange models (sequencing, gwas, microarrays, proteomics, metabolomics)

Workpackage 4: Scalable and flexible data management and processing platform (3+0.5+0.5 fte)

This work package is cross-cutting across working group 1-3 addressing the software engineering aspects of large scale data and processing infrastructure and underlying hardware. The remit of this workpackage is to produce a general platform to store all files, all metadata (so data on samples, annotations and so on), have efficient and flexible methods to assemble computational workflows, and have a catalog of all tools underlying WP1-WP3. Experience shows that none of the existing tools will provide a turn-key solution but we expect to save much time and effort by building on existing toolboxes. To ensure that this workpackage really serves the needs of workpackages 1-3 there should be at least one participant in each workpackage who is also in workpackage 4. The infrastructure aspects will be filled-in in collaboration with national initiatives of NBIC, NPC, NMC, technical BBMRI-EU workpackages, national BBMRI hubs and ELIXIR and hardware organizations as CIT/Groningen, SARA/Amsterdam and BigGRID.

Task 4.1: flexible toolboxes to assemble data management, processing and web access tools

Given the dynamics of the life science, the first challenge of workpackage 4 is to produce the necessary software infrastructure such that it can be quickly adapted to new protocols and data (i.e., in this project, rapidly evolving methods used by NGS and GWAS efforts). We therefore will adopt an alternative software engineering strategy, as outlined in our recent review (Swertz & Jansen, 2010), that enables generation of such software efficiently using three main components: 1) a compact and extensible 'standard' model of data and software; 2) a high-level domain-specific language (DSL) to describe in a simple manner biology-specific customizations to this software; and 3) a software code generator to automatically translate models and extensions into all low-level program files of the complete working software, which builds upon reusable elements from national and international initiatives, collaborations and open source projects (like MOLGENIS for data and Galaxy/Taverna for workflows). This way, this task will provide WP1-WP3 with (MOLGENIS-like) methods to efficiently assemble all data models, compute services and pipelines together with suitable user interfaces for biologists and programmatic interfaces for bioinformaticians.

Task 4.2: scalable data infrastructure

Advances in GWAS and sequencing technology over the last decade have transformed biology in an information rich science. In parallel to this, the storage and analytical requirements have grown beyond what commodity database systems and computing infrastructure can provide. Several projects have been addressing these life science specific issues with some success, such as XGAP, EGA, the ABEL set of programs, GRIMP, 1KG, ENGAGE, ObiBaand so on, from which we expect to assimilate the best practices. This task will build on, and tailor, these efforts to BBMRI-NL to produce a suitable, scalable data platform to deal with the GWAS and NGS data within data archives, algorithms and data exchange. In this work, we also aim to bridge towards other national

infrastructure efforts such as the String of Pearls Initiative and future data sharing initiatives in BBMRI-NL (see appendix 3).

Task 4.3: scalable compute infrastructure

Next-generation sequencing analyses and GWAS imputation requires vast compute infrastructure which are non-trivial to operate. Large genome centers are forced to deal with such gargantuan challenges setting up large compute pools, automated data management systems and analysis pipelines. Such setup should incorporate support staff working under the same roof who can create software tailored to the needs of researchers and prepared to solve computational problems. We are fortunate in the Netherlands to have strong compute centers, SARA in Amsterdam and CIT in Groningen connected by Big Grid, with whom we collaborate to develop life science friendly compute services. Therefore, this task will build on this collaborations to produce a flexible platform and user-friendly interfaces to use publicly available or custom-developed genetic software for large scale analyses such as imputation and variation analysis on large biobank populations (workpackages 1-3) using scalable super-computing clusters, grid or cloud infrastructures

Appendix 2 – Organization

Suggested participating experts

- Cisca Wijmenga, Eline Slagboom, Cornelia van Duijn, Jasper Bovenberg, GJ van Ommen (GvNL)
- Morris Swertz, Hans Hillege, Lude Franke, Noortje Festen, Jan Jongbloed, George Byelas, (UMCG, LifeLines)
- Johan den Dunnen, Kai Ye , Judith Boer, Bas Heijmans, Joost Kok (LUMC)
- Dorret Boomsma, Jouke Jan Hottenga, Marleen de Moor (VU FPP NTR)
- Paul de Bakker (UU)
- Yurii Aulchenko en Fernando Rivadeneira (EMC)
- dr. Henk van Kranen (RIVM)
- Leon Mei, Morris Swertz, Christine Chichester, Rob Hooft, Barend Mons (NBIC)

Suggested staff

- Erik Roos (UMCG) – as 3yr technical project coordinator
- Mathijs Kattenberg (NTR, VU – FPP, NBIC/Biobanking platform)* - as 4 year PhD candidate
- Freerk van Dijk (NBIC/sequencing platform) – as 3yr engineer sequencing pipelines
- Despoina Antonakaki (NBIC/biobanking) – as 3yr engineer biobank catalog

National and international collaborations:

- Netherlands Bioinformatics Center, in particular BioAssist Biobanking, Sequencing, Interoperability and e-Science task-forces.
- 1KG project, Sanger & European Bioinformatics Institute, Paul Fliceck & Helen Parkinson
- Centre for Genomic Regulation, Ivo Gut?
- UW Genome Center, Deborah Nickerson
- ABEL software suite, coordinated by Yurii Aulchenko
- GRIMP, Fernando Rivadeneira-Karol Estrada
- TriTyper, Lude Franke
- Parelsnoer, Gerard van der Hoorn en Joost Schalken
- LifeLines, Hans Hillege
- MOLGENIS and XGAP (eXtensible Genotype And Phenotype platform), UMCG/EBI/GEN2PHEN, Morris Swertz en Joeri van der Velde
- EGA (European Genotype And Phenotype archive), EBI, Ilkka Lappainen
- DataSHIELD - Data Aggregation Through Anonymous Summary-statistics from Harmonized Individual levEL Databases, Paul Burton
- EU-BioSHARE, Ronald Stolk & Paul Burton
- GEN2PHEN data formats, Tony Brookes and Helen Parkinson
- R-cloud/farm, European Bioinformatics Institute, Misha Kapusheshky
- Center for Information Technology Groningen, Hans Gankema
- SARA and NBIC eScience, Machiel Jansen, Joost Kok, Timo Breit
- Target initiative for astronomical large computing and storage, RUG, Edwin Valentijn

Appendix 3 – What is not covered by this proposal is phenotype harmonization

At the steering committee meeting of February 15, 2010 Morris Swertz was asked to assemble a project plan for BBMRI-NL on 'data management and analysis'. The results conform this request are described above with one modification: an additional work packages was suggested to deal with 'phenotype variable harmonization and integration', that is, ask the question if particular variables can be harmonized between biobanks. With the installation of the GvNL rainbow project a huge amount of new data management and processing work has been added to what was originally proposed. After discussion with BBMRI-NL steering committee members it was decided to take this work package out of the current proposal to make sure it gets the resources needed in dedicated project. We have added a summary of the 'phenotype variable harmonization' work needed below:

Phenotype variable harmonization work still needed beyond this project

The Netherlands can excel at harmonizing phenotype descriptions. There are technical and content-related aspects to this. Technically, phenotype descriptions (variables, observations, protocols) should be unambiguously defined within each biobank and be related and converted between biobanks. This "inter-operabilization" should be tackled partly centrally and partly for each biobank individually. Concerning content, we can make a distinction between clusters of clinical biobanks and population biobanks, and the definition of minimal data sets needed for the analysis of specific diseases and phenotypes. In its program of data harmonization, such project will need to concentrate on formalizing the core set of subcomponents which need to be uniform – or nearly so - between biobanks. For example, universally recognized core object models and ontologies/terminologies are needed to ensure datasets can be adjoined, and data exchange formats with certain required fields are necessary to enable data to pass easily between systems whilst simultaneously tracking the allowed uses, consents, and researcher contributions. In order to immediately create as much added value as possible, harmonization should be driven by projects that are related to practice. Opportunities and needs in scientific research of multi-factorial diseases should be at the core. Examples are combining biobanks with specific phenotypes in order to obtain enough samples to study the occurrence and progression of a certain disease or converting free text clinical descriptions to structured annotations with unequivocal variable definitions. Critically, harmonization is context specific. Some of the information available from one study may be similar enough to that from another to allow pooling for a given scientific question, while other items of information from those same two studies may not be suitable for pooling. Working hypothesis is that we can reuse much of the data exchange methods developed from prospective data collection in the Parelsnoer initiative including (parts of) its Reference Information Model (PRIM), pseudonimization services and phenotype variable harmonization methods, and from the Mondriaan project including trusted-third-party linkage of healthcare databases. Such future project should cooperate with international consortia like Concept Web Alliance, P3G/DataSHaPER and BioSHARE-EU.

Suggested partners

All participating BBMRI-NL biobanks
Gerard van der Hoorn, Parelsnoer
Marc Rietveld, Mondriaan
Hans Hillege, LifeLines, UMCG
Ronald Stolk, BioSHARE-EU
Isabel Fortier, P3G
Barend Mons, Concept Web Alliance