Publishable Executive Summary
January 2008 to December 2008

ENGAGE - European Network for Genetic and Genomic Epidemiology FP7-HEALTH-201413

ENGAGE concept and objectives:

This is an exciting time in human genetics and genomics as a stream of novel disease-susceptibility genes have recently been identified through the application of technologies that look for disease associations across the whole genome (genome wide association studies). These studies are providing an insight into the biological pathways underlying the major causes of human morbidity and mortality. It is clear, however, that the power of single cohort studies to detect causative genetic variants is limited to relatively large effect sizes by common alleles. Therefore, in order to identify the full range of genetic variation contributing to common disease and to uncover the effects of the complex interactions of genes, environment and lifestyle factors on disease risk, a wider scale epidemiological approach is required.

Collectively the ENGAGE consortium partners have access to an extensive range of well phenotyped and catalogued population cohorts representing >600,000 subjects and including a number of ethnically homogeneous population sets. Genome wide association data (GWA) are available for >100,000 of these subjects and an early goal of the ENGAGE project is to bring together these datasets to perform large scale integrated genetic asso-

Project Coordinator:
Professor Leena Peltonen
Research Director, Institute for Molecular Medicine Finland FIMM, Nordic EMBL Partnership for Molecular Medicine, Biomedicum Helsinki 2U, 00251 Helsinki, Finland
email: contact@euengage.org
ENGAGE Structure:

ENGAGE activities are organized through nine work packages

- WP1 Genome Wide Data Integration
- WP2 Novel sources of Genomewide Variation
- WP3 Novel Phenotypes
- WP4 Informatics and Bioinformatics
- WP5 Genetic Refinement of Identified Loci
- WP6 Epidemiology and Joint Effects
- WP7 Clinical Translation
- WP8 Societal Aspects
- WP9 Training and Dissemination

All work packages have been operational in the first 12 months with major progress around activities supporting the sharing of data for large scale integration studies and the identification of disease susceptibility genes through the meta-analysis of GWA datasets from the ENGAGE cohorts.

Association analyses. The best represented phenotypes include cardiovascular and metabolic disease related traits, but also include a diversity of behavioural traits relevant for disease risk. Adopting this approach allows the consortium to identify novel disease-susceptibility variants that would not be detectable in lower powered individual cohort studies. A key ENGAGE objective is to evaluate the clinical and public health relevance of the novel disease and trait-susceptibility genes that we identify and to demonstrate that these findings can be used as diagnostic indicators for common diseases helping us to better understand risk factors, disease progression and why people differ in responses to treatment.

ENGAGE will extend our integrated genetic analyses to additional sources of genome variation (e.g. structural variation) as methods improve for the large-scale collection and analysis of these data types, and to novel phenotypes (e.g. trait clusters, transcriptomic, proteomic and metabolomic data) as such datasets become available from ENGAGE partners. We will also explore key methodological questions relevant to European research in genetic and genomic epidemiology (including for example, the consequences of ethnic and environmental heterogeneity for gene discovery efforts and the allelic architecture of common disease) and develop novel statistical approaches for data analysis.

Key to the success of the consortium in risk marker identification and clinical translation are the ENGAGE objectives for data sharing and harmonisation. We will develop new computational approaches supporting data sharing and the harmonization of cohort phenotypes whilst establishing protocols for managing the ethical aspects of sample and data sharing according to informed consent, local ethical approval and the governance structures of each ENGAGE partner. ENGAGE will work collaboratively with other research consortia including P3G, EGA, ELIXIR and BBMRI to ensure the compatibility of ENGAGE approaches with existing European and global initiatives in this area.

Disease focus: Initially ENGAGE will concentrate our efforts around metabolic and cardiovascular disease phenotypes, due to their importance for European and global health, but the methods developed and lessons learned will be applied to a wider range of disease areas including behavioural and psychiatric phenotypes during the course of the project.

Data sharing and integration: During 2008 the WP1 and WP4 teams worked closely to identify the data submission and exchange requirements needed to support the large scale integrated analyses of ENGAGE GWA data. The data submission system deployed for the project (SIMBioMS www.simbioms.org) is comprised of two components, AIMS and SIMS which enable ENGAGE partners to submit and share standardised GWA data and phenotypic meta-data within the consortium. Data access rights can be set up in accordance with study requirements and with the ENGAGE data access policy established by WP8. From a standardisation perspective the system is compatible with data export to major public data archives (e.g. the European Genotyping Archive (EGA), ArrayExpress and PRIDE (for proteomic data)). To ensure that collaborative data sharing efforts in ENGAGE are operating within an appropriate ethical framework, the WP8 team have collected ethics and consent forms for all partici-
pating cohort studies and have established a coordinating committee to prepare a project policy for managing the release of project generated data and knowledge to the scientific community.

To date summary GWA data from 230 ENGAGE cohort and phenotype combinations has been uploaded to AIMS and ENGAGE has led or contributed to meta-analysis efforts around several phenotypes resulting in the successful identification of over 60 disease susceptibility loci during 2008. These include loci affecting several well-established risk factors for cardiovascular diseases and T2D (serum glucose and lipid levels, blood pressure, height, BMI and weight). In the next reporting period we anticipate the completion of gene identification from a series of ongoing meta-analysis studies coordinated through WP1 and WP7 (clinical translation) looking at T2D complications, pulse pressure, urate levels, liver and lung function tests and behavioral traits such as smoking and alcohol and handedness. We will also further test observations from the ENGAGE lipid meta-analysis that the genetic risk profiles associated with serum lipid levels improve our ability to identify individuals at higher risk for developing dyslipidemia.

**Funding and participants:**

The ENGAGE project is funded with 12 million Euros through the EU 7th Framework programme and is coordinated from the University of Helsinki. The ENGAGE consortium is comprised of 24 partners, from Europe, Canada and Australia, including twenty-two from Universities and Research Institutes and two commercial partners. More information about the key scientists involved in the project at each partner site can be found on the project website.

**ENGAGE PARTNERS**

1. University of Helsinki, Institute for Molecular Medicine Finland (FIMM), Finland
2. University of Oxford, The United Kingdom
3. European Bioinformatics Institute (EMBL-EBI), The United Kingdom / European Molecular Biology Laboratory, Germany
4. Queen’s University Belfast, The United Kingdom
5. King’s College London, The United Kingdom
6. Illumina, Cambridge LTD, The United Kingdom
7. Leiden University Medical Centre Centre for Medical Systems Biology (CSMB) / Vrije Universiteit Amsterdam, The Netherlands
8. Erasmus Medical Center, The Netherlands
9. German Research Center for Environmental Health (former GSF), Germany
10. Karolinska Institutet, Sweden
11. Lund University, Sweden
12. Uppsala University, Sweden
13. Royal Institute of Technology, Sweden
14. Norwegian Institute of Public Health, Norway
15. University of Tartu Estonia Genome Project (ECP), Estonia
16. deCODE genomics, Iceland
17. Ontario Institute for Cancer Research, Canada
18. Université de Montréal, Canada
19. Centre for Genomic Regulation, Spain
20. University of Leicester, The United Kingdom
21. University Lübeck/University Medical Center Schleswig-Holstein, Germany
22. Imperial College London, The United Kingdom
23. Queensland Institute of Medical Research, Australia
24. The Wellcome Trust Sanger Institute, The United Kingdom

**Novel phenotypes and sources of genetic variation:** The maturity and relatively lower experimental cost of the technology platforms generating GWA datasets have meant that these datasets have been the first to be available on a large scale across the ENGAGE cohorts. The consortium has moved quickly to translate meta-analysis of GWA data into novel genetic discoveries, but is cognizant of the opportunity provided for risk biomarker discovery from additional sources of both genomic variation and novel molecular ‘omic’ phenotypes emerging for the ENGAGE cohorts. WP2 are leading the efforts to utilise copy number variation (CNV) and deep resequencing data in large-scale integration studies over the course of the project. Early efforts have focused on evaluating existing algorithms for studying CNV data and establishing a common protocol for data standardisation across platforms. Current public initiatives such as the 1000 genomes project and WTCCC will soon report on large scale CNV studies and ENGAGE will use the observations from these efforts to focus and inform our CNV activities over the coming period. During year 1 resequencing platforms have been installed at several ENGAGE facilities and efforts targeted to the set up and testing of the systems to produce standard protocols and a pipeline for analysis. A flagship re-sequencing project focused on a subset of the loci identified from the ENGAGE led meta-analysis around lipid and glucose levels has been designed and will be implemented in 2009.

During 2008 WP3 has focused on identifying existing metabolomic and proteomic datasets from ENGAGE cohorts and identifying biofluid sample resources for use in focused studies, for example looking at extreme phenotype or trait ends. Steps have also been taken towards evaluating novel statistical methods to allow cross platform data integration.

**Refinement studies of identified loci:** To identify the causative genetic variants that are driving ENGAGE disease susceptibility signals we need to define the full allelic architecture of the associated genomic region in addition to evaluating the relationship of the effect on related phenotypes or co-morbid conditions. ENGAGE WP5 is leading the consortium efforts to develop efficient strategies for genetic fine mapping of SNPs and comprehensive resequencing in genomic regions harbouring associations. Refinement by resequencing is costly and initial efforts by WP5 partners have focused on evaluating strategies that reduce the experimental requirements and maximise the information available from existing GWA data. During year 1 ENGAGE partners have successfully demonstrated that sample pooling prior to resequencing is a viable alternative to sequencing individual samples when combined with genotyping of the new variants identified. Additionally a method of determining long-range haplotypes using existing GWA data without additional genotyping, has been developed for use in populations where at least 1% of the individuals have been genotyped with >300,000 SNPs. This method can be used to prioritise regions for resequencing and applies also to at least five ethnically homogeneous populations within ENGAGE.
Data harmonization: Genotype and phenotype data for the ENGAGE cohorts has been generated using a range of technology platforms and collected through cohort specific questionnaires. This complexity can be a barrier to pooling of phenotypes and genotypes across cohorts; in order to maximize the number of ENGAGE resources that can be utilised in integrated studies the consortium has set out to define harmonised descriptions and formats for traits of interest. Initial efforts by WP4, WP6, WP3 and WP8 have focused on establishing a harmonised vocabulary for the metabolic syndrome. A strategic collaboration has been established between ENGAGE and the Public Population Project in Genomics (P3G) Consortium, who have provided expertise on the assessment of consistency and quality of the mappings between individual cohort parameters and the harmonized vocabulary for the metabolic syndrome. ENGAGE WP4 has developed a web based solution for storing these mappings for ENGAGE cohorts, the sample availability system, (SAIL). This platform will greatly support future ENGAGE efforts by facilitating the reporting of meta-data for summary level meta-analysis and supporting WP6 efforts to identify cohorts with the relevant lifestyle and environmental phenotypes to include in studies to identify the joint effects of genes and lifestyle. Data harmonisation of the ENGAGE efforts will be closely integrated with the ongoing ESFRI activities like BBMRI, ELIXIR and EATRIS.

Training and dissemination: The results of the first round of ENGAGE led meta-analyses in lipids and glucose have been published in high profile journals and have raised interest in the mainstream press, resulting in a dissemination of key messages to the general public.

Plans for the future: A series of high-impact ‘flagship projects’ are planned for the next 18 months of the ENGAGE project. These include a deep resequencing effort to refine a subset of the genomic regions identified from the early WP1 led integrated GWA meta-analyses and containing susceptibility loci affecting glucose and lipid levels, blood pressure, height, BMI, weight and T2D. Similarly ENGAGE will perform large scale epidemiology studies by genotyping a selection of the SNP markers showing association from these studies in 100,000-plus DNA samples from across the range of ENGAGE cohorts.

Early meta-analysis efforts in ENGAGE have used summary GWA data from partners. During the last year ENGAGE has formulated informatics, data access, harmonization and ethical solutions for sharing genotype and phenotype data at the individual level. These data will support future efforts of ENGAGE with respect to the analysis of complex multivariate phenotypes. Data harmonization efforts will be continued and coupled with the processes for sharing individual lifestyle and environment data will enable the consortium to design experiments to address questions of gene and environment interactions for the ENGAGE disease susceptibility loci.

ENGAGE on the web:
Follow project news and learn about ENGAGE-sponsored training events and publications and progress in research activities at the project website:

www.euengage.org

Early successes: ENGAGE led meta-analysis studies investigating genetic variants with an influence on serum glucose and lipid levels culminated in two high profile publications for the project in period 1.
