



Publishable Summary (January 2008 to June 2010-reporting periods 1 & 2)

Project acronym: *ENGAGE*
Project title: *European Network for Genetic and Genomic Epidemiology*
Grant Agreement No: *201413*
Instrument: *Seventh Framework Programme, Collaborative project*
Project Website: *www.euengage.org*

The 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.

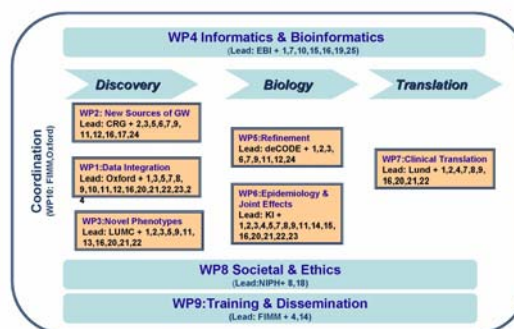
Collectively the ENGAGE consortium partners have access to an extensive range of well phenotyped and catalogued population cohorts representing >600,000 subjects and including a number of ethnically homogeneous population sets. Genome wide association data (GWA) are available for >100,000 of these subjects and an early goal of the ENGAGE project is to bring together these datasets to perform large scale integrated genetic 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 encompass additional sources of genome variation (e.g. Epigenetic variation) as methods improve for the large-scale collection and analysis of these data types, and to additional phenotypes (e.g. trait clusters, transcriptomic, proteomic and metabonomic 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 have developed 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 Structure and Management: ENGAGE activities are organized through ten 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
- WP10 Coordination



All work packages have been operational in the first half of the 60 month project with major progress around scientific 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.

Data sharing and integration: During period 1 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)). SIMBIOMS is widely used in ENGAGE and by the end of period 2 over 800 GWA datasets had been uploaded to support meta-analysis studies in WP1 and WP7. To ensure that collaborative data sharing efforts in ENGAGE are operating within an appropriate ethical framework, the WP8 team reviewed consent forms for all participating cohort studies and published their guidance on how to handle issues around retrospective consent as viewed through the ENGAGE experience (Tassé et al., 2010).

Discovery of Susceptibility Loci: ENGAGE has both led and joined forces with other consortia (e.g. MAGIC, CHARGE, GIANT, HaemGen, TAG, OX-GSK, CGASP, SpiroMeta, EGG, CARDIOGRAM, SUNLIGHT) in a series of meta-analysis studies leading to the identification of over 150 genetic variants with an influence on: anthropometric traits (heights, obesity and body composition), serum glucose, lipid levels, coronary artery disease (CAD), blood pressure, lung function, early growth phenotypes (e.g. fetal growth, birth weight), urate levels, haematological parameters, vitamin D deficiency, and smoking and nicotine dependence. These analyses have culminated at least 22 high profile publications published in journals such as Nature, Lancet, Nature Genetics for the project in period's 1 and 2, for example:

- Inga Prokopenko et al, 'Variants in MTNR1B influence fasting glucose levels'. *Nature Genetics* 41, 77 - 81 (2009, epub 2008).
- Yurii S Aulchenko et al, 'Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts'. *Nature Genetics* 41, 47 - 55 (2009, epub 2008)
- Gudmar Thorleifsson et al. *Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nature Genetics* 41, 18 - 24 (2009, epub 2008)
- Newton-Cheh et al. *Genome-wide association study identifies eight loci associated with blood pressure. Nature Genetics* 41, 666 - 676 (2009).
- Soranzo et al. *A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nature Genetics* 41, 1182- 90 (2009)
- Repapi et al. *Genome-wide association study identifies five loci associated with lung function. Nat Genet* 42, 36-44 (2010)
- Tobacco and Genetics Consortium. *Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet. 2010 May;42(5):441-7*
- ThorgeirThorgeirsson et al., 'Sequence variants at CHRN3-CHRNA6 and CYP2A6 affect smoking behavior'. *Nature Genetics* 42, 448-453(2010.)
- Wang et al. *Common genetic determinants of vitamin D insufficiency: a genome-wide association study. Lancet*, 376(9736):180-8 (2010)
- Teslovich et al.. *Biological, clinical and population relevance of 95 loci for blood lipids. Nature* 466(7307):707-13 (2010)

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), epigenomic and deep resequencing data in large scale integration studies for ENGAGE. During period 2 the message emerging from the large public initiatives charged with investigating the potential for large scale studies of CNVs (e.g. WTCCC and GSV projects) has indicated that these markers offer less immediate utility as genetic risk factors than was initially hoped. Therefore ENGAGE has deprioritised these efforts and have refocused our efforts on flagship activities in the epigenomic analysis of ENGAGE twin resources, in studies of telomere repeat length and its relationship to complex disease and through WP3 also in the generation of new metabolomics data from ENGAGE cohort samples.

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. In year 1 of the project refinement by resequencing was costly and early efforts by WP5 partners focused on evaluating strategies that reduce the experimental requirements and maximise the information available from existing GWA data. ENGAGE partners 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. Re-sequencing platforms have been installed at several ENGAGE facilities and during period 2 efforts have focused on set up and testing of the systems to produce standard protocols and a pipeline for analysis at an affordable cost. 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 late 2010.

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. In period 2 ENGAGE WP4 has developed a web based solution for storing these mappings for ENGAGE cohorts, the sample availability system (SAIL; <http://www.ebi.ac.uk/Tools/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. Considerable challenges remain in data harmonization of environmental variables due to multiple developmental, social, cultural and linguistic effects in creating comparable measures for studies that were not originally developed for this purpose. Data harmonisation of the ENGAGE efforts will be closely integrated with the ongoing ESFRI activities like BBMRI, ELIXIR and EATRIS to ensure the compatibility of ENGAGE approaches with existing European and global initiatives in this area.

Training and dissemination: WP9 has organised a series of workshops and training courses open to both ENGAGE and external participants. These included training courses in data analysis and management, workshops on statistical methods development and the joint P3G/CSG/ENGAGE Summer Institute on "Genetics, ethics and clinical translation". ENGAGE partners have published 45 manuscripts relating to project funded activities in the first half of the project. These include a number of project led disease gene identification studies in high profile journals and have raised interest in the mainstream press, resulting in a dissemination of key messages to the general public. A project website (<http://www.euengage.org>) has been established, with a news section where developments of interest from the consortium are targeted to a wider audience.

Current developments: A series of high-impact 'flagship projects' are being performed during 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. These aims to evaluate the population impact of the SNPs in unrelated population cohorts and explore the joint effects of genes and environment.

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 more complex questions of gene and environment interactions for the ENGAGE disease susceptibility loci.

ENGAGE funding and participants: The ENGAGE project is funded with 12 million Euros through the EU 7th Framework programme and is jointly coordinated from the University of Oxford and the University of Helsinki. The ENGAGE consortium is comprised of 25 partners, from Europe, Canada and Australia, including 23 from Universities and Research Institutes and 2 commercial partners. More information about the key scientists involved in the project at each partner site can be found on the project website: www.euengage.org

Current Project Coordinator Contact details:

Professor Mark McCarthy

Robert Turner Professor of Diabetes

Oxford Centre for Diabetes, Endocrinology and Metabolism (OCDEM)

Churchill Hospital, Old Road

Headington, Oxford, OX3 7LJ, UK

email: contact@euengage.org

The ENGAGE project was led by the world-renowned human geneticist Professor Leena Peltonen from FIMM, University of Helsinki for the first 26 months of the project, until mid March 2010 when she sadly passed away after a long illness. The project co-coordinator, Professor Mark McCarthy from University of Oxford has assumed the project leadership and will continue to lead ENGAGE towards the project end. University of Helsinki continues as EC contractual coordinator in this project.

