Project No. - LSM-CT-2004-005268
Project acronym - EuroClot
Project title – Genetic Regulation of the End-Stage Clotting Process that leads to Thrombotic Stroke
Instrument – Sixth Framework Programme Priority 1
Thematic Priority – Life Sciences, Genomics and Biotechnology for Health

Final Activity Report

Period covered: 1st January 2005 – 30th June 2008
Date of preparation:
Start date of project: 1st January 2005
Duration: 42 Months
Project coordinator name: Prof Tim Spector
Project coordinator organisation name: King's College London
Version number: 3.0

Version control

Version No	Date	Name	Amendment
1	1/6/08	Tim Bowden	Creation of report
2	7/10/08	Tim Bowden	Major revision with change in document structure
3	17/12/08	Tim Bowden	Major additions

Coordinator Contact Detail

Professor Tim Spector
St Thomas' Hospital
Twin Research Unit
Lambeth Palace Road
London SE1 7EH
United Kingdom

Tel: +44 207 188 6765
Fax: +44 207 188 6718
Email: tim.spector@kcl.ac.uk

Contractors Involved

No.	Participant	Principal investigators
1	King's College London	Prof Tim Spector
2	University of Leeds	Prof Peter Grant
3	Leiden University Medical Centre	Prof Frits Rosendaal
4	University of Helsinki	Prof Aarno Palotie
5	Queen's University Belfast	Prof Alun Evans
6	National Centre of Epidemiology Rome	Dr Maria Antonietta Stazi
7	Karolinska Institute	Prof Nancy Pederson
8	Hospital Santa Crui, Barcelona	Prof Jose Manuel Soria
9	Guy's and St Thomas' NHS Trust	Prof Tim Spector

Thrombotic stroke is a disabling condition - affecting an estimated 650,000 Europeans annually, with considerable mortality and costing over €30 billion/yr. Genetic factors account for a substantial component of the incidence & mortality of stroke. There is little effective therapy. EuroClot aims to identify and validate potentially therapeutically useful genes associated with thrombotic stroke using a novel approach. Stroke is a complex end-point disease involving the interaction of many pathologic processes, such as vessel wall atheroma, hypertension, platelet function & coagulation. EuroClot focuses on uncovering the genes that control the end-stage of the coagulation process that leads directly to the production of the thrombus (clot) that causes vascular obstruction and tissue death. Clinical studies indicate that alterations in fibrin structure and/or function create a prothrombotic phenotype, which increases vascular risk. Twin studies have shown a substantial genetic component to levels of activation peptides and the final common pathway of thrombus (fibrin structure/function).

Activation of the coagulation cascade results in thrombin generation and cleavage of fibrinogen to form fibrin monomers, which polymerise and are cross-linked by thrombin-activated factor XIII, to form the structural scaffold for thrombus formation. Fibrinolysis of fibrin clots is influenced by various factors including the fibrin binding characteristics of tPA/plasminogen, local

concentrations of fibrinolysis inhibitors and the structure of the clot. For example, studies *in vitro* have shown that dense clots composed of thinner fibres lyse more slowly than less dense clots formed from thicker fibres. Alterations in clot structure/function, including increased clot density and decreased clot lysis times, have been observed in subjects with arterial and venous thrombosis. We have shown that genetic factors contribute to variance in turbidimetric measures of clot structure/function and numerous studies have demonstrated genetic influences on proteins involved in coagulation and fibrinolysis. Furthermore, genetic factors have been estimated to account for approximately 60% of the risk of thrombosis[8]. Consequently, the identification of genetic loci influencing clot structure/function may further our understanding of the underlying factors predisposing to occlusive vascular diseases.

The aim of Euroclot was to identify the major genes involved in variations of the end-stage clotting process and investigate the role of these novel genes (and existing candidate genes) in the pathogenesis of stroke across Europe. EuroClot would study stroke intermediate phenotypes in over 3000 twins from GenomEUtwin project involving 8 countries and 700 subjects from extended families from the GAIT2 (Spain) and EuroHead (Finland) studies. Genes would be validated in 1000 stroke cases including those from the large European prospective MORGAM study. Cross-European differences in allelic frequencies would be examined along with their relative impacts. Phenotyping would be standardised and harmonised and a European database established. Progress of the project overall has been satisfactory with all of the deliverables have been met.

Workpackage 1 involved the collection and redistribution of blood samples and patient data for the project. A standardized approach to collecting plasma citrate and serum from 400 mono and dizygotic twins was the main focus during the first and second years of the project. Sample and data collection began in January 2005 for London and Helsinki, October 2005 for Stockholm and December 2005 for Rome. In parallel, similar samples have been collected from extended families from Leeds and Spain. There were considerable delays in sample collection at all locations, which resulted in late arrival of samples at the Leeds phenotyping centre. These delays were largely due to the bombings that took place in London.

The phenotyping work was performed within Workpackage 2. The following phenotypic data was recorded for each sample:

- d-dimer
- turbidity
- max absorbance
- lysis 1, 2 & 3

Biochemical risk factors were also collected including:

- lipids
- cholesterol and triglycerides
- glucose

- insulin
- CRP

Finally, demographic data was also collected:

- Sex
- Age
- Height
- Weight
- Blood pressure
- Smoking history
- Alcohol
- Exercise
- Drug history

D-dimer is a fibrin degradation product, a small protein fragment present in the blood after a blood clot is degraded by fibrinolysis. D-dimer levels were determined according to the manufacturers' instructions using TintElize(r) D-dimer ELISA kit, (Biopool, Umea, Sweden). Clot formation, morphology and lysis were assessed ex-vivo using high throughput turbidimetric assays. The following variables were analysed: $Lag_{C,}$ (which represents the time at which sufficient protofibrils have formed to enable lateral aggregation, was taken as the time point at which an exponential increase in absorbance occurred) $MaxAbs_{C,}$ (a measure of clot density reflected by the absorbance at which 3 consecutive readings were identical corrected for the Lag_C absorbance) $Lys50_{t0}$ (calculated as the time from initiation of clot formation to the time at which a 50% fall in absorbance from $MaxAbs_L$ occurred) and AUC (area under the curve, reflecting the balance between coagulation and fibrinolysis).

The EuroCLOT consortium examined D-dimer and turbidimetric measures of fibrin clot structure/function to determine the heritability of these phenotypes in the Leeds Family Study before going on to confirm heritability estimates and to identify linkage to particular regions of the genome using genome-wide markers in twins. In the Leeds Family Study the turbidimetric clotting and lysis variables were moderately to strongly correlated. Lag times, maximum absorbancies and clot rates measured in the presence and absence of tPA showed correlations of >0.75. The lysis time variables and lysis rate were highly correlated; Lys_T was the variable most sensitive to variance in PAI-1 and least sensitive to variance in fibrinogen, however Lys_T also displayed the highest inter-assay CV, consequently $Lys50_{t0}$ may be the most suitable lysis variable for analysis in clinical studies. Maximum absorbance was only modestly correlated with the lysis time variables, however AUC (area under the curve) was highly correlated with maximum absorbance and lysis times, lending support for AUC as a measure of the balance between clot formation and clot lysis. The phenotypes LR, $MaxAbs_C$, CR_C and CR_L are similar across the countries; Italy, Spain, Denmark, Finland, United Kingdom and Sweden. The lysis phenotypes ($Lys50_{t0}$, $Lys50_{Lag}$, $Lys50_{MA}$), Lag_C and Lag_L are larger for the northern countries (Finland and Sweden) when compared to the southern countries (Italy and Spain). The phenotype AUC is on average the same for Finland, Sweden, Spain and Denmark (340 – 349) while Italy has the lowest (179) and the United Kingdom has the highest (422). The

variable $MAXAbs_L$ is on average similar (about 0.3) for all countries except for Italy which has a lower average at 0.2. (Definitions as follows: Turbidimetric clotting assay variables: lag time (Lag_C) was taken as the time point at which an exponential increase in absorbance occurred; maximum absorbance ($MaxAbs_C$) was taken as the absorbance at which 3 consecutive readings were identical, corrected for the Lag_C absorbance; crude rate of clot formation (CR_C) was derived from time and absorbance values for Lag_C and $MaxAbs_C$. Turbidimetric lysis assay variables: lag time (Lag_L); maximum absorbance ($MaxAbs_L$), taken as the highest absorbance value, adjusted for Lag_L absorbance; crude rate of clot formation (CR_L); $Lys50_{t0}$, taken as the time from initiation of clot formation to the time at which a 50% fall in absorbance from $MaxAbs_L$ occurred; $Lys50_{MA}$, taken as the time from $MaxAbs_L$ to the time at which a 50% reduction in absorbance occurred; Lys_T, taken as the time from $MaxAbs_L$ to the time for the absorbance values to return to baseline; crude lysis rate (LR), derived from time and absorbance values for $MaxAbs_L$ and the point at which absorbance values returned to baseline; area under the curve (AUC)).

Genotyping took place within workpackage 3 and was completed using several different methodologies applied to data held by centres in London and Denmark. Both these centres obtained linkage genotyping on their twin cohorts. Using bioinformatics the two data sets were combined to provide a single high-powered linkage analysis of these novel phenotypes that pertain to coagulation.

Genome-wide association scans (GWAS) were used to identify novel SNPs associated with plasma coagulation factors. These were combined with other SNPs and tested for replication in the Leeds Family Study. Originally, several genotyping methods were under consideration. A decision was made to use the Illumina 317K chip because this contains tagSNPs, which offer more cost effective coverage of the haplotypes in the genome than the other chips under consideration. GWAS using the Illumina 317K chip were available on a number of the UK twins. In order to optimize chances of publishing any of our findings it was decided to split the twin sample into 2 groups, one for discovery and one for replication. In the discovery phase GWAS were examined for association in both the turbidity and lysis phenotypes and in plasma coagulation factors. The top 50 SNPs identified in each of these two groups were then replicated in twins and in the Leeds Family Study. From this the most significant 20 SNPs were taken through into clinical studies in the MORGAM stroke cases and controls, and in the Leeds Stroke cohort.

For the UK sample, genotyping was performed using 2231 polymorphic genetic markers - 737 microsatellite markers from the ABI Prism set (Applied Biosystems, Foster City, CA) and 1494 SNP markers from the HuSNP GeneChip linkage mapping set (Affymetrix Inc. Santa Clara, CA), as described previously. The estimated genotyping error rate was <1%. Allele frequencies were estimated from the whole sample of genotyped subjects. The map positions were taken from Rutgers combined linkage physical map (MAP-O-

MAT). The genetic locations of markers not on the Rutgers maps were interpolated from their physical position. The Danish twins were genotyped as part of the GenomEUtwin consortium.

Complete data were available from the UK cohort on 1814 healthy Caucasian female twins (447 DZ pairs and 460 MZ pairs), and from the Danish cohort on 199 unselected DZ twin pairs.

Heritability (proportion of phenotypic variance attributable to additive genetic effects) was estimated using the UK twins as this cohort had fibrin-related intermediate phenotypes available for both MZ and DZ twins. Heritability estimates were in the range 17- 46% and indicate a significant contribution of additive genetic factors to variance in fibrin clot structure and function.

Combining UK and Danish twin cohorts for this linkage study of fibrin phenotypes revealed 6 chromosomal regions with significant linkage peaks (LOD scores >3.0) and a further 9 regions of suggestive linkage (LOD scores 2.0 – 3.0). The peaks having significant linkage did not harbour any identifiable obvious candidate genes that might influence fibrin phenotypes, but a number of possible candidates have been identified. The highest LOD scores were observed for Lag_c and AUC (4.57 and 3.51) on chromosomes 16 and 17 respectively. Whilst no clear candidates were observed under the peak related to Lag_c, a number of putative candidate genes for AUC under the wide 17q23.3 peak were identified.

The top 100 SNPs identified in the GWAS and were taken from two sources: coagulation and fibrinolysis factors which had been obtained previously and the turbidity and lysis plus d-dimer phenotyping from this project. The top 50 SNPs that were considered to represent true positive finds in each of the phenotyping exercises were then identified.

The 100 SNPs were examined in both a second group of twins and in the LFS and replication in either of these groups was obtained in 27 SNPs. These 27 SNPs were identified across phenotypes from measures of coagulation and fibrinolysis *in vivo* and measures of clot behaviour, or turbidity and lysis, *in vitro* . Twenty-six SNPs replicated in twins, with just a single SNP (rs 2144039 on chromosome 14) replicating in the LFS. These results indicate that a further selection of genes, which may be considered candidates for a role in the endstage of coagulation, were identified and these were taken through to examination in two clinical cohorts, the MORGAM study and the Leeds Stroke study.

The availablilty of the GWA data provided access to surrogates for investigation of geograpical origin from the Twins. The most easily identified markers are those for lactose intolerance. The relationship between these markers and the replicated loci for clot formation was explored further however there were no significant interactions. There were also no differences

between means of overlapping clotting factors between Denmark, UK and Spain.

Analysis of both the phenotype and genotype data was carried out within workpackage 4. Genotype and phenotype data were available on 447 UK DZ pairs and 199 DZ Danish pairs. Phenotype data (D-dimer, Lag_C, $MaxAbs_C$, $Lys50_{t0}$ and AUC) were analyzed using R. Data were transformed to optimize closeness to normality using a Box-Cox transformation using the box.cox.power function in the car package, to reduce the type I errors while preserving power for phenotypes having skewed distribution. Phenotype values falling outside the mean ± 3 standard deviations were considered outliers and excluded from the analysis.

When modelling for heritability estimates the phenotypic variance is assumed to be due to three latent factors namely additive polygenic effects (A), common environment (C) and specific individual effects and measurement error (E). Data were modelled using structural equation models implemented in Mx to obtain estimates for the parameters of the ACE model and its sub-models AE, CE and E. The full model (ACE) was compared to submodels AE, CE and E to find the best fitting model with comparison of the fit made using a chi-square difference test. This test evaluates the difference in the chi square and the difference in degrees of freedom of the full versus the sub-model. If this test statistic is not significant, the reduced model is accepted as the more parsimonious explanation of the data, otherwise the full model is retained. In all analyses, age was included as a covariate.

To evaluate jointly the UK and DK linkage results which partially overlap in their markers, multipoint identity-by-descent (IBD) probabilities were calculated on a 1 cM grid using Merlin for each sample separately, using their specific genetic map and allele frequencies. The IBD estimates were collected in a single IBD file and the two samples were pooled together in a joint linkage analysis using the variance component engine implemented in QTDT. Age, sex and country of origin were included as covariates. Approximate support intervals (SI) were generated using a -1 LOD approach.

The most easily identified markers are those of lactose intolerance which is much more common in Southern Europe than northern Europe. Mutations in the gene have been considered responsible for the evolutionary onset of milk drinking in Northern migrations. The relationship between these markers and the replicated loci for clot formation was explored – using the marker as a surrogate for genotyping across the North and South Europe countries. This comparison of means was done in order to see if there is a trend with regard to the North-South differences. The clotting factor used in the mean comparisons included Ddimer, Lag turb, maxabsturb, clotrateturb, lag_lys, maxabslys, clotratelys, Lysis1, Lysis2, Lysis3 and Lysis Area. For most of the phenotypes the means significantly differed between the three countries. There is no clear trend with regard to whether the North has consistently

lower or higher means on the phenotypes. The pattern seems to be random no major geographical variations.

The lactase gene was used to test SNPs for lactose intolerance and stratify gene frequencies of the replicated loci. SNP rs1042712 on chromosome 2q21 from the LCT gene was identified. This SNP has alleles C and G. Assuming that the lactase SNP rs1042712 is recessive for the G allele, SN genotypes coded as 1 if genotype = GG and 0 if genotype was CG or CC were recoded. This renders a binary variable. Subsequently, whether this SNP is associated with the 14 SNPs that replicated in the genome-wide association analyses was tested. The results suggest that in general the replicated SNPs are not associated with the lactase SNP.

All phenotypes were log transformed prior to regression analysis. As can be seen in the table none of the phenotypes is significantly associated with the lactase SNP at a 1% level of significance.

The covariates for BMI, glucose and smoking were added to the regression analyses for the 14 replicated SNPs mentioned above to see if adjusting for covariates made a difference to the results. It appeared that for most of the analyses, the respective SNPs were not significantly associated with the clotting factors after adjusting for the covariates. However, it should be noted that the sample size drops dramatically (from around 1200 to 350) when covariates are included. In such a case it is not clear whether the fact that a SNP is not significantly associated with a particular phenotype is a real result or whether it is due to the small sample size. In other words, there wasinsufficient sample size to be able to identify clearly whether adding covariates to the regression analyses involving the clotting factors and the SNPs makes a difference.

Gene-environment interactions were explored using linear regression of the replicated loci for variations with the major environmental factors- smoking, BMI and fasting insulin and glucose levels and no clear signifciant interactions were found. This study was however not sufficiently powered to observe small to moderate gene- environment interactions, given the modest effect sizes of the genetic risk factors found in GWA studies and no major gene-environment interaction was detected.