EuroBATS

PROJECT FINAL REPORT

Grant Agreement number: 259749

Project acronym: EUROBATS

Project title: Identifying Biomarkers of ageing using whole transcriptome sequencing

Funding Scheme: FP7 HEALTH-2010

Period covered: from 01/01/2011 to 30/06/2014

Name of the scientific representative of the project's co-ordinator[1],

Title and Organisation: Professor Tim Spector, King's College London

Tel: +44 207 188 6735

Fax:

E-mail: tim.spector@kcl.ac.uk

Project website address: www.eurobats.eu

[1] Usually the contact person of the coordinator as specified in Art. 8.1. of the Grant Agreement.

Executive Summary

The "**EuroBATS**" (Identifying Biomarkers of Ageing using whole Transcriptome Sequencing) project aims to identify biomarkers of ageing and improve our understanding of the genetic mechanisms of ageing. EuroBATS undertook a multidisciplinary project combining cutting edge RNA sequencing technology and novel high throughput telomere measurement in multiple tissues. EuroBATS utilized both systems biology and genetic epidemiology approaches to explore this dataset **and identify biomarkers of ageing at both the tissue and global systemic level.** This project is exceptional in delivering all this data in the same deeply phenotyped 850 individuals; making these subjects the world's best phenotyped and genotyped group for further investigation of the role of biomarkers in cellular senescence.

The major work carried out in this project includes:

- Generation and analysis of whole transcriptome shotgun sequencing data via RNAseq from multiple tissues in ~ 800 individuals. With this dataset we identified greater than 10,000 regulatory variants, including those that effect total expression levels, differential splicing and allele specific expression. This represents the largest publically available multiple tissue RNA sequencing resource and is of outstanding value to the genomics community as a whole.

- Generation of a novel telomere length dataset consisting of both longitudinal samples from lymphocytes and a multi-tissue measurement of cross-sectional telomere length derived from skin biopsies, ultilizing two state of the art laboratory techniques.

- A comprehensive investigation of the genomics of ageing, including development and application of novel statistical methodology to identify interactions between age, genotype and expression. These methods allowed us to identify candidate genes which are likely to affect an individual's physiological age (biomarkers of ageing) in multiple tissues.

- Identification of organ-system level estimates of biological age and development and application of factor analysis methods to identify biological pathways implicated in systemic ageing. We found, however, little evidence of large numbers of systemic markers for every organ, concluding than age has system specific effects that require mode detailed studies.

- Study of the relationship between multilevel phenotypes to understand the mechanism of action of the biomarkers of ageing, providing for the first time knowledge of the changes in the relationship of different phenotypes with age.

Summary Description of project context and objectives

By 2050 the number of people in the EU aged 65+ will increase by 70% and the 80+ age group will increase by 170% in the same period. If healthy life expectancy evolves broadly in line with the change in age-specific life expectancy, then the projected increase in spending on healthcare due to ageing would be halved (The impact of ageing on public expenditure - DG ECFIN 2006, p. 133). A healthy, active ageing population can be supported through effective health policy across the lifecycle. Such a policy requires an understanding of the ageing process. The aim of this research is to define robust cellular markers of ageing including the identification of robust markers of cellular senescence and investigation of their role in ageing. This research further aims to characterize how the genome and transcriptome jointly interact with the aging process in order to identify genotype-specific ageing responses.

This ground-breaking project has for the first time investigated in great detail the transcriptome of a large cohort of extensively phenotyped twins for the study of ageing. This unique data set allowed us to derive robust markers of cellular senescence which can be correlated with ageing phenotypes to investigate ageing. We hypothesised that obtaining and analysing specific age related RNA sequencing data from skin, blood and fat will provide major insights into the ageing process in other biological systems. This allowed us to develop biomarkers of ageing that reflect generalised ageing; potentially identifying targets for anti-ageing interventions.

Telomeres are nucleoprotein structures capping and protecting the ends of chromosomes. Because of the “end-replication problem”, telomeres shorten with each cell division and leukocyte telomere length has been shown to decrease with age at a rate of 20-40 base pairs per year. Telomere attrition is enhanced by inflammation and oxidative stress and short telomere length has been associated to age-related diseases as well as to cellular senescence, the loss of a cell’s ability to proliferate. Ageing in humans is not a consistent process; this is due to both genetic heterogeneity and a variable environment. Biological age estimates the functional status of an individual in reference to his/her chronological peers and may help identify individuals at risk for age-related disorders, predict disability in later life and mortality independent of chronological age. In humans, studies are often limited by the necessity to measure telomeres in leukocytes, which is a far from ideal situation, and does not allow individual specific predictions in other cell types

We sought to address the lack of ageing biomarkers and improve our understanding of the genetic mechanisms of ageing. We have completed a multidisciplinary project combining cutting edge RNA sequencing technology and novel high throughput telomere measurement in multiple tissues. This unique data set has allowed to investigate the role of genomics in ageing. We have utilized both systems biology and genetic epidemiology approaches to explore this unique twin dataset. This has allowed us to identify a transcriptional signature of ageing that reflects generalised ageing as well as genotype-specific signatures of ageing; potentially identifying targets for targeted anti-ageing interventions.

The primary objectives of the EuroBATS overall project were to:
1. Obtain whole transcriptome shotgun sequencing data from skin, fat whole blood and lymphoblastoid cell lines (LCL) on 800 well phenotyped individuals.
2. Perform association analysis between sequence variation and RNA sequencing (eQTL) on the largest available multiple tissue RNAseq (WP1).
3. Obtain HT Q-FISH telomere length determination at 2 time points on snap frozen peripheral blood lymphocytes taken from the same individuals.
4. Generation of topographic telomere length maps on histological sections (telomaps) in a subset of 200 existing skin (epidermal keratinocytes, dermal fibroblasts and melanocytes) biopsy samples to further explore ageing-associated differential changes in the different cell types
5. Explore genetic factors (expression and genomic) that are correlated with both telomere length and change in telomere length over time
6. Develop variables for both skin ageing and global systemic ageing. Using a combination of factor analysis and principal component analysis; from extensive existing phenotypic data held at Department of Twin Research.
7. Develop genetic biomarkers of ageing in skin, fat whole blood and lymphoblastoid cell lines.
8. Ascertain those biomarkers of ageing common to all tissues. In the future common biomarkers from skin or blood maybe used for their predictive value in other less accessible tissues.
9. Generate the largest available multiple tissue RNA shared sequencing resource; allowing future comparison with data derived from diseased tissue.

Description of the main S&T results/foregrounds

WP1 - RNA sequencing

This workpackage sought to obtain and analyse whole transcriptome shotgun sequencing data via RNAseq from multiple tissues in ~ 800 individuals. The primary aims and achievements were the generation and quality control of the large sequencing dataset, the comparison of the sequencing data in comparison to previous microarray data and a comprehensive interrogation of the genetic regulation of the transcriptome data, including regulatory variants that effect total expression as well as splicing specific regulation. The RNAseq data and regulatory variants identified in this workpackage have been directly incorporated into analysis in subsequent workpackages in this Project. This work has generated the largest publically available multiple tissue RNA sequencing resource, which will allow future comparison with data derived from diseased tissue and is of outstanding value to the genomics community as a whole. All the data will be made available for other researchers via public repositories or the project web page (www.eurobats.org).

We present the significant results and details for each task within this package:

Task 1: RNAseq data generation

The project is using existing biopsies from skin, fat as well as blood samples (for generation of lymphocytes cell lines (LCLs) collected from a maximum of 856 twins (154 monozygotic twin pairs, 232 dizygotic twin pairs and 84 singletons) aged 38.7–84.6 years from the well-characterised TwinsUK Resource. The samples were

collected and made available by the MuTHER project (Multiple Tissue Human Expression Resource, http://www.muther.ac.uk/). Whole genome transcription profiles using microarray technology (illumina HT-12 array) has been already generated (Grundberg et al. 2012) and made available for validation of the whole transcriptome RNA sequencing produced by EuroBATS. ~400 Whole Blood samples from the same individuals were added to the project following suggestions arising from the midterm review.

We assayed LCLs, fat, whole blood and skin RNA samples. For each RNA sample, the mRNA fraction was selected and sequenced using IlluminaTrueSeq technology and a HiSeq2000 sequencer with 49 base paired-end reads. We obtained an average of 36.3 million reads per sample in fat, 43.6 million reads per sample in LCL and 33.7 million reads per sample in skin (Table 1). We excluded samples that failed in the library prep or sequence process. We also excluded samples with less than 10 million reads sequenced and mapped to the exons. Finally we excluded samples in which the sequence data did not correspond with the actual genotype data. We ended with 766 samples for fat, 814 for LCL, 716 for skin and 384 for whole blood.

Task 2: Analysis of raw sequence data

Read mapping and exon quantification

Reads were mapped to the reference human genome (GRCh37/hg19) using BWA v0.5.9 (allowing 2 mismatches in the first 32 bases). Reads were subsequently filtered to include only those which were called as properly paired, had a mapping quality score greater than or equal to 10 and overlap with known exons from GENCODE annotation (version 10). This yielded an average of 22.4 million reads per sample in fat, 22.4 million reads per sample in LCL and 19.3 million reads per sample in skin (Table 1). We looked at the distribution of counts per gene per sample and we observed that more than 12000 genes are very well covered with more than 100 reads per sample in all four tissues. Other 5000 genes are covered with between 10 and 100 reads per sample.

Allelic Specific Expression (ASE)

We took advantage of the digital expression profiles obtained by RNAseq to define genome-wide allele-specific expression profiles in all individuals and tissues. which will provide a comprehensive view of the transcriptome in multiple individuals and multiple tissues. We assessed statistically significant ASE sites using a binomial test. We did a test for each heterozygous SNP in every individual to detect the presence of statistically significant allelic imbalance. For each site-individual we counted the number of reads covering each allele and calculated a binomial test comparing the observed proportion of reference allele counts with the expected proportion. In theory, this expected proportion should be 0.5 but mapping bias can change it a little bit. To correct for systematic bias in allelic ratios we calculated the overall reference to total allele ratio for each individual for each SNP base combination. These ratios were then used as the expected ratios in the binomial test. We called significant ASE sites using a 10% FDR threshold and found 3136 genes with significant ASE in fat, 3956 in LCL and 3911 in skin.

Task 3: Genetics of gene expression

eQTL discovery

To look for ciseQTLs in the three tissues we used a linear regression approach with SNPs in a 1Mb window each side of the TSS for each gene. We identified 9166 significant ciseQTLs in fat, 9551 in LCLs, 8731 in skin and 5313 in whole blood (1% FDR) (Table 2, Figure 1).

Genotying and imputation. Samples were genotyped on a combination of the HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo 1M Illumina arrays. Samples were imputed into the 1000 Genomes Phase 1 reference panel (data freeze, 10/11/2010) using IMPUTE2 (Howie et al. 2009)and filtered (MAF<0.01, IMPUTE info value < 0.8).

Exon quantifications: All overlapping exons of a gene were merged into meta-exons with identifier of the form "geneID_start.pos_end.pos". We counted a read in a meta-exon if either its start or end coordinate overlapped a meta-exon.

Normalization: All read count quantifications were corrected for variation in sequencing depth between samples by normalizing the reads to the median number of well-mapped reads. We used only exons quantified in more than 90% of the individuals. We removed the effects of technical covariates regressing out the first 50 factors from PEER (Parts et al. 2011)including BMI and age in the model to preserve important biological sources of variation.

eQTL association: Since our data samples are twins, they are not independent observations and we needed to take that into account in our models. We used the two-steps strategy described by Aulchenko et al. (Aulchenko et al. 2007). First we kept the residuals of a mixed model that removed the effects of the family structure using the implementation in GenAbel R package. We then transformed those residuals using a rank normal transformation. Finally, we performed a linear regression of the transformed residuals on the SNPs in a 1Mb window around the transcription start site for each gene, using MatrixeQTL R package (Shabalin 2012). We did the association at the exon level and we kept the best association per gene.

Permutations: We permuted the quantifications of each exon 2000 times, keeping the best p-value per exon from each round. From these data, we adjusted the empirical FDR to 1% according to the most stringent exon of each gene, stratifying the analysis on the number of exons for a given gene.

Comparison with microarray results

We repeated the same analysis using the same individuals and the same SNPs but expression measures derived from microarrays (Grundberg et al. 2012) instead of RNAseq counts. At the same threshold of statistical significance (FDR 10%) we found 1593 ciseQTL in fat, 2363 in LCL and 1470 in skin. That means that RNAseq measures of expression allow a substantial increase in the discovery of ciseQTLcompared with microarrays in the three tissues. To measure the degree of replication of eQTL discovered with both sets of data we calculated the association in the microarray data for SNP-Gene pairs found significant at different pvalue thresholds in the RNAseq data and calculated Pi1 as an estimate of the proportion of true

positives (Storey and Tibshirani 2003). We found that as the significance threshold increases, the degree of replication increases, supporting the idea that we are getting similar results using both sets of data

eQTL browser

We added the eQTLs as custom tracks to the UCSC browser. In this way the users could see the eQTLs in the context of all the other annotation information shown in this widely used genomic browser. These tracks will be accessible to the general public once the main paper describing the eQTL is accepted for publication. The links to access the eQTL information in the browser will be in the paper itself and in the EuroBATS web page.

Variance components of gene expression

cis-eQTL are only a small part of the genetic effects that affect gene expression. By exploiting the twin structure of our sample, we dissected the proportions of gene expression variation which is due to different genetic and non-genetic causes. We observed that, on average, common cis-eQTL only explained about a 20% of the heritability of gene expression while other genetic variants in cis (mainly rare variants or common variants with small effects) explained about 30% of heritability. The remaining 50% of the heritability was explained by genetic variants in trans.

Alternative Splicing Analysis and asQTL discovery

Genetic variation may also affect gene expression by modifying mRNA splicing processes. HalitOngen in our lab has developed a novel method for the relative quantification of splicing events (Ongen, 2014, *under review*). The method uses the paired-end nature of the RNA-seq experiment. When one read maps to one exon and its mate to a different exon, we count a "link" between two exons. For a given exon, we calculate the fraction of links that forms with every other exon respect to the total.

We used these exon-exon link fractions as our phenotype to measure alternative splicing and calculated the association with cis SNPs following the same pipeline described for eQTL discovery. We identified 2481 asQTL in fat, 4102 in LCL and 1566 in skin.

Genetic architecture of allele-specific expression (ASE)

ASE may be caused by genetic or epigenetic / environmental factors. To measure the relative contribution of the underlying causes of allelic expression we estimated the variance components of the ASE ratios using the identity-by-descended status (IBD) of the twin pairs at the ASE site and the identity-by-state status (IBS) at the best eQTL. We found that about 40% of the variance in ASE is due to the effect of the best eQTL , 17% to the additive effect of the other genetic variants in cis, 23% to the interaction between cis and trans variants and 20% to the individual environment. The additive trans and the shared environmental effects were negligible. There were small differences among tissues. The sum of all the genetic effects gives an average heritability estimate of 80%. Our results show a complex genetic architecture for allelic expression that identifies GxG and putative GxE effects. We utilized the twin structure of our sample to look for examples of GxE interactions. Since MZ twins are genetically identical, differences in allelic expression in a MZ pair are determined by non genetic

effects. For every site, we calculated the association between allelic expression differences within MZ pairs and SNPs around the site and found examples of potential GxE interactions. One example in fat tissue was found for ADIPOQ, a gene that codifies for adiponectin, whose expression has been observed to be affected by environmental factors such as diet and physical exercise.

In summary, we propose a model that best fits the data is one where ASE requires genetic variability in cis, a difference in the sequence of both alleles, but where the magnitude of the ASE effect depends on trans genetic and environmental factors that interact with the cis genetic variants.

Tables

Table1. Average number of reads per sample and per tissue

	Total reads	Exonic and good quality reads
FAT	36,343,383	22,433,747 (61%)
LCL	43,637,726	22,368,953 (51%)
SKIN	33,721,164	19,273,820 (57%)
BLOOD	30,829,602	15,877,864(51%)

Table 2. Number of regulatory variants identified per tissue

	CiseQTLs	Splice QTLs	ASE
FAT	9166	2481	3136
LCL	9551	4102	3956
SKIN	8731	1566	3911
BLOOD	5313	*	4379

* the calculation of the splicing eQTLs is in progress

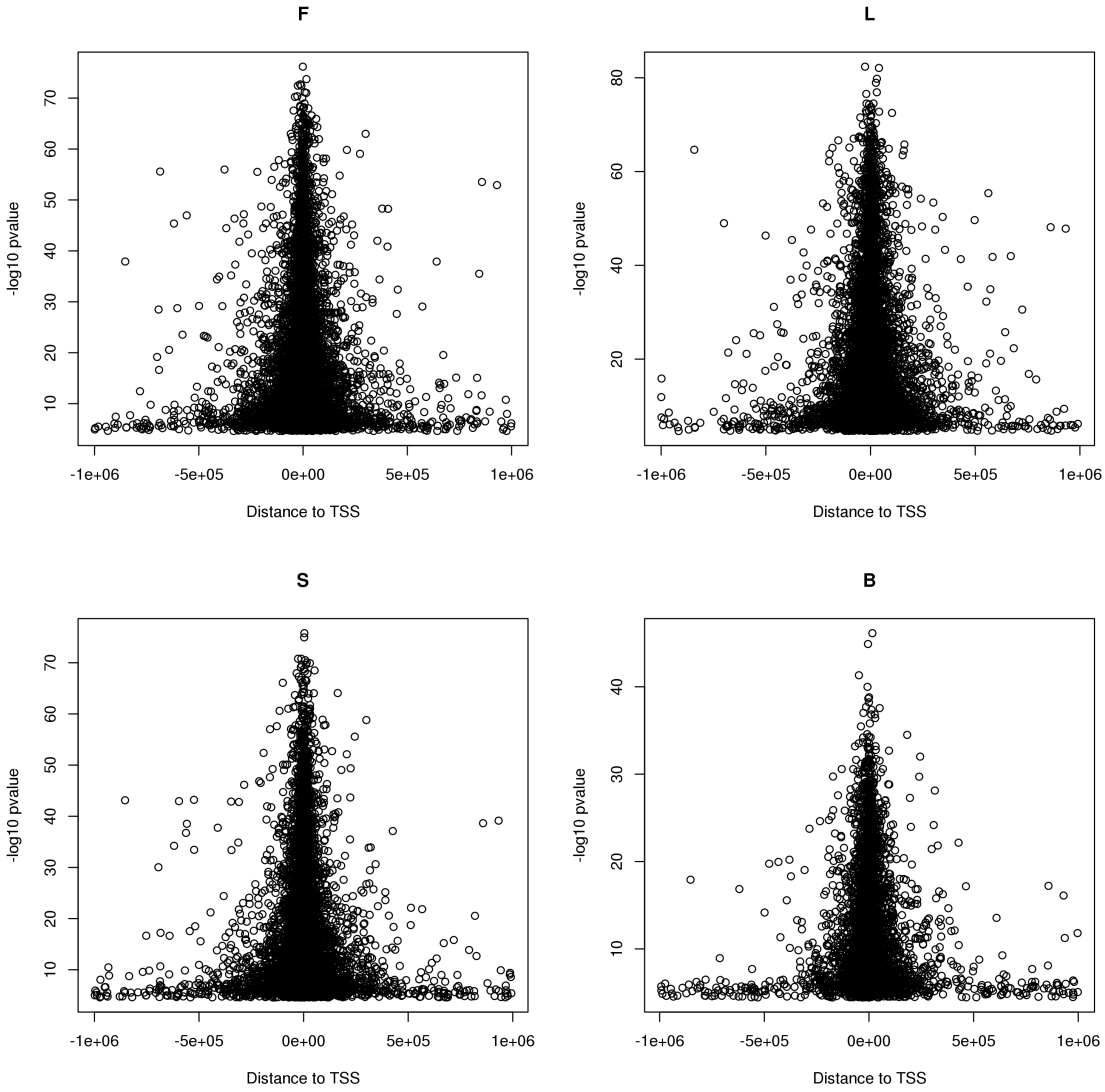

Figure 1. eQTLs related to its distance to the TSS of the gene (F for fat, L for LCL, S for skin and B for blood).

WP2 - HT Q-FISH telomere length measurement and histological telomapping

This workpackage obtained a novel telomere length dataset consisting of both longitudinal samples from lymphocytes and a multi-tissue measurement of cross-sectional telomere length derived from skin biopsies. This work package employed two state of the art laboratory techniques (HT Q-Fish and Telomapping) and generated a unique dataset with which to investigate the role of cellular senescence in aging, and its interplay with tissue specificity.

HT Q-FISH measurement of telomere length in lymphocyte

Overview: Life Length´s participation in this project was focused on measuring telomere length of peripheral blood mononuclear cells (PBMC) by its proprietary TAT technique based on Q-FISH. The analyzed samples belong to a collection of snap frozen PBMCs taken at two time points (1999 and 2009) of healthy UK twins. A total of 724 samples were analyzed in 2013. Five replicates were assessed per sample in a single plate. Other 824 samples were analyzed in 2014. The same number of replicates per sample were included for this second set but each sample was plated in two independent plates (hence, there was a total of 10 replicates per sample).

To measure median telomere length in peripheral blood lymphocytes, Life Length used a high-throughput (HT) Q-FISH technique. This method is based on a quantitative fluorescence in situ hybridization method modified for cells in interphase (Canela et al, ProcNatlAcadSci U S A. 2007 Mar 27;104(13):5300-5). In brief, telomeres are hybridized with a fluorescent Peptide Nucleic Acid probe (PNA) that recognizes three telomere repeats (sequence: Alexa488-OO-CCCTAACCCTAACCCTAA, Panagene). Images of nuclei and telomeres are captured by a high-content screen system (see below). The intensity of the fluorescent signal from telomeric PNA probes that hybridize to a given telomere is linearly proportional to the length of the telomere. Intensities of fluorescence are translated to telomere lengths, by comparing the obtained intensities of fluorescence versus a standard regression curve built with control cell lines of known telomere length.

Control cell lines and Southern blot: Life Length's control cell lines C0126, C0154, C0106 are immortalized human B cells purchased from European Collection of Cell Culture (ECACC). Lymphoblastoidtumoral cell lines REH and RAJI were purchased from ATCC (CRL-8286, CCL-86). Cellular stocks were prepared and kept in liquid nitrogen. Telomere length of these cell lines was determined by a non-radioactive TRF (Southern blot) assay following a protocol as described in Kimura et al. (Kimura et al., Nat Protoc. 2010 Sep;5(9):1596-607.)

Sample Preparation for HT Q-FISH: On processing day, samples and control cell lines were thawed at 37ºC and cell counts and viability were determined. Aliquots with viability lower than 80% were invalidated as well as samples contaminated with erythrocytes. Cells were seeded in a clear bottom black-walled 384 well plates at the density of 25.000 cells/ well with 5 replicates of each PBMC sample and 8 replicates of each control cell line. Two identical independent plates were prepared for each set of samples. Cells were fixed with methanol/acetic acid (3/1, vol/vol). On the next day, fixed cells were treated with pepsin to digest cytoplasm and nuclei were processed for *hybridization in situ* with the PNA probe. After few washing steps adding DAPI for DNA staining, the plate was filled up with mounting medium and kept overnight at 4ºC.

HT Microscopy: Quantitative image acquisition and analysis were performed on a High Content Screening Opera System (Perkin Elmer), using the Acapella software, Version 1.8 (Perkin Elmer). Images were captured, using a 40x 0.95 NA water immersion objective. UV and 488 nm.excitation wavelengths were used to respectively detect the DAPI and A488 signals. With constant exposure settings, 15 independent images were

captured at different positions for each well. After image acquisition, the nuclei image was used to define a region of interest for each cell measuring telomere fluorescence intensity in the A488 image in all of them. Results of intensity for each foci identified were exported from the Acapella software (Perkin Elmer). The telomere length distribution and median telomere length were calculated with Life Length´s proprietary program.

A total number of 1,548 blood samples were processed, of which 1,211 passed quality control checks. Of those samples, 742 correspond to repeated extractions from the same individuals (two time points), which allow longitudinal study of telomere attrition. Figure 2 show the difference in year between samples. The long bar on the left indicate that approximately half of the individuals of the study had only one time point available.

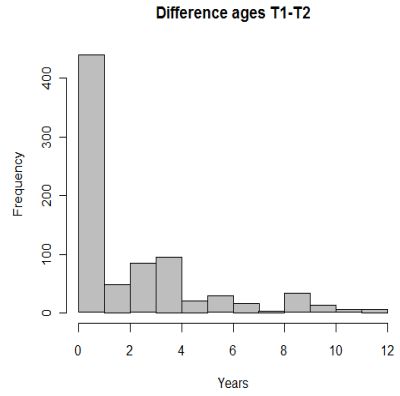

Figure 2.Difference in year between longitudinal HT –QFISH samples

Telomapping of histological sections of skin biopsies taken from a subset of 200 twins.

For the analysis of telomere length of cells in a tissue section Beneficiary No. 4 has developed telomapping, a method for the generation of topological maps of telomere length in which confocal telomere Q-FISH is performed directly in tissue sections coupled to a single-cell high-throughput image analysis platform Telomapping of human healthy skin sections will allow monitoring of changes in telomere length in the different cell types of the ageing skin and will advance the understanding of both the human skin ageing process and of the functionality of the ageing biomarkers uncovered in the project.

For Q-FISH analysis on skin tissue samples, OCT sections were hybridized with a PNA-tel Cy3-labeled probe, and telomere length was determined as described (Zijlmans et al. 1997;Gonzalez-Suarez et al. 2000; Samper et al. 2000; Muñoz et al. 2005, Flores et al. 2008). DAPI and Cy3 signals were acquired simultaneously into separate channels using a confocal ultraspectral microscope Leica TCS-SP5 and maximum projections from image stacks were generated for image quantification.

For image acquisition we used a new tool for intelligent screening named “intelligent matrix screening remote control (iMSRC)” developed at CNIO. iMSRC application manages a first fast scan with low-resolution settings, generating one image per sample of the whole tissue and later localizes the areas of interest, extracting their coordinates and surface area. With the spatial information, the iMSRC application

interacts with the microscope and load high-resolution settings, scanning automatically just the areas of interest.

Quantitative image analysis of telomere fluorescence intensity was performed on confocal images using the Definiens Developer Cell software (Definiens Developer XD). The DAPI image was used to define the nuclear areas that were separated by a Cellenger-Solution. After defining the nuclear areas a predefined Ruleset was used for the quantification of telomere fluorescence intensity (Cy3 image).

Fluorescence intensities were measure together with L5178Y-S cells as calibration standards. Using the controls, telomere measurements were converted into kb. Telomere fluorescence values were normalised by dividing mean telomere fluorescence intensities of all nucleus per sample by the corresponding mean telomere fluorescence intensity of the control sampleA total of 166 samples were processed for telomapping measurement. Of those, 133 passed all quality controls and were suitable for further analysis. The data were analysed as part of the WP4, and are fully described in that section.

WP3 - Ageing Phenotype Refinement

Phenotypic data has been cleaned to develop variables for both tissue specific ageing and global systemic ageing from data held at DTR. The phenotype data has been analysed and modelled to report measured and inferred quantitative ageing traits for use in WP4.

Ageing-related measurements from the Twins UK cohort were selected which covered a number of different systems in the body. Preference was given to phenotypes that had at least two repeated measures per individual, but categorical data, such as age of menarche or diseases prevalence, which does not change over time but affects ageing, has also been collected in order to establish a high quality replicable dataset of ageing related phenotypes.
The work here reported involved collection and cleaning of data from self-reported and clinically measured phenotypes. Beneficiary 1 has been collecting an extensive set of high quality data since 1992; however, over this considerable period of time inevitably some data quality issues and missing information have been reported. To guarantee the highest possible quality of the data, quality checks (QC) process were established to document phenotypes and also to identify and report any potential issues within each selected phenotype (e.g.: coding change, typos, questionnaires changes, etc). The data validation process based on data issues reported during the QC process included retrieval and confirmation of original values recorded on paper or electronic format when necessary.

Longitudinal data cleaning of Twins UK cohort

The EuroBATSproject is using existing biopsies from skin, fat as well as blood samples (for generation of LCLs) collected from a maximum of 856 twins aged 38.7–84.6 years from the TwinsUKcohort. Initially, objectives for the WP involved collection of phenotypic information for these 856 individuals with available

biopsies. The list of phenotypes here reported differs slightly from the list of selected phenotypes on the grant agreement due to the limitations on the data from the subgroup individuals with available biopsies. Phenotypes were selected according to availability of two or more values at different time-points for a majority of individuals with biopsy. However, maximising the sample size is a critical factor in producing accurate models for phenotypic change with both chronological and biological age.

For this reason, all the available individuals ever recorded in Twins UK (>8000) were included in the cleaning and QC process for each selected phenotype. Moreover, identifying and cleaning phenotypic data, especially longitudinally, is not a trivial task and must be undertaken meticulously to ensure reproducibility and accuracy. Therefore, we chose to focus our efforts on reduced number of longitudinal phenotypes (Table 3) and confounding variables (Table 4), while establishing collaboration with other researchers working with Beneficiary 1 to increase the set of clean longitudinal data. The phenotypes, currently available for all beneficiaries,have been utilized for the different projects within WP4 to develop better understanding of physiological function changes with age and the ageing process.

Projects in WP4 include association with fat-related phenotypes (e.g.: lipids and glycamina); gene expression affected by age and association with age-related phenotypes; or estimation of biological age and association with inferred phenotypes (see WP4 report for further details).

System	Phenotypes	Nr. Individuals 1 data-point	Nr. Individuals 2 data-point	Max. number of repeated measurements
Respiratory	FEV1/FVC	7120	7005	5
Anthropometrics	BMI	8006	8003	4
	Weight	8003	8003	12
Cardiovascular	ECG	5535	1862	3
Renal	Creatinine	6937	1929	3

Table 3: Set of ageing-related selected phenotypes with longitudinal information list of cofounding variables

Confounding variables	Number individuals
Smoking	10117
COPD	8127
Asthma	7710
Asthma medication	1117

Menopause status	11630

Table 4: List of confounding variables and the number of individuals with available data.

Phenotype data from EuroBATS individuals

Beneficiaries 3 and 4 successfully generated high resolution molecular phenotypes (RNA sequencing and telomere length measurements) from 856 individuals with the available biopsies and blood samples (WP1, WP2). Projects in WP4 require high quality phenotypic information from those individuals. From the longitudinal data cleaning of TwinsUK cohort previously described, we have extracted the corresponding phenotypic information during the longitudinal data cleaning of Twins UK (Table 5). Also, we have extracted cross sectional phenotypic information from 40 phenotypes covering multiples organ systems at the closest available date to biopsy extraction. This valuable information allows complex association studies between genome, transcriptome and age-related phenotypes in a multilevel approach.

System	Phenotype	Nr individuals 1 measurement	Nr individuals 2 measurements	Nr individuals 3 measurements
Respiratory	FEV1	850	739	438
Anthropometrics	BMI	857	701	324
	Weight	857	701	324
Cardiovascular	ECG	817	573	225
Renal	Creatinine	787	342	90

Table 5: List of longitudinal phenotypes data available for EuroBATS individuals and number of individuals with repeated measurements per phenotype.

Inferred ageing variables

Phenotypes data have been analyzed to measure and define indicators of their strength of association with age. As an example, we present here results from the descriptive analyses of the respiratory system. Pulmonary function is often measured and assessed with two correlated measurements: force vital capacity (FVC) and forced respiratory volume in 1 second (FEV1). Our current data set for these measurements include approximately 7,000 individuals with up to five repeated measurements over the last 20 years. The age of the individuals range from 15 to 85 years old (Figure 3, left). Lung function is known to increase during childhood until 20-25 years of age, when an age-related decline on lung function affects the pulmonary function (Kerstjens et al. 1997).This trend can be significantly affected by environmental variables such as smoking or physical

activity (exercise), or respiratory diseases such as asthma or chronic obstructive pulmonary disease (COPD). By considering all available data 2273 individuals, out of 7005, can be identified as “lung healthy” (never reporting from asthma or COPD) and people who have never smoked. These individuals may be considered healthy control cases for further analysis of genetic and environmental factors affecting lung function. The remaining individuals (4621) either reported to be asthmatic, diagnosed with COPD or current or ex-smokers and were therefore considered to have compromised lung functions. Our current work aims to determine the degree of impact on lung function by different covariates and their influence on the decline of the phenotypes with age.

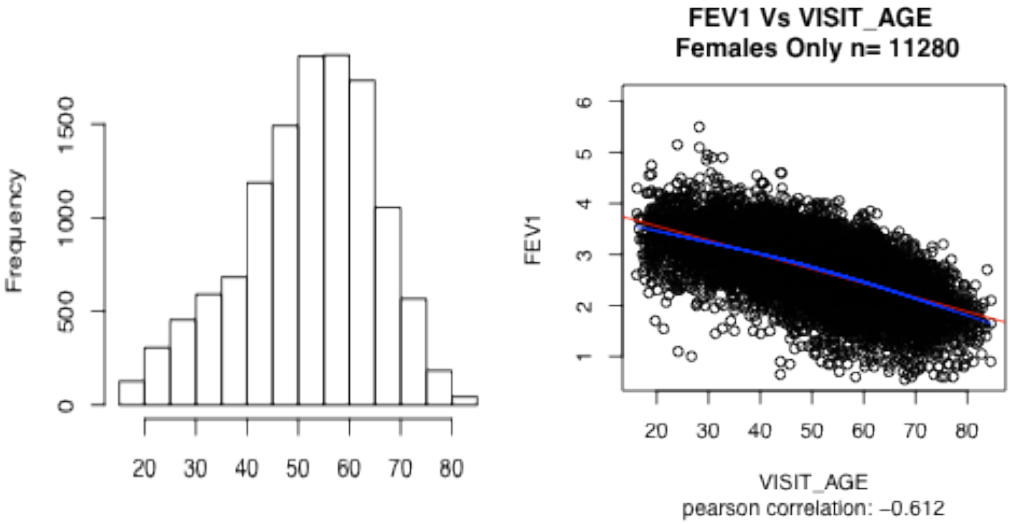

Figure 3: Descriptive plot for pulmonary function. The left-hand figure shows the age distribution of the twins with recorded measurements for FEV1. The right shows a scatter plot of over 11,000 measurements for FEV1 over 20 years and its strong negative Pearson correlation with age (r=-0.612).

In conclusion, we have produced a set of age-related phenotypes from the Twins UK cohort and a collection of confounders/phenotypes known to affect them. The phenotypes cleaned were available for all beneficiaries involved in WP4 and other collaborators for further studies. Cross sectional phenotypes have been utilized to develop measurements of biological age (see projects WP4) and estimations of physiological function changes with age to be used in combination with other molecular phenotypes (WP1 and WP2) to infer a causative model of ageing. The phenotypes have been also tested for specific association with genes expression provided by WP2 as well as association with genetics markers to better understand the genetic control of age related phenotypes.

WP4 - Analysis

The aim of Work Package 4 was to identify markers of cellular senescence to investigate the role these markers had in ageing and to develop descriptive models of ageing by integrating genomic (WP1), telomeric (WP2) or phenotypic data (WP3) data. In addition we used the analyses in this workpackage to explore the validity of various hypotheses of mechanisms of ageing

We have employed several strategies to meet this aim, which are detailed below. Overall we have 1) Completed a comprehensive investigation of the Genomics of Ageing, including development and application of novel statistical methodology to identify interactions between age, genotype and expression, 2) Used factor analysis methods to derive novel summary phenotypes to identify biological pathways implicated in systemic ageing; 3) Used models of biological age at the level of the organ system to inform risk of disease and co-morbidity; 4) integrate multiple omics phenotypes and different phenotypes association in the identification of biomarkers of ageing.

Genomics of Ageing

This project represents the largest genome-wide analysis of gene expression with age in humans and the first study utilising RNAseq data from the same individuals in multiple tissues. We have analyzed differential expression at multiple levels with age to identify candidate genes which are likely to affect an individual's physiological age (biomarkers of ageing). In addition, the regulation of gene expression also varies as the organism gets older. Previous studies have suggested that the regulation of gene expression tends to decrease with age, leading to more (stochastic) variation in gene expression. Therefore, regulation of gene expression is assumed to decrease with age due to an increase in random loss of fidelity in molecules interactions. To investigate this hypothesis, we dissected the causes of variation in gene expression with age to describe global regulatory changes in gene expression which occur as an individual gets older. In parallel to the RNA-seq analysis (**WP1**), the causes of variation were dissected by investigating the age influence on eQTL and heritability of expression.

In this way, we investigated genotype and age influences in gene expression simultaneously in multiple tissues (the first time such a study has been carried out using human samples). Moreover, we have investigated the change in gene expression regulation by analysing the generation of splice variants. mRNA splicing allows a single gene to code for more than one transcript (isoform) which may have a different function in a highly regulated process. The production of aberrant splice products has been linked to the etiology of several diseases. Genes containing splice variants which are sensitive to ageing are likely candidate genes to increase susceptibility to age related diseases.

Ageing is known to affect expression of genes, but tissue specificity of age-related sources of variation in expression is largely unknown. Using the Eurobats RNAseq dataset generated in **WP1**, we first identified genes with an age-related component to expression. We found that 34% genes in all tested tissues changed in expression with age with age (Figure 4A). Of those 5,224 genes affected by age, 8.3% were significant

in two tissues with only 5 genes in common among the three primary tissues. However, *P*value enrichment analysis comparing the values per tissues indicated shared age related effects from 21% to 60% between primary tissues. In addition, we found that 59.7% and 32% of genes with multiple exons have also signs of age effect in splicing, including genes associated with age-related diseases like *APOE, LMNA, SIRT2, AKT1* and *AKT2* (Figure 4D). With the increased power of RNAseq and our results, we conclude that ageing effects on gene expression are to be less tissue specific than implied by microarrays results (Glass et al. 2013).

Increased variation of gene expression during aging is often assumed to be the result of decreased gene expression regulation and assumed to play a role in the ageing process. We aimed to identify genes in which age had an effect in their variance, rather than in the mean effects, as these genes would be markers for change in gene regulation with age. We used the established framework for identifying variance-eQTL and developed by beneficiary 2 (WT) (Brown et al. 2014). A Spearman correlation test identified evidences for an age effect on variance in gene expression in all tissues and found 3,112,19 and one genes for which the variance in gene expression changed with age in Fat, Skin, Blood and LCLs respectively (Figure 4B) To our surprise,a majority of significant genes showed a decrease in variance with age.

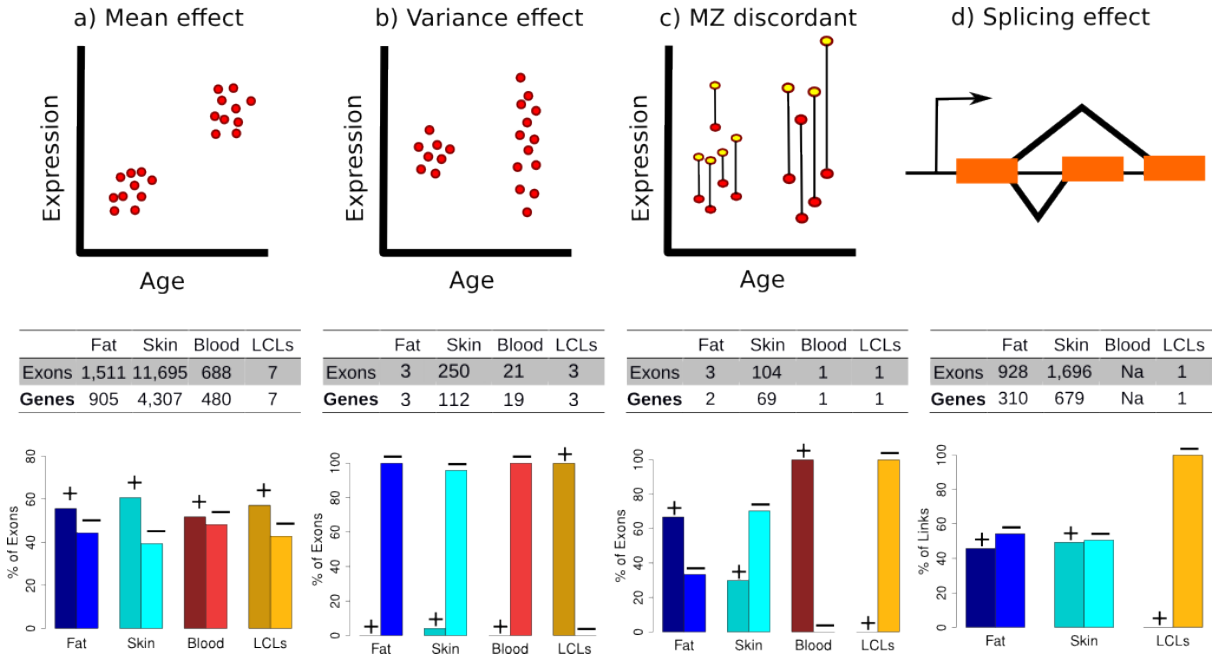

Figure 4: Schematics of the possible effects of ageing in gene expression a) mean, b) variance, c) MZ discordance and d) splicing. Bellow each diagram we show the actual number of exons and genes with a significant effect in a genome wide analysis for age effects. The last row shows the percentage of exons with positive (+) or negative (-) age effect in expression, variance, discordance and splicing. Discordant analysis only

compared monozygous twins. For the splicing analysis, only links (reads between two exons) were considered.

To further understand whether these changes in gene expression were due to gene by environment interactions (GxE) or environmental factors we took advantage of the unique genetic characteristics of twins to investigate changes in discordance of expression within monozygous twin (MZ) pairs with age. We found 2 genes in adipose, 69 in skin, and 1 in blood and LCLs where discordance was age-dependent (Figure4C). As MZ twins are genetically identical, age-dependent differences must be due to a changing environmental component. Due to the genetic relationship between the individuals in the dataset, we where also able to perform a decomposition of variance affecting gene expression.

Development of novel statistical methodology to identify gene x age interactions on gene expression as biomarkers of ageing

We showed that age related genes had a larger genetic component, and that the sources of variation were highly tissue specific. While this could be due to increased stochasticity, it is plausible that some of this effect is due to the eQTL being modified as the individual ages. To identify such SNPs whose effects on gene expression changed with age we performed a genome-wide scan looking for age-genotype interactions (GxA) using the framework developed by beneficiary UNIGE for **WP1**. We hypothesise that genetic variants interacting with age would be relevant to explain the progression and onset of age related diseases. One gene, CD82, was genome-wide significant in fat, showing a concrete example of how genetic control of expression is modified over time. Interestingly this gene, showed increased expression with increasing age in individuals with a particular, potentially protective, allele (Figure 5).The gene is a metastasis suppressor so could have an important role in age-related cancers, suggesting that the identification of GxA may be a good approach to identify relevant age related variants.

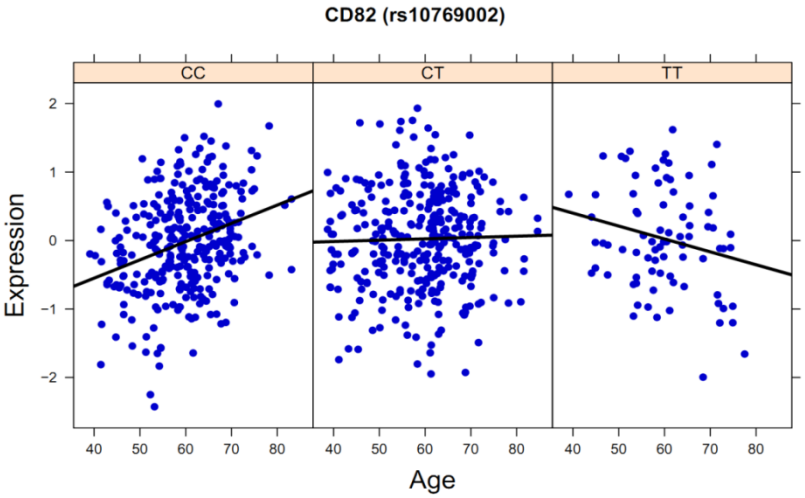

Figure 5 – Genotype x Age interaction on expression levels of CD82

Furthermore, in published work (Brown et al. 2014), we have looked for the presence of non-additive interactions between genetic variants, or epistasis, affecting gene expression. This is a possible explanation for the gap between heritability of complex traits and the variation explained by identified genetic loci. Filling this gap may help to identify the underlaying causes of age-related diseases. Therefore, interactions give

rise to genotype dependent variance, and the identification of variance quantitative trait loci can be an intermediate step to uncovering epistasis. Using RNA-sequence data from lymphoblastoid cell lines (LCLs) from **WP1**, we identify a candidate set of 508 variance associated SNPs (variance-eQTL). Further investigation of these loci reveals 57 epistatic interactions that replicated in a smaller dataset, explaining on average 4.3% of phenotypic variance. In 24 cases, more variance is explained by the interaction than their additive contributions. Using molecular phenotypes in this way may provide a route to uncovering genetic interactions underlying more complex traits. This study has also demonstrated that gene expression has sufficient power to discover aspects of genetic architecture which would require sample sizes that are orders of magnitude larger with other complex traits. This has motivated further work in which we have looked for cases in which environments accumulated during the lifespan of the individuals can modulate the effects of genetic variants, in particular when we see that the age of an individual can change the effects of DNA.

Such cases are known as gene-environment interactions (GxE). Because GxE also affects trait variance in a genotype dependent fashion, we used the strategy of prioritising SNPs that associated with variance in expression (v-eQTLs) when looking for interactions. Similarly, SNPs associated with a change in discordance of expression between monozygotic twin pairs (d-eQTLs), would indicate presence of GxE. We found 1198 v-eQTLs in LCLs, 620 in fat, 368 in skin and 39 in blood, and 73, 211, 63 and 1 d-eQTLs in those tissues. Functional overlap analysis using ENCODE data revealed that in LCLs v-eQTLs were significantly depleted in transcriptionally repressed regions (odds ratio, 0.82) and enriched in enhancers (OR 1.71); d-eQTLs were enriched in promoters (OR 5.29). Skin d-eQTLs were enriched in H3K36me3 regions (OR 4.02), a mark of active transcription.

To find environments involved in GxE signatures, we tested all v- and d-eQTLs for interactions with age, BMI and 20 expression principal components (PCs), having previously shown that the PCs can be highly heritable. We observed 4 interactions with age affecting expression of *HLA-DRB5* in LCLs, *COX20* in blood and *SLFNL1* and *ARID4B* in skin. There were three Bonferroni significant interactions between genotype and BMI observed in fat expression ($p<1.94e-5$). We saw large numbers of interactions with PCs: 2 in blood, 10 in fat, 39 in skin and 66 in LCLs ($p<9.70e-7$). Analysis of separate dermis and epidermis data suggested that some skin d-eQTLs are cell specific eQTLs. In summary, we detect widespread variance effects in gene expression and observe that d-eQTLs consistently have more success at mapping GxE with phenotypes, PCs and tissue composition measures.

In summary, we have produced a comprehensive description of how aging affects expression and its genetic control, observing that these effects are frequently tissue specific. Genes commonly affected by age in multiple tissues, are strong candidates for systemic biomarkers of ageing.

Factor analysis methods to derive novel summary phenotypes to identify biological pathways implicated in systemic ageing.

Statistical factor analysis methods have previously been used to remove noise components from high dimensional data prior to genetic association mapping, and in a

guided fashion to summarise biologically relevant sources of variation. We demonstrated how the derived factors summarize pathway expression can be used to analyse the relationships between expression, heritability and ageing, and to derive new biomarkers of ageing. We used the skin gene expression data from TwinsUK and applied factor analysis to concisely summarise patterns of gene expression, both to remove broad confounding influences and to produce concise pathway-level phenotypes. We derived 930 "pathway phenotypes" which summarised patterns of variation across 186 KEGG pathways (five phenotypes per pathway).

We identified 69 significant associations of age with phenotype from 57 distinct KEGG pathways at a stringent Bonferroni threshold (P<5.38x10^-5). These phenotypes are more heritable (h^2=0.32) than gene expression levels. On average, expression levels of 16% of genes within these pathways are associated with age. Of the 57 significant pathways, we frequently see four types of pathways, all of which have been previously linked with ageing: i) insulin signalling; ii) sugar and fatty acid metabolism; iii) xenobiotic metabolism; and iv) cancer related pathways. We have demonstrated that factor analysis methods combined with biological knowledge can produce more reliable phenotypes with less stochastic noise than the individual gene expression levels, which increases our power to discover biologically relevant associations. Finally, our analysis reveals pathways that have been seen to be important in longevity from a number of previous studies, as well as novel pathways that can be further investigated [ref Brown and Ding, *under review G3*]

Using models of biological age at the level of the organ system to inform risk of disease and co-morbidity

Biological age has been studied mainly at the whole body level, but the complexity of the phenotypes used, and the lack of reliable data in a sufficient number of individuals have made very difficult the reliable estimation of a systemic global biological age in humans. On the other hand, at the respiratory system level, an estimate of biological age has a potentially more powerful type of information which has been used in motivating smokers to quit. Several studies have investigated the possibility of estimating biological age of the respiratory system by using as biological age of a smoker the chronological age of a non-smoker of same height, gender and average FEV1 obtained from predictive equations. Therefore we decided to used spirometry tests results to evaluate lung function collected from The TwinsUK Registry (Chiappa et al. 2013). The data were investigated and cleaned by beneficiaries 1 and 3 for **WP3** and are described in detail in their section. For this study, we considered female individuals with spirometry data collected between 1992 and 2010 and with recorded height. Males were excluded as their number was too small to enable reliable estimation of model parameters and for consistency within the EuroBATs project that produced RNAseq data only from females.

We propose a probabilistic model that expresses the effects as number of years added to chronological age or, in other words, that estimates the biological age of the lungs. In our model, chronological age and other factors such a smoking and reported respiratory diseases affecting the health status of the lungs generate biological age, which in turn generates lung function measurements. This structure enables the use of

multiple types of measurement to obtain a more precise estimate of the effects and parameter sharing for characterization over large age ranges and of co-occurrence of factors with little data. Furthermore, we use the model to investigate the effects of smoking, asthma and COPD on the TwinsUK Registry.

Our results suggest that the combination of smoking with lung disease(s) has higher effect than smoking or lung disease(s) alone, and that, in smokers, co-occurrence of asthma and COPD is more detrimental than asthma or COPD alone. Moreover, we proposed that our model or other models based on the estimation of specific organs biological age could be of help in improving the understanding of factors affecting the organs function by enabling characterizations over large age ranges and of co-occurrence of factors with little data and the use of multiple types of measurement. The software implementing the model is available here: http://silviac.yolasite.com/ and it has the potential to be used to motivate life style changes in individuals under serious risk of suffering failures on the respiratory system due to hazardous exposures.

Integrating multilevel molecular phenotypes to reveal their putative mechanism of action under a causative model of ageing

Identifying age biomarkers can help to predict and monitor age-related physiological decline and disease, and, importantly, it can also provide molecular insights into the aging process and into early developmental processes that influence aging. Many of the genes and genetic markers (SNPs) identified have the potential to become reliable biomarkers of ageing. In order to better understand the relationship of multiple molecular phenotypes and the effect of ageing on them,we employed available datasets to investigate the relationship between expression and other molecular phenotypes in relation with ageing in an attempt to identify the causes of age related changes in expression.

Age and telomeres

Telomeres are nucleoprotein structures capping and protecting the ends of chromosomes. Telomeres shorten with each cell division and leukocyte telomere length has been shown to decrease with age at a rate of 20-40 base pairs per year. Telomere attrition has been associated with age-related diseases and expression regulation of genes near the ending of the chromosomes, but little is known of their regulation of effect on genes expression regulatory changes with age or tissue specificity effect.

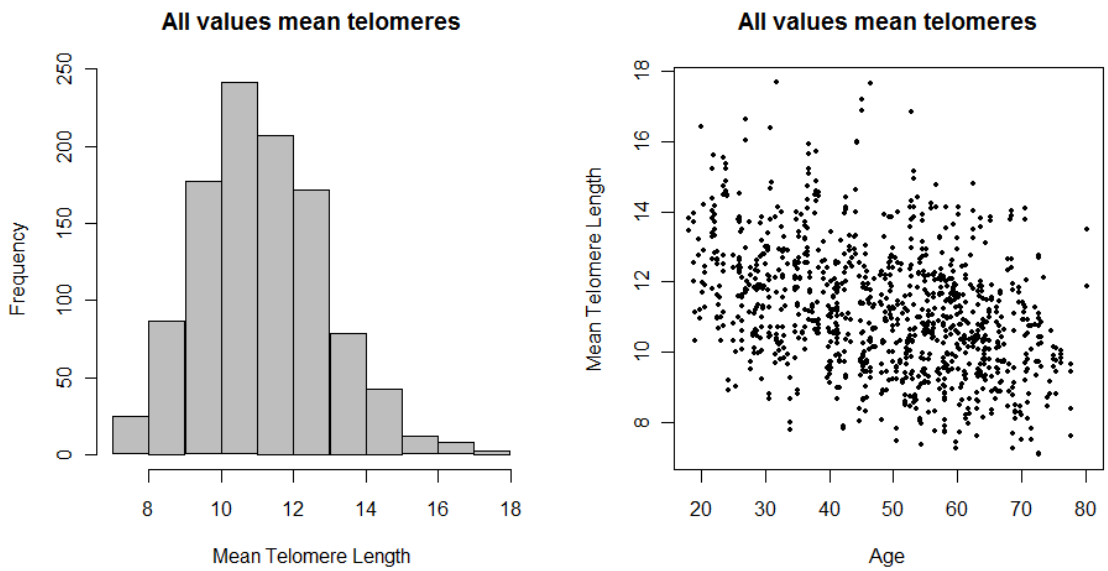

Figure 6: Blood telomeres data correlate with age.

Using telomapping measurements of human healthy skin sections provided by beneficiary 4 (CNIO) in **WP2**, we monitored the changes in telomere length in the different skin layers in an ageing population. We observed a very low correlation of telomere attrition with age in the skin telomapping measurements (spearman correlation = 0.06). However, it is not clear whether this is attributed to the small sample size or the high turnover of stem cell renewing the skin cell composition. To test this, we performed a genome wide association of telomere length measure in three skin layers, outer layer, basal layers and fibroblast cells, measure by telomapping. Although the small sample size did not allow us to find any significant association for specific genes, we were able to investigate potential significant results by *P*value enrichment analysis. From the expressed genes in skin, up to 20% of them were associated with telomere length in the basal layer, 15% in outer layer and only 8% in fibroblast. These different results between layers suggested that some of those associations were cell-type specific, suggesting that telomere length may act by regulating genes in a cell type specific manner.

Using Quantitative fluorescence in situ hybridization (Q-FISH) in a high throughput platform developed by beneficiary 5 (LL), which has potential advantages over PCR and Southern Blot methods, we obtained quantitative information on telomere length distributions from whole blood samples in 1703 samples from 800 individuals. Our analysis indicates that telomere length correlated negatively with chronological age (Spearman correlation = -0.249, Pvalue< 2.2e-16) (Figure 6).

In conclusion, we aimed to investigate the potential of telomere analysis for the development of biomarkers of skin ageing. However, our results indicated that although relevant to understanding the process of cellular senescence in skin, telomeres may not be the best marker for the study of systemic ageing in skin and are unlikely to have clinical potential. Telomeres in blood have been proved to be a useful tool for their association to difference diseases and their relationship with ageing and

expresison, however, different measurement methods tend to produce different results with no clear gold standard with strong clinical use potential.

Age and methylation

Multiple studies have found that age modifies methylation patterns; however it is not yet clear if these changes have consequences in the transcription of genes or if they only act as markers of ageing. We used fat methylome data available for 552 of the Eurobats twins with RNAseq data (Grundberg et al. 2013)(MuTHER Multiple Tissue Human Expression Resource, http://www.muther.ac.uk/). From 370,731 CpGs maker tested, 39,092 were significantly affected by age (10.54%), compared to the 1,511 exons affected by age (1.49%) indicating a wider effect of age in methylation levels than in expression in fat tissue. Applying Bayesian Networks we tested whether changes in expression with age were mediated by epigenetic markers, but in most cases we found little evidence that epigenetic markers were involved in differential expression, suggesting the greater effect of age in methylation does not necessarily translate into changes in gene expression.

Age and proteomics

Studies on aging using high-throughput proteomics identified proteins whose plasma levels and cerebrospinal fluid (CSF) levels substantially change with increasing age. Available subproteome targeted data by the SOMAscan assay was profiled in blood samples from 202 females from the TwinsUK cohort. Eleven proteins were associated with chronological age and were replicated at protein level in an independent population. Of those proteins, we found that the coding genes for three of them were also associated with age in their expression. We conclude that the relationship with age is the same both for mRNA and protein levels, although it is not significant with mRNA. There are many processes between transcription and translation, which result in a weak correlation between protein levels and mRNA levels and protein stability is a big factor (Menni et al. 2014)

Age related phenotypes and genotype

It has been established that the study of the respiratory system, which functionally changes with age, has the potential to provide relevant information to understand the ageing process, but the function of the respiratory system has not previously been linked to the genetics of individuals. In order to identify genotype-specific signatures of ageing and potentially identifying targets for anti-ageing interventions in the respiratory system we, performed genome-wide association study meta-analysis of force vital capacity (FVC) in collaboration with other groups(Loth et al. 2014). The collaborative study included 52,253 individuals from 26 studies and followed up the top associations in 32,917 additional individuals of European ancestry The study found six new regions associated at genome-wide significance ($P < 5 \times 10^{-8}$) with FVC in or near *EFEMP1*, *BMP6*, *MIR129-2–HSD17B12*, *PRDM11*, *WWOX* and *KCNJ2*. Two loci previously associated with spirometric measures (*GSTCD* and *PTCH1*) were related to FVC. We also investigated the association of available telomere length measurements from circulating leukocytes and respiratory disease (COPD and asthma), and the spirometric indices described above. We observed negative associations between telomere length and COPD (β=-0.0676, p=0.018) as well as asthma (β=-0.0452,

p=0.024) with stronger effects in women compared to men. These results indicate that lung function may reflect biological aging primarily due to intrinsic processes which are likely to be aggravated in lung diseases. Shortened telomeres in lung disease suggest that aging processes are involved in the pathogenesis of COPD and asthma with some genetics variants playing an important role on the progression of the age related decay (Albrecht et al. 2014).

Age related phenotypes and expression

Finally, we investigated the association between multiple ageing and health related phenotypes and concurrently measured expression levels across all four tissues (blood, LCL, adipose, skin). We found strong tissue-specific correlations between expression levels and multiple traits (Table 6). Notably, we find very few associations between phenotypes and LCL expression, indicating that cell lines are not the best model to capture in vivo relationships between phenotypes and expression.

Domain	Phenotype	Fat	Skin	Blood	LCLs
Lipids	Triglycerides	11,268	6	180	1
	HDL	13,002	50	5	2
	LDL	4,272	4	3,409	0
	Total Cholesterol	9	10	1,095	1
Glycaemina	Glucose	1,275	5	2	0
	Insulin	8,289	19	6	0
	Adiponectin	10,848	0	0	0
Anthropometric	BMI	16,816	9,216	6,640	0
	Hip	26	---	---	---
	Waist	3,897	5	2,420	---
	WHR	7,232	1	8	---

Table 6: Genes within each tissue with expression levels associated to a range of cross-sectional phenotypes (significance cutoff = FDR 5%)

In conclusion, the projects in Work Package 4 have successfully studied age related effects in the genome in a way that fully exploits the data generated in WP1, WP2 and WP3.

References

Albrecht E, Sillanpaa E, Karrasch S, Alves AC, Codd V, Hovatta I, Buxton JL, Nelson CP, Broer L, Hagg S et al. 2014. Telomere length in circulating leukocytes is associated with lung function and disease. *The European respiratory journal* **43**(4): 983-992.

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. 2007. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**(10): 1294-1296.

Brown AA, Buil A, Vinuela A, Lappalainen T, Zheng HF, Richards JB, Small KS, Spector TD, Dermitzakis ET, Durbin R. 2014. Genetic interactions affecting human gene expression identified by variance association mapping. *eLife* **3**: e01381.

Chiappa S, Winn J, Vinuela A, Tipney H, Spector TD. 2013. A probabilistic model of biological ageing of the lungs for analysing the effects of smoking, asthma and COPD. *Respiratory research* **14**: 60.

Glass D, Vinuela A, Davies MN, Ramasamy A, Parts L, Knowles D, Brown AA, Hedman AK, Small KS, Buil A et al. 2013. Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome biology* **14**(7): R75.

Grundberg E, Meduri E, Sandling JK, Hedman AK, Keildson S, Buil A, Busche S, Yuan W, Nisbet J, Sekowska M et al. 2013. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *American journal of human genetics* **93**(5): 876-890.

Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, Keildson S, Bell JT, Yang TP, Meduri E, Barrett A et al. 2012. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature genetics* **44**(10): 1084-1089.

Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**(6): e1000529.

Kerstjens HA, Rijcken B, Schouten JP, Postma DS. 1997. Decline of FEV1 by age and smoking status: facts, figures, and fallacies. *Thorax* **52**(9): 820-827.

Loth DW Artigas MS Gharib SA Wain LV Franceschini N Koch B Pottinger TD Smith AV Duan Q Oldmeadow C et al. 2014. Genome-wide association analysis identifies six new loci associated with forced vital capacity. *Nature genetics* **46**(7): 669-677.

Menni C, Kiddle SJ, Mangino M, Vinuela A, Psatha M, Steves C, Sattlecker M, Buil A, Newhouse S, Nelson S et al. 2014. Circulating Proteomic Signatures of Chronological Age. *The journals of gerontology Series A, Biological sciences and medical sciences*.

Parts L, Stegle O, Winn J, Durbin R. 2011. Joint genetic analysis of gene expression data with inferred cellular phenotypes. *PLoS genetics* **7**(1): e1001276.

Shabalin AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**(10): 1353-1358.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**(16): 9440-9445.

WP6 – Dissemination

Dissemination for the project included publications in peer reviewed journals, presentations, talks and posters at conferences, a public symposium, a scientific blog and a website which contained both public and private areas. All activities and publications have been uploaded to the portal.

Presentations at Meetings and Conferences

Participation at conferences, workshops and meetings have continued throughout the EuroBATS project, with members of the Consortium and young scientists encouraged to attend, and to submit abstracts.

Conference based dissemination activities (oral and poster presetations) were made at the following major meetings:
Title: Genetic Variation of Gene Co-Expression Networks in Three Tissues
Place: International Congress of Human Genetics. American Society of Human Genetics. Montreal 11-15th October 2011.
Alfonso Buil, University of Geneva

Title: Population genetics and genomics of cellular phenotypes.
Place: International Congress of Human Genetics. American Society of Human Genetics. Montreal 11-15th October 2011.
Manolis Dermitzakis, University of Geneva

The Fourth International
Conference of Quantitative Genetics: Understanding Variation in Complex Traits - Edinburgh International Conference Centre, 17-22 June 2012
Attended by Ana Vinuela, Andrew Brown, Sanger Institute

Selected talk presentation in the RoSyBA: Rostock Symposium on Systems Biology and Bioinformatics in Ageing Research. 15th-17th September 2011.
Title : Genome-wide transcriptome analysis with age and the effect of natural genetic variation in humans
Ana Viñuela, King's College London, UK.

European Society of Human Genetics, Paris 2013
Cold Springs Harbour, USA, May 2013
American Society of Human Genetics, Boston, USA 2013
American Society of Human Genetics, Montreal, 2012

American Society of Humans Genetics, San Francisco, November 6-10, 2012
21th Annual International Genetic Epidemiology Society. Stevenson, Washington, USA. October 18-20, 2012

In addition presentations and talks were given at many smaller meetings, including at Wellcome Trust Sanger Institute, Genomic Medicine in the Mediterranean, Rostock Symposium on Systems Biology in Ageing Research, Genetic variation and human health, Basel, Switzerland.

Abstracts have been accepted for the American Society of Human Genetics in San Diego, 2014.

In addition EuroBATS scientists have attended conferences and meetings including EMBL, Heidelberg (October 9-12 2013), Genomics of Common Diseases, Oxford, 26/5/13, Conference on Ageing Genetics, Edinburgh, 17-21/6/12. 1000 Genomes and Beyond, Cambridge, 23/6/14.

Symposium: Genomics of gene expression and ageing, a one day symposium at the Royal College of Physicians, London, May 2 2014. Presentations from EuroBATS Consortium members and invited speakers. The full programme is available on the EuroBATS website www.eurobats.eu/news

Collaborations

EuroBATS has close collaborations with the MuTHER Consortium, http://www.muther.ac.uk

EurHEALTHAGEING (FP7 277849) and EpiTwin, http://www.epitwin.eu

and will continue to provide useful data to other projects such as BPomics and MRC projects at the Department of Twin Research.

Website

The project website, www.eurobats.eu has been running since the beginning of the project and is a repository for members' publications, and meeting minutes etc, in addition to having a public face and general information. In order to maximise the impact of eurobats dissemination the website is being redesigned and will provide a platform for future collaborations and a repository for eurobats data. All documents and papers relating to EuroBATS will be held on the website, either in the public or password protected areas, depending on their nature.

All publications and dissemination activites have been uploaded to the SESAM reporting portal.

Blog

http://sangerinstitute.wordpress.com/2014/06/12/the-search-for-epistasis/
http://sangerinstitute.wordpress.com/2014/06/12/en-busca-de-la-epistasia/

Published in both English and Spanish.

Publications

Hypermethylation in the ZBTB20 gene is associated with major depressive disorder Matthew N. Davies, Ph.D et al, Genome Biology, 15/4 1 Jan 2014 R56

A probabilistic model of biological ageing of the lungs for analysing the effects of smoking, asthma and COPD, Silvia Chiappa et al, Respiratory Research, 14/60 30 May 2013 epub

Omics technologies and the study of human ageing, Ana Valdes et al, Nature Reviews Genetics 14, 13/8/13 601-607

Gene expression changes with age in skin, adipose tissue, blood and brain D. Glass et al, Genome Biology, 14/R75, 26/7/13 epub

Genetic interactions affecting human gene expression identified by variance association mapping. A.A. Brown et al. elife, Vol 3/0 1/1/13

Circulating proteomic signatures of chronological age, Cristina Menni et al, Journals of Gerontology, epub 14/8/14 doi 10.1093/gerona/glu121

Telomere length in circulating leukocytes is associated with lung function and disease, Schultz H, et al, European Respiratory Journal, Vol 42, epub 1/9/2013