

Informatics

Challenge of linking data whilst maintaining confidentiality

Catherine M Rice

Sharing Human Tissue: New opportunities,
new horizons. National Motorcycle Museum,
Birmingham, 15th September 2010



Overview

- Introduction to Sanger Institute's science
 - On a large scale
 - integrated projects
 - Some involving confidential data
- Data integration and handling
 - Farm of compute power necessary
 - Format exchange
 - Variety of types

Overview

- Confidentiality challenges
 - Risk to patients
 - Risk to institute
 - Allowing appropriate analysis
- Examples of recent EU collaborations
- and EMBL- EBI - data repository

Science & Research

- Involves a variety of data types
 - Sequence, genotype (genome-wide association study: GWAS), functional
 - Collaborative

Uncovering the heart of genetic influence on cholesterol and triglyceride levels in the blood
04.08.10: Worldwide research study reveals 95 genetic risk factors that alter levels of blood fats in multiple human populations [more ▶](#)

Decoding diabetes
28.06.10: 12 new genes linked to type 2 diabetes [more ▶](#)

Largest study of genomes and cancer treatments releases first results
15.07.10: UK-US collaboration building up a database for personalised cancer treatment [more ▶](#)

Celebrating a 'decade of discovery' since the Human Genome Project
24.06.10: New project launched to sequence 10,000 genomes in three years [more ▶](#)

Kymab and Wellcome Trust sign £20 million financing agreement
12.07.10: UK biopharmaceutical company will be built based on human monoclonal antibody technology from the Sanger Institute [more ▶](#)

1000 Genomes Project releases data from pilot projects on path to providing database for 2,500 human genomes
24.06.10: Freely available data supporting next generation of human genetic research [more ▶](#)

Sharing Human Tissue: New opportunities, new horizons. National Motorcycle Museum, Birmingham, 15th September 2010

Integrating the Data

- Handling the DNA samples

SANGER PLATE ID	WELL	SANGER SAMPLE ID	SUPPLIER SAMPLE NAME	COHORT	GENDER	MOTHER (optional)	FATHER (optional)	SIBLING (optional)	IS CONTROL?	COUNTRY OF ORIGIN	GEOGRAPHICAL REGION	ETHNICITY	DNA SOURCE
-----------------	------	------------------	----------------------	--------	--------	-------------------	-------------------	--------------------	-------------	-------------------	---------------------	-----------	------------

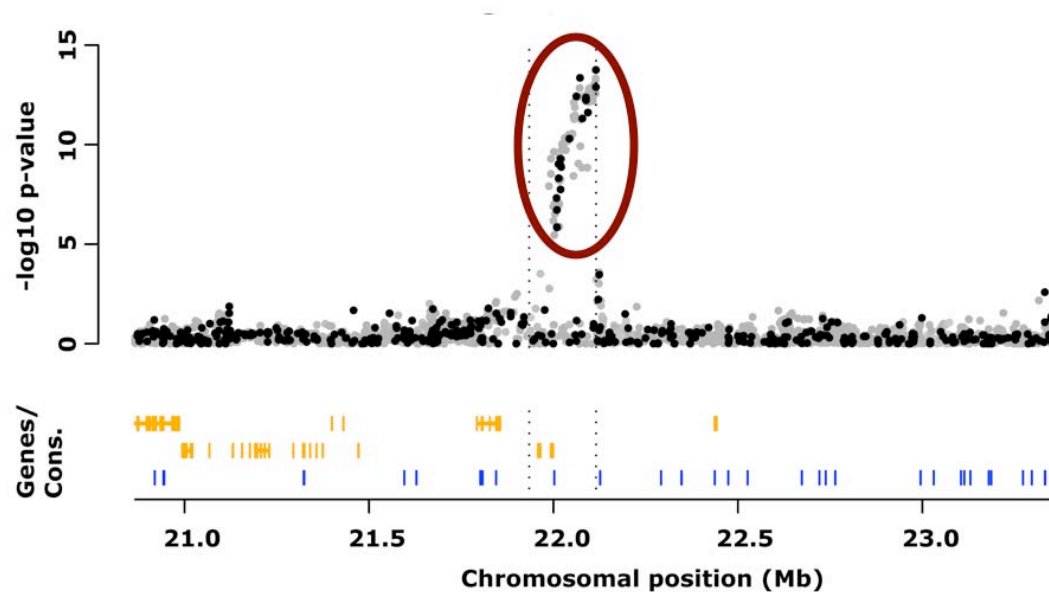
DATE OF SAMPLE COLLECTION (MM/YY or YYYY only)	DATE OF DNA EXTRACTION (MM/YY or YYYY only)	DNA EXTRACTION METHOD	DNA STORAGE CONDITIONS	SAMPLE PURIFIED?	PURIFICATION METHOD	VOLUME (µl)	CONC. (ng/µl)	CONCENTRATION DETERMINED BY	Priority (1 to 5)
--	---	-----------------------	------------------------	------------------	---------------------	-------------	---------------	-----------------------------	--------------------

- Storing phenotype data

Field Name	Short Name	Data Type	Unit
 <u>high blood pressure</u>	High blood pressure?	character	
 <u>high blood pressure age years</u>	Age at onset of AH (years)	integer	
 <u>high blood pressure year</u>	Year of diagnosis of Hypertension	integer	

Integrating the Data - more

- Generation of the experimental data



- lead SNP >100 kb away from known genes
- limited evolutionary conservation
- functional relevance not known

Integrating the Data - more

- Movement of the intervening data
 - Unix files
 - Download directories

cardiogenics WPS Data Access Area

You are logged in as user **cmr** ([log out](#)).

You have been assigned the following access levels; level001 levelAdmin levelAllelic_imbalance levelCNV_analysis levelCamData levelGWAS_GWE_analysis levelSCG_analysis levelSeraya levelUserAdmin levelWp5analysis level_eQTL testcmr

Home QueryDB **Files** Admin UserAdmin

File access tool; levelCamData

Currently viewing: /

Index file index.xml can be read

Index file validates OK

Name	Size	Date	Action
Human1-2M-DuoCustom_v1_A.csv.gz	122597891	23-Nov-2009 22:07	Download
Human670-QuadCustom_v1.txt	34455632	23-Nov-2009 20:54	Download
Human670-QuadCustom_v1_A.csv	275169705	23-Nov-2009 20:54	Download
cardio-1200.tar.gz	136823896	18-Jan-2010 16:44	Download
cxma04-670_only.tar.bz2	61259459	24-Nov-2009 15:22	Download

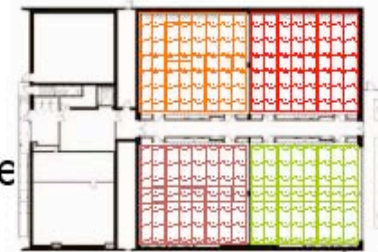
Mouse over file and directory names to view more information (requires Javascript)

Sharing Human Tissue: New opportunities,
new horizons. National Motorcycle Museum,
Birmingham, 15th September 2010

Data generation, analysis and handling

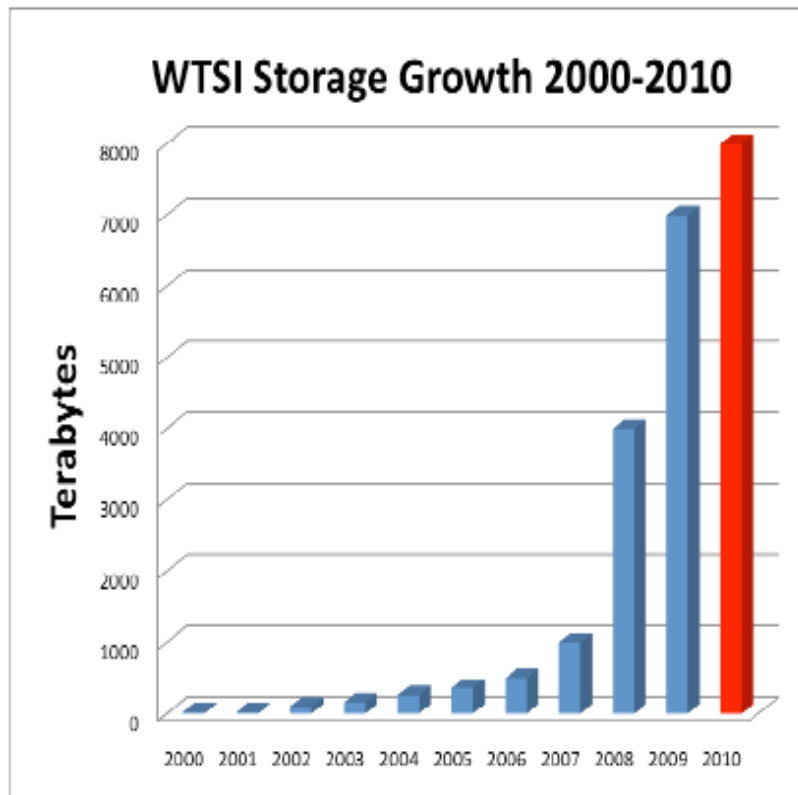
The Wellcome Trust Sanger Institute Data Centre

- 1,000 m² of data centre space (4 x 250m²)
- Hosts Sanger and EBI's compute and storage
- Combined totals > 13,000 cores of compute
> 13 Petabytes of storage
- Rivals the largest high performance compute facilities in the UK
- Possibly the largest Life Science data centre in Europe



Data generation, analysis and handling

By 2005 we had accrued
300 Terabytes of data



With the introduction of new
sequencing technologies
this has grown > 20 fold In
the last 2-3 years

We now have **>8 Petabytes**
of storage capacity
(8,000 Terabytes)

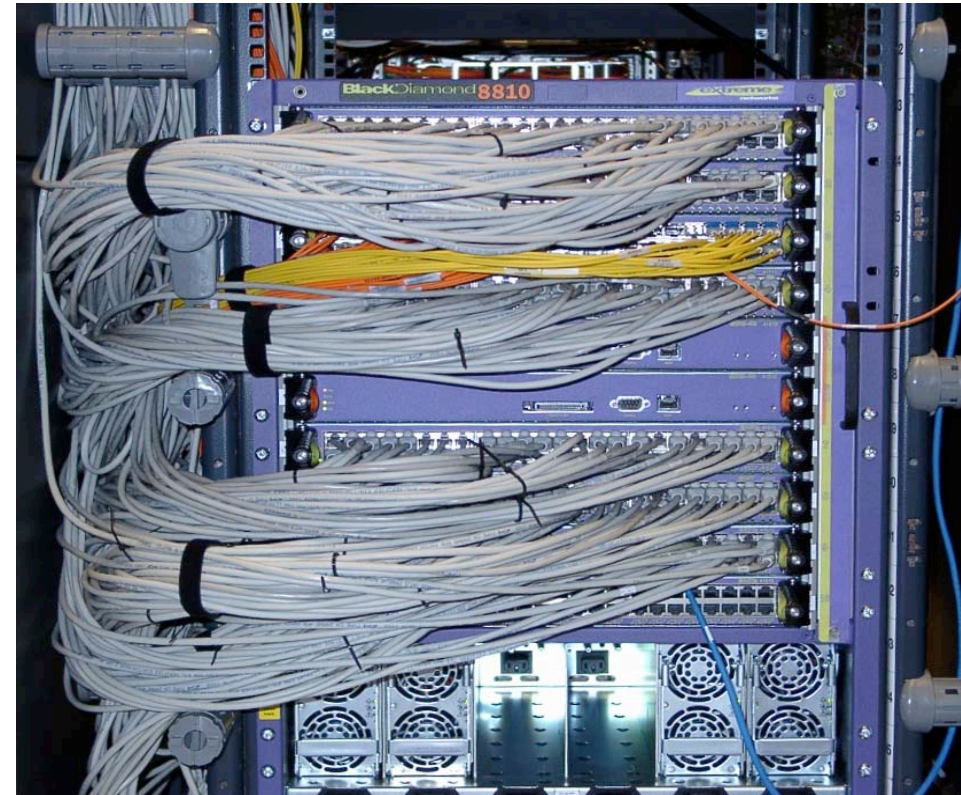
Sharing Human Tissue: New opportunities,
new horizons. National Motorcycle Museum,
Birmingham, 15th September 2010

Slide courtesy of Head of IT



IBM Blade computers
and DDN high performance storage

Sharing Human Tissue: New opportunities,
new horizons. National Motorcycle Museum,
Birmingham, 15th September 2010



Data cabling for the storage
infrastructure

Slide courtesy of Head of IT

Confidentiality issues

- Assessment of the situation
 - Wellcome Trust Sanger Institute is open access for data
 - Make publicly/securely available at data repository on site: EGA
 - Where does the risk lie if we get it wrong?
- Risk to the patients/donors
 - Sensitive information may get released
 - Trust in us/others to keep data safe may be lost
 - Possible loss of volunteers
- Risk to the Institute
 - Reputation
 - Hackers - daily make *failed* attempts

Confidentiality issues

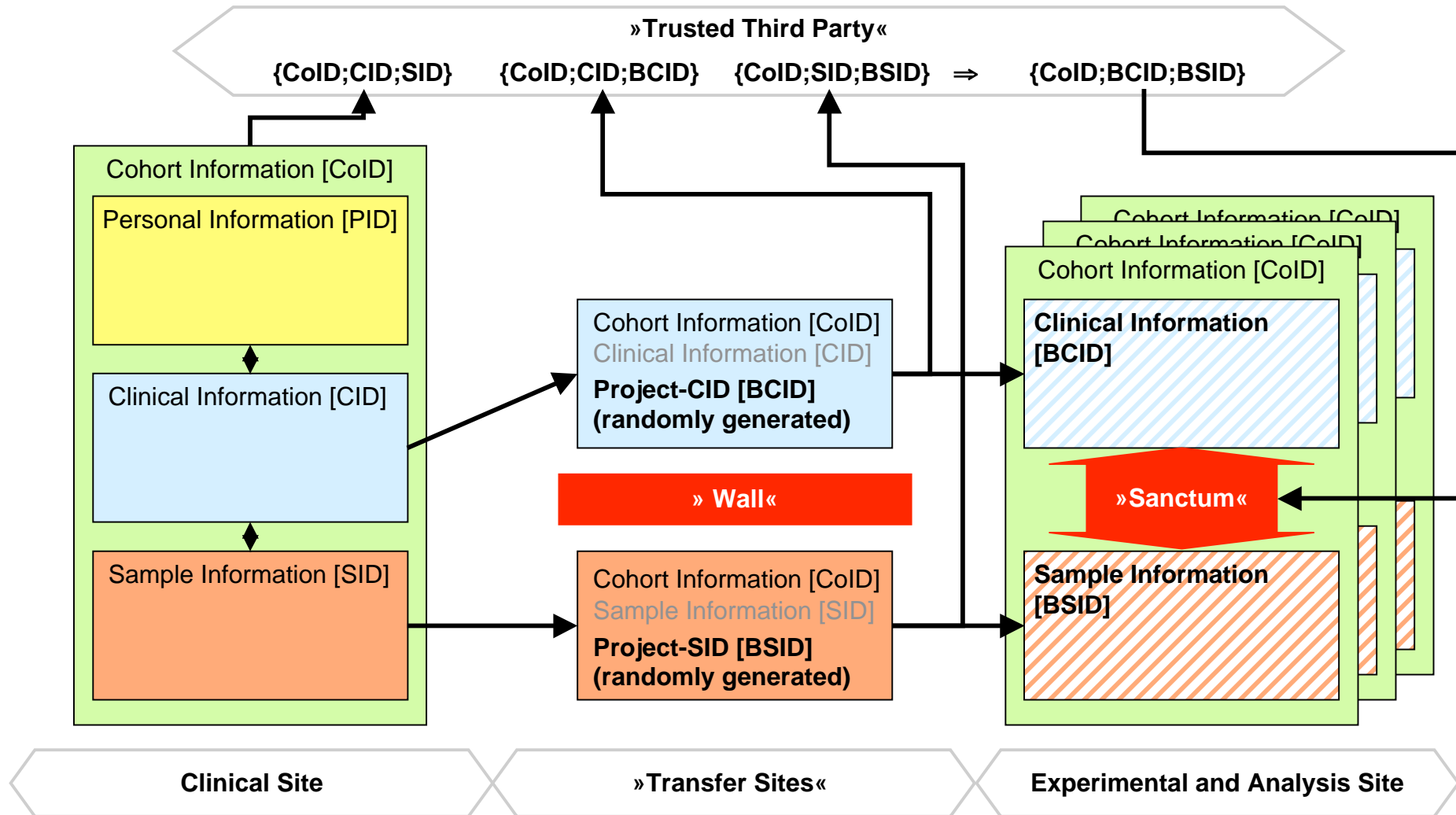
- What was the agreement with the donors (c.f. 1000 genomes, HapMap)
 - Open access
- Do we hold phenotypic data of very detailed sort or is that held elsewhere
 - Or only geographical region/year of birth, gender
- How do we restrict access to the appropriate people for analysis
 - Statisticians, Bioinformaticians
- We need to ensure appropriate levels of security for each of our wide variety of projects

Bloodomics and Cardiogenics EU Projects

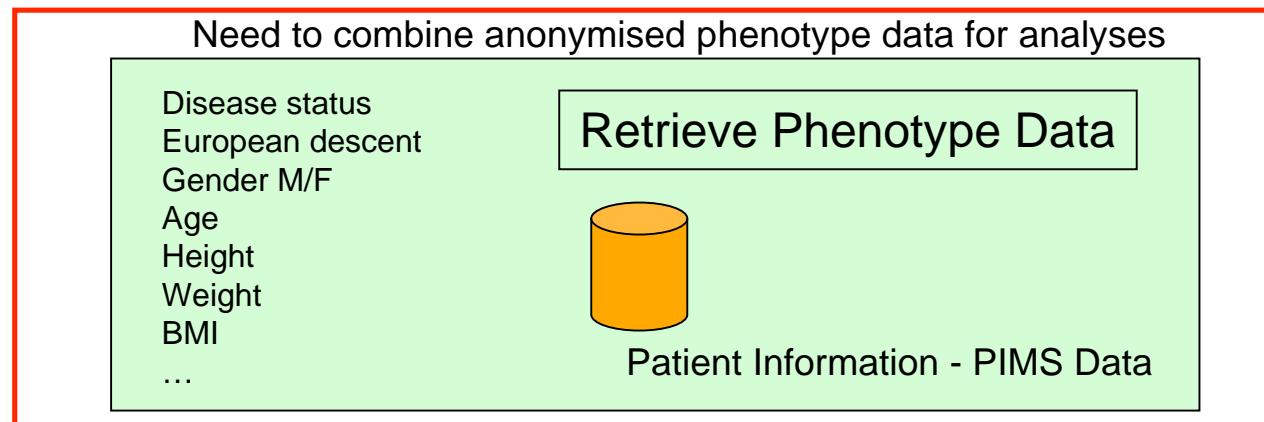
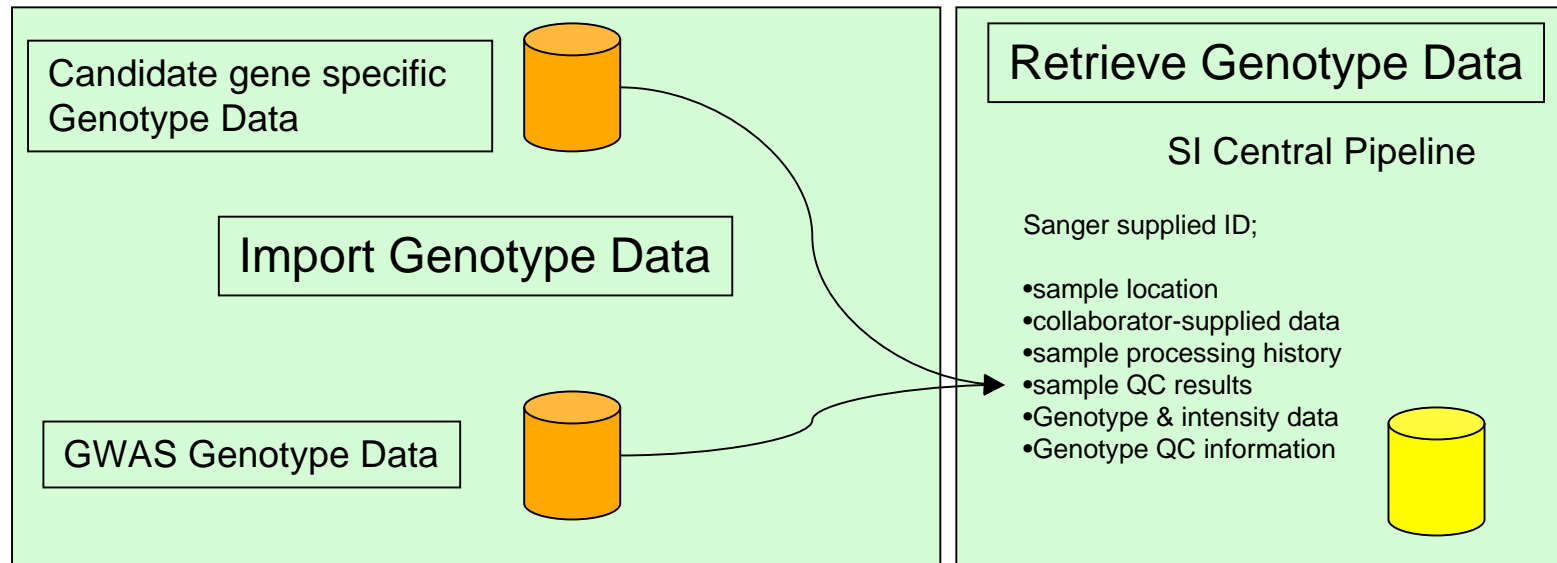
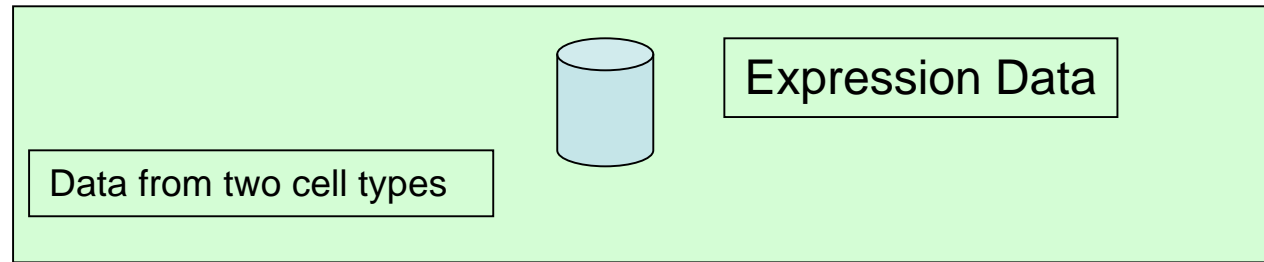
Phenotype

	Case1	Case2	Case3	Control1	Case5
N	161	102	126	463	74
Disease status	PMI	PMI	PMI	Healthy	CAD
European descent (%)	100	100	100	100	100
Gender M/F	120/13	80/12	101/21	190/262	61/10
Age (years; mean±sd)	~55	~55	~55	~53	~65
Height(years; mean±sd)	Etc...				
Weight(Kg; mean±sd)					
BMI (Kg/m ² ; mean±sd)					
Sys BP (mm Hg; mean±sd)					
Dia BP (mm Hg; mean±sd)					
Diabetes (%)					
Hyperlipidemia (%)					
Non-current smokers (%)					
Alcohol consumption (%)					

Confidentiality and Anonymisation



Project specific data Integration



Bloodomics security

WWW

- Public data: public website

LIMS

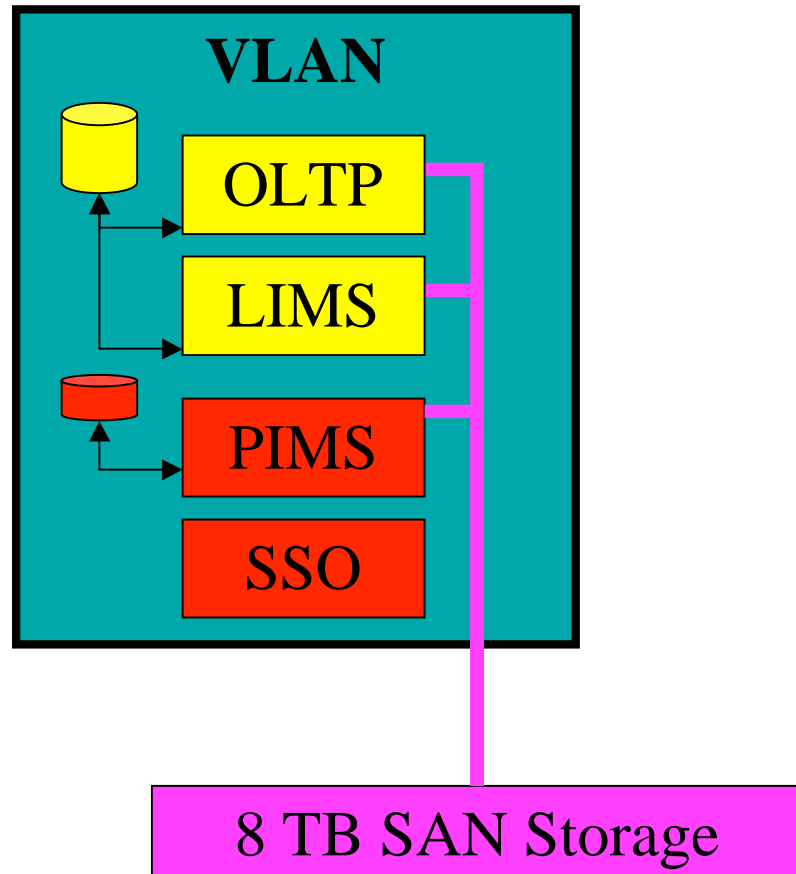
- Project data: presentations, common documents (Intranet)
- Project private data: workgroup shared documents
- Sample data (genotypes, gene expression, sequence, ...)

Sanctum

- Patient data (phenotypes)

Servers Overview

Virtual LAN



Local Area Network



Sharing Human Tissue: New opportunities,
new horizons. National Motorcycle Museum,
Birmingham, 15th September 2010

The European Genome-phenome Archive



- Secure storage and authorised access to all types of data sets that might be generated in the context of research into molecular medicine

- Sequence; Genotypes
- Transcriptomics; Proteomics
- Phenotype data

- Enable the collection of larger cohorts and maximisation of resource use
 - Sequencing capacity is increasing dramatically
 - Analysis capacity is increasing more slowly

Sharing Human Tissue: New opportunities, new horizons. National Motorcycle Museum, Birmingham, 10-11 September 2010

The screenshot shows the EMBL-EBI website interface. At the top, there is a search bar with 'EB-eye Search' and 'All Databases' dropdown. Below the search bar are navigation tabs: Databases, Tools, EBI Groups, Training, Industry, About Us, and Help. The main content area is titled 'The European Genome-phenome Archive' and contains a description of the archive, a user login form, and a list of research projects. The footer includes 'Terms of Use', 'EBI Funding', 'Contact EBI', and '© European Bioinformatics Institute 2009. EBI is an Outstation of the European Molecular Biology Laboratory.'

Slide courtesy of Paul Flicek, EBI



EGA Data Acceptance and Access



- Access decisions will remain with the data generating body
 - Distributed model
 - Transparency to the data generators
 - EGA manages the access granted
 - Users can also be restricted to particular collections within a study
- EGA is the European peer database to dbGAP (NCBI)
 - dbGAP has adopted a more centralised model of data access decisions
 - We plan data exchange of meta data and more extensive discussions are on going to increase data discoverability
 - Working toward a common application for both databases to lower administrative burden

Community Benefits of EGA

- Data subject to access controls is a burden and it limits the number of researchers that will reuse the resources
 - This may slow the pace of science and prevent serendipitous discovery
- However...
 - Five years ago accessing this type of data was impossible
 - Now it is just incredibly difficult
 - This is real progress
- Complicated or overly onerous data access agreements are more likely to be ignored

EGA Consortium Page for WTCCC

EMBL-EBI EB-eye Search All Databases Enter Text Here Go Reset Advanced Search Give us feedback

Databases Tools EBI Groups Training Industry About Us Help Site Index

- EGA Home
- Information
- Jobs
- Contact

View White Paper

User Login

Username:

Password:

Login

[I forgot my password](#)

EBI > The European Genotype Archive > Wellcome Trust Case Control Consortium

Wellcome Trust Case Control Consortium

Details

Description	The Wellcome Trust Case Control Consortium (WTCCC) is a collaboration of 24 leading human geneticists, who will analyse thousands of DNA samples from patients suffering with different diseases to identify common genetic variations for each condition.
URL	http://www.wtccc.org.uk
Abstract	The WTCCC has now searched for the genetic variation associated with coronary heart disease, type 1 diabetes, type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder and hypertension. The research was conducted at a number of institutes throughout the UK, including the Wellcome Trust Sanger Institute, Cambridge University and Oxford University. Researchers will have analysed over 14,000 DNA samples - two thousand patients for each disease and three thousand control samples - searching for important genetic differences between people who do and don't have each disease.

Studies

- Bipolar Disorder (BD)
- Coronary Artery Disease (CAD)
- Crohn's Disease (CD)
- Hypertension (HT)
- Rheumatoid Arthritis (RA)
- Type 1 Diabetes (T1D)
- Type 2 Diabetes (T2D)
- Ankylosing Spondylitis (AS)
- Autoimmune Thyroid Disease (ATD)
- Multiple Sclerosis (MS)
- Breast Cancer (BC)

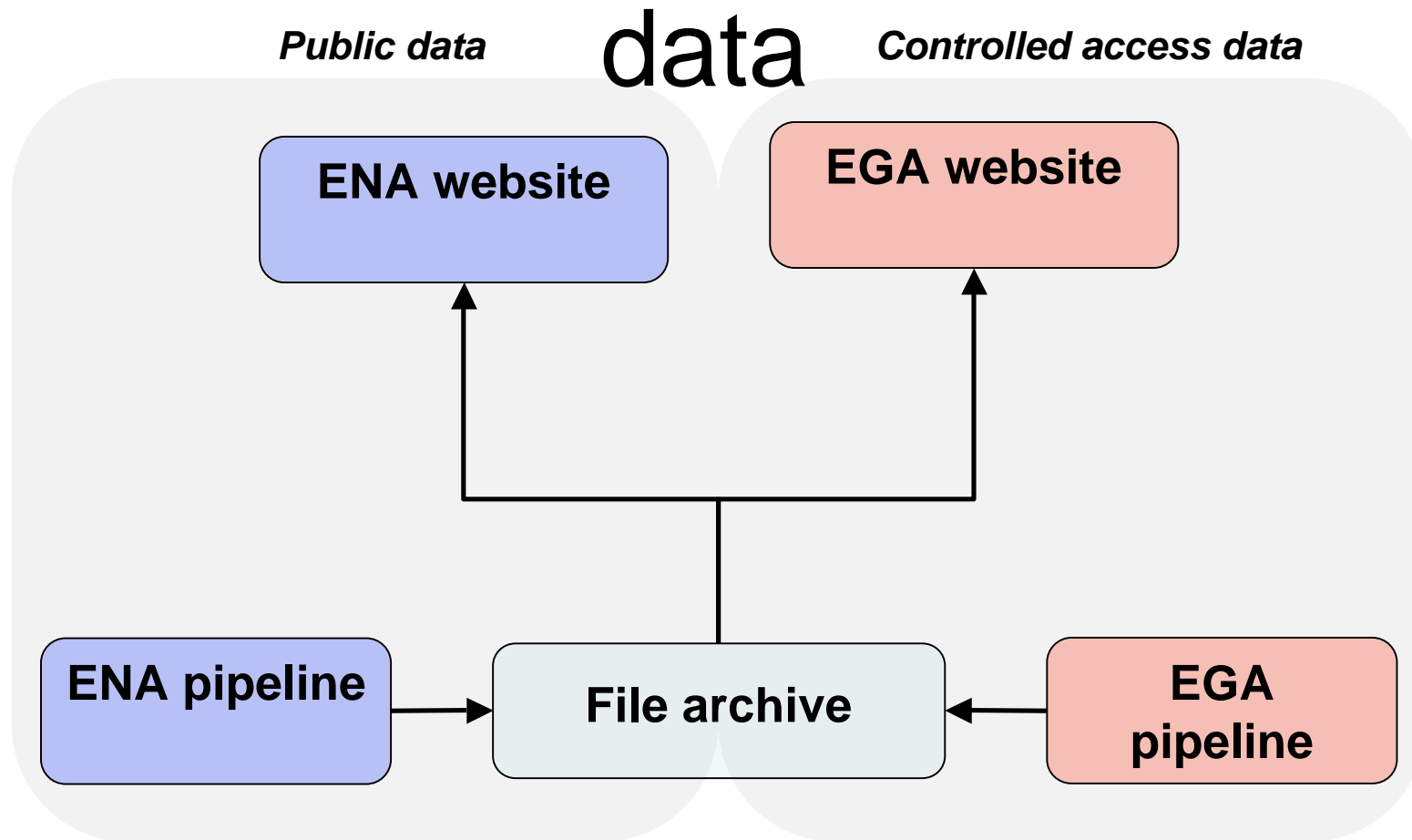
[Terms of Use](#) | [EBI Funding](#) | [Contact EBI](#) | © European Bioinformatics Institute 2006-2007. EBI is an Outstation of the [European Molecular Biology Laboratory](#).

Sharing human genome: new opportunities,
new horizons. National Motorcycle Museum,
Birmingham, 15th September 2010

Slide courtesy of Paul Flicek, EBI



Designing Archives for both public and controlled access



Conclusions

- Security
 - A few *specified* levels of security
 - Appropriate levels to the individual project
 - Memorable standard operating procedures for each level of access
 - Who and how
- Linking
 - Stable systems
 - Appropriate and varied technical support
 - Flexible approach to development
 - Reusable code