MEDI6215

UNIVERSITY OF Southampton

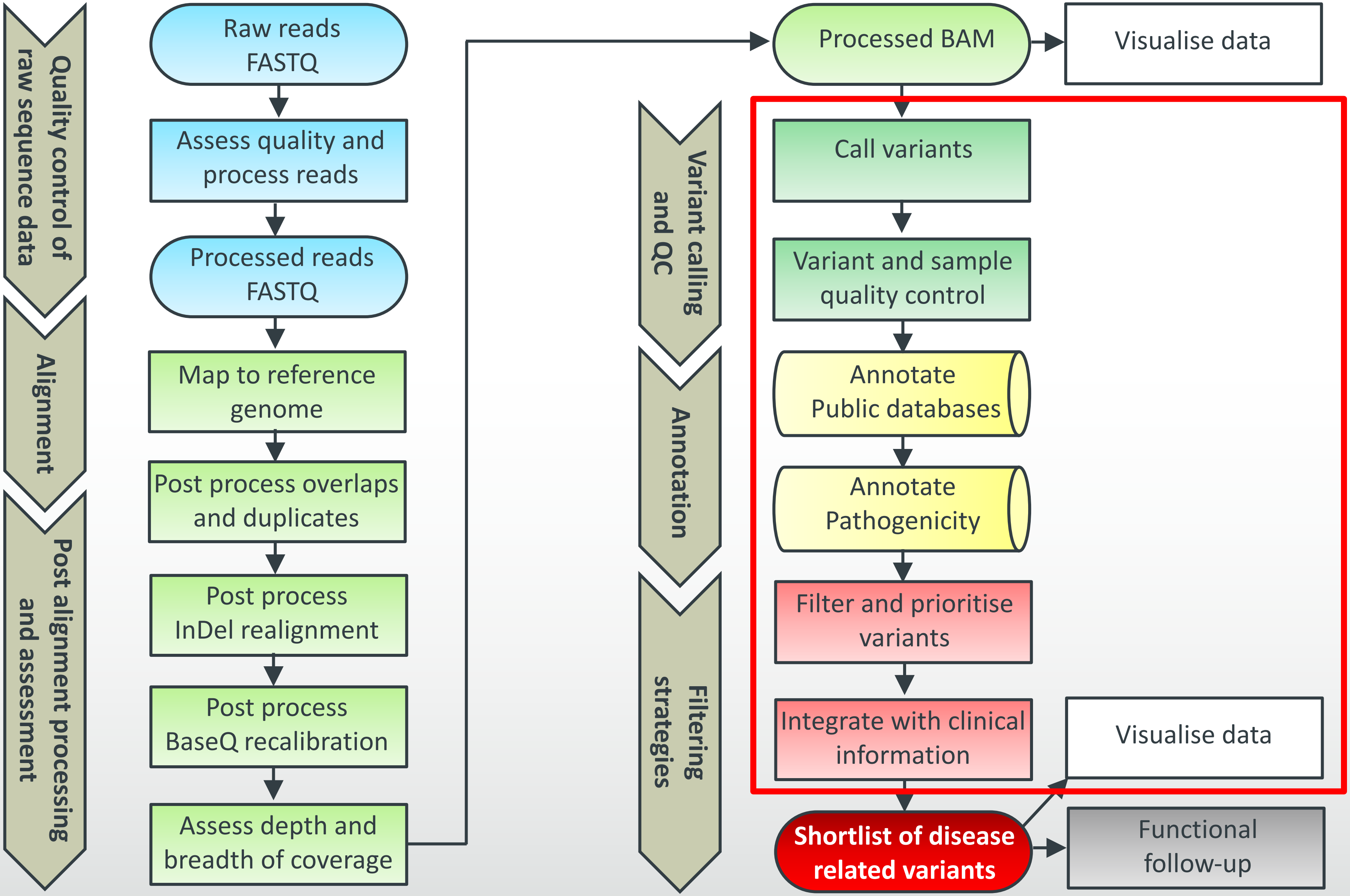
The Training Embassy

Bioinformatics, Interpretation, and Data Quality Assurance in Genome Analysis

Christopher Woelk, PhD

December 16th, 2015

Analysis workflow


Quality control of raw sequence data
Alignment
Post alignment processing and assessment
Raw reads FASTQ
Assess quality and process reads
Processed reads FASTQ
Map to reference genome
Post process overlaps and duplicates
Post process InDel realignment
Post process BaseQ recalibration
Assess depth and breadth of coverage
Variant calling and QC
Annotation
Filtering strategies
Processed BAM
Visualise data
Call variants
Variant and sample quality control
Annotate Public databases
Annotate Pathogenicity
Filter and prioritise variants
Integrate with clinical information
Visualise data
Shortlist of disease related variants
Functional follow-up

OUTLINE

- **LEARNING OUTCOME**

  Practice in examples of analysis of genomic data in the Training Embassy within the Genomics England Data Centre.

- **INDICATIVE CONTENT**

  Gain practical experience of the bioinformatics pipeline through the Genomics England programme.

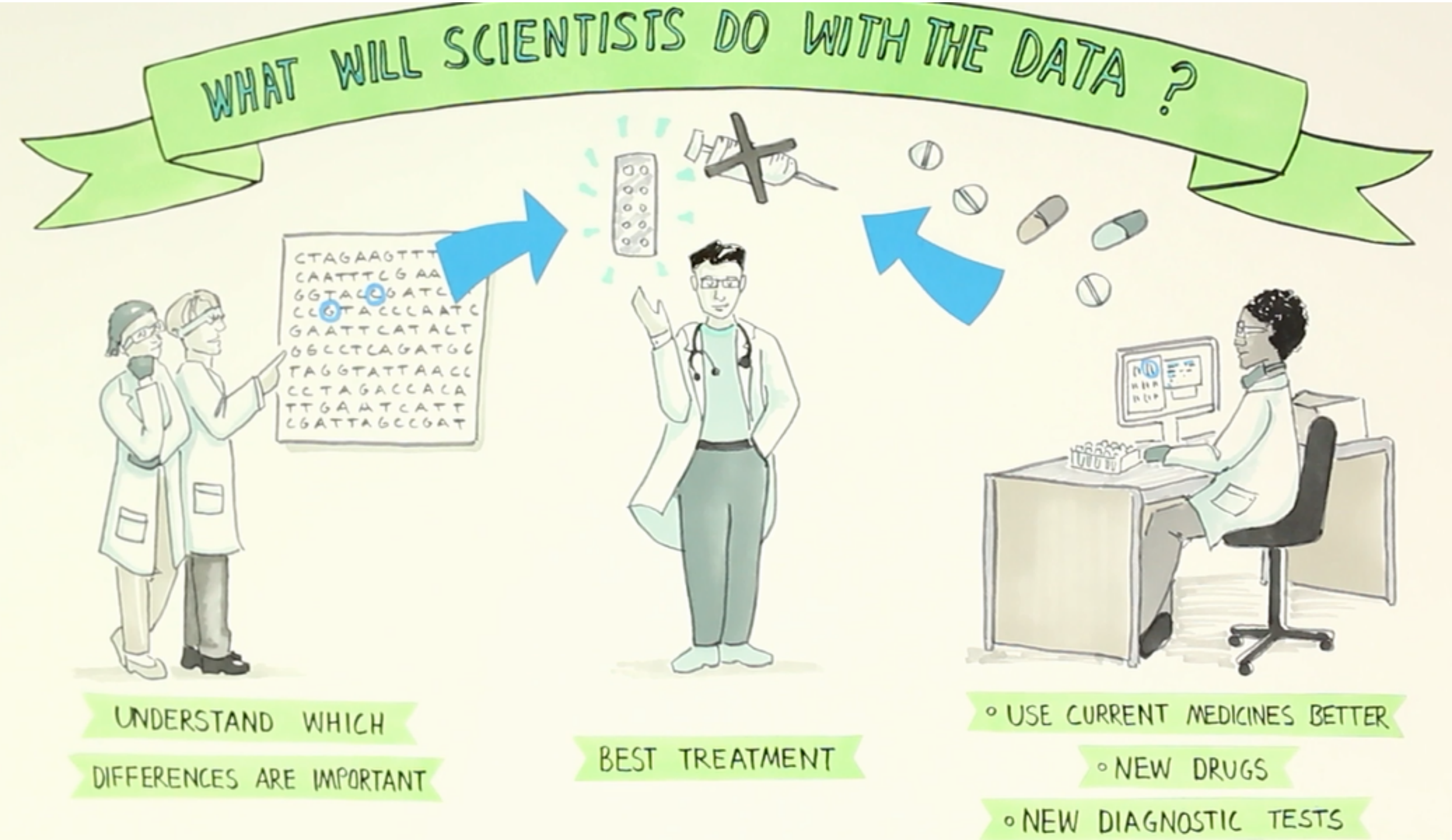
What is a “Training Embassy”?

- Domains will have specific ‘embassies’ within the Genomics England data centre, which comprise of the sub-section of the data related to that domain and components of the Genomics England computing infrastructure relevant to analysis of the data.
- These embassies are intended to be seen as areas in which healthcare professionals, researchers, trainees and pre-competitive industry partners undertake their work.
- Work undertaken within the embassies is subject to the governance and terms and conditions of Genomics England.

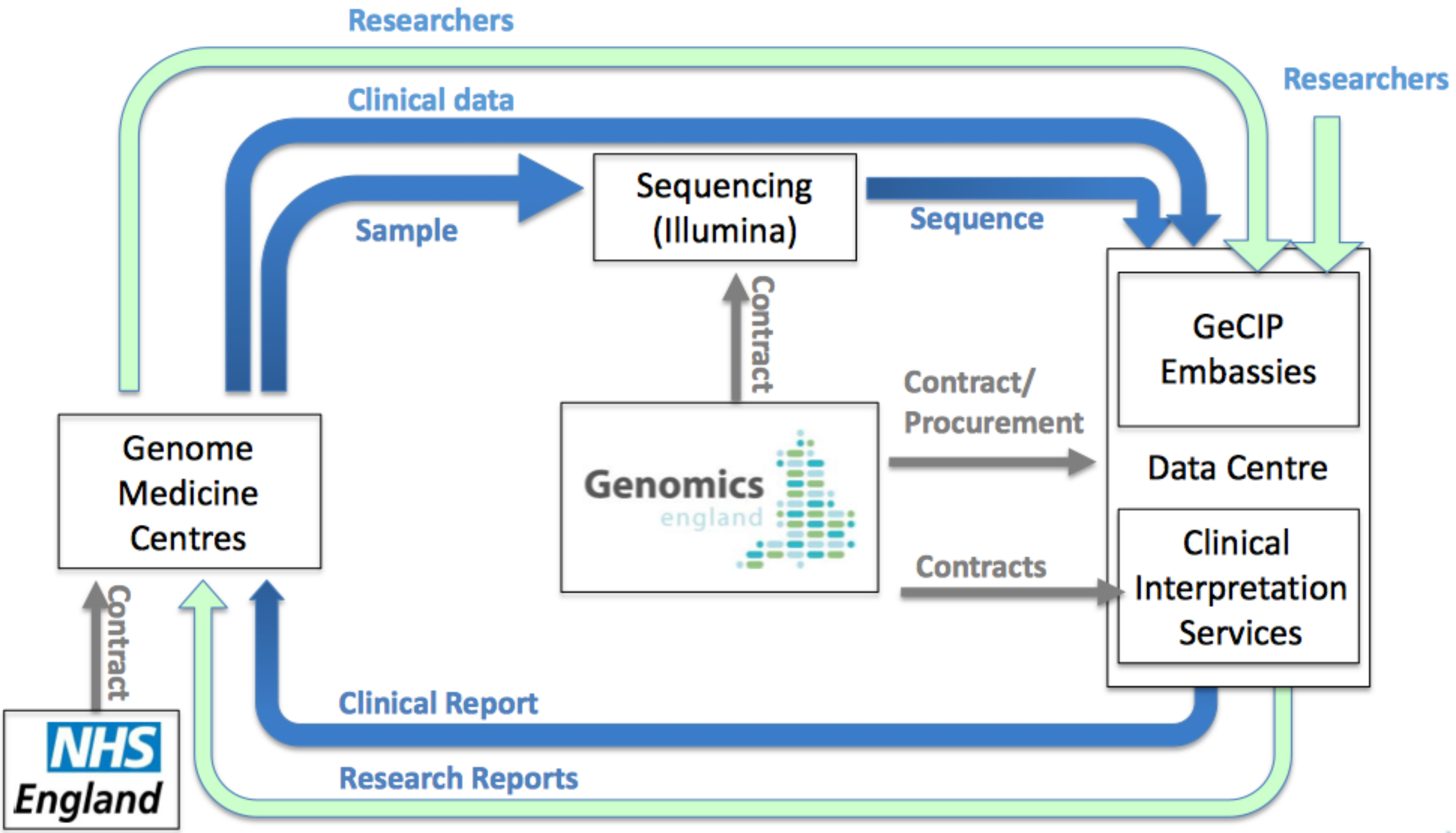
Vision

- As part of the 100,000 genomes project, Genomics England will provide a Training Embassy with access to genome sequence data gathered in a real clinical context.
- These data will be made available to all trainees, who are part of the HEE training programmes and can be used by trainers to develop some exemplars of genomic interpretation and its application in the clinic, thereby providing the hands-on practical training needed for NHS staff.
- We strongly recommend that training developed by the Health Education England takes advantage of this facility.

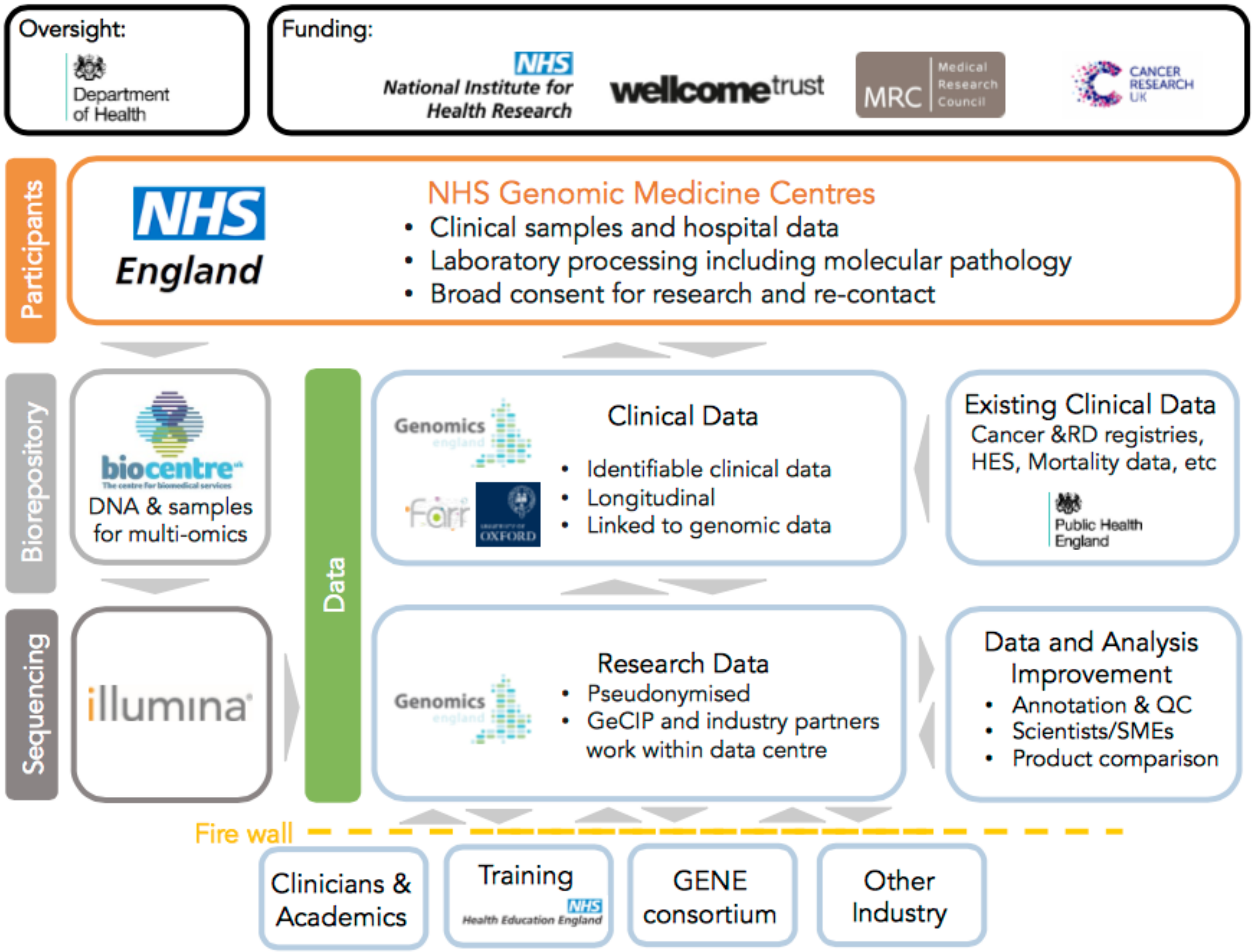
Public Perception



More Detail …


Researchers
Clinical data
Researchers
Sample
Sequencing (Illumina)
Sequence
Contract
Contract/ Procurement
GeCIP Embassies
Data Centre
Genome Medicine Centres
Genomics england
Contracts
Clinical Interpretation Services
Contract
Clinical Report
NHS England
Research Reports

Even More Detail …



Oversight:

Department of Health

Funding:

NHS National Institute for Health Research

wellcometrust

MRC Medical Research Council

CANCER RESEARCH UK

Participants

NHS England

NHS Genomic Medicine Centres

- Clinical samples and hospital data
- Laboratory processing including molecular pathology
- Broad consent for research and re-contact

Biorepository

biocentre

The centre for biomedical services

DNA & samples for multi-omics

Sequencing

illumina

Data

Genomics england

Farr

UNIVERSITY OF OXFORD

Clinical Data

- Identifiable clinical data
- Longitudinal
- Linked to genomic data

Existing Clinical Data

Cancer &RD registries, HES, Mortality data, etc

Public Health England

Genomics england

Research Data

- Pseudonymised
- GeCIP and industry partners work within data centre

Data and Analysis Improvement

- Annotation & QC
- Scientists/SMEs
- Product comparison

Fire wall

Clinicians & Academics

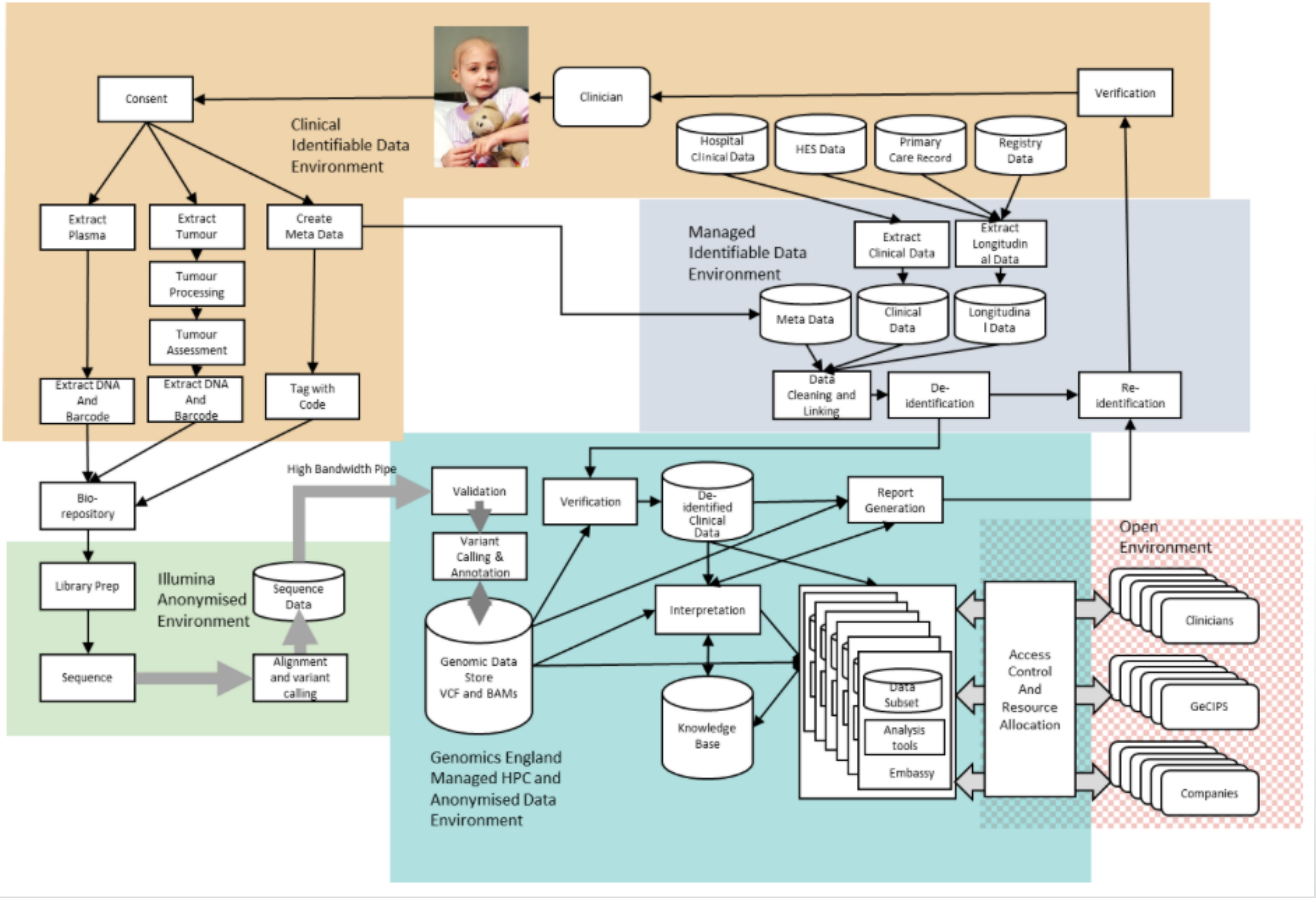
Training

NHS Health Education England

GENE consortium

Other Industry

Even Even More Detail …



GeCIPs

Genomics England Clinical Interpretation Partnerships

GeCIP Domains are UK-led consortia of researchers, clinicians and trainees.

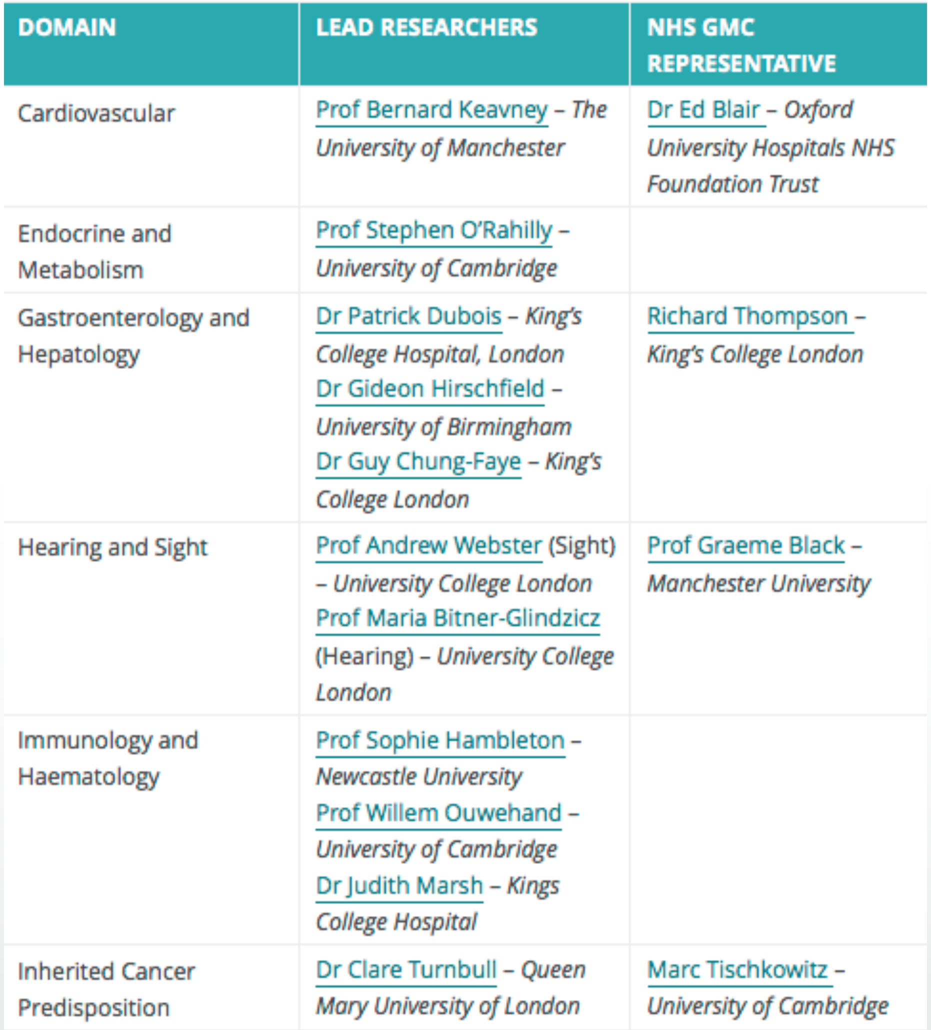
Each domain will work on improving the clinical application and interpretation of the data in the 100,000 Genomes Project.

OBJECTIVES:

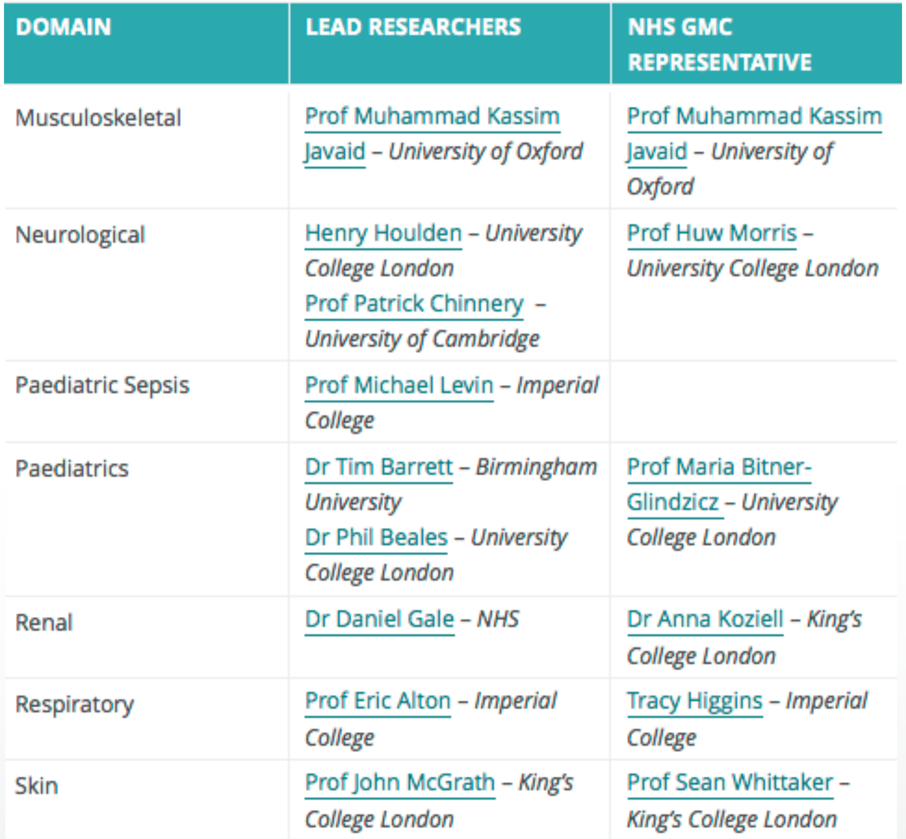
1) Improve our understanding of genomic medicine and its application to healthcare
2) Improve our understanding of disease
3) Lead the way to developing new diagnostics and treatments.

The MSc programme also includes a core research module (30 ECTS) with opportunities to access the emerging data from the 100,000 Genomes Project through the GeCIP training domains.

Rare Disease GeCIPs

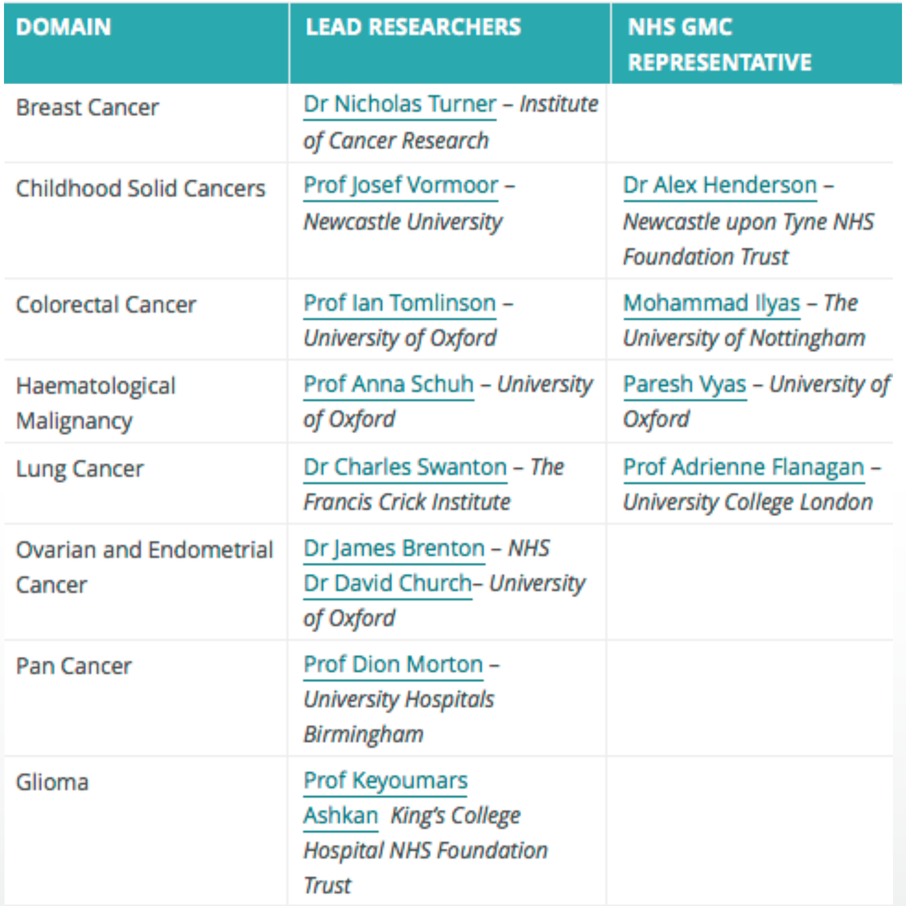


DOMAIN	LEAD RESEARCHERS	NHS GMC REPRESENTATIVE
Cardiovascular	Prof Bernard Keavney – *The University of Manchester*	Dr Ed Blair – *Oxford University Hospitals NHS Foundation Trust*
Endocrine and Metabolism	Prof Stephen O'Rahilly – *University of Cambridge*	
Gastroenterology and Hepatology	Dr Patrick Dubois – *King's College Hospital, London* Dr Gideon Hirschfield – *University of Birmingham* Dr Guy Chung-Faye – *King's College London*	Richard Thompson – *King's College London*
Hearing and Sight	Prof Andrew Webster (Sight) – *University College London* Prof Maria Bitner-Glindzicz (Hearing) – *University College London*	Prof Graeme Black – *Manchester University*
Immunology and Haematology	Prof Sophie Hambleton – *Newcastle University* Prof Willem Ouwehand – *University of Cambridge* Dr Judith Marsh – *Kings College Hospital*	
Inherited Cancer Predisposition	Dr Clare Turnbull – *Queen Mary University of London*	Marc Tischkowitz – *University of Cambridge*

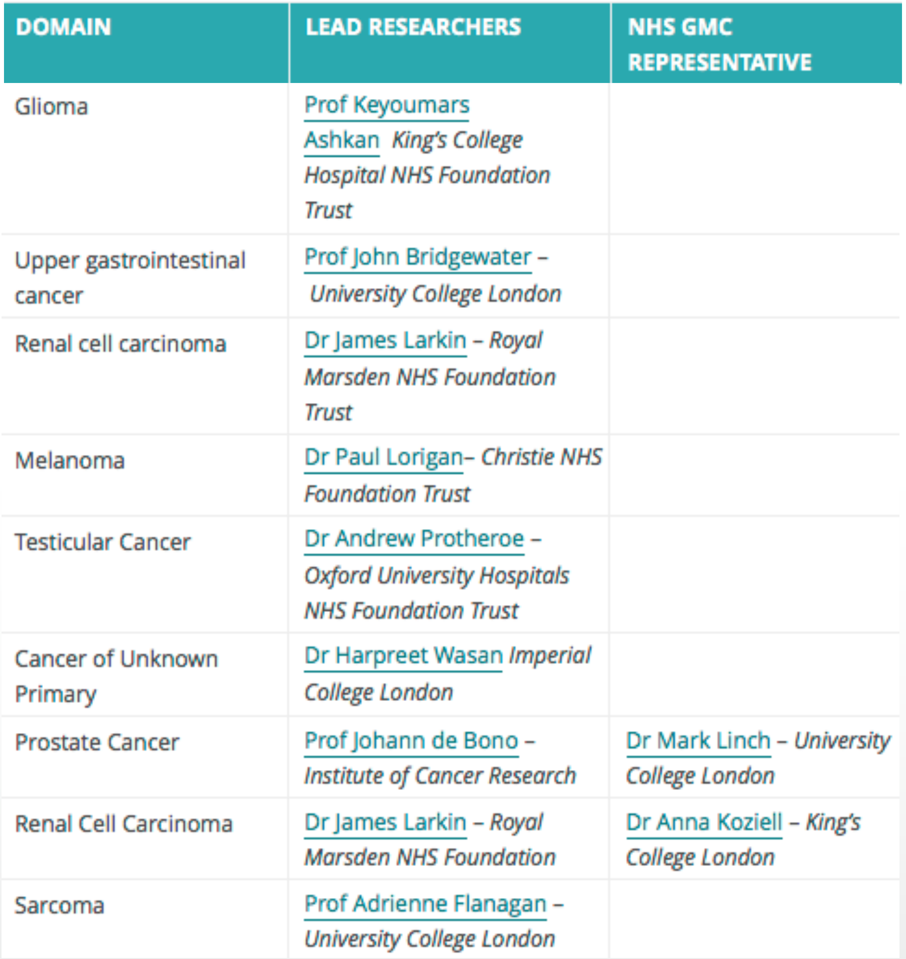


DOMAIN	LEAD RESEARCHERS	NHS GMC REPRESENTATIVE
Musculoskeletal	Prof Muhammad Kassim Javaid – *University of Oxford*	Prof Muhammad Kassim Javaid – *University of Oxford*
Neurological	Henry Houlden – *University College London* Prof Patrick Chinnery – *University of Cambridge*	Prof Huw Morris – *University College London*
Paediatric Sepsis	Prof Michael Levin – *Imperial College*	
Paediatrics	Dr Tim Barrett – *Birmingham University* Dr Phil Beales – *University College London*	Prof Maria Bitner-Glindzicz – *University College London*
Renal	Dr Daniel Gale – *NHS*	Dr Anna Koziell – *King's College London*
Respiratory	Prof Eric Alton – *Imperial College*	Tracy Higgins – *Imperial College*
Skin	Prof John McGrath – *King's College London*	Prof Sean Whittaker – *King's College London*

Cancer GeCIPs

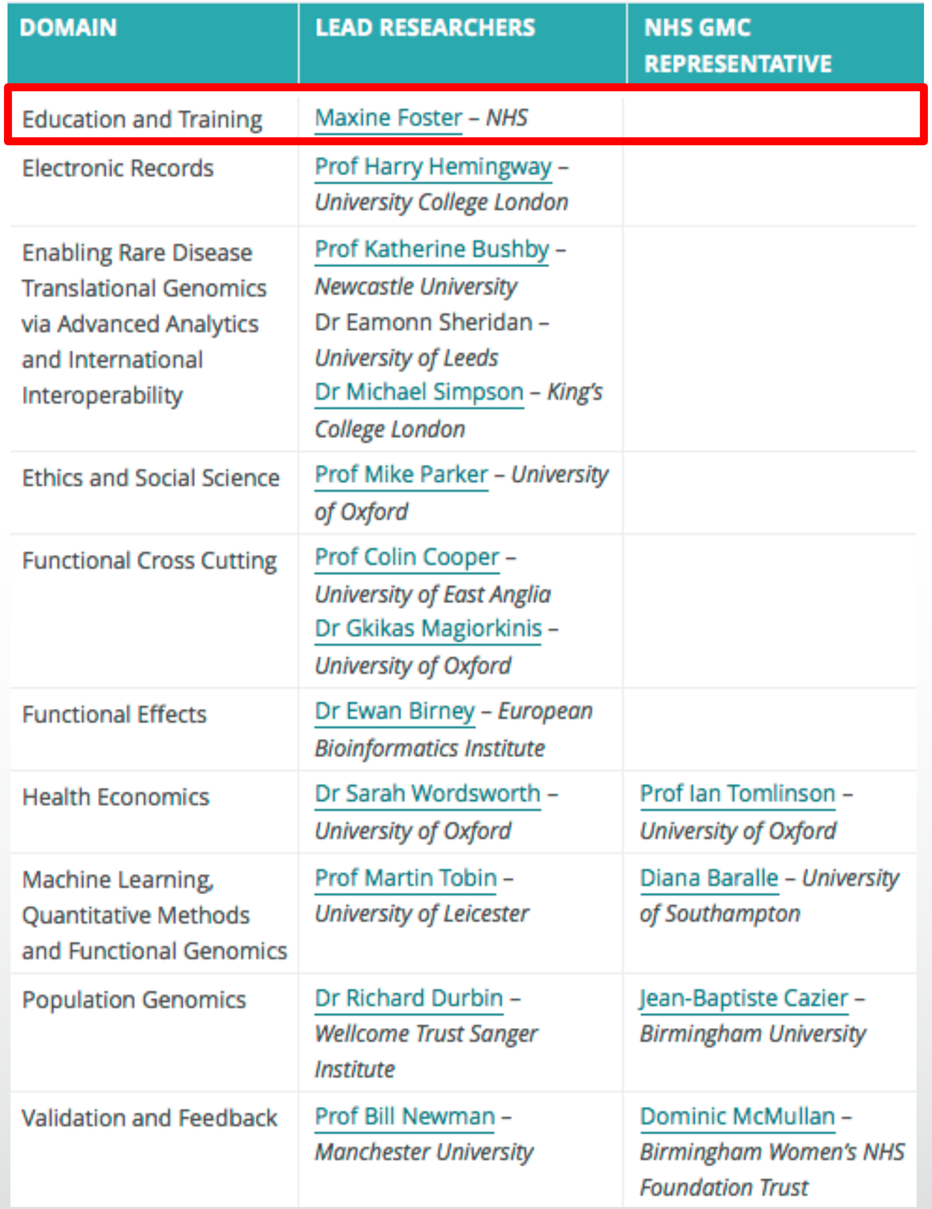


DOMAIN	LEAD RESEARCHERS	NHS GMC REPRESENTATIVE
Breast Cancer	Dr Nicholas Turner – *Institute of Cancer Research*	
Childhood Solid Cancers	Prof Josef Vormoor – *Newcastle University*	Dr Alex Henderson – *Newcastle upon Tyne NHS Foundation Trust*
Colorectal Cancer	Prof Ian Tomlinson – *University of Oxford*	Mohammad Ilyas – *The University of Nottingham*
Haematological Malignancy	Prof Anna Schuh – *University of Oxford*	Paresh Vyas – *University of Oxford*
Lung Cancer	Dr Charles Swanton – *The Francis Crick Institute*	Prof Adrienne Flanagan – *University College London*
Ovarian and Endometrial Cancer	Dr James Brenton – *NHS* Dr David Church– *University of Oxford*	
Pan Cancer	Prof Dion Morton – *University Hospitals Birmingham*	
Glioma	Prof Keyoumars Ashkan *King's College Hospital NHS Foundation Trust*	



DOMAIN	LEAD RESEARCHERS	NHS GMC REPRESENTATIVE
Glioma	Prof Keyoumars Ashkan *King's College Hospital NHS Foundation Trust*	
Upper gastrointestinal cancer	Prof John Bridgewater – *University College London*	
Renal cell carcinoma	Dr James Larkin – *Royal Marsden NHS Foundation Trust*	
Melanoma	Dr Paul Lorigan– *Christie NHS Foundation Trust*	
Testicular Cancer	Dr Andrew Protheroe – *Oxford University Hospitals NHS Foundation Trust*	
Cancer of Unknown Primary	Dr Harpreet Wasan *Imperial College London*	
Prostate Cancer	Prof Johann de Bono – *Institute of Cancer Research*	Dr Mark Linch – *University College London*
Renal Cell Carcinoma	Dr James Larkin – *Royal Marsden NHS Foundation*	Dr Anna Koziell – *King's College London*
Sarcoma	Prof Adrienne Flanagan – *University College London*	

Cross Cutting GeCIPs



DOMAIN	LEAD RESEARCHERS	NHS GMC REPRESENTATIVE
Education and Training	Maxine Foster – *NHS*	
Electronic Records	Prof Harry Hemingway – *University College London*	
Enabling Rare Disease Translational Genomics via Advanced Analytics and International Interoperability	Prof Katherine Bushby – *Newcastle University* Dr Eamonn Sheridan – *University of Leeds* Dr Michael Simpson – *King's College London*	
Ethics and Social Science	Prof Mike Parker – *University of Oxford*	
Functional Cross Cutting	Prof Colin Cooper – *University of East Anglia* Dr Gkikas Magiorkinis – *University of Oxford*	
Functional Effects	Dr Ewan Birney – *European Bioinformatics Institute*	
Health Economics	Dr Sarah Wordsworth – *University of Oxford*	Prof Ian Tomlinson – *University of Oxford*
Machine Learning, Quantitative Methods and Functional Genomics	Prof Martin Tobin – *University of Leicester*	Diana Baralle – *University of Southampton*
Population Genomics	Dr Richard Durbin – *Wellcome Trust Sanger Institute*	Jean-Baptiste Cazier – *Birmingham University*
Validation and Feedback	Prof Bill Newman – *Manchester University*	Dominic McMullan – *Birmingham Women's NHS Foundation Trust*

Urgent announcement about Genomics England Training Embassy

On 2nd December, Genomics England told us that as students in the Genomic Medicine programme, you can gain access to the Training Embassy for the 100KGP. We hope to take a look round it during your next teaching days for informatics. (In case you are wondering what an embassy is: it's the extremely secure virtual desktop which will be used by everyone accessing 100KGP. It contains clinical and genomic data, and informatic and other tools. All MSc personnel are automatically members of a Training and Education GECIP. The Embassy structure and contents are being finalised right now so we don't know much more detail, but here's a link for it… http://www.genomicsengland.co.uk/faqs-about-gecip/)

Access to the embassy is granted uniquely to each individual, and tracked by HEE.

The details needed are

First Name / Surname

Institutional Email address (your student email address)

UK Mobile number

The Faculty of Medicine office holds the mobile numbers you gave when you first joined the programme. If you are agreeable, the office can just capture those and send them en bloc to HEE, and you will then be contacted individually to join the Embassy.

Please contact GenomicMedicine@southampton.ac.uk by 9am Weds 9th December IF YOU DO NOT WANT YOUR DETAILS PASSED ON, OR IF YOUR MOBILE NUMBER HAS CHANGED.

If we don't hear from you we will assume you are content.

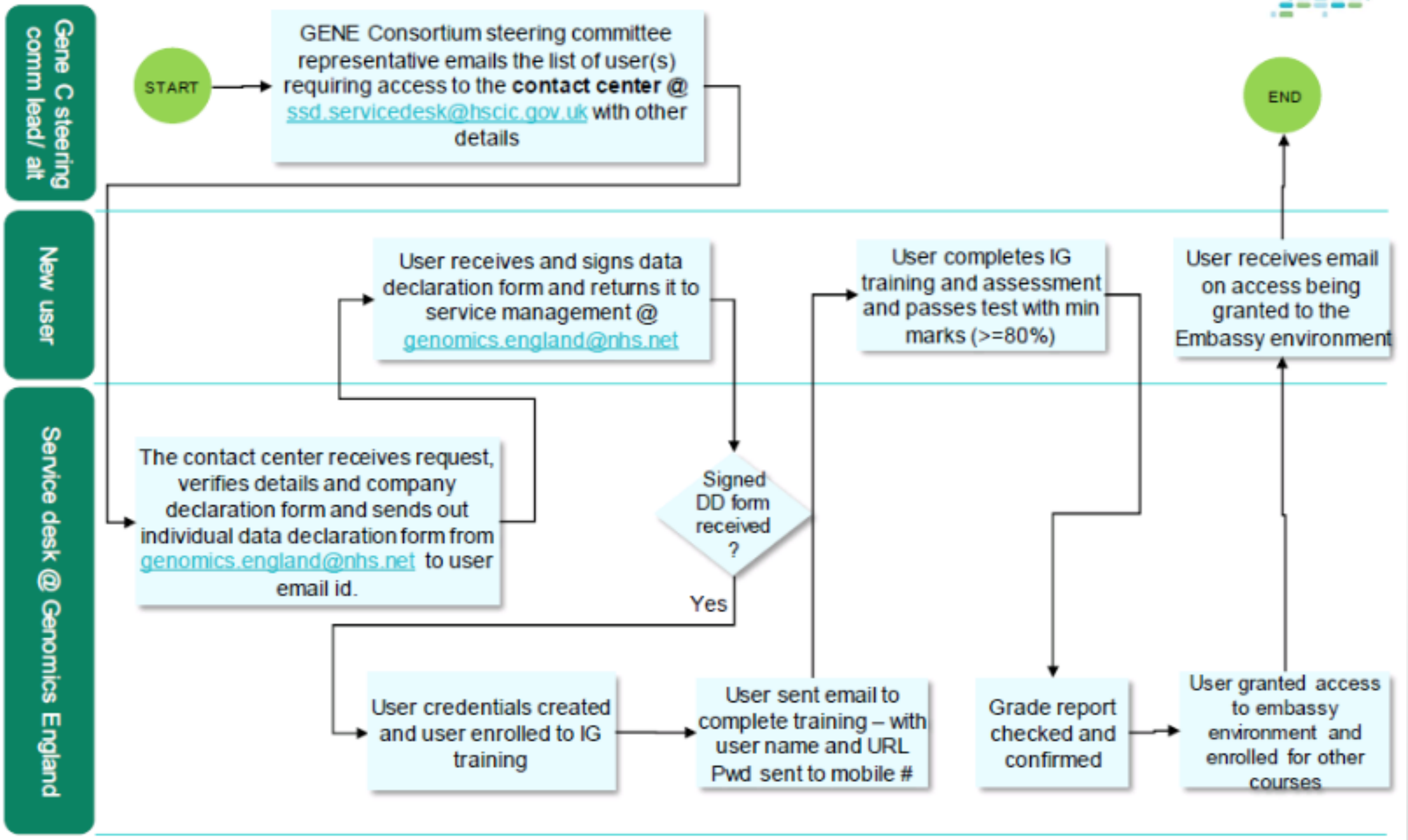
(Just so you know – new arrivals must complete some additional documents and training; details are online and should be presented when you first log on.)

with best wishes

Deborah

New user Access request process

Genomics england



Gene C steering comm lead/ alt

New user

Service desk @ Genomics England

START

GENE Consortium steering committee representative emails the list of user(s) requiring access to the **contact center @** ssd.servicedesk@hscic.gov.uk with other details

The contact center receives request, verifies details and company declaration form and sends out individual data declaration form from genomics.england@nhs.net to user email id.

User receives and signs data declaration form and returns it to service management @ genomics.england@nhs.net

Signed DD form received ?

Yes

User credentials created and user enrolled to IG training

User sent email to complete training – with user name and URL Pwd sent to mobile #

User completes IG training and assessment and passes test with min marks (>=80%)

Grade report checked and confirmed

User granted access to embassy environment and enrolled for other courses

User receives email on access being granted to the Embassy environment

END

Access to the Embassy – process flow

1. Validated list of students and instructors (users) sent to Service Desk via email to ssd.genomics@nhs.net .

2. Service desk issues Embassy log in details (username/password) via e-mail. These details are used to access all tools and applications within the Embassy (e.g. LabKey, Confluence, JIRA).

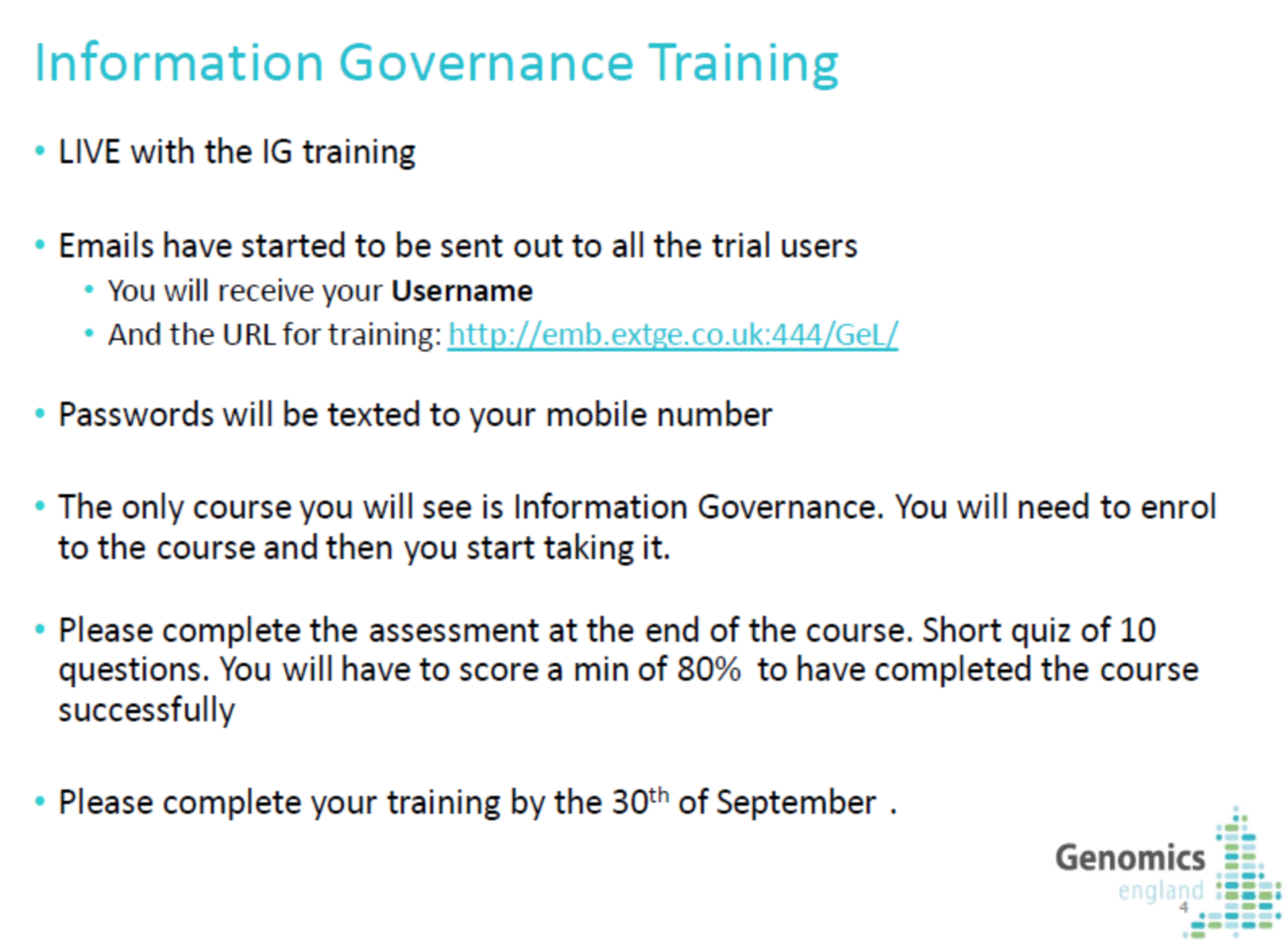
3. Log in details allow users to complete the Information Governance (IG) training.

4. All users must notify Service Desk of completed IG training via email to ssd.genomics@nhs.net.

5. Service desk grants access to the Embassy. Users can log in.

6. On entering embassy, users open LabKey and sign in using provided login details (not email address). At this point no data is visible.

7. Users need to alert Service Desk that they have logged into LabKey (ssd.genomics@nhs.net). Service desk then grant permission to view all LabKey project data.



Information Governance Training

- LIVE with the IG training
- Emails have started to be sent out to all the trial users
  - You will receive your **Username**
  - And the URL for training: http://emb.extge.co.uk:444/GeL/
- Passwords will be texted to your mobile number
- The only course you will see is Information Governance. You will need to enrol to the course and then you start taking it.
- Please complete the assessment at the end of the course. Short quiz of 10 questions. You will have to score a min of 80% to have completed the course successfully
- Please complete your training by the 30th of September .

Genomics england

Information Governance

Data Access and Acceptable Uses Policy

1. All principles expressed in this document are based on the current version of the 100,000 Genomes Project Protocol and may be amended as the programme evolves.

2. Activity within the 100,000 Genomes Project is based on the donation of samples and associated consent for sample storage, DNA sequencing, and other analysis of the participants' tissue and clinical data for the purposes of processing by research and commercial users.

3. Genomics England will process and deliver decisions on data sharing and access requests keeping in mind the public interest, scientific utility, the corresponding consent, and wider Genomic England policies.

4. Genomics England intends to offer comprehensive access to enable leading edge outputs from analysis of the data that are expected to have both scientific research value and clinical value particular to the participants whose data is involved.

5. For each request, Genomics England will determine the appropriate cohort of genomic and associated data to make available within the approved trusted environment.

6. Data access will only be granted to users validated by Genomics England, using traceable IP addresses, who have a data-sharing contract or are approved staff of Genomics England.

7. Genomics England will have procedures for monitoring user activities and interaction with the data, and for the management and escalation of suspected non-compliance.

8. Decisions relating to data access and data sharing will be made considering the policy provisions given in this document, as well as further detailed in corresponding SOPs.

9. This policy and associated SOPs shall be made available to all Genomics England staff; and corresponding training and awareness activity relevant to their role shall be provided as appropriate.

10. Genomics England shall create and operate an Access Review Committee (ARC). This shall include an independent chair, and appropriate independent experts in areas such as genomic medicine and ethics. It will receive advice and input as needed from internal committees of Genomics England such as the Ethics Advisory Committee and the Data Advisory Committee. A member of Genomics England staff will support the Committee including a Senior Information Risk Owner who may call upon the advice of a Caldicott Guardian for all data held.

11. Genomics England will process access applications and provide decisions via the independent ARC on behalf of the Genomics England Executive Board.

12. Genomics England will utilise best practice guidance and privacy enhancing technologies to provide de-identification of data, meet anonymisation standards, and minimise the risks of inadvertent disclosure.

13. Compliance with the policy shall be audited and verified as required. Non-compliance shall be reported to the Genomics England Programme Board and appropriate contractual and legal action taken. Actionable decisions in the case of a breach will be under the remit of the Caldicott Guardian.

14. Arrangements for archiving of the data when it is no longer required will be set out in the Genomics England Information Quality and Records Management Policy.

15. Parties will sign legally binding data access contracts with Genomics England that outline the terms of access and processing information, which includes as a breach of contract any attempts to re-identify participants.

16. Users will have a facility to introduce new data sets into the Data Centre for wider research purposes as defined in Section 8 of the Genomic England Protocol document. This data will then be subject to the same rigorous data access, sharing and acceptable uses procedure as set out in this document. Genomics England will ensure compliance with governance requirements across the pipeline regardless of data source to safeguard data confidentiality and participant privacy.

Caldicott Principles

1) Justify the purpose(s)

Every single proposed use or transfer of patient identifiable information within or from an organisation should be clearly defined and scrutinised, with continuing uses regularly reviewed, by an appropriate guardian.

2) Don't use patient identifiable information unless it is necessary

Patient identifiable information items should not be included unless it is essential for the specified purpose(s) of that flow. The need for patients to be identified should be considered at each stage of satisfying the purpose(s).

3) Use the minimum necessary patient-identifiable information

Where use of patient identifiable information is considered to be essential, the inclusion of each individual item of information should be considered and justified so that the minimum amount of identifiable information is transferred or accessible as is necessary for a given function to be carried out.

4) Access to patient identifiable information should be on a strict need-to-know basis

Only those individuals who need access to patient identifiable information should have access to it, and they should only have access to the information items that they need to see. This may mean introducing access controls or splitting information flows where one information flow is used for several purposes.

5) Everyone with access to patient identifiable information should be aware of their responsibilities

Action should be taken to ensure that those handling patient identifiable information - both clinical and non-clinical staff - are made fully aware of their responsibilities and obligations to respect patient confidentiality.

6) Understand and comply with the law

Every use of patient identifiable information must be lawful. Someone in each organisation handling patient information should be responsible for ensuring that the organisation complies with legal requirements.

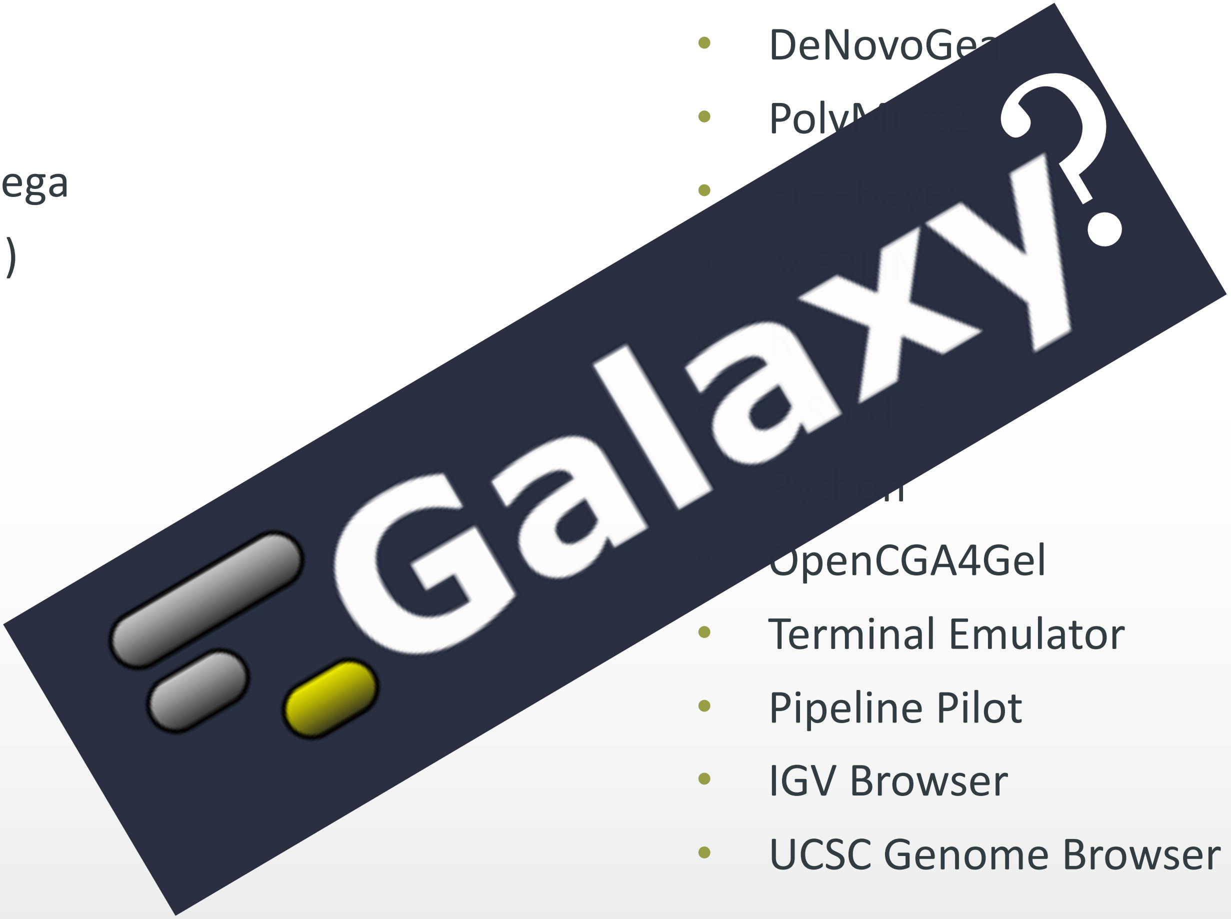
7) The duty to share information can be as important as the duty to protect patient confidentiality

Professionals should in the patient's interest share information within this framework. Official policies should support them doing so.

Content

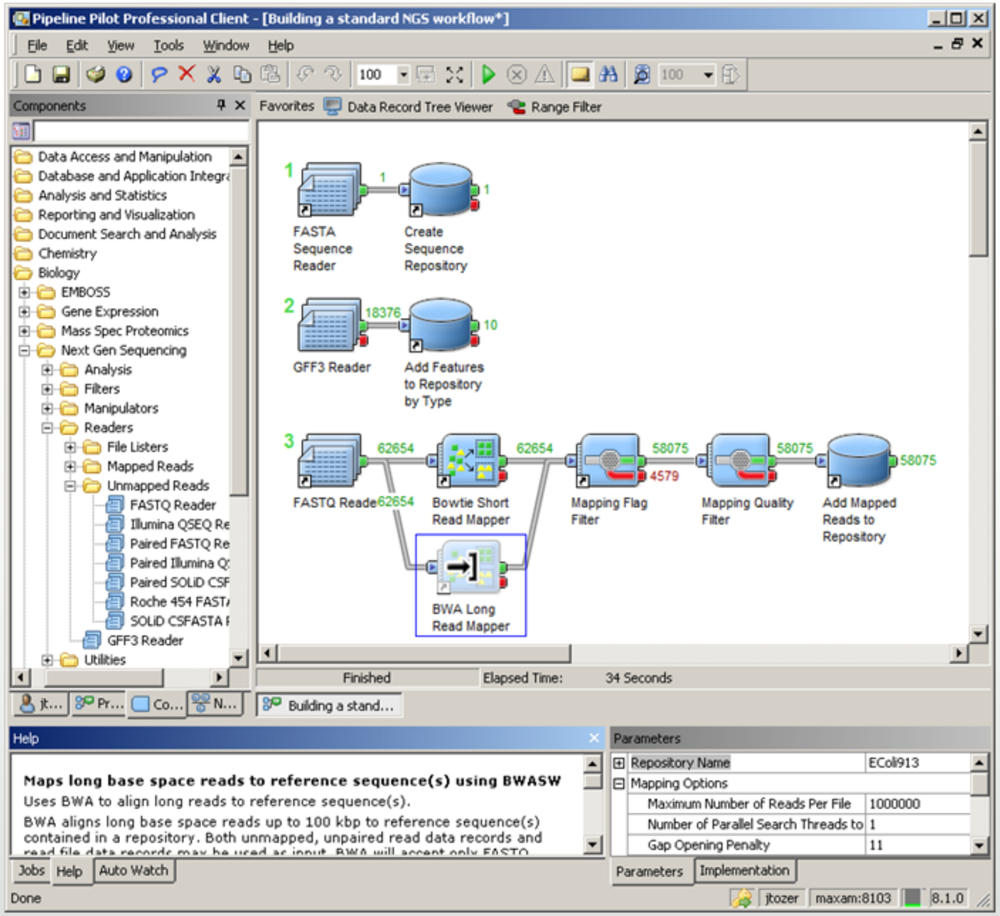
- GATK
- Picard
- Samtools
- Clustal Omega
- Blast (NCBI)
- bfast
- HMMER
- vcftools
- LoFreq
- BWA
- bedtools
- Genscan
- FASTQC
- xBrowse

- ANNOVAR
- DeNovoGear
- PolyMutt2
- FreeBayes
- MERLIN
- R
- RStudio
- Python
- OpenCGA4Gel
- Terminal Emulator
- Pipeline Pilot
- IGV Browser
- UCSC Genome Browser



Pipeline Pilot

- Pipeline Pilot is a commercially developed workflow environment (like Galaxy) that will be incorporated into Genomic England.
- Pipeline Pilot enables scientists to create, test and publish scientific services that automate the process of accessing, analyzing and reporting scientific data.
- Anyone can create scientific protocols that provide access to research data locked in silos (like Genomic England). These protocols automate the scientific analysis of the data and enable researchers to rapidly explore, visualize and report results.
- Pipeline Pilot allows sharing workflows



Finished Elapsed Time: 34 Seconds

Maps long base space reads to reference sequence(s) using BWASW

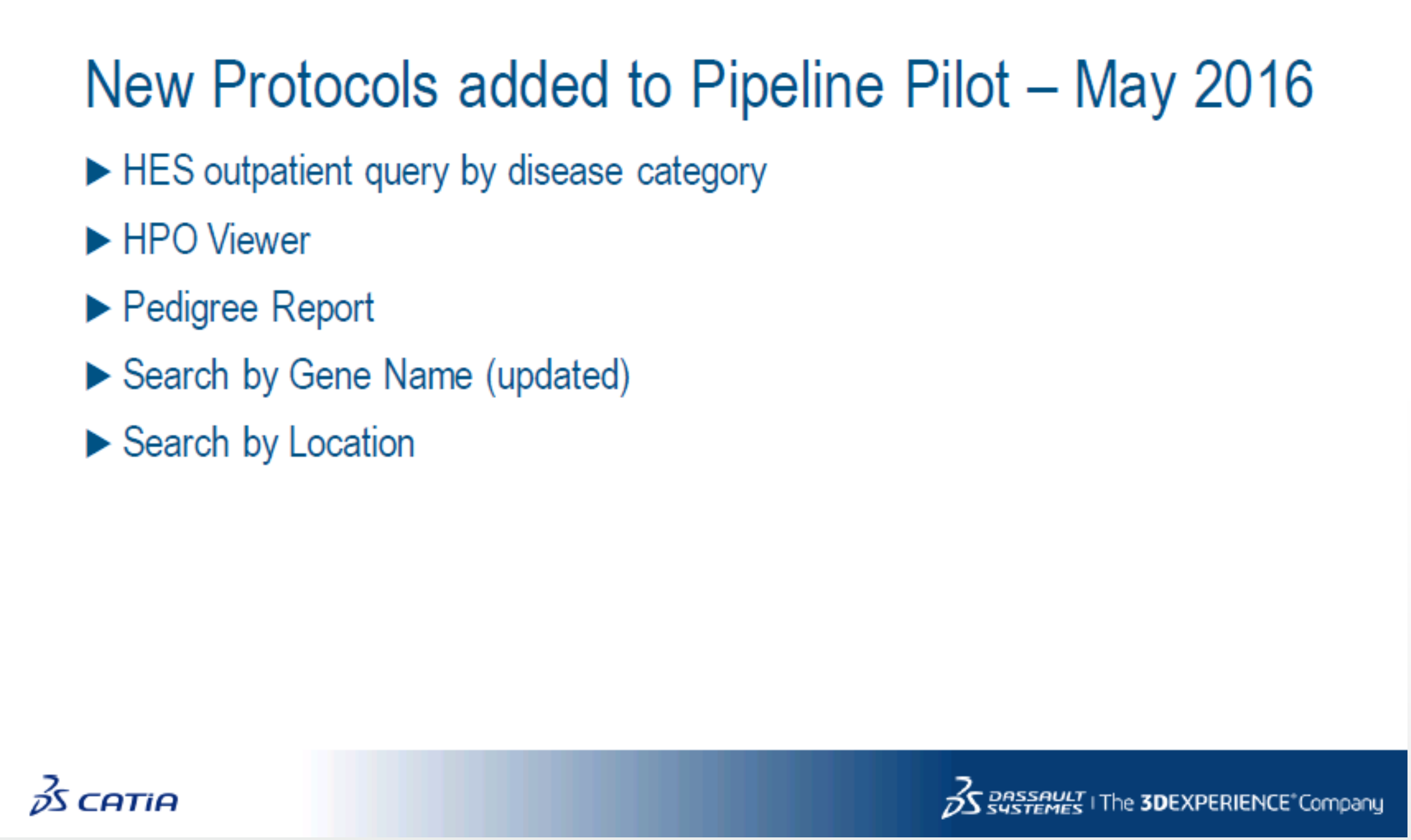
Uses BWA to align long reads to reference sequence(s).

BWA aligns long base space reads up to 100 kbp to reference sequence(s) contained in a repository. Both unmapped, unpaired read data records and read file data records may be used as input. BWA will accept only FASTQ

⊞ Repository Name	EColi913
⊟ Mapping Options	
Maximum Number of Reads Per File	1000000
Number of Parallel Search Threads to	1
Gap Opening Penalty	11

Done jtozer maxam:8103 8.1.0

Pipeline Pilot Protocols



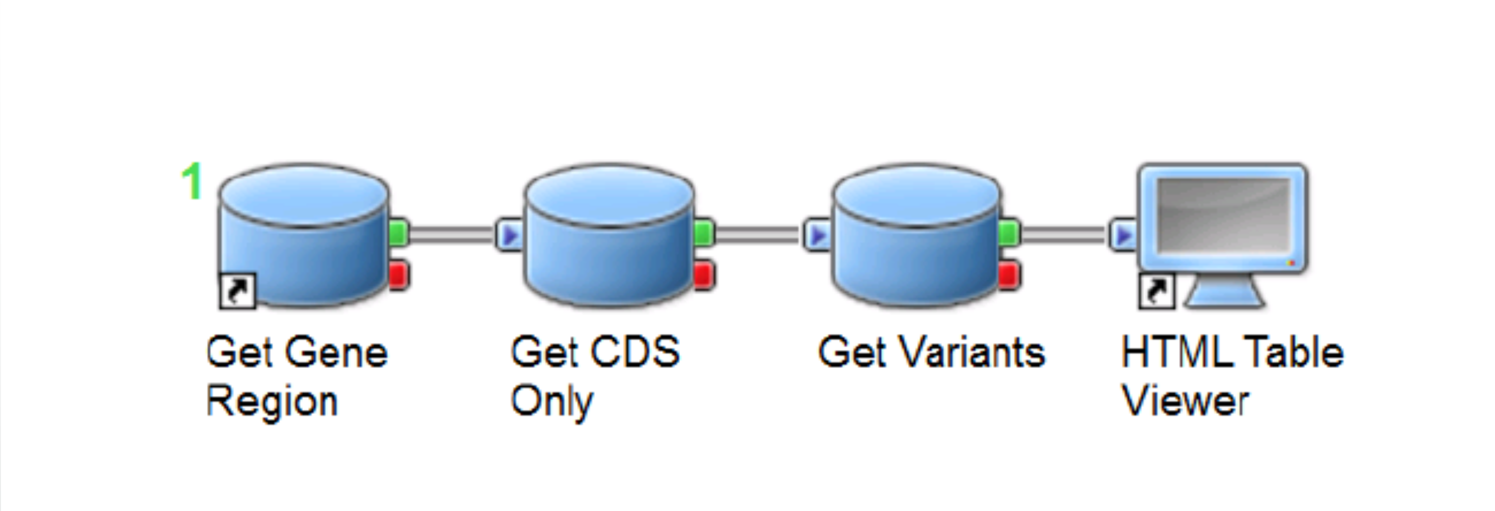
New Protocols added to Pipeline Pilot – May 2016

- HES outpatient query by disease category
- HPO Viewer
- Pedigree Report
- Search by Gene Name (updated)
- Search by Location

Pipeline Pilot Query

Basic NGS Query Example

- Want to get the variants from within a single gene
- But only want those within exons



Industry Partners

- AbbVie
- Alexion Pharmaceuticals
- AstraZeneca
- Biogen
- Dimension Therapeutics
- GSK
- Helomics
- Roche
- Takeda
- UCB

PROBLEMS

- **Known Knowns**

  Difficulty in getting access to the Training Embassy.

  Galaxy not currently loaded as a software option.

- **Known Unknowns**

  Each GeCIP member will have a finite number of Central processing unit (CPU) hours per month. The exact allocation is yet to be defined, and may change as the project and datacentre mature. Use of CPU beyond this allocation will require payment.

  How to access patient data?

- **Unknown Unknowns**

  What happens with IP and publication when academic/clinician and industry partner are independently working on the same data?

SUMMARY

- **LEARNING OUTCOME**

  Practice in examples of analysis of genomic data in the Training Embassy within the Genomics England Data Centre.

- **INDICATIVE CONTENT**

  Gain practical experience of the bioinformatics pipeline through the Genomics England programme.

UNIVERSITY OF Southampton

THE END