The 100,000 Genomes Project: Engaging the research community

Tim Hubbard @timjph
Genomics England
King’s College London, King’s Health Partners
Wellcome Trust Sanger Institute

DNAdigest’s Symposium:
How can you implement open science in human genomics
22nd November 2014
Linking Health data to Research

Clinical Data World

Electronic Health Record

Healthcare Professional

Phenotype
Electronic Health Records

Genotype and Phenotype relationship capture

Human sequence data repositories

Genomic Biology Data World

Reference genome sequence
~3 gigabytes

EBI: repositories (petabytes of genome sequence data)
Sanger: sequencing (1000 genomes, uk10K)
Value of cross disease analysis

Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients

Anders Boeck Jensen\textsuperscript{1,2}, Pope L. Moseley\textsuperscript{2,3}, Tudor I. Oprea\textsuperscript{1,3,4}, Sabrina Gade Ellesøe\textsuperscript{2}, Robert Eriksson\textsuperscript{1,2}, Henriette Schmock\textsuperscript{5}, Peter Bjødstrup Jensen\textsuperscript{2}, Lars Juhl Jensen\textsuperscript{2} & Søren Brunak\textsuperscript{1,2}
Figure 3 (b): Cerebrovascular cluster with epilepsy as key diagnosis

Steps in UK towards E-Health Research, Genomic Medicine

• Health data to Research
  – 2006 Creation of OSCHR
    • Increase coordination between funders: MRC and NIHR
  – 2007 OSCHR E-health board
    • Enable research access to UK EHR data
    • Build capacity for research on EHR data

• Genomics to Health
  – 2009 House of Lords report on Genomic Medicine
  – 2010 Creation of Human Genomic Strategy Group (HGSG)
2011: UK Life Sciences Strategy

Linking Health data to Research

Clinical Practice Research Datalink (CRPD)

Farr Institute

Genomic Biology Data World

EBI: repositories (petabytes of genome sequence data)

Sanger: sequencing (1000 genomes, uk10K)

Reference genome sequence ~3 gigabytes

Human sequence data repositories

Genotype and Phenotype relationship capture

Clinical Data World

Electronic Health Record

Healthcare Professional

Phenotype

Electronic Health Records
2012: Human Genome Strategy Group report
UK Life Science Strategy Update; 100K Genomes

DH: http://www.dh.gov.uk/health/2012/01/genomics/
BIS: http://www.gov.uk/office-for-life-sciences/
Linking Health data to Research

Clinical Data World

Electronic Health Record

Farr Institute

Phenotype
Electronic Health Records

Genotype and Phenotype relationship capture

Human sequence data repositories

Healthcare Professional

Genomic Biology Data World

EBI: repositories (petabytes of genome sequence data)
Sanger: sequencing (1000 genomes, uk10K)

Reference genome sequence
~3 gigabytes

Clinical Practice Research Datalink (CRPD)
Genomics England launched, mapping DNA to better understand cancer, rare and infectious diseases

http://www.genomicsengland.co.uk/
Genomics England - mission

• 100,000 whole genome sequences in NHS patients with rare diseases and cancers from the NHS in England
• Generate health and wealth
• Legacy of infrastructure, human capacity and capability
• Enable large scale genomics research
Scale compared to existing WGS

- 1000 genomes and UK10K
  - low coverage genomes (~4x illumina)
- Limited number of ‘clinical grade’ WGS
  - TCGA: ~700
  - ICGC: ~700
  - WGS 500: 500
Now is the moment to commit to WGS

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Large-scale structural changes</th>
<th>Balanced translocations</th>
<th>Distant consanguinity</th>
<th>Uniparental disomy</th>
<th>Novel/known coding variants</th>
<th>Novel/known non-coding variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targeted gene sequencing</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>SNP array</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Array CGH</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Exome</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Whole Genome</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
WGS500 Results

- 7 Novel genes for disease
- 6 Novel phenotypes for known genes
- 2 pathogenic regulatory variants in or downstream of known candidate genes
- 6 genes missed by prior Sanger Sequencing

WTCHG, Oxford: http://www.well.ox.ac.uk/wgs500

MENDELIAN

Of 95 families, to date
- 23 families have new clinical diagnosis
  - NB pre-screened for known genes
  - result will increase with follow-up
- 74 families in follow up studies
- Over 50% of these have strong lead candidate
Rare Diseases – inclusion criteria

- Rare disease with residual unmet diagnostic need with a proband from the NHS in England
- Evidence of previous genetic testing
- Prospective collection but will include legacy collections
- Family structures: ideally parent-offspring trios
- Provision of Genomics England informed consent
- Availability of clinical and further phenotypic data
- Blood samples for DNA extraction and multi-omic samples
Cancer – inclusion criteria

• Lung, breast, colon, prostate, ovary, some leukaemias
• Rare and childhood cancers, unknown primary
• Provision of Genomics England informed consent
• Availability of clinical and further phenotypic data
• Tumour DNA primarily from FFPE samples
Feedback to the NHS

• Diagnostic reports that are accessible and meaningful
• Dynamic serial reporting - evolving findings

**Primary** findings:
  • Known pathogenic and expected pathogenic variants on known genes

**Secondary** “looked for” findings (currently for 10 conditions):
  • Strong cancers predisposition and familial hypercholesterolemia
  • For example Lynch syndrome, BRCA1/2, multiple endocrine neoplasia (MEN1), VHL

**Carrier states** of reproductive importance (currently for 12 conditions):
  • Thalassemia, sickle cell, hemophilia A, …

• Read and variant level data accessible to NHS referring teams
• Patients can request genomic data files from Genomics England
• Patients are consented to be contacted up to four times a year

Feedback to the NHS
Genomics England Pilots

• **Phase 1** - Sequencing and Annotation Competition
  - 4 providers 15 samples (5 tumour – normal pairs and 5 germline)
  - Testing Sequencing QA and annotation

• **Phase 2a-2000 Rare Inherited Disease WGS** - 30x depth – over 2014
  - Partnering NIHR BioResource and Translational Research Collaborative
  - 5 centres - 928 samples since end of November- 1st 96 are in sequencing.

• **Phase 2b- 3000 Cancer Patients** (Lung, Breast, Ovary, Prostate & Colon)
  - Somatic (?50-80x) and germline (30-40x) – tendering now
  - Optimise Molecular Pathology pipeline
  - 11 CRUK Centres and BRCs

• **Pathogens will be with Public Health England**
Process Overview

Sample DNA → Sequence (BAM) → Variants (VCF) → Candidate Variants → Clinical Interpretation → Clinical Action
Process Overview

Sample DNA → Sequence (BAM) → Variants (VCF) → Procured Sequence

Variants (VCF) → Candidate Variants → Clinical Interpretation

Clinical Interpretation → Procured Annotation

Procured Annotation → Sequence Validation → Clinical Action

GeL Database

NHS
Sequencing and Annotation assessment

• Sequencing bake-off
  – Samples sent to participants; returned sequence assessed
  – Evaluation on quality and coverage
  – Informed sequencing contract

• Annotation bake-off
  – Sequence sent to participants (BAM+VCF)
    • Rare diseases: trio
    • Cancer: germline + tumour
  – Harder than assessing sequencing
  – Gold standard less well defined
  – Lack of established data standards
Implementation of Main Programme

- **Sequencing**: contract signed with Illumina; Wellcome Trust Sequencing Centre Building at Hinxton
- **Samples**: Tender for NHS Genomic Medicine Centres
- **Data**: Award from MRC to build datacentre for GeCIP
- **Research**: Genomics England Clinical Interpretation Partnership (GeCIP) launched

- Biorepository to be established
- Expected WGS volume: 20k in 2015; 30k in 2016; 40k in 2017 (Elasticity in the pipeline)
Genomics England – Proposed data flows

Patient Consent

Clinical Genetics, Cancer & Public Health.
NHS Trusts, Patients & Public

Pilots: Selected Centres, CRUK, BRCs

Main Program: Genomic Medicine Centres

Sequencing Centres -> Sample repository

Refreshable identifiable Clinical Data, linked to anonymised Whole Genome Sequence

Annotation ‘Apps’

Safe haven: Anonymised Clinical data and DNA sequence

Clinicians & Academics -> Industry

Sample

EHR Primary Care Hospital episodes

Clinical Report

GeCIP

Fire wall
Patient data stays on NHS side

Only processed results pass outside
Call for Expressions of Interest

The Genomics England Clinical Interpretation Partnership

Genomics England invites ‘expressions of interest’ from UK led consortia of clinicians, researchers, analysts and those in training to propose disease specific domains in the areas of rare inherited disease, cancer and infectious disease. The Genomics England Clinical Interpretation Partnership will lead research to enhance the clinical interpretation of whole genome sequences and support the delivery of healthcare transformation from the 100,000 Genomes Project.

This will be the route by which Genomics England will engage with the UK academic and healthcare community and their international collaborators to discover new biological insights into disease, elucidate functional impact, develop novel analytical approaches and create high cadre expertise in genomic medicine.

The overall aim is for the Genomics England Clinical Interpretation Partnership to create thriving, sustainable communities of research and clinical (NHS) disease experts to interrogate the 100,000 whole genome sequences. The domains within Genomics England Clinical Interpretation Partnership will have three primary roles:

• **Research:** Harnessing opportunities for research and discovery enabled by the 100,000 Genomes Project with the intention of further enhancing our understanding of genomic medicine and its application in healthcare.
Genomics England Clinical Interpretation Partnership - GECIP

- Drive up the fidelity of clinical interpretation of WGS
- UK-led and self-organised into domains
- Partnership with researchers, the NHS and Trainees.
- Possible formation of a precompetitive consortium of industry partners.
- Designed to accelerate academic/industry partnership and development of diagnostics and therapies.
- All data generated contributes to the Genomics England Dataset and are available to all.
- Recognises that to get to a therapy will require significant additional R&D which we aim to stimulate in the UK.
<table>
<thead>
<tr>
<th>Disease-specific domains</th>
<th>Rare Inherited Diseases</th>
<th>Function-specific domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancers</td>
<td>Rarer Inherited Diseases</td>
<td>Advanced analytical methodologies</td>
</tr>
<tr>
<td>Breast</td>
<td>Cardiovascular disease</td>
<td>Interpretation, Validation and Feedback</td>
</tr>
<tr>
<td>Lung</td>
<td>Respiratory</td>
<td>Ethics and Social Science</td>
</tr>
<tr>
<td>Ovarian</td>
<td>Endocrine and metabolism</td>
<td></td>
</tr>
<tr>
<td>Colorectal</td>
<td>Gastroenterology</td>
<td></td>
</tr>
<tr>
<td>Prostate</td>
<td>Immunological diseases</td>
<td></td>
</tr>
<tr>
<td>Haematological malignancies</td>
<td>Neurological and degenerative disease</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Musculoskeletal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Renal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skin</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-malignant Haematology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Paediatric</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inherited Cancers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Severe response to infection</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rheumatology</td>
<td></td>
</tr>
</tbody>
</table>
Data Sharing

• Open to all
  – Human Genome Projects where subject consented: Hapmap, 1000 genomes
  – Repository: Genbank, ENA, DDBJ (INSDC)

• Managed distribution (must be *bona fide* researcher)
  – Genetic data for disease cohorts, with phenotypes
  – Repository: DbGaP, EGA (Encrypted distributions etc.)

• Managed access, no redistribution
  – Genomics England datasets
  – Repository: GeL Datacentre
A future with closed datasets

• Multiple sets of Hospital/National datasets with no redistribution policies

• Value for research in generating statistics across this global set
Global Alliance for Genomes and Health
http://genomicsandhealth.org/
National Information Board 13/11/14 report

https://www.gov.uk/government/organisations/national-information-board

Personalised Health and Care 2020
Using Data and Technology to Transform Outcomes for Patients and Citizens
A Framework for Action

November 2014
Genomics England

• 100,000 WGS of NHS patients

• Working with NHS, academics and industry to drive Genomic Medicine into the NHS

• Support that with education

• Leave a legacy of NGS Centres, sample pipeline and biorepository, large-scale data store that makes this usable by the NHS

• New diagnostics and therapies and opportunities for patients

• By end of 2017