Progress with Rare Diseases in the Genomics England 100,000 Genomes Project

Dr Anna C Need
GeCIP Team Lead, Genomics England
The 100,000 Genomes Project

Announced by the former Prime Minister in December 2012
An Olympic Legacy

Genomics England announced by Secretary of State for Health in speech during NHS 65th Anniversary Celebrations, July 2013

Opening of new Sequencing Centre in 2016

CMO’s Generation Genome and the Life Sciences report in 2017
The 100,000 Genomes Project

- **100,000** genomes
- **70,000** patients and family members
- **21** Petabytes of data. 1 Petabyte of music would take 2,000 years to play on an MP3 player.
- **13** Genomic Medicine Centres, and **85** NHS Trusts within them are involved in recruiting participants
- **1,500** NHS staff (doctors, nurses, pathologists, laboratory staff, genetic counsellors)
- **2,500** researchers and trainees from around the world

24 May 2018
The infrastructure for delivery

- Nationwide network of 13 NHS Genomic Medicine Centres – each serving ~3-5 million population
- Includes over 85 hospitals across England
- Integrated with genetic laboratories, genetic services and local pathology laboratories
- Scotland, NI and Wales also now part of the Project
How the 100,000 Genomes Project works

Patient consent

Samples + Clinical Data + Longitudinal Data

Biorepository

Sequencing Centre

Genomics England Informatics Architecture

Clinicians

GeCIP
Scientific and Clinical Users

Discovery Forum
Industry Users

24 May 2018
What information can be fed back?

• Information about a patient’s main condition

• Information about additional ‘serious and actionable’ conditions (optional)

• Carrier status for non affected parents of children with rare disease (optional)
Primary clinical data collection

• Core clinical data set:
  • Disease status (bespoke)
  • Pedigree data (Panogram)
  • Human Phenotyping Ontology
  • ICD10, SNOMED CT, OMIM

• Clinical test data where relevant and not captured by HPO
  • Using established standards wherever possible

Additional data emulating test request in clinical practice

Gene panels
Penetrance settings
Other bespoke elements
  e.g. >1 analysis for family
PanelApp
https://panelapp.genomicsengland.co.uk/

Genomics England PanelApp
A crowdsourcing tool to allow gene panels to be shared, downloaded, viewed and evaluated by the Scientific Community

PanelApp Update 30th October 2017
Thank you to all our Reviewers & Curators

Email: panelapp@genomicsengland.co.uk
The variant allele is not commonly found in the general healthy population.
Allelic state matches known mode of inheritance for the gene and disorder.
Familial segregation (where applicable)

- Yes
  - Known Pathogenic
    - Is the variant in a gene in the Virtual Gene Panel (green list) for that disorder?
      - Yes: Tier 1
      - No: Tier 3
  - Protein truncating
    - Is the variant in a gene in the Virtual Gene Panel (green list) for that disorder?
      - Yes: Tier 2
      - No: Tier 3

- No
  - Other consequence
Semi-automated Interpretation pipeline

- DNA
- Phenotypes & Pedigree
  - Patient/family
  - Annotated VCFs
  - Gene Panel
    - Variant filtering
- Tiered variants
- Review
- Clinical reporting
  - PanelAssigner
  - PanelApp
  - Report QA
  - Decision support tool
  - Validation Outcomes

WORKFLOW MANAGER

DATA DISTRIBUTION FRAMEWORK

PanelAssigner
PanelApp
Report QA
Decision support tool
Validation Outcomes

Annotation Companies

FABRIC
CONGENICA
WuxiNextCODE
Genomics England
Interpretation at the hospitals

1. ‘Tiering’
   – **Automated** focused initial analysis set up prior to interpretation
   – **Aims to mirror standard diagnostic analysis**
   – ‘Tier 1 or 2’ – variants in 0-5 genes within prescribed panel(s) (median = 1.2)
   – Non-penetrance pipeline can be run

2. **Broader analysis**
   – Decision support software allows hospitals further, bespoke analysis
   – ‘Tier 3’ – standard GeL pipeline but not restricted to gene panels; 20-100s variants (median = 285)
   – Other tools, dependent on which CIP system

3. **Lab-clinical team curate variants and record outcomes**
   – Curations (including ACMG classifications) saved to central knowledge base
   – Record clinical impacts of diagnostic result
Progress to date

**Samples**
- **85,898** Samples collected from NHS GMCs
  - **19,451** cancer
  - **66,447** rare disease

**Genomes**
- **60,679** Genomes sequenced
  - **11,407** cancer
  - **49,272** rare disease

**Analysis and Reports**
- **11,883** Reports for families sent to NHS GMCs
- **24,596** Equivalent to genomes
- **20-25%** actionable findings

- **5,677** genomes since last month

Figures as at 09/05/2018

24 May 2018
Recruitment

- Average ~500 participants / week
- Rare disease recruitment expected to end 30th September 2018
- Mean family size = 2.3
Recruitment patterns

Recruitment pattern by age

Age profile of probands
44% of probands are <18 years of age
Recruitment - West Midlands

- Cardiovascular disorders (716)
- Ciliopathies (39)
- Dermatological disorders (92)
- Dyssomorphic and congenital abnormality syndromes (55)
- Endocrine disorders (116)
- Gastroenterological disorders (24)
- Growth disorders (13)
- Haematological and immunological disorders (78)
- Hearing and ear disorders (91)
- Infectious diseases (0)
- Intellectual disability (565)
- Metabolic disorders (66)
- Neurology and neurodevelopmental disorders (418)
- Ophthalmological disorders (75)
- Psychiatric disorders (0)
- Urinary tract disorders (1037)
- Respiratory disorders (10)
- Rheumatological disorders (53)
- Skeletal disorders (103)
- Tumour syndromes (99)
- Ultra-rare disorders (301)

- Recruited probands
- Median recruitment
- Best-recruiting GMC
- Other GMC
- At or below median recruitment
- Higher than median recruitment
Recruitment - West Midlands

Weekly proband consent rate for West Midlands from 1st Jul 2016 onwards

Weekly proband consent rate to the RD programme from 1st Jul 2016 onwards
## Diagnostic yield in different disease areas

<table>
<thead>
<tr>
<th>Disease</th>
<th>SINGLETONS No. families</th>
<th>SINGLETONS Diagnostic yield (%)</th>
<th>TRIOS No. families</th>
<th>TRIOS Diagnostic yield (%)</th>
<th>ALL No. families</th>
<th>ALL Diagnostic yield (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intellectual disability</td>
<td>8</td>
<td>25</td>
<td>38</td>
<td>39.5</td>
<td>63</td>
<td>36.5</td>
</tr>
<tr>
<td>Rod-cone dystrophy</td>
<td>9</td>
<td>11.1</td>
<td>32</td>
<td>53.1</td>
<td>52</td>
<td>42.3</td>
</tr>
<tr>
<td>Renal tract calcification</td>
<td>33</td>
<td>6.1</td>
<td>1</td>
<td>0</td>
<td>38</td>
<td>5.3</td>
</tr>
<tr>
<td>Non-CF bronchiectasis</td>
<td>21</td>
<td>4.8</td>
<td>7</td>
<td>0</td>
<td>32</td>
<td>3.1</td>
</tr>
<tr>
<td>CAKUT</td>
<td>25</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>29</td>
<td>6.9</td>
</tr>
<tr>
<td>Multiple endocrine tumours</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>Severe multi-system atopy</td>
<td>4</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>21</td>
<td>0</td>
</tr>
</tbody>
</table>
Impact of family structure on numbers of tiered variants

Autosomal recessive

<table>
<thead>
<tr>
<th>1 member</th>
<th>2 members</th>
<th>3 members</th>
<th>4 members</th>
</tr>
</thead>
</table>

Autosomal dominant

<table>
<thead>
<tr>
<th>1 member</th>
<th>2 members</th>
<th>3 members</th>
<th>4 members</th>
</tr>
</thead>
</table>

24 May 2018
Improving annotation of candidates

Tiering:
Classification (filtering)
• >90% precision for tier 1 variants (some ascertainment bias)
• >80% recall for tiered variants (1+2+3)

Exomiser:
Prioritisation
• Recovers 65% of the tier 3 diagnoses
• Recovers 57% of the untiered diagnoses

Variants highlighted by tiering and Exomiser have >90 sensitivity and various levels of specificity
A pipeline for diagnostics

Validation, QA and accreditation

- Major focus is the validation, QA and accreditation to ready out platforms for use as part of the NHS England Genomic Medicine Service from October 2018
- This involves analysis of samples already held as part of the programme
- Plus sequencing of additional positive control samples
- Existing pipeline components
- New components
  - STRs
  - CNVs and SVs
  - Low level mtDNA variants

ATXN7 – SCA7 comparison

Samples from NHNN

- EH estimates
- Experimental estimates

Repeat size (repeat units)
# Diagnostic STR results

18 diagnostic loci examined; first ~5,000 families

<table>
<thead>
<tr>
<th>Disease</th>
<th>Gene</th>
<th>Expansion consistent with phenotype</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fragile-X syndrome</td>
<td>FMR1</td>
<td>2</td>
<td>1 validated; 1 in progress</td>
</tr>
<tr>
<td>ALS/MND</td>
<td>C9orf72</td>
<td>1</td>
<td>Validated</td>
</tr>
<tr>
<td>Huntington Disease</td>
<td>HTT</td>
<td>3</td>
<td>2 validated; 1 in progress</td>
</tr>
<tr>
<td>SCA6</td>
<td>CACNA1A</td>
<td>1</td>
<td>Validated</td>
</tr>
<tr>
<td>SCA12</td>
<td>PPP2R2B</td>
<td>1</td>
<td>Validated</td>
</tr>
<tr>
<td>DRPLA</td>
<td>ATN1</td>
<td>1</td>
<td>In progress</td>
</tr>
<tr>
<td>Kennedy disease</td>
<td>AR</td>
<td>1</td>
<td>In progress</td>
</tr>
</tbody>
</table>

- Collection of up to 180 positive controls in process across a range of allele sizes
Complementary Clinical Datasets

NHSD success: annual agreement, received quarterly, matched 98.2% participants
April HES delivery 2.3m episodes on 31,781 participants, increases 400K / quarter
Death data received on 430 participants, other data sets arriving now

Life course data: Secondary sources

**NHS Digital**
- Hospital Episodes
- ONS death details
- Diagnostic Imaging
- Patient recorded outcomes
- Mental health & intellectual disability

**Public Health England**
- Cancer registry & datasets (COSD, SACT, RTDS, DID)
- Other disease registries
- Clinical audit
- Screening programmes

**GP data**
- Prescribing/dispensing
- Reports/letters
- Notes (free text)

**GeL genomic results**
- Interpretation
- Validation
- Clinical application
- Germline
- Somatic
- Exit questionnaire

**GMC clinical data**
- Interpretation data:
  - RD: HPO terms, Pedigree
  - Ca: Diagnosis & staging
- Comprehensive clinical data:
  - Data models for key data
  - Including lab test results
  - EHR data dump
  - Treatments & Investigations

**GMC registration**
- Demographics
- Consent status
- Additional findings
- Registration
- Sample management

24 May 2018
### 3rd Main Programme data release: April 2018

<table>
<thead>
<tr>
<th>Genomes</th>
<th>42,700 genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMC Clinical data</td>
<td>61,500 participants 380+ data fields</td>
</tr>
</tbody>
</table>

### Secondary data
- Hospital Episode Statistics (HES)
- Diagnostic Imaging Dataset (DID)
- Patient Reported Outcome Measures (PROMs)
- Mental Health Services Data Set (MHSDS)
- Office for National Statistics (ONS) – mortality data and cancer flagging

### Tiering data
- Tier 1, 2 and 3 variants from interpretation pipeline
- Facilitate GeCIP interpretation of Project cases

### Headline tables
- Key information from different LabKey tables, merged and filterable
- Merged with QC data
- Will facilitate cohort-building and project feasibility assessment

24 May 2018
# The Research Environment at a glance

## Data and documentation

<table>
<thead>
<tr>
<th>Genomes stored in by date folders on Isilon share</th>
<th>Clinical data stored in LabKey</th>
</tr>
</thead>
</table>

Confluence contains:
- data release notes
- user guides
- workaround instructions

## Tools and analysis

- Virtual desktop interface provides GUI and security
- Terminal allows command line querying of the data
- R and Rstudio allows statistical analysis of the data
- Firefox browser allows access to whitelisted sites
- Access to modules and the submission node to run large scale analysis

## Collaboration and data flow

- Domain-specific and shared storage for files
- Social media platform for communication
- Research registry – to promote collaboration and enforce publication moratorium

---

24 May 2018
Going Forward
The Research Environment
Going Forward

The Research Environment
Going Forward

The Research Environment
<table>
<thead>
<tr>
<th>Participant ID</th>
<th>Sex</th>
<th>Ethnic Group</th>
<th>Participant type</th>
<th>Germline Genome build 37</th>
<th>Germline Genome build 38</th>
<th>Tumour genome build 37</th>
<th>Tumour genome build 38</th>
<th>Tiered Data</th>
<th>Disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000001</td>
<td>M</td>
<td>Black or Black British</td>
<td>Proband</td>
<td>Passed QC</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>Intellectual disability</td>
</tr>
<tr>
<td>0000002</td>
<td>M</td>
<td>Black or Black British</td>
<td>Mother</td>
<td>Passed QC</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>0000003</td>
<td>F</td>
<td>Black or Black British</td>
<td>Father</td>
<td>Passed QC</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>0000004</td>
<td>F</td>
<td>White British</td>
<td>Proband</td>
<td>NA</td>
<td>Passed QC</td>
<td>NA</td>
<td>Passed QC</td>
<td>NA</td>
<td>Ductal</td>
</tr>
<tr>
<td>0000005</td>
<td>M</td>
<td>Other</td>
<td>Proband</td>
<td>NA</td>
<td>Passed QC</td>
<td>NA</td>
<td>Passed QC</td>
<td>NA</td>
<td>Endometrioid adenocarcinoma</td>
</tr>
</tbody>
</table>
The Discovery Forum
A driver of translational research

- **Exploring** the business value of genomic medicine data.
- **Connecting** industry stakeholders to the Genomics England community.
- Providing a **gateway** to our Research Environment and dataset.
- Leading to **discovery** and development of precision methods, diagnostics, and therapeutics.
Thank you!

Stay in touch

@genomicsengland  #genomes100k

Like the ‘Genomics England’ page

Follow ‘Genomics England’

www.genomicsengland.co.uk
If a participant agrees, we look for changes in the following genes:

**Bowel cancer predisposition:**
- *MLH1* (adult only), *MSH2* (adult only), *MSH6* (adult only), *MUTYH* (adult only)
- *APC* (adult and child)

**Breast and ovarian cancer predisposition:**
- *BRCA1* (adult only), *BRCA2* (adult only)

**Other cancer predisposition:**
- *VHL* (adult and child), *MEN1* (adult and child), *RET* (adult and child)

**Familial hypercholesterolaemia:**
- *LDLR* (adult and child), *APOB* (adult and child), *PCSK9* (adult and child)

Currently, if a participant agrees, we look for changes in the following genes: *CFTR* (Cystic fibrosis).