Sharing genotype and phenotype data between stakeholders: The UK experience

Willem H Ouwehand FMedSci
Professor of Experimental Haematology
University of Cambridge; Wellcome Sanger Institute and NHS Blood and Transplant
who1000@cam.ac.uk
The UK story of data sharing started in 2005 with the Wellcome Case Control Consortium and the first ever comprehensive Genome Wide Association Study.
Hence translating the philosophy of data sharing into the NHS does builds on a solid foundation with common principles

- Sharing of patient genotype and phenotype data between stakeholders is only legitimate if the patient has given informed consent.

- The data owner (i.e. the NHS) accepts responsibility for the safe-keeping of the data and granting access to stakeholders.

- Stakeholders in Hospitals, other Hospital Authority entities, and researchers in the Public and Private sector agree to come to the library to read the ‘book’, but the book is not to be taken home.
A story of how we are tackling two challenges

- The perils of centralisation
- New analysis methods
- Interpretation bandwidth mismatch
Perils of centralisation

Centralisation has many benefits but comes with great responsibilities
Benefits

• Centralised systems are easier to build; in converse, creating a federated rich data ecosystem that enables large scale analyses is really tough

• More cost effective to build and operate – build once, run once

• Standardisation of processes and data is easier to achieve and enforce

• Facilitates data accumulation

• Machine learning needs lots of data to be trained effectively
Pitfalls of centralisation

- Stifles innovation

- Can seem like a black whole to stakeholders - creates insiders vs outsiders
Some common technological solutions

- **Data models**: Well-modelled data so that systems can exchange rich information in computable ways and humans can understand what it is about.

- **Open APIs***: programming interfaces that are open and can be used to programmatically interact with the system (as opposed to point and click graphical user interfaces).

- **Modular architecture**: Possibility to bring third party analytics to the data.

* Application Programme Interface
Interpretation Platform API
to seamlessly dock in other applications

- clinical-report
- docs
- download-token
- exit-questionnaire
- file
- get-token
- health
- interpretation-request
- interpreted-genome
- qc-outcome
- status
- tiering-qc-outcome
- workspace-groups
New analysis methods

No statistical methods were available at the outset of the 100,000 Genomes Project to search for the ‘needle in the haystack’ which is causative of unexplained rare diseases.
• **Family studies:** The ‘tried and tested’ approach of studying large families has become challenging with ‘no. of children/nuclear family’ being <2

• **Time consuming:** Developing new methods takes time and requires large datasets for methods training

• **Statistics:** Bayesian statistics provides the flexibility to model the complexity of different modes of, e.g. inheritance, variable penetrance, inability to accurately predict the consequence of a mutation on gene function

Greene et al, AJHG 2017
Strong evidence for 99 genetic associations between rare variants and grouping of patients with similar phenotype by applying BeviMed

Of these 99 associations, 61 are consistent with firmly established evidence and a further 18 have been reported in the literature since 2015, either by us or by others.

Greene et al, AJHG 2017
Interpretation bandwidth mismatch

The focus must be on how to help clinical teams make the best decisions in the most efficient way.
Modular architecture for interpretation services

• Rather than “hard coding” algorithms as part of an analytics pipeline, define interpretation services that can be called as required
• By defining this modular architecture everyone can bring a service to be applied on the data

We support:
• Tiering (classification)
• Exomiser (prioritisation)
• BeviMed (gene discovery) – in progress
• Others to come
A joint responsibility

• Genomics England currently analyses genomes **about 5 times faster** than the NHS labs can interpret them.

• Genomics England could just add more compute and it would go faster, but

• Genomic Laboratory Hubs cannot simply add more ‘trained’ people – it is not like there is a pool of experts waiting to be hired.

• Everyone fails if the coupling between central analysis power and local ability to complete the job (issue a MDT-reviewed pathology report) is not carefully monitored.
The aim is therefore clear

Serve useful information for clinical genomics by providing analytics, content, services and systems to improve genome interpretation

Where improvement is measured as increased sensitivity and specificity of the interpretation process for time and cost spent by medical professionals looking at a case

One way we can improve interpretation is by building a knowledge base from the information that is accumulated in the 100,000 genomes project and the Genomic Medicine Service
What we currently capture

- Variant prioritisation and classification results (results of the interpretation services)
- Interpretation interactions (data entered by users into the systems)
- Summary of findings (the variants selected by the lab to report)
- Outcomes (what variants were confirmed, fed back and why)
Welcome to the Clinical Variant Ark

A knowledge base built from the 100,000 Genomes Project and NHS Genomic Medicine Service

Search for a case or variant

<table>
<thead>
<tr>
<th>Variant</th>
<th>Gene</th>
<th>Type</th>
<th>Consequence type</th>
<th>Interpreted by</th>
<th>Zygosity</th>
<th>Proband affected</th>
<th>Cases with variant in primary findings</th>
</tr>
</thead>
</table>
| 16:2115373:C:T | AC0090656  
ENSG000000261240  
PKD1  
ENSG00000008710  
AC0090653  
ENSG00000260447 | SNV               | splice region variant | Exo, Tier Inc. in SOF |                |         |                 |                                        |
| 3:50841714:G | DOCK3  
ENSG00000088538 | SNV               | frameshift variant       |                |         |                 |                                        |
| 14:220207908:A:C | TRAV18  
ENSG000000211798  
TRAV19  
ENSG00000211700 | SNV               | missense variant         | Exo, Tier     |                |         |                 |                                        |

2,322

Variants
Learning from the interpretation process

Clinically relevant variants

Clinical Variant Ark

All variants

opencga
A roadmap for delivering analytics to improve interpretation

- Train machine learning to offer more discriminatory power to the human
- Offer users rich ways of accessing the data for them to come to conclusions
- Build databases to store the captured data
- Establish systems to capture data from the interpretation process
Conclusions

• The successful delivery of the 100,000 Genomes Project was critically dependent on the timely delivery of an integrated informatics platform

• Not a single UK hospital has the compute power to process and analyse the equivalent of the content of 500,000 mobile phones

• Errors have been made on the way, e.g.: with the solutions to capture clinical data (too clunky), interpretation mismatch

But

• These errors can be overcome if you have a Team of Smart People with the right mix of expertise of physics, math and computer science
• To recruit & retain these people you will need to provide competitive levels of remuneration