ELIXIR and IMI Oncotrack: Providing long term storage and data sharing

Dylan Spalding

www.elixir-europe.org
Introduction: The data challenges

- Data production sites increasing across Europe
- Data growth in life sciences
- Secure access and governance of human data
- Open data mandates of National and European funders
Introduction: Outline

- EGA
  - Statistics
  - Metadata model / Data Model
- IMI Oncotrack
  - Overview
  - Data collection / Types / Infrastructure
- Study
  - Aims
  - Results
- Conclusion
European Genome-phenome Archive

- Joint service from the European Bioinformatics Institute (EMBL-EBI), UK and the Centre for Genomic Regulation (CRG), Spain
- Secure and permanent archive for all type of genetic, -omics and phenotypic data from humans
- ALL data consented for research use
- Data access managed through application and encrypted data delivery
- Submissions include raw data from genome sequence, transcriptome or epigenome experiments
  - Called variants and genotypes
  - Sample phenotypes
About EGA: What is controlled access data anyway?

- **Informed consents specify controlled release requirements**
- Human, personally identifiable data types
  - ‘raw’, processed, phenotypic
- Affiliated to bio-medical research or consortium projects
- Access by formal application procedure to Data Access Committee (DAC)
About EGA: What is controlled access data anyway?

Open & public archives

Controlled access archive
About EGA: Context and metrics

- Launched 14\textsuperscript{th} July 2008
- Role: \textbf{Secure} archive for \textbf{controlled} distribution of \textbf{consented} genetic & phenotypic data
- Almost \textbf{9000} data access accounts; over \textbf{450} submission accounts
- \textbf{\sim 3.7PB} available for download; \textbf{\sim 2939} datasets; \textbf{3.8PB} distributed last year
- +\textbf{200} contacts to Helpdesk/month
About EGA: Data Growth

Month

Data Volume TB

Volume
Cumulative
Studies by Technology

- Epigenetics: 118
- Exome Sequencing: 321
- GWAS: 224
- Genotyping: 242
- Resequencing: 324
- Single-Cell Sequencing: 22
- Transcriptome: 202
- Whole Genome Sequencing: 298
Studies by type

- **Cancer**: 620 studies
- **Cardiovascular**: 130 studies
- **Infectious**: 35 studies
- **Inflammatory**: 44 studies
- **Neurological**: 64 studies
- **Other**: 221 studies
About EGA: Download Volume
• Serve +220 institutes worldwide
• Single publication submissions to large project/consortiums
About EGA: Consortia Data archived at EGA

International Cancer Genome Consortium
https://icgc.org/

Cancer Research UK
http://www.icr.ac.uk/

Wellcome Trust Case Control Consortium
http://www.wtccc.org.uk/

biobank.uk
http://www.uk10k.org/

HipSci
http://www.hipsci.org/

RD Connect
http://rd-connect.eu/
EGA metadata ‘object’ requirements

ENA and EGA shared requirements

• Study/Project – Information about project or study
• Samples – describe each sample (donor ID, gender, phenotype)
• Experiments – type, platform, library information
• Run/Analysis – references raw data (Run) or processed file (Analysis)

Data Access Committee (DAC) – defines the data access authority
• Policy – Data Access Agreement – ‘terms and conditions’ of data use
• Dataset – package files into ‘units’ for access & distribution

EGA specific requirements
EGA metadata ‘object’ requirements
EGA Accessions

- Study/Project – EGAS00001000XXX (publication accession)
- Sample – EGAN00001000XXX
- Experiment – EGAX00001000XXX
- Analysis – EGAZ00001000XXX
- Run – EGAR00001000XXX
- Dataset – EGAD00001000XXX (or EGAD00010000XXX for non-NGS)
- Policy – EGAP0000X000XXX
- Data Access Committee (DAC) – EGAC00001000XXX
Meta-data

- Meta data since 2013 stated as being public
  - Object attributes
  - Donor ID, Gender, Phenotype
- Currently consented sample meta data archived as a file
  - Low – level phenotype, age etc.

A Typical EGA submission consists of one study linked to multiple samples, with each sample linked to a separate experiment and a run. The runs are then linked to a single dataset, pointing to a single DAC and policy.
EGA Data Access

- Access to data controlled by individual Data Access Committees (DAC)
  - Authorisation to access data determined by DAC
- EGA plays no role in authorisation to grant access
- Organised by Study and Dataset
- Access granted on a per dataset level
  - Subject to a specific data access agreement
Why Submit to the EGA?

• Maximise benefit from your data
• Funders & journals require researchers to have data sharing plan
  • Wellcome Trust "Policy on data management and sharing”
    • http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.html
  • Nature "Availability of data and materials”
    • http://www.nature.com/authors/editorial_policies/availability.html
  • Public Library of Science (PloS) "Sharing of Materials, Methods, and Data”
    • http://www.plosone.org/static/policies.action#sharing
• Utilise EGA distribution technology and access tools
Why Submit to the EGA?

• Make data FAIR
  • Findable
  • Accessible
  • Interoperable
  • Re-useable
IMI Oncotrack

- IMI – large public-private partnership for life science
  - EFPIA and the European Commission
- Oncotrack: Academic – Industry translational research project
  - Goal of identifying biological markers that will help our understanding of tumors
- Funded for a specific timescale
- Data includes:
  - Phenotype, Genome sequencing data, Computer modeling data, Imaging, Xenograft data
  - tranSMART subset of OncoTrack database
Drug development frequently fails as does biomarker validation

Extensive research has resulted in modest impact on precision medicine

- Approximately 200 pharmacogenomic biomarkers listed in FDA labels for licensed drugs
  - Approx 50% for oncologic indications
  - Only a small number selected using a companion diagnostic
Drug development frequently fails as does biomarker validation

Extensive research has resulted in modest impact on precision medicine

**Premise:** Source data do not represent the complexity of the tumor well enough to allow the identification of novel biomarkers for the stratification of cancer patients

- Lack of data

**Strategy:** New methodology to data collection allowing a more complex data set to be used for biomarker discovery and validation

- Surgical specimens of tumor, metastases, blood and normal tissue all collected at time of diagnosis
- Patients followed up and samples taken in cases of disease progression
IMI Oncotrack – Data Collection

- Tumor sample sequencing:
  - Genome, transcriptome, methylome
- Tumor sequences compared to normal colon or liver tissue, or germ-line DNA from peripheral blood lymphocytes
- Xenograft and organoids derived from tumor samples
  - Assess drug response
  - Transcriptome analysis and confirmatory mutation analysis
- Proteomic mass spectrometry measurements, reverse-phase arrays, proximity ligation assays
- Conventional bioinformatics approaches and systems biology in silico modelling
IMI Oncotrack – Data Management

- Data is pseudonymised
  - Patient feedback
  - Personal data – data protection legislation
- Large volume of data
- Diverse data sets / formats
  - Analysis of these diverse sources
IMI Oncotrack – Data Infrastructure

Clinical data

Blood samples
- Circulating Tumor Cells

Tissue samples
- Primary cell lines
- Xenograft models

Next generation sequencing

Proteomic Analysis

Molecular Pathology

Drug response experiments

In silico modelling

Collaboration with eTRIKS

OpenClinica
Open Source for Clinical Research

tranSMART
vers. 1.2

oncoTRAK
DB

web interface

Data management and Data Re-use in Collaborative Research Projects, 2016

David Henderson, Bayer Pharma AG
ELIXIR – IMI OncoTrack

• Scoping study to begin understand:
  • Long-term knowledge management requirements
    • Data storage
    • Data governance - Consent and Access management
  • Metadata mapping
    • Public and Private
  • Generate a use-case for other IMI Consortia

User access through DACs

Long-term storage and management

Genomic data
Animal studies data
Data from cell culture studies
Clinical data
...
ELIXIR – IMI OncoTrack

- Pseudonymised data
  - Consent for oncology research only
- Sustainable long-term data distribution
ELIXIR – IMI OncoTrack

• Pseudonymised data
  • Consent for oncology research only
• Sustainable long-term data distribution
ELIXIR – IMI OncoTrack

- Pseudonymised data
  - Consent for oncology research only
- Sustainable long-term data distribution
ELIXIR – IMI OncoTrack

- Pseudonymised data
  - Consent for oncology research only
- Sustainable long-term data distribution
ELIXIR – IMI OncoTrack

- Pseudonymised data
  - Consent for oncology research only
- Sustainable long-term data distribution
  - Oncotrack DAC in place
  - DAA drawn up and under review
ELIXIR – IMI OncoTrack

- Pseudonymised data
  - Consent for oncology research only
- Sustainable long-term data distribution
  - Oncotrack DAC in place
  - DAA drawn up and under review
- Costs estimated – vary depending on access frequency
ELIXIR – IMI OncoTrack

- Pseudonymised data
  - Consent for oncology research only

- Sustainable long-term data distribution
  - Oncotrack DAC in place
  - DAA drawn up and under review

- DAC costs estimated – vary depending on access frequency

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Average Requests per Month</th>
<th>Estimated Hours per Request</th>
<th>Hours per Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGAD00000000000002</td>
<td>15</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>EGAD000001000741</td>
<td>11</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>EGAD000001000786</td>
<td>0.2</td>
<td>2</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Estimated hours per month staff time to administer access to a range of EGA datasets administered by the Wellcome Trust Sanger Institute
ELIXIR – IMI OncoTrack

- Pseudonymised data
  - Consent for oncology research only
- Sustainable long-term data distribution
  - Oncotrack DAC in place
  - DAA drawn up and under review

- DAC costs estimated – vary depending on access frequency
- Hardware costs at EGA minimal due to relative size of Oncotrack data vs EGA data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Average Requests per Month</th>
<th>Estimated Hours per Request</th>
<th>Hours per Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGAD00000000000002</td>
<td>15</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>EGAD00001000741</td>
<td>11</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>EGAD00001000786</td>
<td>0.2</td>
<td>2</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Estimated hours per month staff time to administer access to a range of EGA datasets administered by the Wellcome Trust Sanger Institute.
ELIXIR – IMI OncoTrack

- Data Upload – metadata mapping
  - Unique mapping per submission
  - EGA Helpdesk and submitter map the metadata
# ELIXIR – IMI OncoTrack: Example metadata mapping

<table>
<thead>
<tr>
<th>Table (Object)</th>
<th>Description</th>
<th>C</th>
<th>Sample to file ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTC</td>
<td>Circulating Tumor Cells. Approx 100 datasets, 20 patients</td>
<td>analysis</td>
<td>1:2</td>
</tr>
<tr>
<td>drugResponse</td>
<td>Drug at specific concentration on a specific sample. In analysis</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>drugResponseValues</td>
<td>Values associated to above</td>
<td>analysis</td>
<td>all:n (n=3;1 patient sample)</td>
</tr>
<tr>
<td>inSitu</td>
<td>HiRes images / textfiles</td>
<td>analysis</td>
<td>1:4</td>
</tr>
<tr>
<td>massSpec</td>
<td>1 file - will change soon</td>
<td>analysis</td>
<td>all:1</td>
</tr>
<tr>
<td>methylation</td>
<td></td>
<td>array</td>
<td>1:3</td>
</tr>
<tr>
<td>modelling</td>
<td>In Silico - will change so leave for the moment</td>
<td>analysis</td>
<td>1:1</td>
</tr>
<tr>
<td>modelling2sequencingAnalysis</td>
<td>relates analysis to insilico model</td>
<td>analysis</td>
<td>NA</td>
</tr>
<tr>
<td>P3D</td>
<td>3D cell lines</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>P3Dimage</td>
<td>Imaging files</td>
<td>NA</td>
<td>1:n (n ~ 1-8)</td>
</tr>
<tr>
<td>patient</td>
<td></td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>PEA</td>
<td>Microarray dataset - similar to rppa</td>
<td>array</td>
<td>all:4 or 1:n (n = 1-4)</td>
</tr>
<tr>
<td>PEAvalues</td>
<td>3 files per batch - raw, normalised and ddCq files</td>
<td>array</td>
<td>n:3 (n ~ 98)</td>
</tr>
<tr>
<td>rppa</td>
<td>Reverse phase protein array, 644 datasets</td>
<td>array</td>
<td>n:1 (n ~ 190)</td>
</tr>
<tr>
<td>rppaRUN</td>
<td>Batches of rppa</td>
<td>array</td>
<td>NA</td>
</tr>
<tr>
<td>rppaValues</td>
<td></td>
<td>array</td>
<td>NA</td>
</tr>
<tr>
<td>Sample</td>
<td></td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>sequencing</td>
<td></td>
<td>Sequence</td>
<td>to be determined</td>
</tr>
<tr>
<td>sequencingAnalysis</td>
<td></td>
<td>analysis</td>
<td>NA</td>
</tr>
<tr>
<td>sequencingAnalysisDeletion</td>
<td></td>
<td>analysis</td>
<td>1:1</td>
</tr>
<tr>
<td>sequencingAnalysisEXPmiRNA</td>
<td></td>
<td>analysis</td>
<td>all:1</td>
</tr>
<tr>
<td>sequencingAnalysisEXPmRNA</td>
<td></td>
<td>analysis</td>
<td>all:1</td>
</tr>
<tr>
<td>sequencingAnalysisEXPRiboZeroRNA</td>
<td></td>
<td>analysis</td>
<td>all:1</td>
</tr>
<tr>
<td>sequencingAnalysisSolution</td>
<td></td>
<td>analysis</td>
<td>1:1</td>
</tr>
</tbody>
</table>
ELIXIR – IMI OncoTrack

- Data Upload – metadata mapping
  - Unique mapping per submission
  - EGA Helpdesk and submitter map the metadata
- Identify areas where EGA submissions do not have adequate support
  - SOP for importing novel consortia data
  - Improved sample data representation
Patient 1

Sample A
  - Tech Rep B1

Sample B
  - Tech Rep B2

Sample C
  - Tech Rep B3
ELIXIR – IMI OncoTrack

- Data Upload – metadata mapping
  - Unique mapping per submission
  - EGA Helpdesk and submitter map the metadata
- Identify areas where EGA submissions do not have adequate support
  - SOP for importing novel consortia data
  - Improved sample data representation
  - Improved ontology / standards support – e.g. CDASH as used by Oncotrack
  - Clearer specification of controlled access and public access attributes:
    - Example: Date of Birth, geographic location vs sample site, high-level phenotype
Conclusion

• Developing partnerships with industry to access data
  • Creating SOPs to facilitate data load at EGA
  • Use EGA to provide long term access to translational data for the community
  • Potential benefits to SMEs

• Costs
  • Metadata mapping and DAC costs
  • Hardware / storage minimal

• Improving meta-data at EGA for translational data
  • Developing sample metadata storage
    • Benefits other consortia such as RD-Connect

• Making consortia data FAIR
Acknowledgements

• Dylan Spalding, Alexander Senf, Jeff Almeida-King, European Bioinformatics Institute (EMBL-EBI) European Molecular Biology Laboratory, United Kingdom
• Jordi Rambla de Argila, Audald LLoret i Villas, Centre for Genomic Regulation, Spain
• David Henderson, Bayer HealthCare Pharmaceuticals, Berlin, Germany
• Reha Yildirimman, Alacris Theranostics GmbH, Berlin, Germany
• Susanna Repo, Niklas Blomberg, Serena Scollen, ELIXIR Hub, Wellcome Genome Campus, Cambridge, United Kingdom