Current and future use of ontologies at Genomics England

Damian Smedley

Director of Genomic Interpretation, Genomics England

Senior Lecturer, Queen Mary University of London

Overview

Genomics England

- Introduction to the 100,000 Genomes Project
- Role of clinical and model organism phenotypes
  - Clinical data collection and panel assignment
  - Automated variant prioritisation
  - Genotype to phenotype knowledgebase

Overview

- Introduction to the 100,000 Genomes Project
- Role of clinical and model organism phenotypes
  - Clinical data collection and panel assignment
  - Automated variant prioritisation
  - Genotype to phenotype knowledgebase

The 100,000 Genomes Project

Genomics England

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.

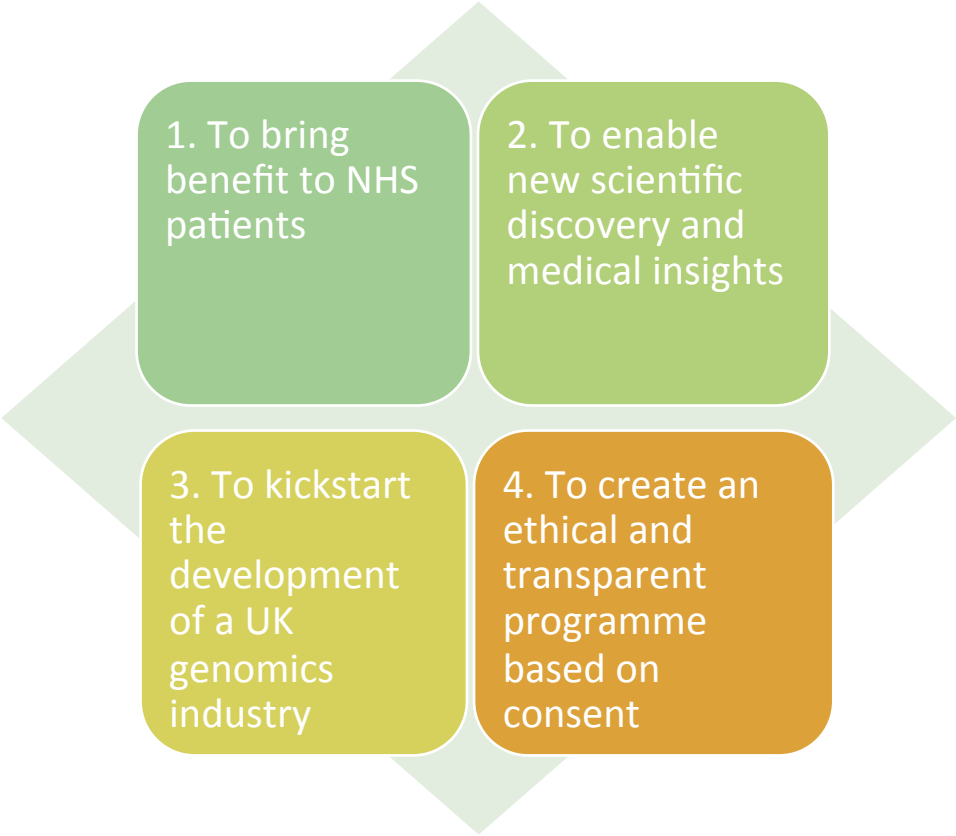
13 Genomic Medicine Centres, and **85** NHS Trusts within them are involved in recruiting participants

1,500 NHS staff
(doctors, nurses, pathologists, laboratory staff, genetic counsellors)

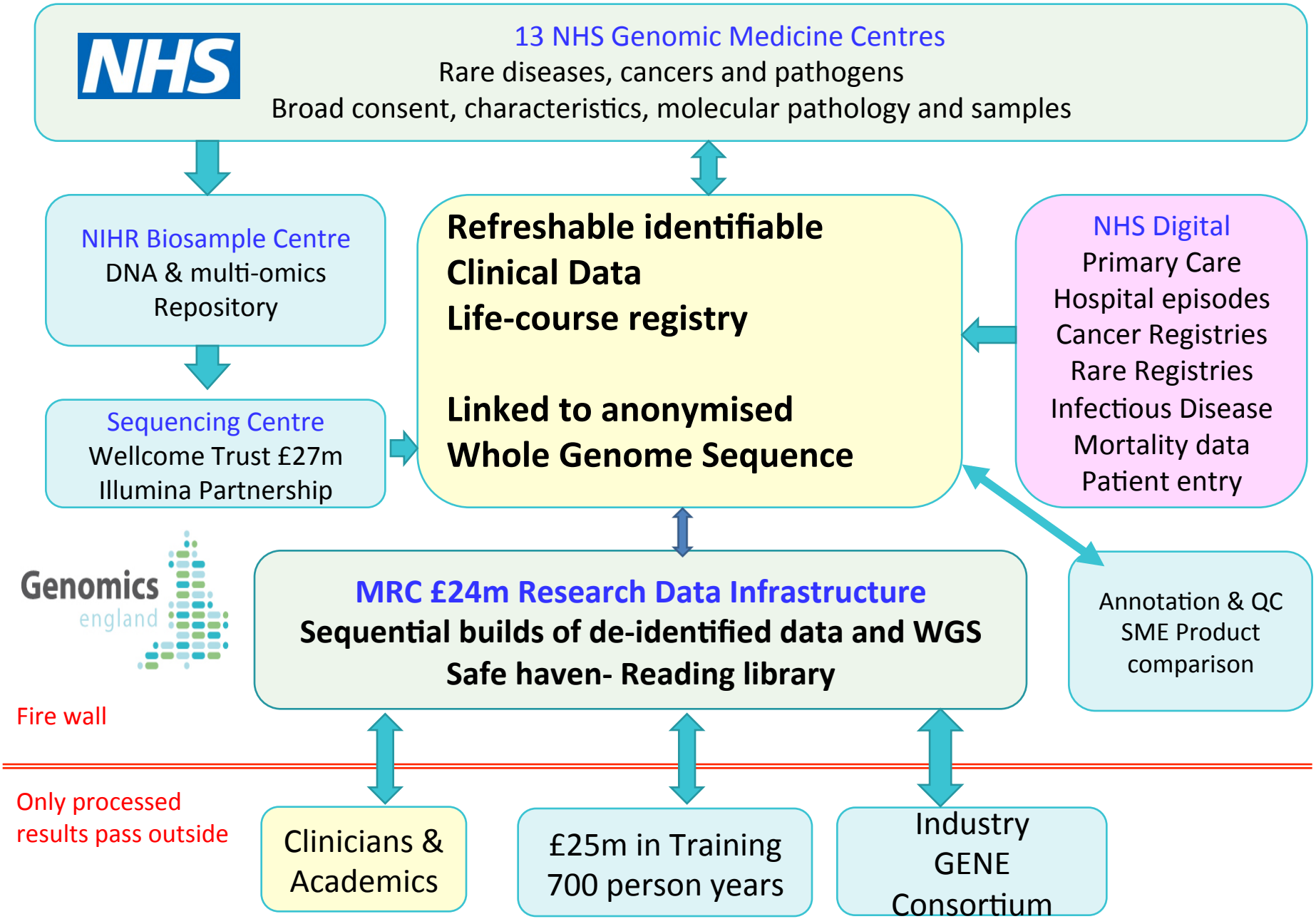
2,500 researchers and trainees from around the world

Goals of the Genomics England project

- Sequence 100,000 genomes
- Cancer and rare genetic disease
- Capture data delivered electronically, store it securely and analyse it within an English data centre (reading library)
- Combine genomes with extracted clinical information for analysis, interpretation, and aggregation
- Create capacity, capability and legacy in personalised medicine for



Organisation of the 100,000 Genomes Project



Genomics England Clinical Interpretation Partnerships (GeCIPs)

Genomics England

The standard way

Start

Genomics Research
Form hypothesis
Get funds and form collaboration
Collect, analyse data and validate results

Publication, dissemination, translation
Publish and disseminate results
Attempt to translate into healthcare

Healthcare adoption and implementation
NHS and NICE evaluation and Guidelines
Education and implementation programme

17 yrs

The GeCIP way

2014

The 100,000 Genomes Project
hypothesis – WGS will enhance diagnosis
Coalition of NHS, academics and trainees
Work together on WGS within GeCIP domains

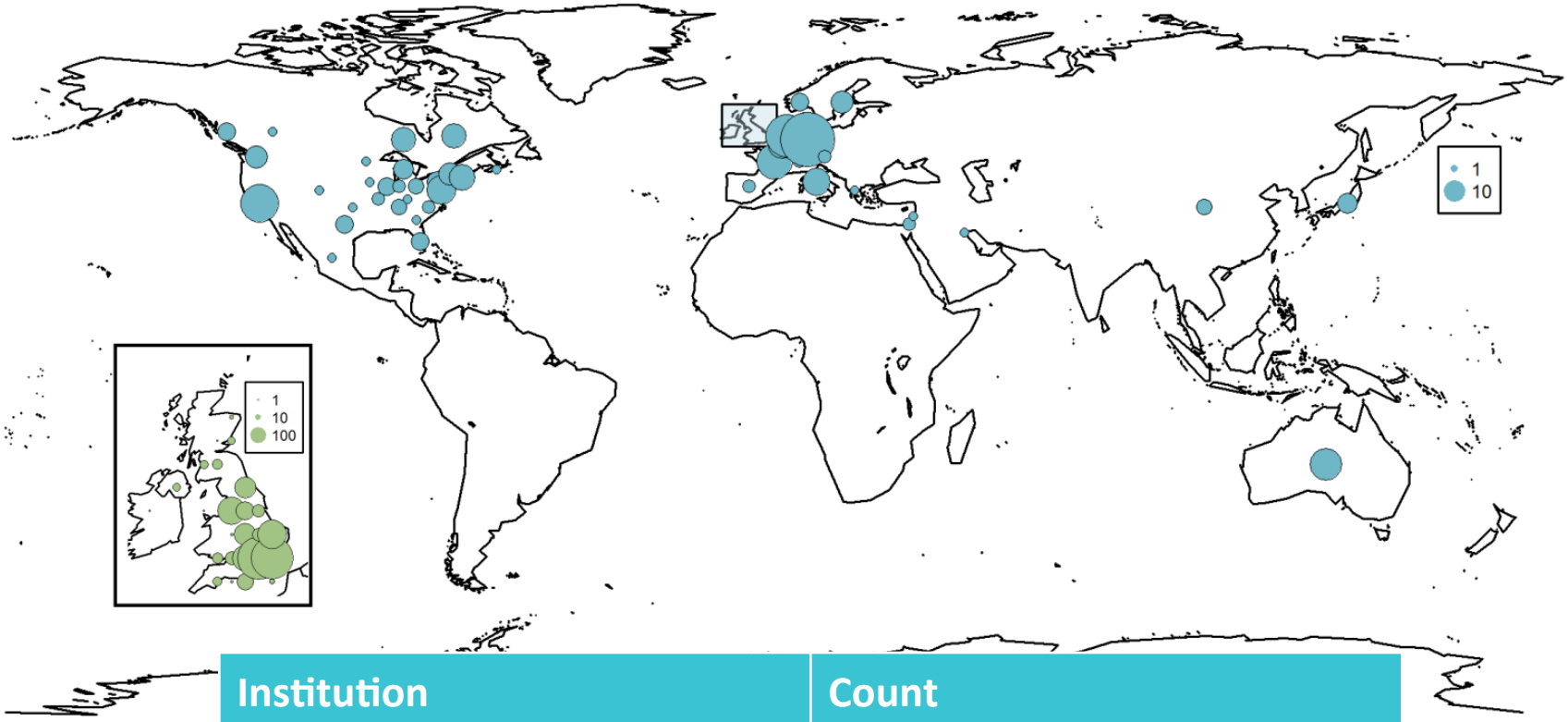
Enhanced interpretation linked to implementation
Validate, publish, educate and translate
The GeCIP Collaborative accelerates Implementation
Evaluate therapeutic innovation potential

Earlier Healthcare adoption and implementation
Accelerated diagnosis and health economic evaluation
Framework for therapeutic innovation

?3 yrs

Genomics England Clinical Interpretation Partnership

- 2,500 prospective GeCIP domain members
- 300 institutions, 24 countries



Institution	Count
UK Academic	1744
NHS Trust	634
International Academic	198
Other	333

GeCIP Domains

Genomics england

Rare

- Cardiovascular
- Endocrine and Metabolism
- Gastroenterology and Hepatology
- Hearing and Sight
- Immunology and Haematology
- Inherited Cancer Predisposition
- Musculoskeletal
- Neurological
- Paediatric Sepsis
- Paediatrics
- Renal
- Respiratory
- Skin

Cancer

- Adult Glioma
- Bladder
- Breast
- Colorectal & upper
- Lung
- Melanoma
- Renal Cell
- Sarcoma
- Testis
- Ovarian
- Prostate
- Childhood Solid Cancers
- Haematological Malignancy
- Pan Cancer
- (Ca of) Unknown primary

Functional

- Electronic Health Records
- Validation and Feedback
- Ethics and Social Science
- Functional Effects
- Health Economics
- Machine Learning, Quantitative Methods and Functional Genomics
- Population Genomics
- Enabling Rare Disease Translational Genomics via Advanced Analytics and International Interoperability
- Functional Cross Cutting
- Education and Training
- Stratified Medicine & Pharmacogenomics

Industry Consortium and Partners

- 13 pharma/diagnostics/SMEs
- Precompetitive consortia
- Work together on 5000 WGS to shape data centre
- Individual company interactions

- Inward investment from Illumina

£50m in new HQ in Cambridge

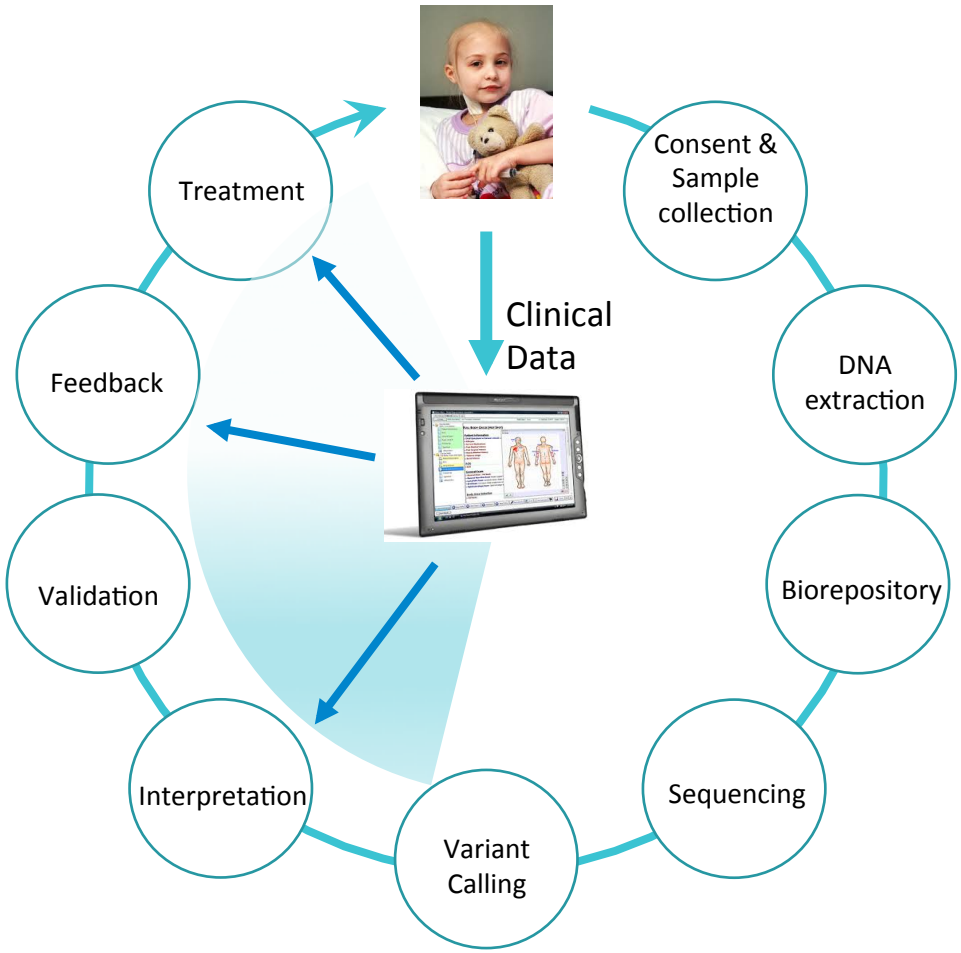
- AbbVie
- Alexion Pharmaceuticals
- AstraZeneca
- Biogen
- Dimension Therapeutics
- GSK
- Helomics
- Roche
- Takeda
- Berg
- Boehringer Ingelheim
- UCB
- Intellia

Overview

Genomics England

- Introduction to the 100,000 Genomes Project
- Role of clinical and model organism phenotypes
  - Clinical data collection and panel assignment
  - Automated variant prioritisation
  - Genotype to phenotype knowledgebase

Genomics England is about helping complete the cycle



- Treatment cycle for just one patient requires a complex chain of operations
- Most of these operations have not been designed or optimised for the purposes of Genomic Medicine.
- So the task is one of catalysing a **Transformation** in Medical Practice, particularly relating to routine use of coordinated data.
- To achieve this one needs to develop/adopt **standards** across the system

Capturing, analysing and sharing clinical data is hard

Genomics england

- **Multiple data types required**
  - Family structures
  - Phenotypic data: evaluations, investigations, treatments

- **No "routine" clinical data collection models**
  - Data collection model is almost unique to each rare disease
  - Often multisystem disorder -> many terms required
  - Many clinicians involved in patient care

- **Implementation within routine health care setting**
  - Time
  - IT structures

Rare disease data models

Genomics England

"The set of clinical features & test results required to interpret WGS data given recruitment to a specific disorder"

- Define what data we want to collect
- By "interpret" we mean more than sufficient for a diagnosis
- For an individual phenotyping may be excessive in some cases, but insufficient in others
- Aim to ensure data is deep, relevant and consistent

Development of models

Get an existing model

McCann *et al. Pediatric Rheumatology* 2014, **12**:31
http://www.ped-rheum.com/content/12/1/31

PEDIATRIC RHEUMATOLOGY

RESEARCH Open Access

Developing a provisional, international Minimal Dataset for Juvenile Dermatomyositis: for use in clinical practice to inform research

Liza J McCann[1*], Katie Arnold[2], Clarissa A Pilkington[2,4], Adam M Huber[5], Angelo Ravelli[6], Laura Beard[2], Michael W Beresford[1,7], Lucy R Wedderburn[2,3,4] and the UK Juvenile Dermatomyositis Research Group (JDRG)[4]

Abstract

Background: Juvenile dermatomyositis (JDM) is a rare but severe autoimmune inflammatory myositis of childhood. International collaboration is essential in order to undertake clinical trials, understand the disease and improve long-term outcome. The aim of this study was to propose from existing collaborative initiatives a preliminary minimal dataset for JDM. This will form the basis of the future development of an international consensus-approved minimum core dataset to be used both in clinical care and inform research, allowing integration of data between centres.

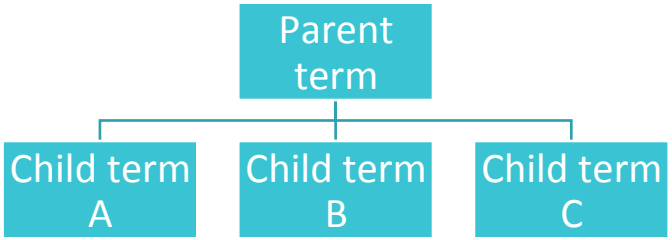
Methods: A working group of internationally-representative JDM experts was formed to develop a provisional

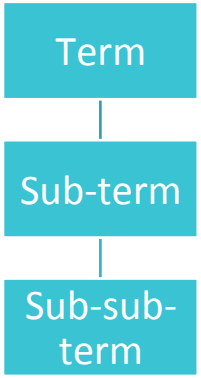
Could also be:

- A project case report form
- Database schema
- A published disease description
- An OMIM model
- A list developed by the clinical expert, possibly using Phenomizer or Phenotips

Revision

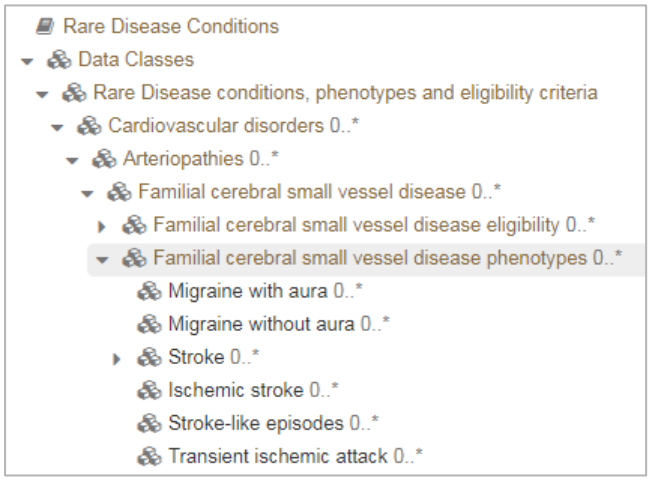
- Clinical revision is vital
- We are aiming to keep models below around 40 terms
- We also wish to present terms in a clinically logical order
- One method of keeping models shorter is replace terms in a sibling relationship with a common parent, or reduce terms in parent/child relationships to one parent term

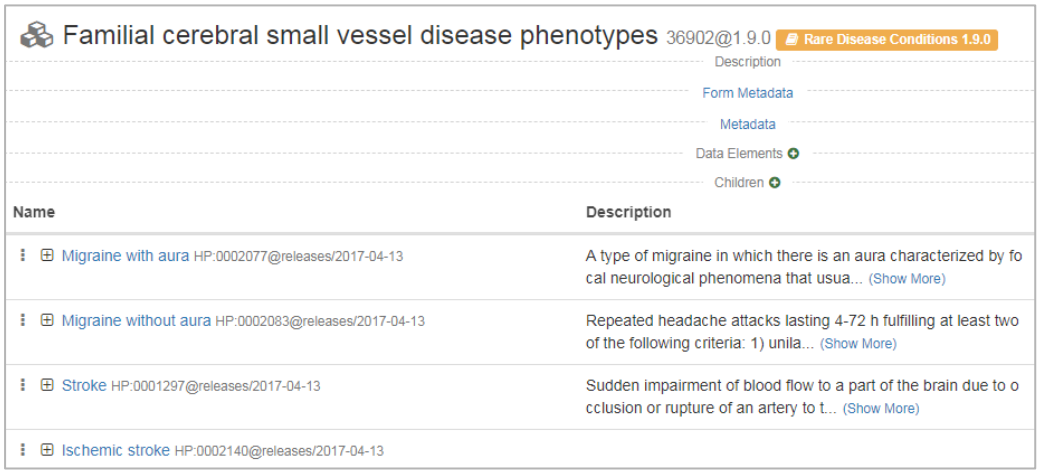




Implement in the Data Model Catalogue

- This is a semantically enabled on-line catalogue which allows hierarchical models to be built and cross-linked
- This ensures that classes and value domains remain independent
- It allows independent version control of models and ontologies





Conclusions from model development

Genomics England

- Revision has been extensive…the models are evolving
- HPO has proved comprehensive but some revision is needed
- There is a need to choose between qualitative and quantitative representation of data – this seems to be a case-by-case decision
- Some models cover more than one disease or set of features, and common observations can be separated out, but many diseases have required their own specific model
- Clinical test data is often used by more than one model, and can be cross mapped to existing standards in some cases

Some other terminologies in GEL clinical data

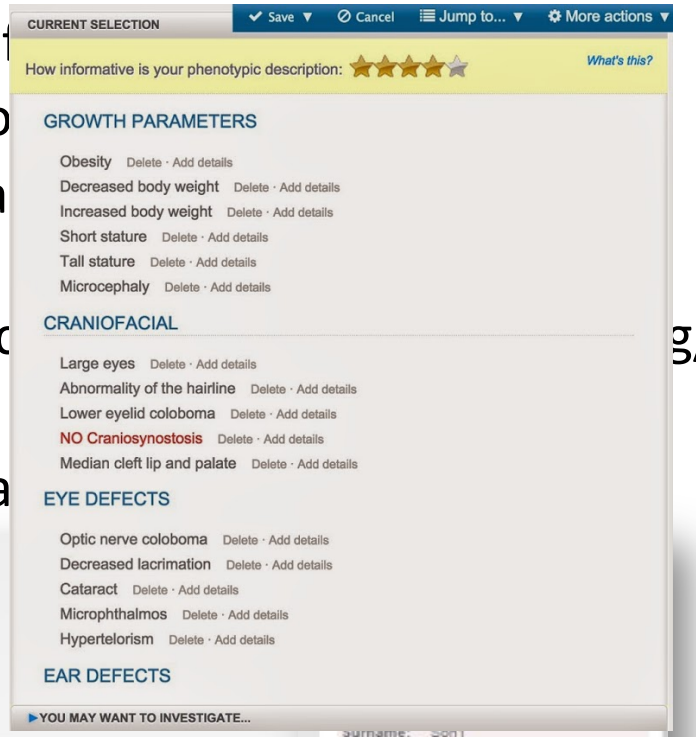
Genomics England

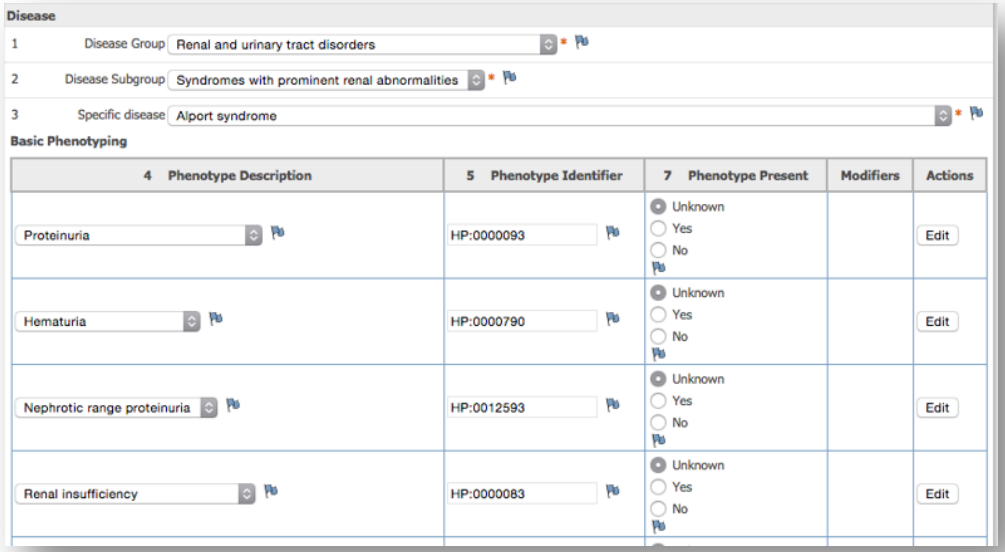
- HES data includes ICD-10 diagnoses and OPCS-4 codes for procedures
- Submitted cancer tumour data includes SNOMED and ICD-O-3 topography and morphology codes
- Diagnostic Imaging Dataset includes SNOMED and NICIP codes for imaging, regions, modalities, systems and morphology

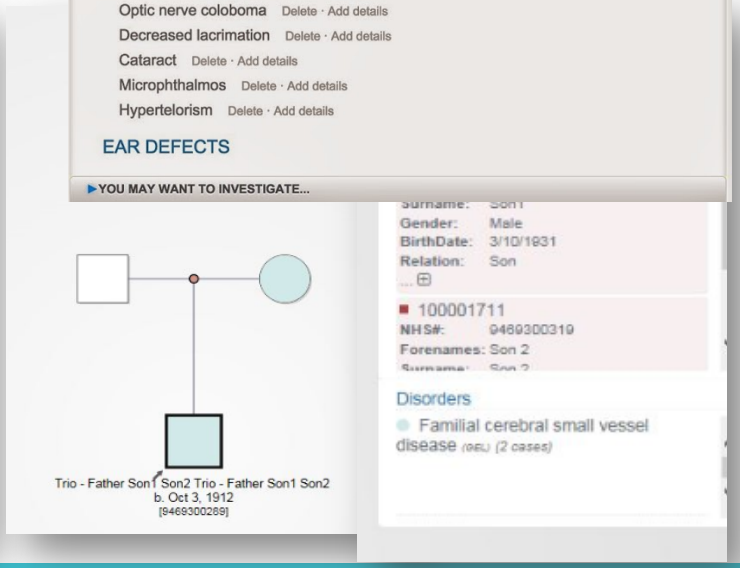
Developing a clinical data capture system for RD genome diagnostics

Genomics England

1. Standardised clinical data questionnaire f
   - OpenClinica or their own systems expo
   - Defined HPO terms for each disease ca negative and additional terms
   - Standardised system for (automatic) ac g/ lab test results
2. Automated pedigree software to ease da







Empowering recruiting clinicians to direct the analysis

Genomics England

3. Presented with a summary of data entered
4. Option to 'control' the genome analysis such as selecting gene panels etc

Taylor, Joseph Timothy | DOB: 23/12/1936

NHS: 1112220098 (Proband, Family Id: 98)

Family Medical Review: No state assigned
Participant Medical Review: Awaiting medical review

Summary | Details | Genetic Tests | Observations | Family | HPO

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Participant Summary

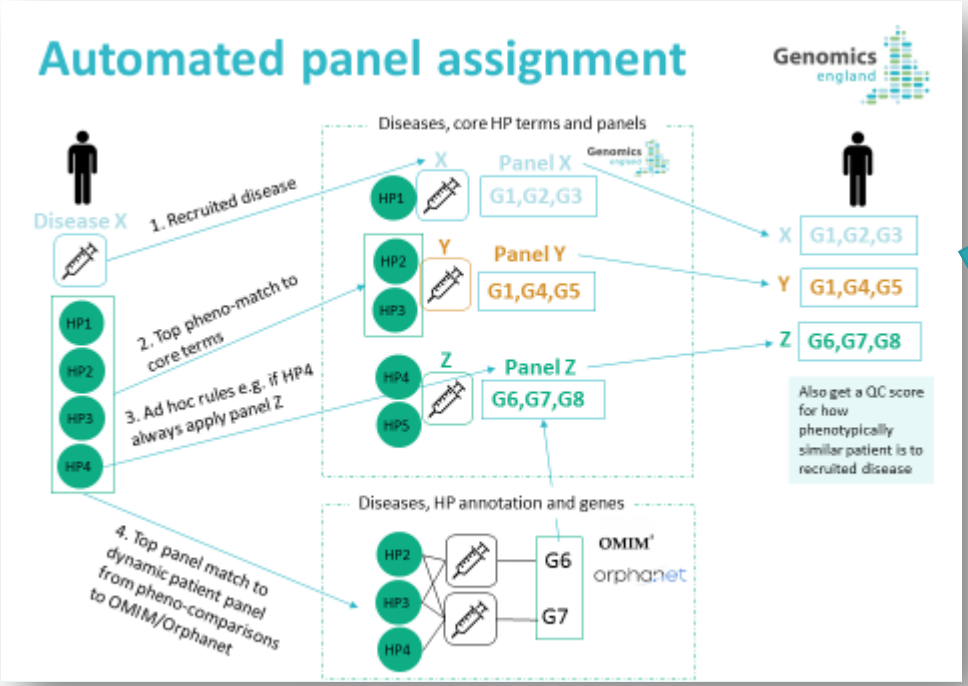
Family Id	98
Participant Id	98
Forenames	Joseph Timothy
Surname	Taylor
Date of Birth	23/12/1936
NHS Number	1112220098
CHI Number	P0098
Person Phenotypic Sex	Male
Relationship	Proband
Disease Status	Erythropoietic protoporphyria, mild variant
Vital status	Alive

Rare Disease Diagnoses

Specific Disease	Age of Onset
Erythropoietic protoporphyria, mild variant	6

Other Diagnoses

Specific Disease	Age of Onset	Code Type
Other SNOMED Disease 5	6	SnomedCT



The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

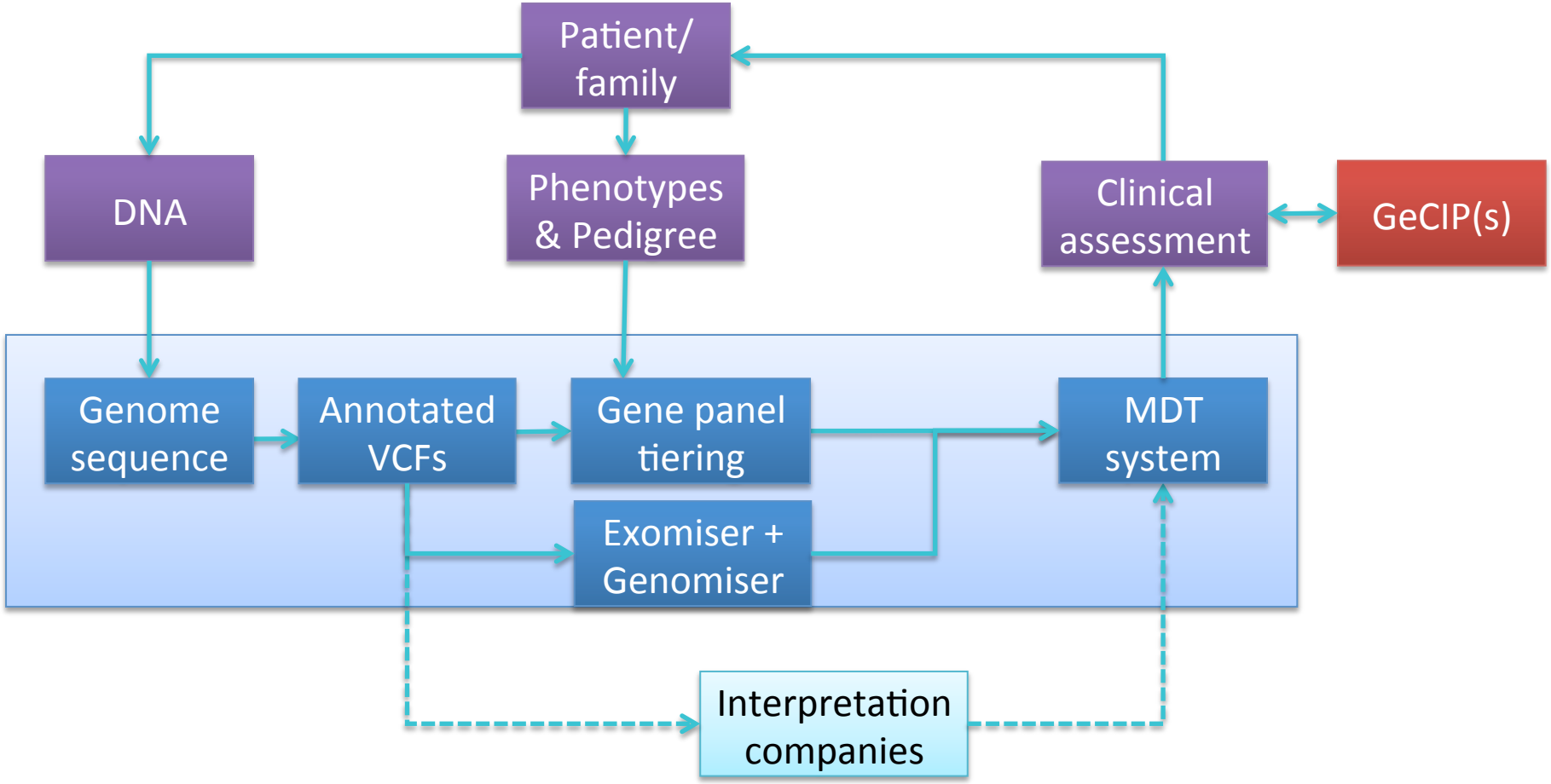
Genomics
england

monarch
INITIATIVE

PanelAssigner
Recovers more than 50% of the diagnoses located outside clinically selected panels.

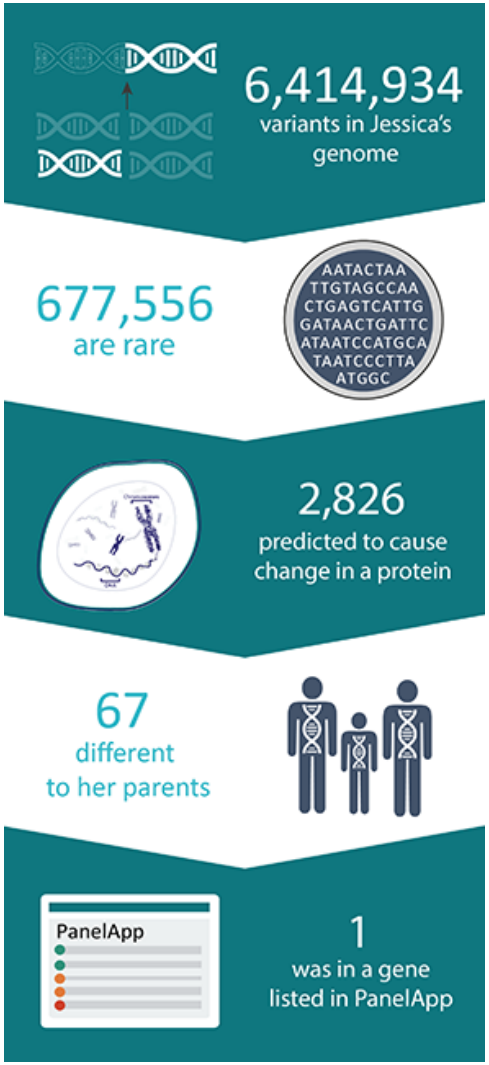
Scalable rare disease diagnostics

Genomics england



omicia CONGENICA WuXiNextCODE

First families diagnosed



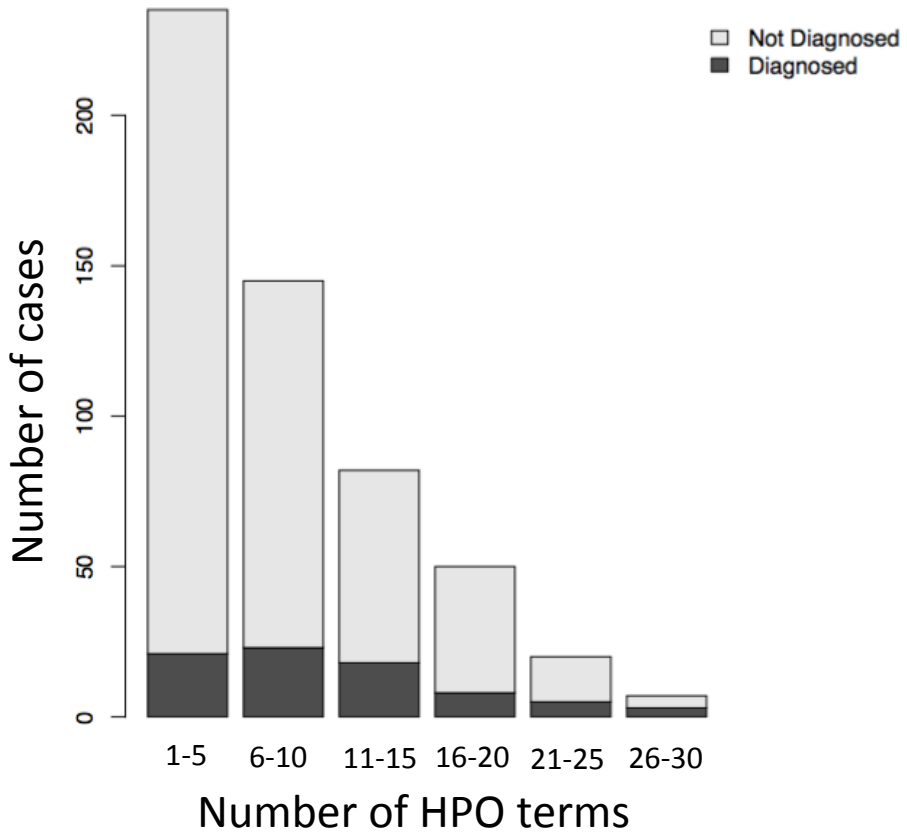
- Jessica (aged 4) has a rare condition which causes epilepsy, affects her movement and developmental delay. Standard genetics tests negative.
- De novo deletion in *SLC2A1* identified as the cause of her Glut 1 deficiency syndrome
- Now being successfully treated with a a ketogenic, low-carb diet
- Low risk for future pregnancies

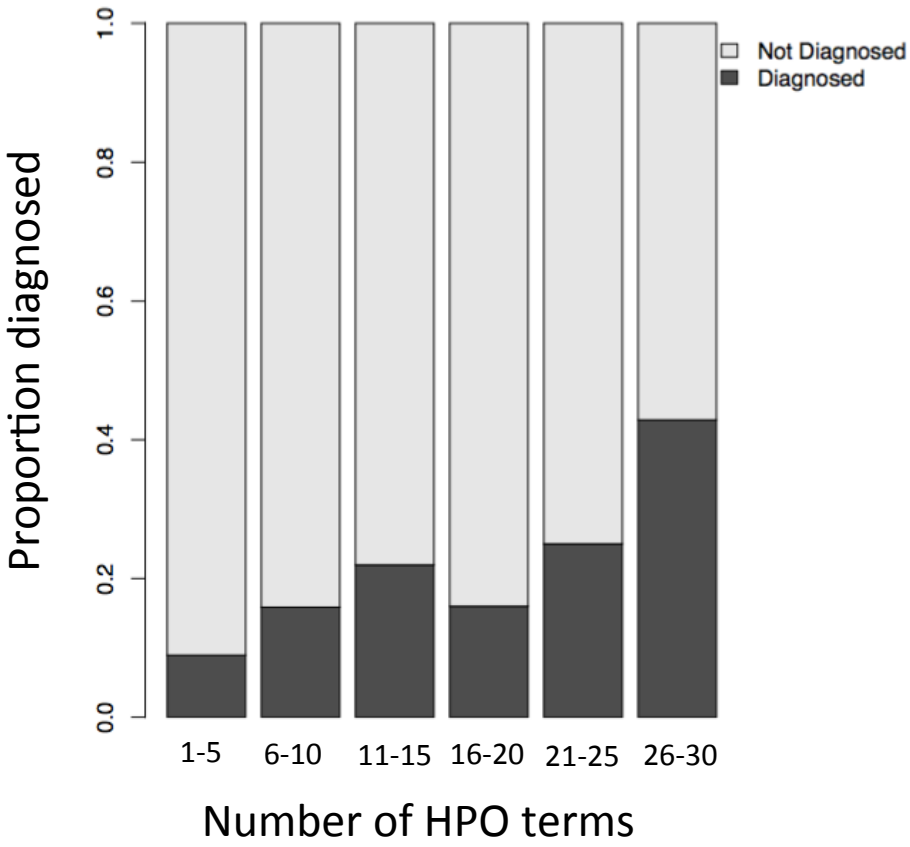
Rare disease pilot results

Genomics England

- ~4800 participants for 170 different conditions
- Standardising eligibility & phenotyping using Human Phenotype Ontology
  - 12,966 positive annotations - presence of a feature
  - 43,088 negative annotations - absence of a feature
- 250,000 hospital episodes associated with participants
- Likely diagnostic rate 20-25%

Diagnostic rate for trios

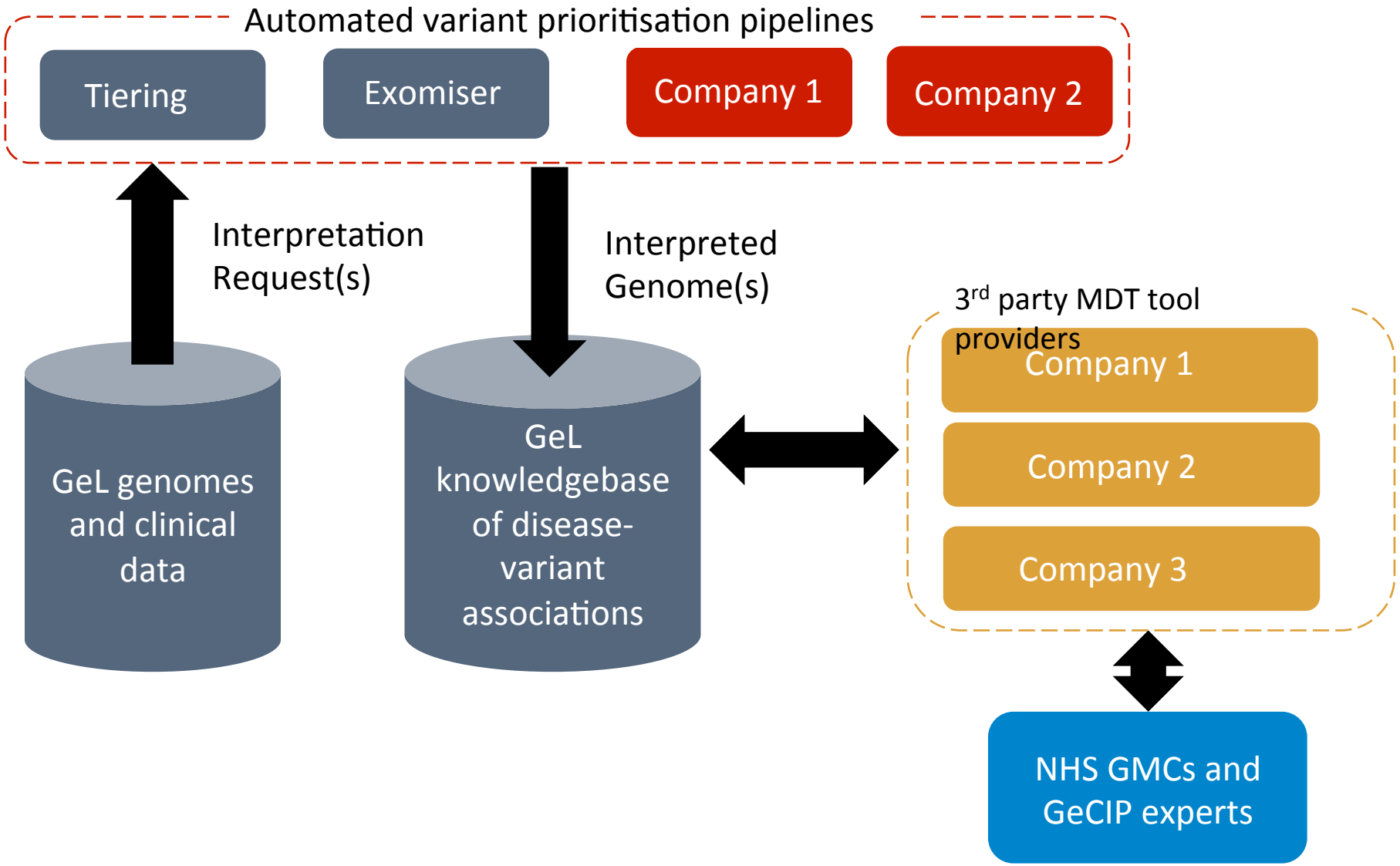




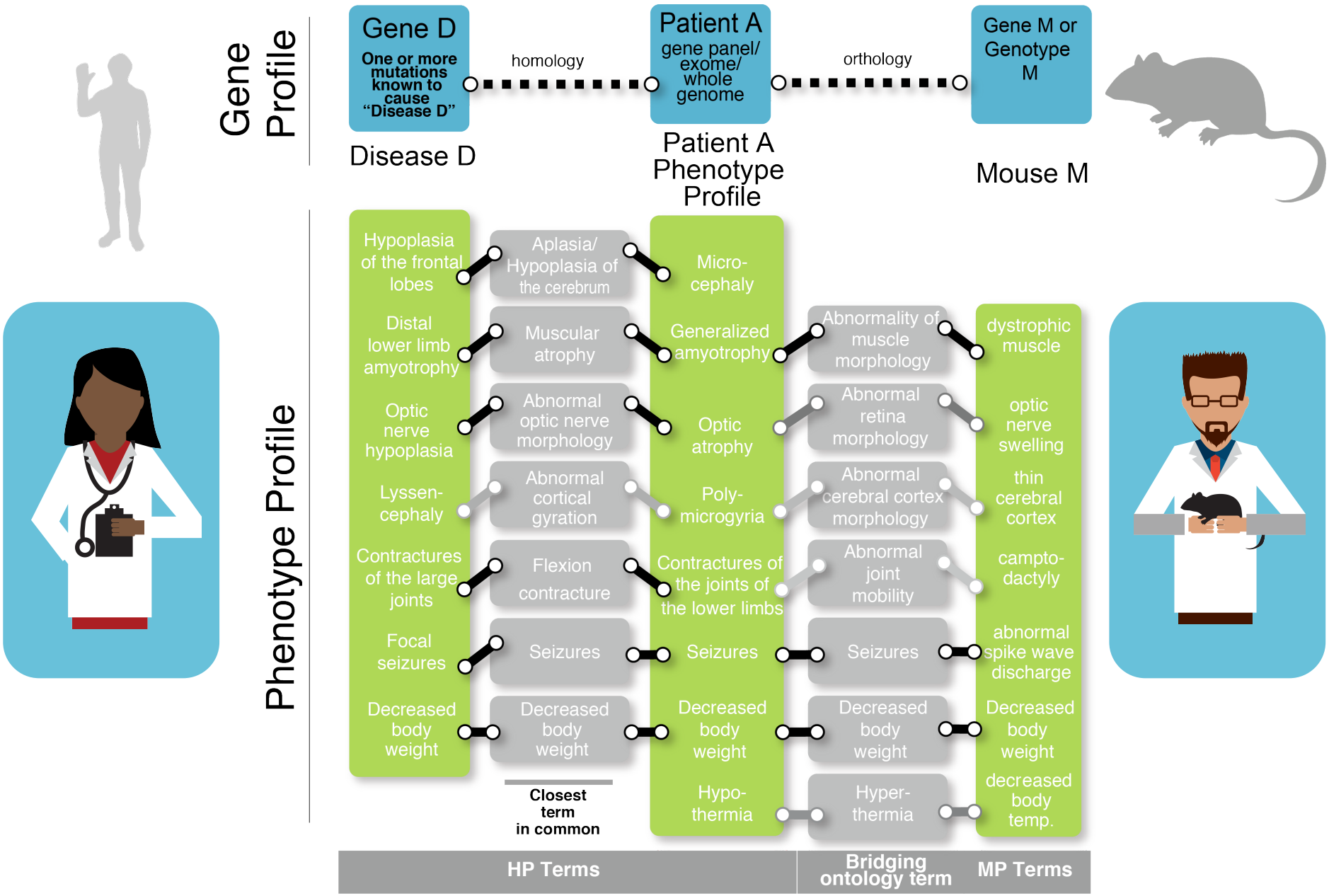
Overview

- Introduction to the 100,000 Genomes Project
- Role of clinical and model organism phenotypes
  - Clinical data collection and panel assignment
  - Automated variant prioritisation
  - Genotype to phenotype knowledgebase

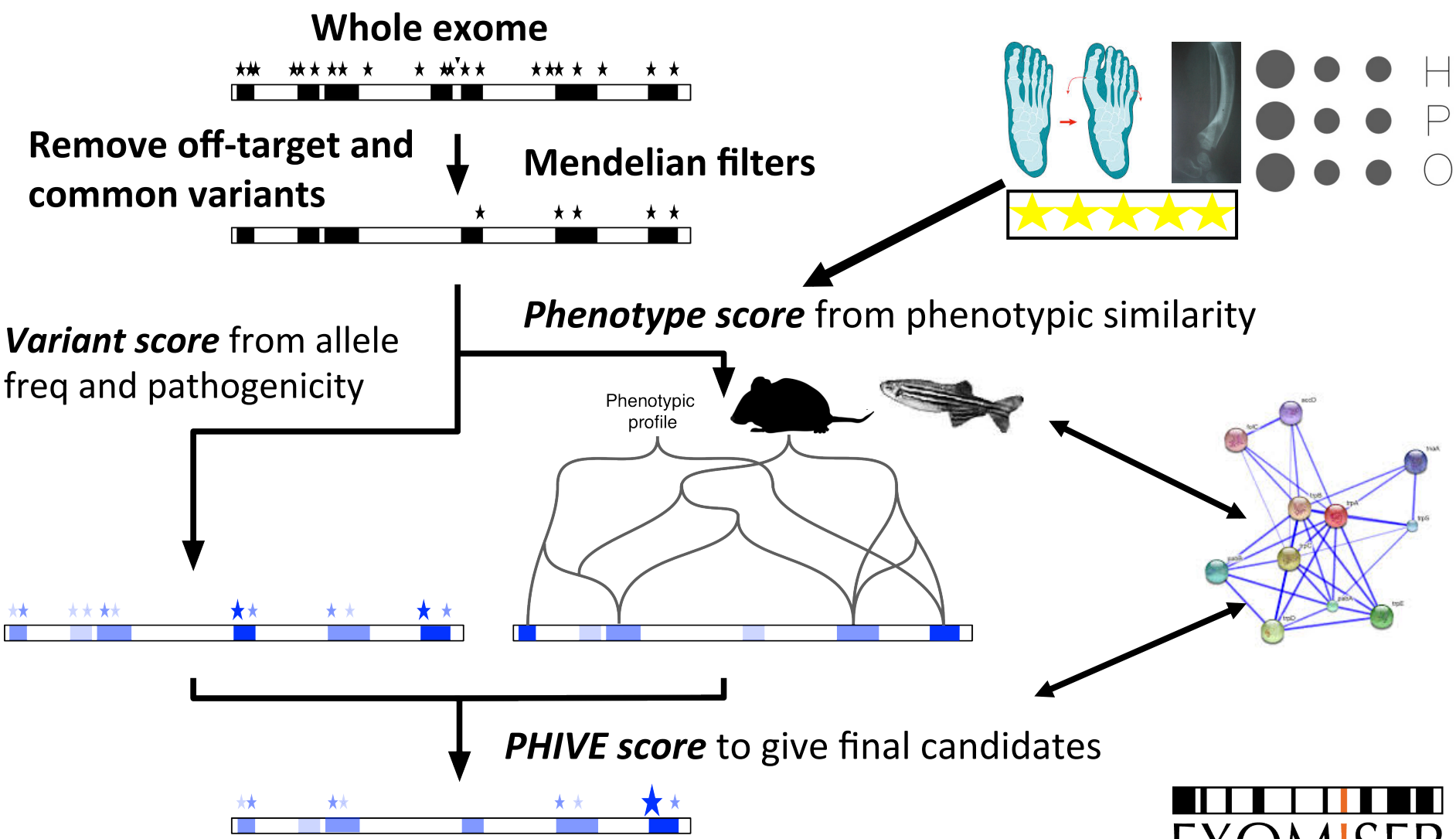
Interpretation ecosystem



Precision fuzzy phenotype matching



Combining G2P data for variant prioritization



EXOMISER

Exomiser software suite

Genomics england

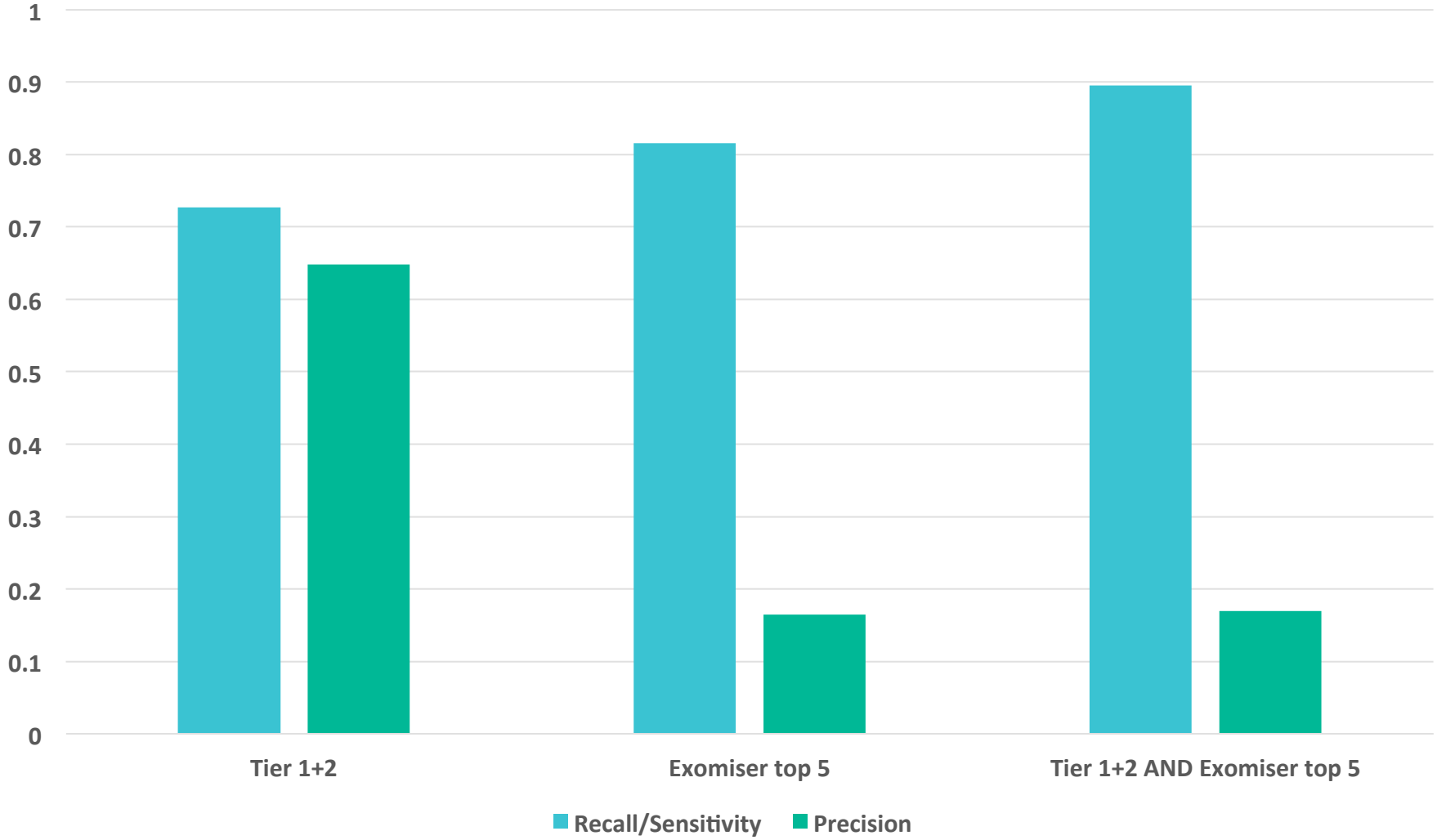
- **How-To guide**: Smedley D et al ***Nature Protocols*** 2015 **10**(12):2004-15

- **PHIVE**: Robinson PN et al. ***Genome Research*** 2014. **24**(2):340-8.
- **hiPHIVE:** Bone W et al. ***Genetics in Medicine*** 2015.
- **Genomiser**: Smedley D et al. ***Am J Hum Genet.*** 2016. **99**(3):595-606.
- **PhenIX**: Zemojtel T et al. ***Science Translational Medicine*** 2014. **6**(252).
- **ExomeWalker**: Smedley D et al. ***Bioinformatics*** 2014 **30**(22):3215.

- **BOQA – Bayesian algorithm**
- **Likelihood scoring algorithm**

Diagnoses by Exomiser and tiering

- Recovers 65% of the tier 3 diagnoses
- Recovers 57% of the untiered diagnoses



Addition diagnoses from phenotypes

- Rare, frameshift deletion in *SORD* for a patient with congenital cataracts
- Not in our panel as limited evidence in OMIM and literature
- Highlighted by Omicia's clinical team and top 5 Exomiser match based on existing mouse (spontaneous mutation removing all functional protein) used as a model of cataract development in diabetes

Omicia Opal | Menu | Clinical Reporter | Genomics England | damian.smedley@genomicsengland.co.uk | Help | Sign Out

Clinical Reports / Variant Selection / Variant Interpretation

Test: Solo
Scoring Rubric: ACMG Mendelian
VAAST Release: 3.0.4.2
Pipeline Version: 6.0.4
Interpreted By: Melanie Babcock

HPO Terms: Congenital cataract

Show/Hide Columns | Reset Filters | Bulk Update | ← Variant Selection | Review Report →

Review Priority	Gene	Position dbSNP	Change	Effect	Zygosity	Quality GQ Coverage	1KG AF GeL AF ExAC AF	Omicia Score	Evidence	Class (Condition)	VAAST gene rank	Phevor gene rank	VAAST Inheritance Model	Tier	Complete Penetrance	Scoring Status	Confirmation Status	Report Section	Latest Classificati (Date Classified) Confirmation Stat
●●●	SORD	chr15:45361216 rs55901542	CG → C c.757delG p.Ala253GlnfsTer27	frameshift	●○	149 99 16 : 10 : 6	- 0.00502 -	0.800		Uncertain Significance (Cataract)	11	13	Dominant			Classified	To Be Confirmed	Not Reported	-

1 Items

Items per page: 25

Research candidates

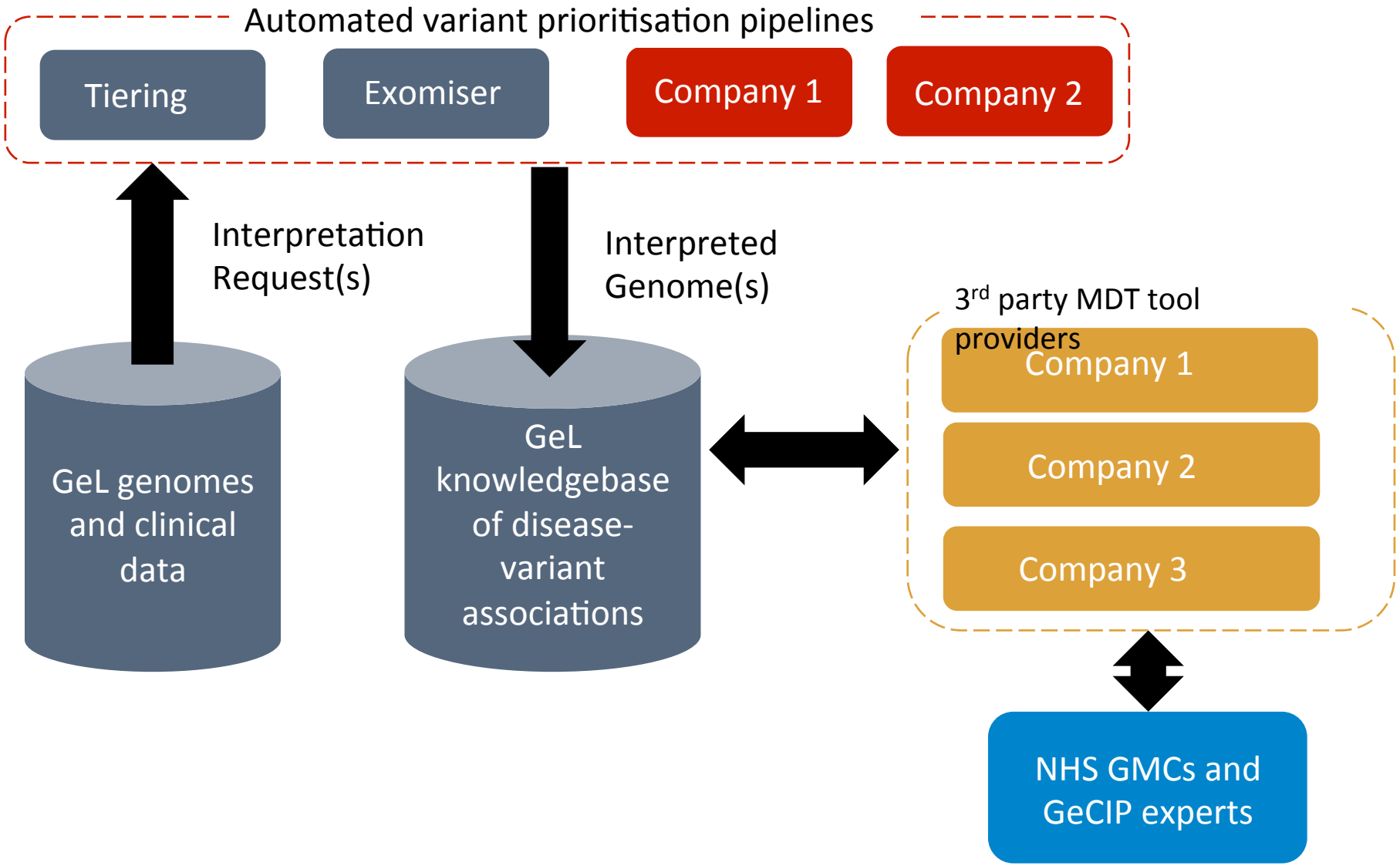
Genomics England

- Likely to have 10-15k rare disease cases without a clear diagnosis from standard pipeline
- Exomiser and Genomiser candidates, especially those based on **model organism phenotypes critical for new disease gene discovery**
- Interactions with GeCIP communities to validate
- Transcriptomics
- **Crispr/Cas9 precision animal models for functional validation and ultimately improved treatment**

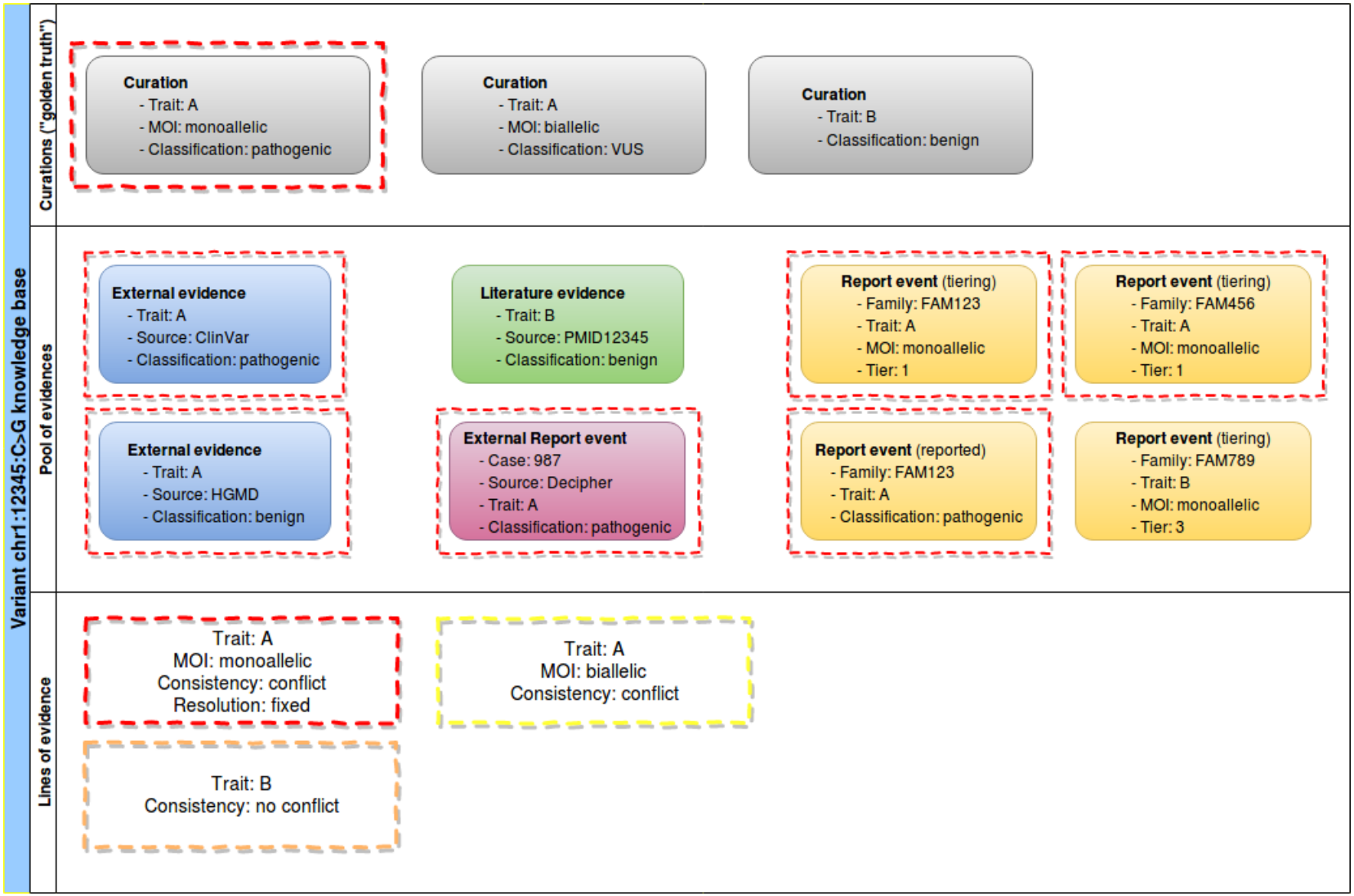
Overview

- Introduction to the 100,000 Genomes Project
- Role of clinical and model organism phenotypes
  - Clinical data collection and panel assignment
  - Automated variant prioritisation
  - Genotype to phenotype knowledgebase

Interpretation ecosystem



Clinical Variant Archive (CVA)



Challenges representing data

- **The ontology problem!** Linking evidences by phenotype, disease, panel name, … non normalised vocabulary. We cannot link together lines of evidence.
  - Storing normalised texts (lower case, no special characters, etc.) and using regex is not enough
  - Monarch Disease Ontology (MonDO) - https://github.com/monarch-initiative/monarch-disease-ontology mapping Orphanet, DO, MESH, OMIM, … MedGen to come.
  - Map phenotypes to diseases. Resources provided by HPO (http://human-phenotype-ontology.github.io/downloads.html).
  - Other approaches…

Variant classification

Variant classification is not standardised

- ACMG classification is the standard de facto for clinical relevance, but it misses other dimensions of variant classification.

Protocols

EvidenceProtocol org.opencb.b...

Types

org.opencb.biodata.models.varian

- AlleleOrigin Enum
- ClinicalSignificance Enum
- Confidence Enum
- ConsistencyStatus Enum
- DrugResponseClassificati...
- EthnicCategory Enum
- EvidenceEntry
- EvidenceImpact Enum
- EvidenceSource
- Evidenc
- Feature
- Genomi
- Heritabl
- ModeO
- Penetra
- Propert
- Somatic
- TraitAss
- Tumorigenesisclassificatio...
- VariantClassification
- VariantFunctionalEffect E...

org.opencb.biodata.models.variant.avro

VariantClassification

The variant classification according to different properties.

Type	Field	Default Value	Description
null | ClinicalSignificance	clinicalSignificance		The variant's clinical significance.
null | DrugResponseClassification	drugResponseClassification		The variant's pharmacogenomics classification.
null | TraitAssociation	traitAssociation		The variant's trait association.
null | TumorigenesisClassification	tumorigenesisClassification		The variant's tumorigenesis classification.
null | VariantFunctionalEffect	functionalEffect		The variant functional effect

org.opencb.biodata.models.variant.avro.**VariantFunctionalEffect**

Variant effect with Sequence Ontology terms.

- SO_0002052 : dominant*negative*variant (http://purl.obolibrary.org/obo/SO_0002052)
- SO_0002053 : gain*of*function*variant (http://purl.obolibrary.org/obo/SO0002053)*
- SO_0001773 : lethal*variant (http://purl.obolibrary.org/obo/SO0001773)*
- SO_0002054 : loss*of*function*variant (http://purl.obolibrary.org/obo/SO0002054)*
- SO_0001786 : loss*of*heterozygosity (http://purl.obolibrary.org/obo/SO_0001786)
- SO_0002055 : null*variant (http://purl.obolibrary.org/obo/SO0002055)*

Enum symbols:

`dominant_negative_variant`, `gain_of_function_variant`, `lethal_variant`, `loss_of_function_variant`, `loss_of_heterozygosity`, `null_variant`

Acknowledgements

- 100,000 Genomes Project
  - The patients and their families
  - NHSE staff
  - Genomics England colleagues
  - UK biobank and Illumina
  - Interpretation companies
  - GeCIP members

Genomics England

monarch INITIATIVE

NHS England

Public Health England

NHS National Institute for Health Research

CANCER RESEARCH UK