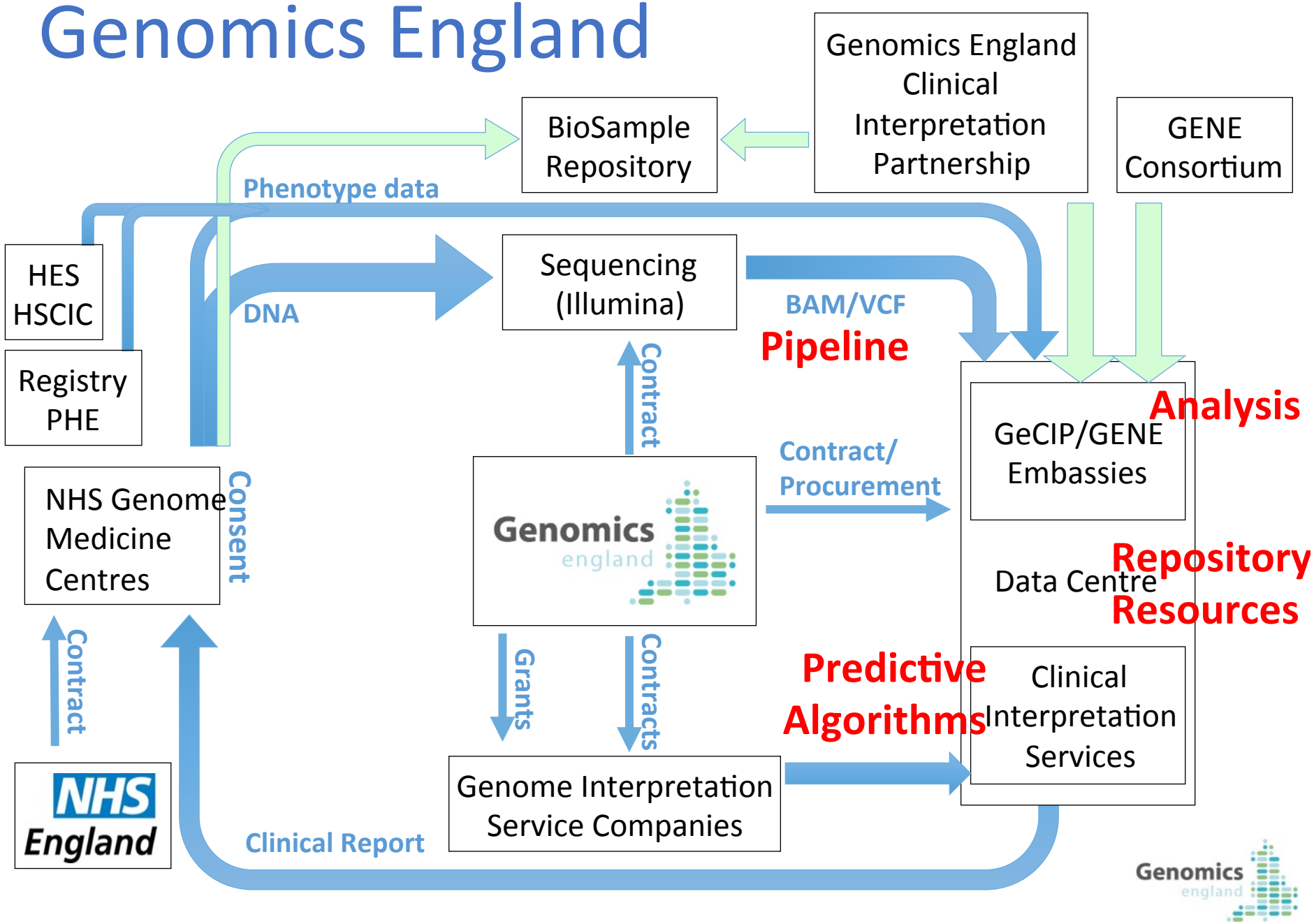
Genomics England Data: Generation, QC & Organisation

Tim Hubbard @timjph
King's College London, King's Health Partners
Genomics England

Bioinformatics, Interpretation and Data Quality in Genome Analysis
MSc in Genomics Medicine
15th February 2016

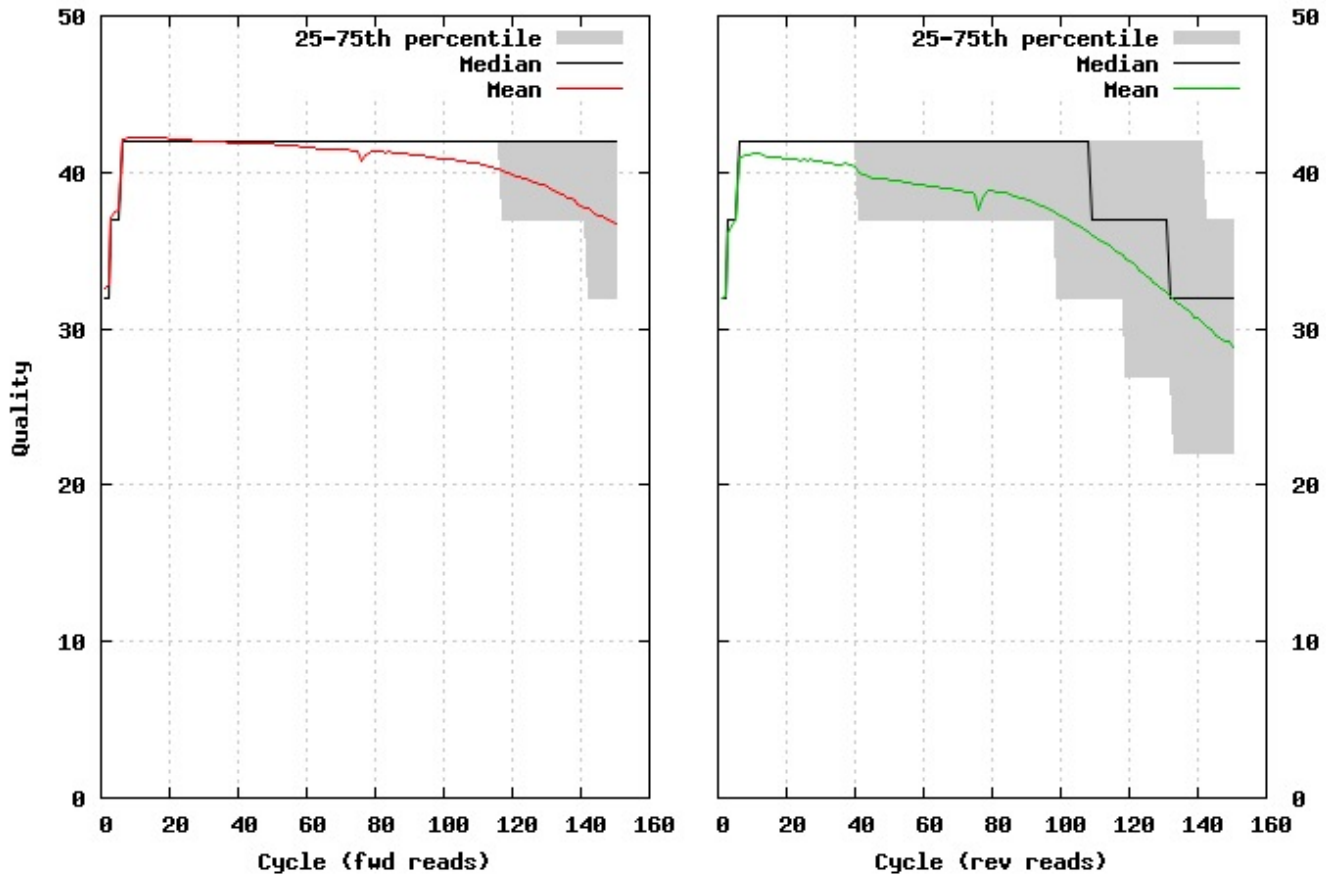
Genomics England



Data generation and processing

- Samples sent to Illumina at Great Chesterford (in future Genome Campus)
  - Sequencing
  - Mapping
  - Calling
- Sequence data sent to Genomics England Data Center by dedicated data link
  - Sequence data processed for QC
  - Variant data injested into openCB variant store
- Clinical data collected at GMCs and transferred to Genomics England Data Centre over secure N3 link
  - Clinical data injested into labkey clinical data store

Standard XTen 2x150bp PCR-free germline genome



As a minimum for a '30X' germline genome:

- 85×10^9 good bases
- **>95% of autosomal genome covered at ≥15X with "good" bases**

Current average stats:

- 97×10^9 bases
- **>97.3% of autosomal genome at ≥15X**

Tumour sequenced at '75X'

Genomics england

Alignment and variant calling

- Currently using ISAAC (HAS) on GRCh37
- Moving to GRCh38 full – ALT by Q4-2015

Pipeline	SNV (0 bp)		Indel (1-50 bp)		Time (hr)**
	Sn (%)	Sp (%)	Sn (%)	Sp (%)	
HAS (Isaac) (H2'15)	96.9	99.8	92.3	98.3	5
GATK 3.2 (BWA mem +HC)	98.1	99.9	88.6	98.9	38

Using 40 CPU, Intel Xeon @ 2.80 GHz, 132 GB RAM

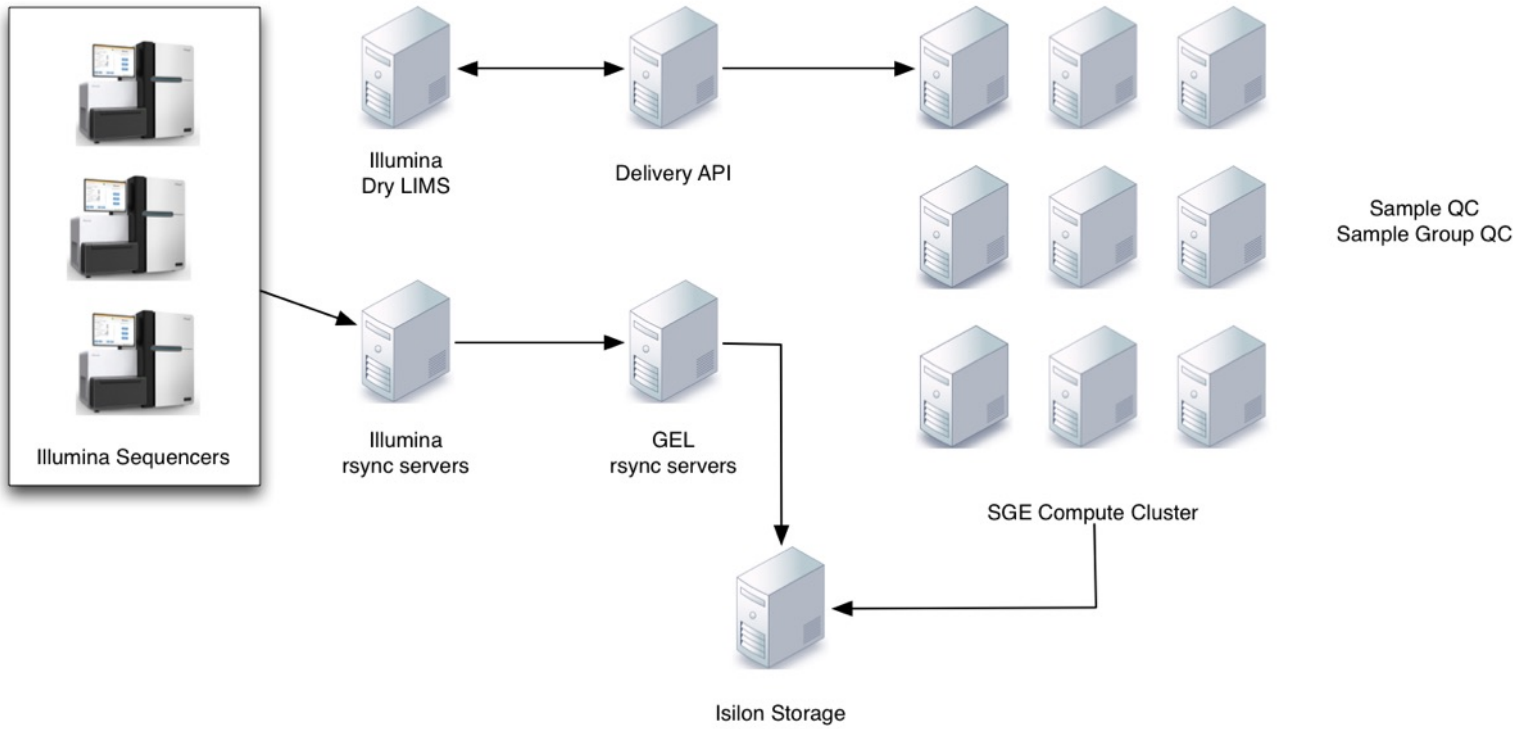
- 100,000 genomes x 5p /CPU-h

  x 200 CPU-h = £1M

  x 1500 CPU-h = £7.5M

Metrics courtesy of Illumina

Data flow from Illumina



Genomics
england

Data received from Illumina (as per V2)

- Single Sample (Germline and Cancer)
  - BAM e.g. *Assembly/LP9006336-DNA_A01.bam* (~60GB)
  - Variants
    - SV+CNV *Variations/LP9006336-DNA_A01.SV.vcf.gz*
    - Small *Variations/LP9006336-DNA_A01.vcf.gz*
    - Small gVCF *Variations/LP9006336-DNA_A01.genome.vcf.gz (~ 500MB)*
  - Metrics + SummaryReport

    Metrics/LP9006336-DNA_A01.Metrics.csv

    Metrics/LP9006336-DNA_A01.baseCompositionPerCycle.csv

    Metrics/LP9006336-DNA_A01.GCDistribution.csv

    Metrics/LP9006336-DNA_A01.insertSizeHistogram.csv

    Metrics/LP9006336-DNA_A01.Qscore_mean_byCycle.csv

    Metrics/LP9006336-DNA_A01.uniformityOfCoverage.csv

Genomics
england

Data received from Illumina (as per V2)

- Paired analysis (Somatic Calls)
  - BAM e.g. *Assembly/LP1000058-DNA_D04.bam* (~200GB)
  - Somatic Variants
    - SV *SomaticVariations/CancerLP1000058-DNA_D04_NormalLP1000059-DNA_C06.somatic.SV.vcf.gz*
    - CNV *SomaticVariations/CancerLP1000058-DNA_D04_NormalLP1000059-DNA_C06.somatic.SV.vcf.gz*
    - Small *SomaticVariations/CancerLP1000058-DNA_D04_NormalLP1000059-DNA_C06.somatic.vcf.gz*
  - Metrics + SummaryReport

    Metrics/CancerLP1000058-DNA_D04_NormalLP1000059-DNA_C06.Metrics.json

    CancerLP1000058-DNA_D04_NormalLP1000059-DNA_C06.SummaryReport.pdf

Closest documentation:
https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/wgs/fasttrack-whole-genome-sequencing-services-user-guide-15040892-d.pdf

Genomics england

QC pipeline

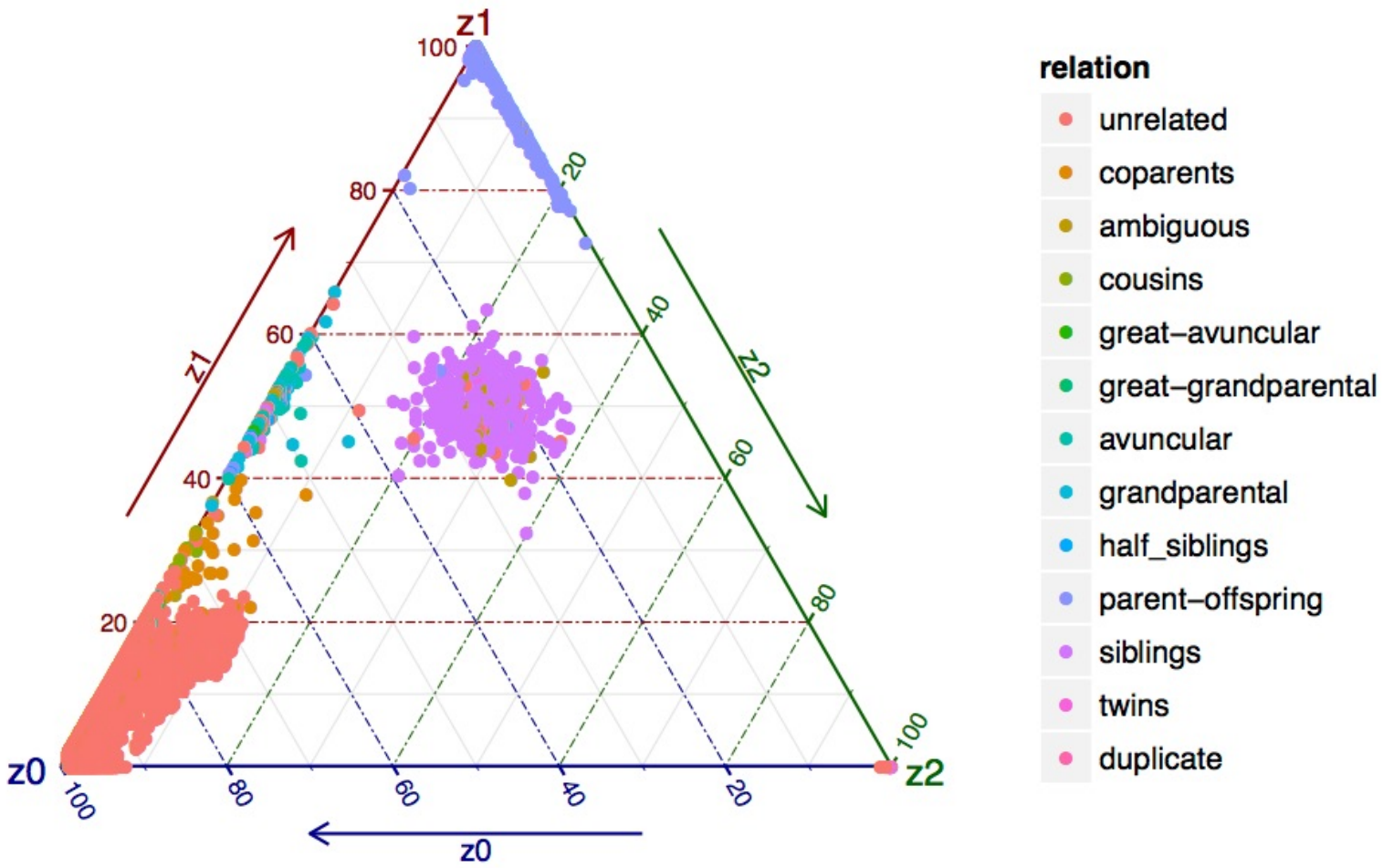
- Automatic stats and checks
  - Delivery integrity – custom Pipeline Pilot protocol
  - BAM and VCF files – picard ValidateSAMFile & bcftools
  - Coverage (perc_bases_ge_15x_mapQ_ge11 > 95%) – samtools stats
  - Number of bases (GbQ30NoDupsNoClip > 85G) – custom pysam based script
  - Samtools stats – samtools stats
  - BCFtools stats – bcftools stats
  - Verifybamid
- Semi-automatic stats and checks
  - Sex
  - Mendelian errors
  - Inbreeding estimates
  - IBD estimation
  - Ancestry

Genomics
england

Genome version

- Currently primary assembly of GRCH37
- Moving to GRCH38+ummaped+unplace+decoy-ALTs by early Jan 2016

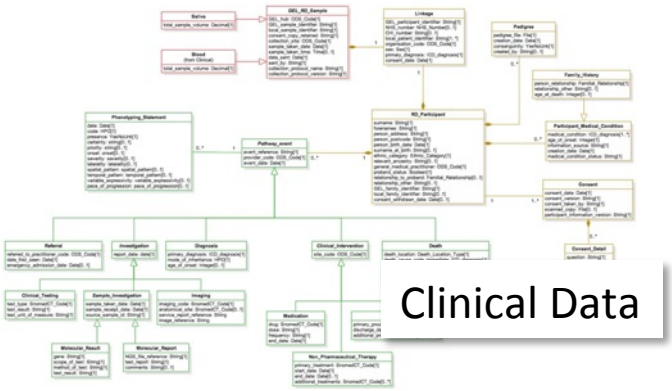
Relatedness checks



Proportion of the genome where 0,1 or 2 alleles are shared between 2 individuals

Most data is modelled

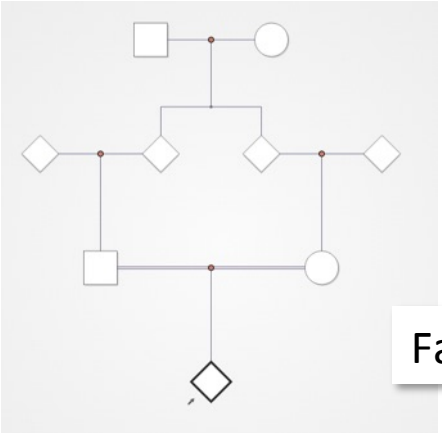
- Model lives separate from the code



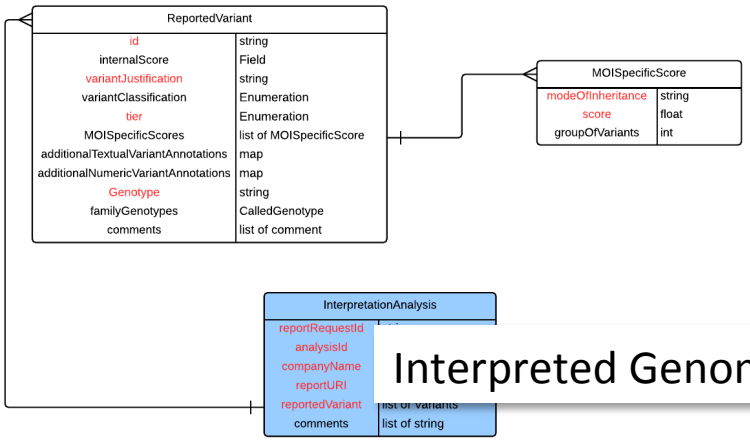
Clinical Data

Promotes:

- Standardisation
- Automation
- API development
- Facilitates system evolution



Family Pedigree



Interpreted Genome

Clinical data models

- Which participants should we recruit?
  - List of conditions: **Eligibility statements**
- What data do we need?
  - Metadata: Demographics, Sample, Consent
  - Clinical data (including family pedigrees and clinical tests): **Data models**
  - Associated genes: **Gene packages**

Genomics England

Data models are specific to each rare disease

Level 1	Level 2	Level 3	Level 4
Rare Disease Conditions and Phenotypes(11144.4)	Cardiovascular disorders(10950.1)	Connective Tissues Disorders and Aortopathies(10951.1)	Familial Thoracic Aortic Aneurysm Disease(11021.1)
		Cardiac arrhythmia(10952.1)	Brugada syndrome(11022.1)
			Long QT syndrome(11023.1)
			Catecholaminergic Polymorphic Ventricular Tachycardia(11024.1)
		Cardiomyopathy(10953.1)	Arrhythmogenic Right Ventricular Cardiomyopathy(11025.1)
			Left Ventricular Noncompaction Cardiomyopathy(15044.1)
			Dilated Cardiomyopathy (DCM)(11026.1)
			Dilated Cardiomyopathy and conduction defects(11027.1)
			Hypertrophic Cardiomyopathy(11028.1)
		Congenital heart disease(10954.1)	Fallots tetralogy(11029.1)
			Hypoplastic Left Heart Syndrome(11030.1)
			Pulmonary atresia(11031.1)
			Transposition of the great vessels(11032.1)
			Left Ventricular Outflow Tract obstruction disorders(11033.1)
			Isomerism and laterality disorders(11034.1)

OpenClinica phenotype entry

Disease

1 Disease Group: Renal and urinary tract disorders *

2 Disease Subgroup: Syndromes with prominent renal abnormalities *

3 Specific disease: Alport syndrome *

OpenClinica
Open Source for Clinical Research

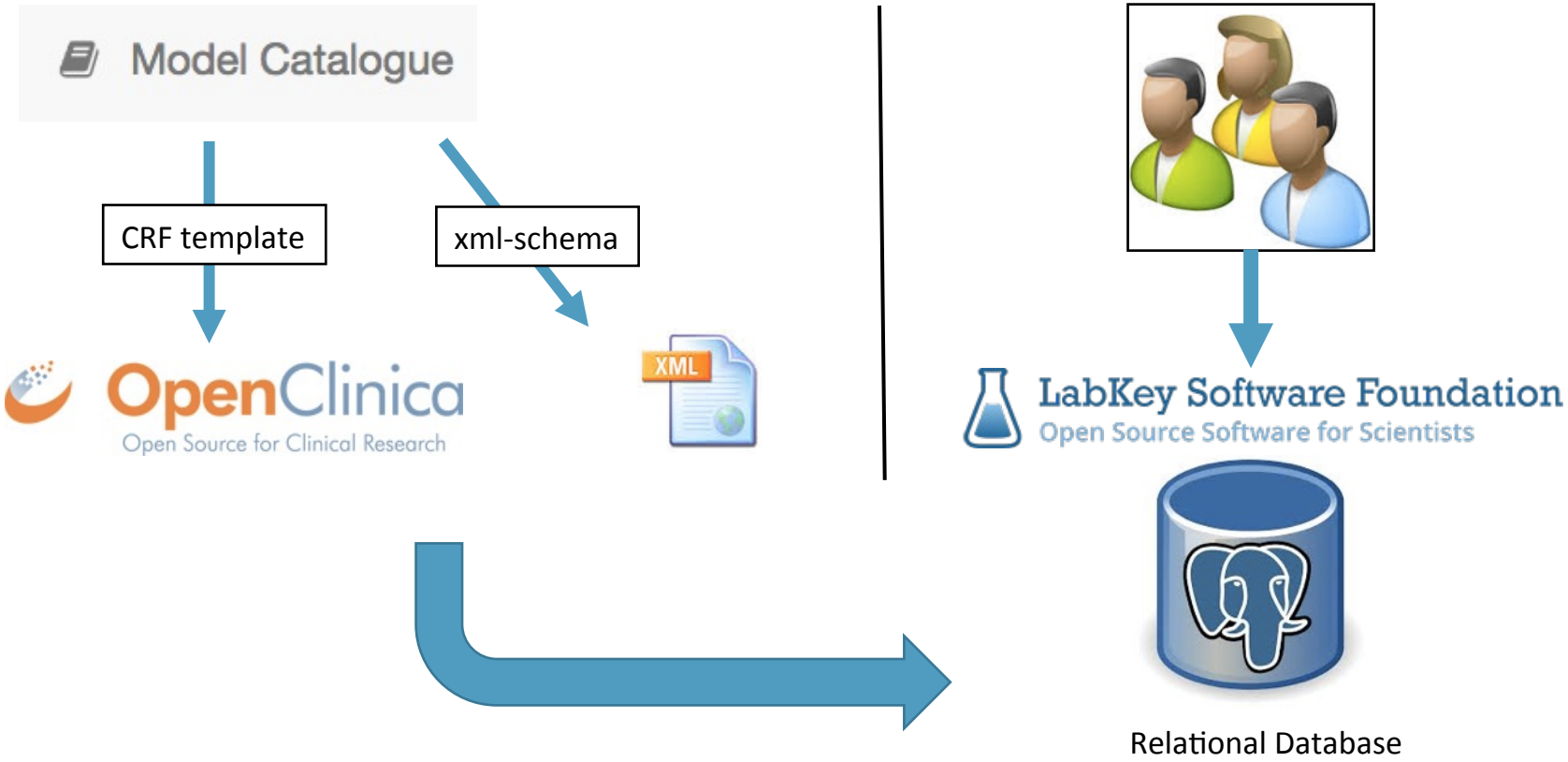
Basic Phenotyping

4 Phenotype Description	5 Phenotype Identifier	7 Phenotype Present	Modifiers	Actions
Proteinuria	HP:0000093	Unknown / Yes / No		Edit
Hematuria	HP:0000790	Unknown / Yes / No		Edit
Nephrotic range proteinuria	HP:0012593	Unknown / Yes / No		Edit
Renal insufficiency	HP:0000083	Unknown / Yes / No		Edit

Additional terms not present in the data model can be naturally added

Genomics England

Information management for clinical data



Genomics England

Processing, analysis and presentation of sequence data

Algorithms for Clinical Interpretation

Genome interpretation challenge

An optimistic estimate

- 25,000 cancer patients + 25,000 families with RD
- 6 reports /person-day x 2.5 years ≅ 15 people

A mixed model for clinical interpretation and reporting

- Genome interpretation produced by third party
- Genomics England performs sanity checks
- GMCs review findings and GECIPs provide expert assesment

Genomics
england

Dual approach to variant prioritisation to enrich and accelerate clinical interpretation

RANKING and ANNOTATION

- Pathogenicity predictors based on knowledge bases (various commercial vendors)
- Population and family based ranking (vaast and p-vaast, etc)
- Phenotypic based ranking (PhenIX, Exomiser, Phevor, etc)
- Consequence and function prediction tools

TIERING

After surviving inheritance and frequency filters

- Tier 1:
  - In gene panel
  - Clear LOF (truncating, splicing, etc)
  - Known pathogenic variants
- Tier 2:
  - In gene panel
  - Missense and other VUS
- Tier 3:
  - Not in panel
  - **Ranking is critical here**

Genomics England

PanelApp

A crowdsourcing tool for gene panels

- A publically-available resource that allows gene panels to be viewed, downloaded and evaluated by the Scientific Community.

- Initial gene panels have been established for all the approved rare diseases (Version 0), and graded using a traffic light system to indicate the number of sources.

- We are seeking expert review of these panels.

Aims:

- Source expert knowledge to establish a **final diagnostic grade gene panel** (or "green list") for each disorder that will be used in the classification of genetic variants to aid clinical interpretation of rare disease genomes (Version 1).

- Engage the Scientific Community, encourage open debate, and begin to establish consensus on gene panels for rare diseases.

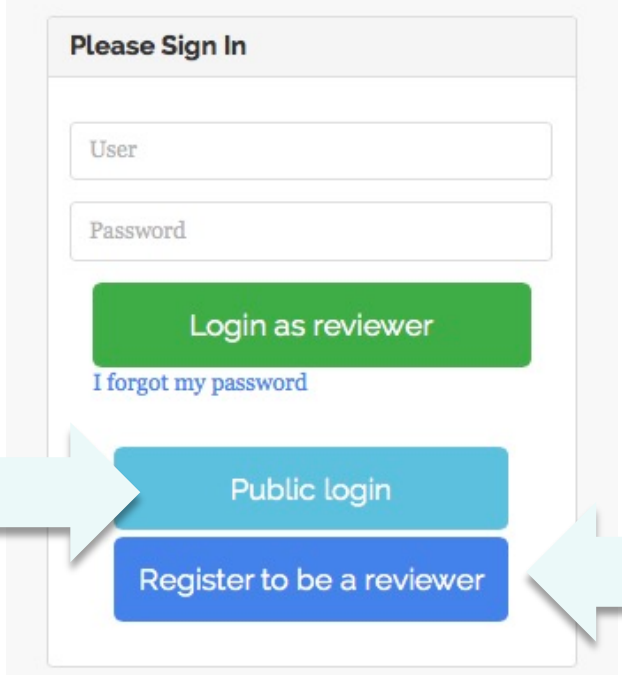
- A mechanism to allow access to the panels, standardisation of terms and collection of gene-disease related information, accumulation of reviews over time, and updated releases (Version 2…).

PanelApp

https://bioinfo.extge.co.uk/crowdsourcing/PanelApp/

Public access

- View and download gene panels.
- View Reviewers' comments.



Register to be a reviewer

- View and download gene panels.
- View Reviewers' comments.

+ Evaluate genes and make comments.
+ Add genes to a gene panel.

Searching Panels

Genomics England

PanelApp Query panels Panels Anonymous Log out

Panels

Panel	Evaluated genes	Reviewers ↑
cardiovascular		
Arrhythmogenic Right Ventricular Cardiomyopathy Level 3: Cardiomyopathy Level 2: Cardiovascular disorders Version 0.0	14 of 14 100%	2 reviewers
Familial Thoracic Aortic Aneurysm Disease Level 3: Connective Tissues Disorders and Aortopathies Level 2: Cardiovascular disorders Version 0.3	28 of 47 59%	2 reviewers
Familial hypercholesterolaemia Level 3: Arteriopathies Level 2: Cardiovascular disorders Version 0.70	41 of 41 100%	2 reviewers
Catecholaminergic Polymorphic Ventricular Tachycardia Level 3: Cardiac arrhythmia Level 2: Cardiovascular disorders Version 0.0	2 of 6 33%	1 reviewer
Fallots tetralogy Level 3: Congenital heart disease Level 2: Cardiovascular disorders Version 0.0	8 of 8 100%	1 reviewer

PanelApp Gene Panel View

Genomics England

17 genes

17 of 17 reviewed

List	Gene	Reviews	Mode of inheritance	Source of Evidence	Phenotypes
Filter genes					
Green	BTK	4 reviews 4 green	X-LINKED: hemizygous mutation in males, monoallelic mutations in females may cause disease (may be less severe, later onset than males)	Illumina TruGenome Clinical Sequencing Services UKGTN Radboud University Medical Center, Nijmegen	Agammaglobulinemia, X-linked; Agammaglobulinemia, X-linked 1, 300755Agammaglobulinemia and isolated hormone deficiency, 307200
Amber	PIK3R1	4 reviews 4 green	Not set	Radboud University Medical Center, Nijmegen UKGTN	Agammaglobulinemia 7, autosomal recessive, 615214SHORT syndrome, 269880
Red	BLNK	3 reviews 3 green	Not set	Radboud University Medical Center, Nijmegen	Agammaglobulinemia 4, 613502

Genome interpretation providers

- Currently contracting a pilot phase for up to 8000 “reports” with four providers

- Genomics England will provide web-based tools to enable the collaboration between GEL, GECIPs and GMCs to analyse, assess, review and validate the clinical interpretation of whole genomes

Genomics
england

Resources to support Research Analysis

Data management and presentation: OpenCB (github.com/opencb)

OpenCGA

- **Catalog**: metadata store
- **Variant**: variant database

APIs

- openCGA4Gel: python API to OpenCGA web services
- R interface in development

Cellbase

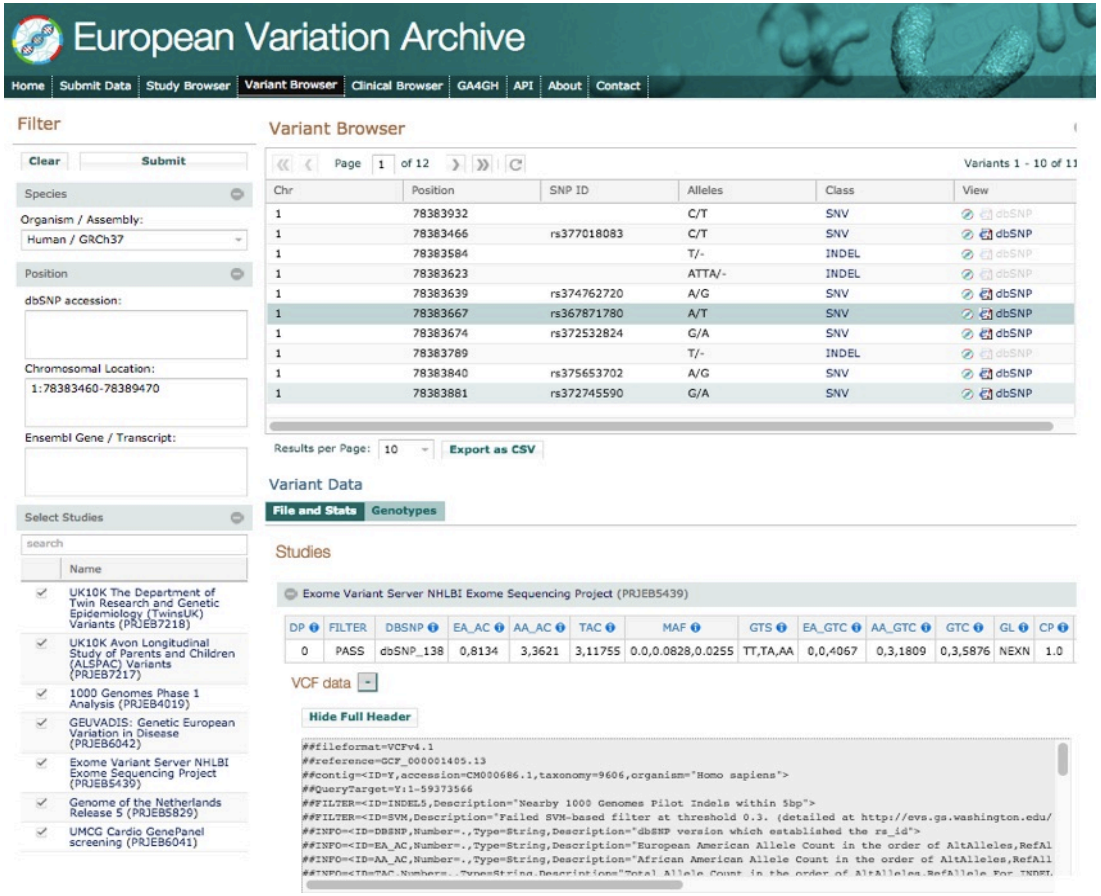
- reference data store

Genomics
england

Variant stores for >1B variants and >1M participants

Support two main use cases:

- Low latency queries from decision support systems and genome browsers
- Large scale processing for cohort analysis



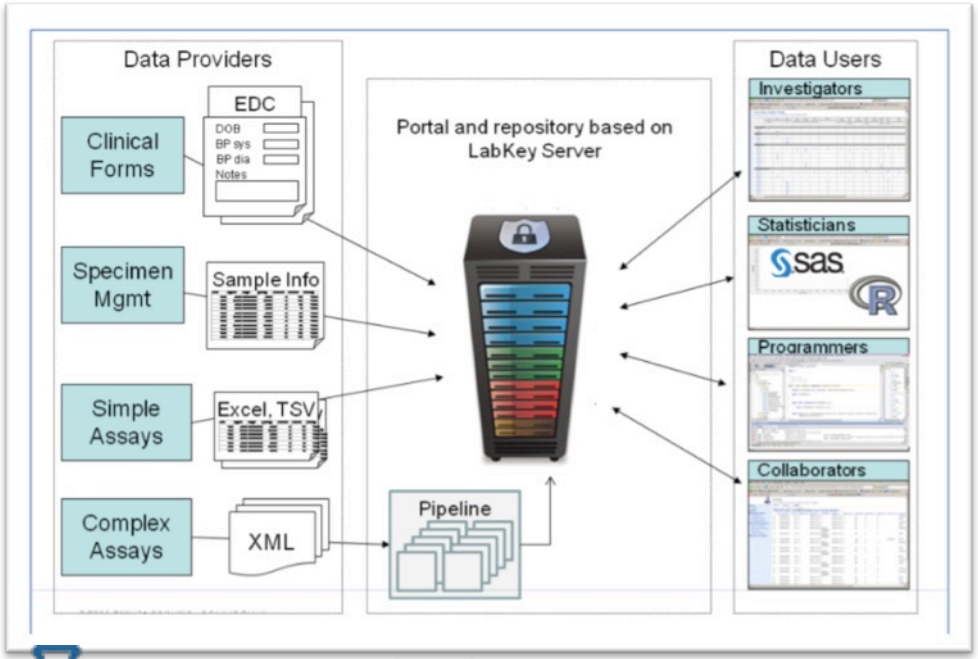
gVCF ➔GA4GH-Avro ➔ Parquet ➔ HBase

Cohort analysis

Low latency

Putting it all together through rich APIs

Clinical Data



LabKey Software Foundation
Open Source Software for Scientists

Genomic Data

+

OpenCB

The users decide how to analyse the data, we just make it easily accessible

Genomics England

Analysis

Matthew Parker

Katherine Smith

Antonio Rueda

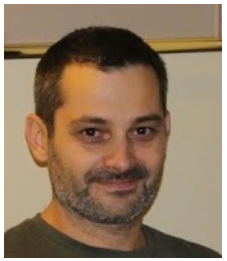
David Montaner

Alona Sosinisky

Curation

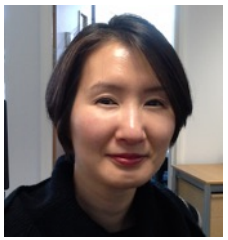
Ellen McDonagh

Pipelines



Razvan Sultana

Duncan Gordon



Mikyung Jang

Software

Ignacio Medina

Jacobo Coll

Pawan Pal

Kalyan Reddy Emani

Clinical Interpretation

Eik Haraldsdottir

Director of Bioinformatics

Augusto Rendon

Genomics england

Acknowledgements

Special thanks

- Cambridge, UCLH, GOSH, Moorfields, Newcastle, Manchester, Guys and St Thomas's, Oxford, Liverpool, Sheffield, Leeds, Birmingham, Royal Marsden, Southampton, UK CLL Consortium, CRUK, RCPath, NHS England, Department of Health, Biobank UK, Sanger, EBI, KCL, UCL and QMUL

All Genomics England Teams:

- Science, Operations, Informatics, **Bioinformatics**, Legal

All advisory committees and working groups:

- Science, ethics, data, cancer, rare diseases, molecular pathology

NHS England

NHS National Institute for Health Research

MRC Medical Research Council

wellcome trust

CANCER RESEARCH UK

genomics england