Introduction to Genomics England

MED676 MSc in Genomic Medicine

Matthew Parker, Ph. D
Lead Bioinformatician
Sheffield Diagnostic Genetics Service

Sheffield Children's NHS
NHS Foundation Trust

Introduction

- Company owned wholly by the department of health
- Delivery of 100k genome sequences – around 40,000 affected individuals
  - Rare disease families
  - Cancer tumour/normal pairs
- NHS transformation
- Education
- Collaborations with Research/Pharma
- Currently around 17,000 genomes sequenced

Genomics england

Genomics England Limited are tasked with delivering the 100KGP

Mission Statement

"Sequence 100,000 genomes from patients with cancer, rare disorders, and infectious disease, and to link the sequence data to a standardised, extensible account of diagnosis, treatment, and outcomes. "

Aims

- **Patient benefit:** providing clinical diagnosis and in time, new or more effective treatments for NHS patients.
- **New scientific insights and discovery:** with the consent of patients, creating a database of 100,000 whole genome sequences linked to continually updated long term patient health and personal information for analysis by researchers.
- Accelerating the **uptake of genomic medicine in the NHS**: working with NHSE and other partners to deliver a scalable WGS and informatics platform to enable these services to be made widely available for NHS patients & creating a mechanism to both continually improve the accuracy and reliability of information fed back to patients and add to knowledge of the genetic basis of disease.
- Stimulating and enhancing **UK industry and investment**: by providing access to this unique data resource by industry for the purpose of developing new knowledge, methods of analysis, medicines, diagnostics and devices.
- Increasing **public knowledge** and support for genomic medicine: delivering an ethical and transparent program which has public trust and confidence and working with a range of partners to increase knowledge of genomics.

Why Genomes?

- Why Sequence Whole Genomes?

Whole Genomes

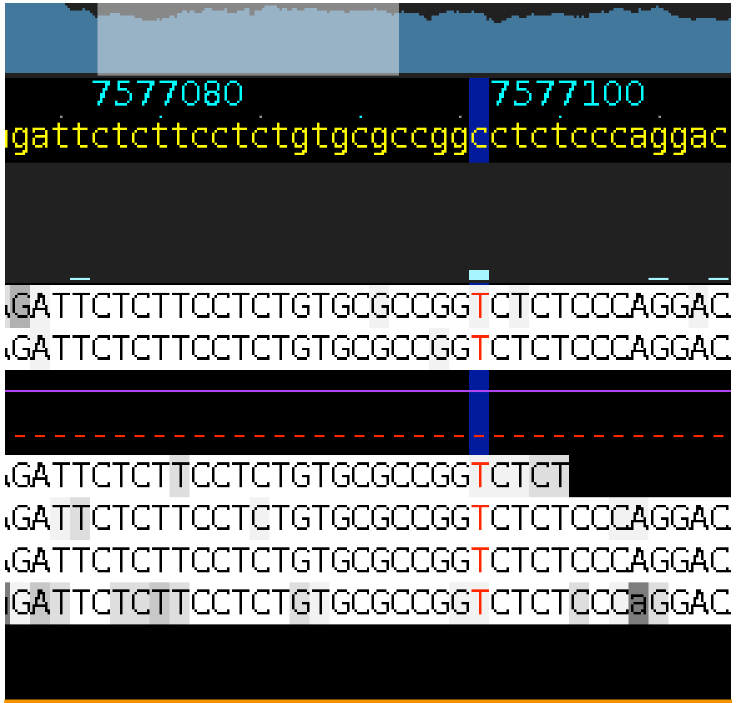
- Patients without the "normal" genetic aberrations associated with their phenotype – where to look?
- Cost reducing
- Non-coding changes could result in disease
- Extra Value:
  - More reliable
    - Small Insertion Deletions Calling
    - Copy Number Calling
    - Structural Variant Calling
- Better representation of true variant allele frequency (cancer)
- Easy to look at "new" causes of disease

HiSeq Xten WGS Sequecning

Can anyone think of any reasons why genomes might not give us all the answers?

Problems With Short Read WGS Data?

- Short reads (150bp) mean Illumina WGS is not great for:
    - Pseudogenes & Gene duplications
    - Repetitive regions
    - Regions of high GC/AT
    - Very complex structural variations
- Incidental findings
- Findings with little evidence of effect
- Slow
- Huge amounts of data that need to be kept for 30yrs


7577080
7577100
gattctcttcctctgtgcgccggcctctcccaggac
GATTCTCTTCCTCTGTGCGCCGGTCTCTCCCAGGAC
GATTCTCTTCCTCTGTGCGCCGGTCTCTCCCAGGAC
GATTCTCTTCCTCTGTGCGCCGGTCTCT
GATTCTCTTCCTCTGTGCGCCGGTCTCTCCCAGGAC
GATTCTCTTCCTCTGTGCGCCGGTCTCTCCCAGGAC
GATTCTCTTCCTCTGTGCGCCGGTCTCTCCCaGGAC

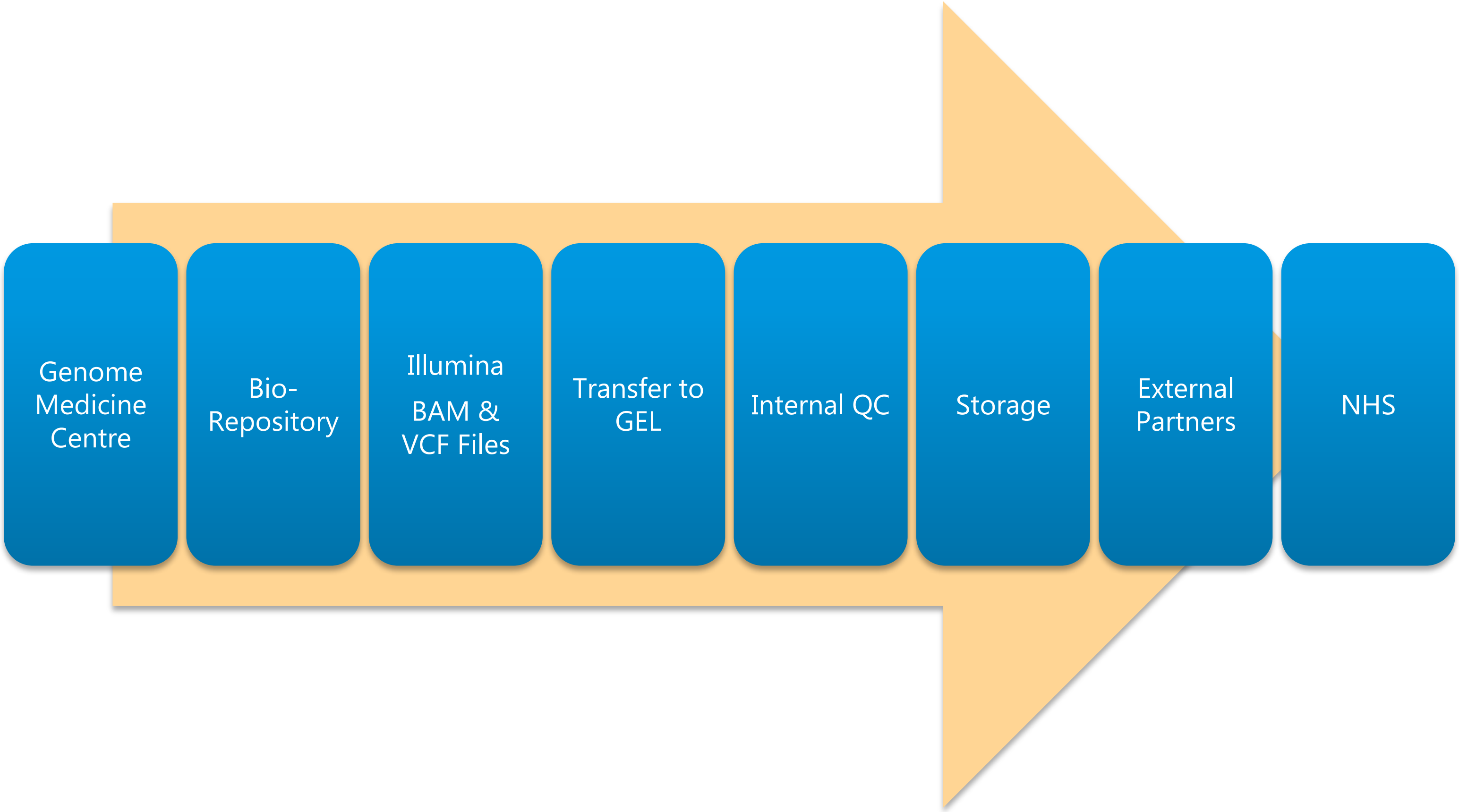
Short reads not always appropriate

Data Generation

- How do Genomics England generate WGS data for the NHS?

Overview of Sample → Result

Sheffield Children's NHS
NHS Foundation Trust



Genome Medicine Centres

- Multi-Disciplinary teams
- Patient recruitment with complete phenotyping information
- Paid per patient recruited – but only if full information provided
- http://compbio.charite.de/hpoweb/showterm?id=HP:0002131
- Reporting to patients

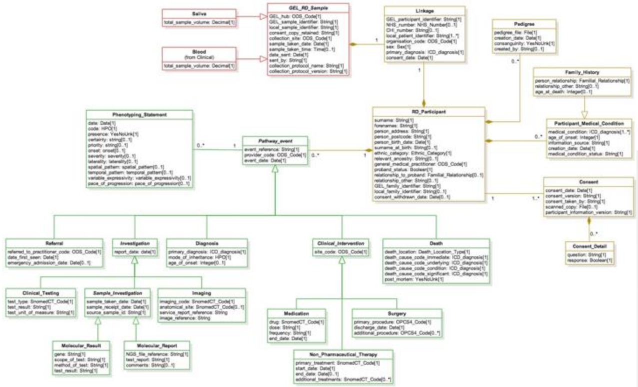
OpenClinica

OpenClinica is used to collect patient information

Key Asset of Genomics England

Sheffield Children's NHS
NHS Foundation Trust

CURATED CLINICAL PHENOTYPES & HEALTH DATA



Saliva
total_sample_volume: Decimal[1]

Blood
(from Clinical)
total_sample_volume: Decimal[1]

GEL_RD_Sample
GEL_hub: ODS_Code[1]
GEL_sample_identifier: String[1]
local_sample_identifier: String[1]
consent_copy_retained: String[1]
collection_site: ODS_Code[1]
sample_taken_date: Date[1]
sample_taken_time: Time[0..1]
date_sent: Date[1]
sent_by: String[1]
collection_protocol_name: String[1]
collection_protocol_version: String[1]

Linkage
GEL_participant_identifier: String[1]
NHS_number: NHS_Number[0..1]
CHI_number: String[0..1]
local_patient_identifier: String[1..*]
organisation_code: ODS_Code[1]
sex: Sex[1]
primary_diagnosis: ICD_diagnosis[1]
consent_date: Date[1]

Pedigree
pedigree_file: File[1]
creation_date: Date[1]
consanguinity: YesNoUnk[1]
created_by: String[0..1]

Family_History
person_relationship: Familial_Relationship[1]
relationship_other: String[0..1]
age_at_death: Integer[0..1]

Phenotyping_Statement
date: Date[1]
code: HPO[1]
presence: YesNoUnk[1]
certainty: string[0..1]
priority: string[0..1]
onset: onset[0..1]
severity: severity[0..1]
laterality: laterality[0..1]
spatial_pattern: spatial_pattern[0..1]
temporal_pattern: temporal_pattern[0..1]
variable_expressivity: variable_expressivity[0..1]
pace_of_progression: pace_of_progression[0..1]

Pathway_event
event_reference: String[1]
provider_code: ODS_Code[1]
event_date: Date[1]

RD_Participant
surname: String[1]
forenames: String[1]
person_address: String[1]
person_postcode: String[1]
person_birth_date: Date[1]
surname_at_birth: String[0..1]
ethnic_category: Ethnic_Category[1]
relevent_ancestry: String[0..1]
general_medical_practitioner: ODS_Code[1]
proband_status: Boolean[1]
relationship_to_proband: Familial_Relationship[0..1]
relationship_other: String[0..1]
GEL_family_identifier: String[1]
local_family_identifier: String[0..1]
consent_withdrawn_date: Date[0..1]

Participant_Medical_Condition
medical_condition: ICD_diagnosis[1..*]
age_of_onset: Integer[1]
information_source: String[*1]
creation_date: Date[1]
medical_condition_status: String[1]

Consent
consent_date: Date[1]
consent_version: String[1]
consent_taken_by: String[1]
scanned_copy: File[0..1]
participant_information_version: String[1]

Consent_Detail
question: String[1]
response: Boolean[1]

Referral
referred_to_practitioner_code: ODS_Code[1]
date_first_seen: Date[1]
emergency_admission_date: Date[0..1]

Investigation
report_date: date[1]

Diagnosis
primary_diagnosis: ICD_diagnosis[1]
mode_of_inheritance: HPO[1]
age_of_onset: Integer[0..1]

Clinical_Intervention
site_code: ODS_Code[1]

Death
death_location: Death_Location_Type[1]
death_cause_code_immediate: ICD_diagnosis[1]
death_cause_code_underlying: ICD_diagnosis[1]
death_cause_code_condition: ICD_diagnosis[1]
death_cause_code_significant: ICD_diagnosis[1]
post_mortem: YesNoUnk[1]

Clinical_Testing
test_type: SnomedCT_Code[1]
test_result: String[1]
test_unit_of_measure: String[1]

Sample_Investigation
sample_taken_date: Date[1]
sample_receipt_date: Date[1]
source_sample_id: String[1]

Imaging
imaging_code: SnomedCT_Code[1]
anatomical_site: SnomedCT_Code[0..1]
service_report_reference: String
image_reference: String

Medication
drug: SnomedCT_Code[1]
dose: String[1]
frequency: String[1]
end_date: Date[1]

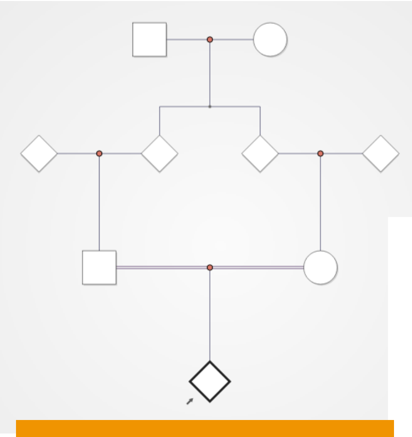
Surgery
primary_procedure: OPCS4_Code[1]
discharge_date: Date[1]
additional_procedure: OPCS4_Code[0..*]

Molecular_Result
gene: String[1]
scope_of_test: String[1]
method_of_test: String[1]
test_result: String[1]

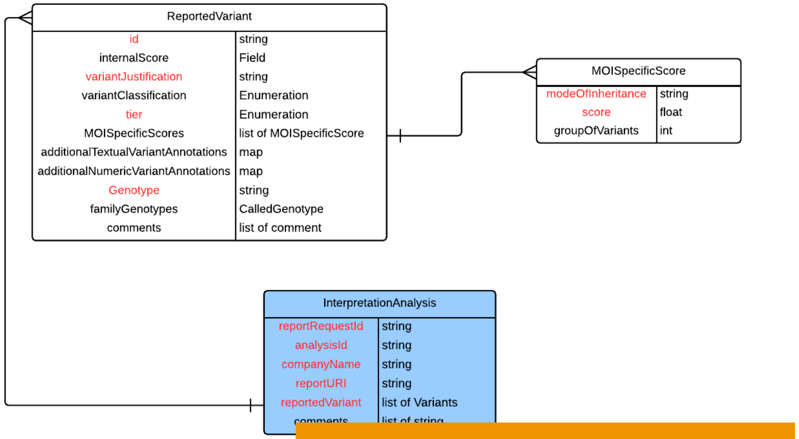
Molecular_Report
NGS_file_reference: String[1]
test_report: String[1]
comments: String[0..1]

Non_Pharmaceutical_Therapy
primary_treatment: SnomedCT_Code[1]
start_date: Date[1]
end_date: Date[0..1]
additional_treatments: SnomedCT_Code[0..*]

Clinical Data



Family Pedigree



ReportedVariant	
id	string
internalScore	Field
variantJustification	string
variantClassification	Enumeration
tier	Enumeration
MOISpecificScores	list of MOISpecificScore
additionalTextualVariantAnnotations	map
additionalNumericVariantAnnotations	map
Genotype	string
familyGenotypes	CalledGenotype
comments	list of comment

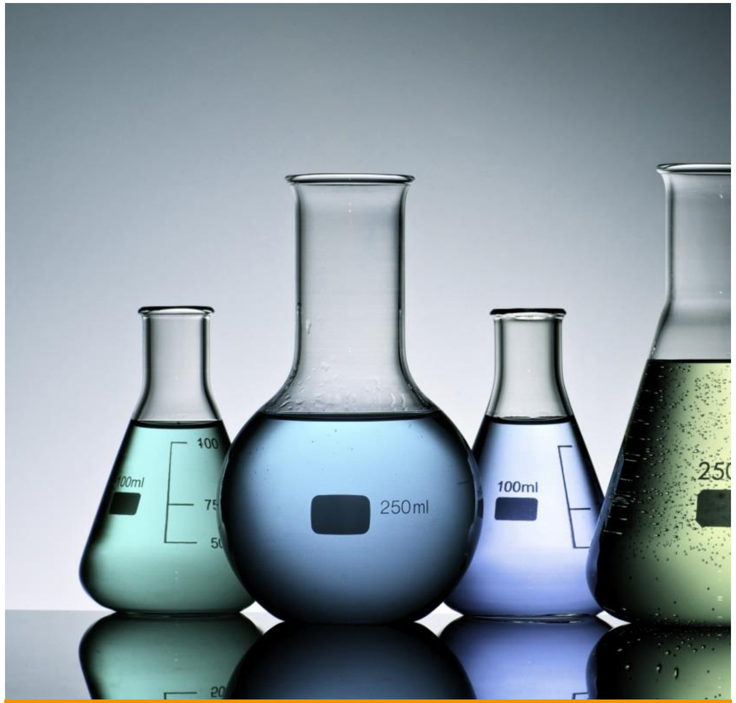
MOISpecificScore	
modeOfInheritance	string
score	float
groupOfVariants	int

InterpretationAnalysis	
reportRequestId	string
analysisId	string
companyName	string
reportURI	string
reportedVariant	list of Variants
comments	list of string

Interpreted Genome

Bio-Repository

- Tasked with sample collection and QC for sequencing
- Act as a barrier for poor quality DNA
- Standardised tubes, labels etc ready for receipt by illumina



Biorepository stores samples

Illumina

Sheffield Children's NHS
NHS Foundation Trust

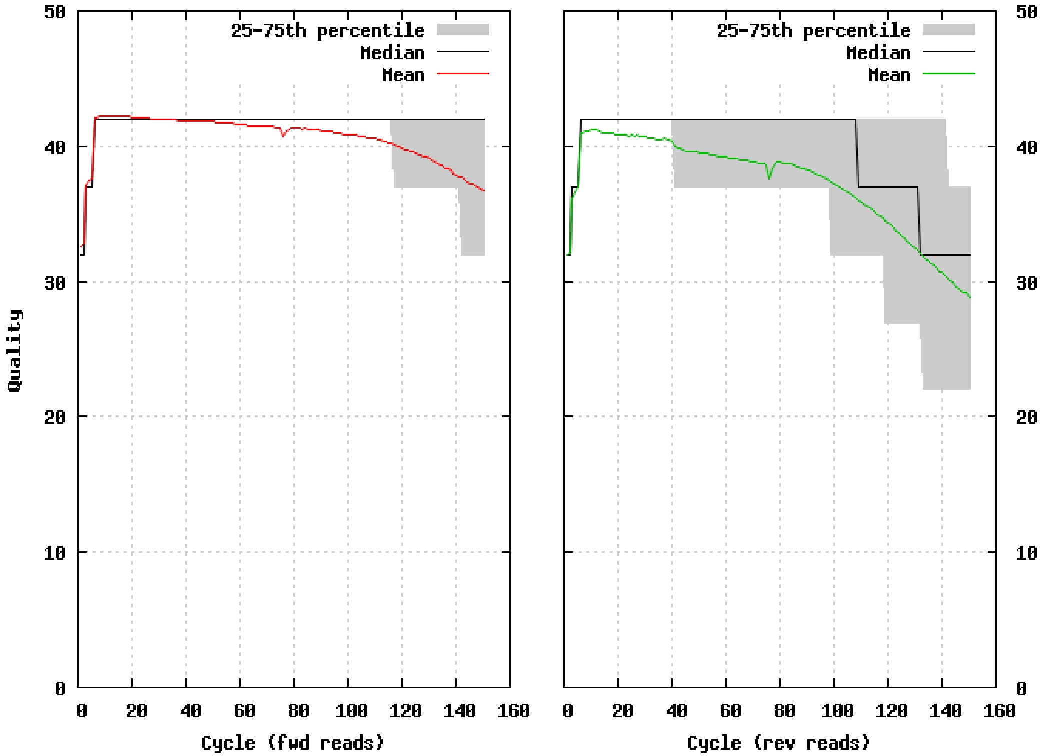
- Chosen as WGS sequencing partner for Genomics England and subsequently the NHS
- Provide Sequencing
- Provide Primary Analysis Pipeline: Codename NORTHSTAR
  - Mapping
  - Variant Calling
- Data delivered to GEL data centre through fast connections

Bridget Ogilvie Building

Illumina XTen Sequencing

Sheffield Children's NHS
NHS Foundation Trust

- 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**
- Tumor sequenced at '75X'



"We're Gonna Need a Bigger Truck"

- 200 Genomes a DAY!!
- Logistically this is a massive challenge

Genomics England

Sheffield Children's NHS
NHS Foundation Trust

Illumina Sends Data to GEL

BERTHA Manages Pipeline

- Illumina sends genomes and associated files like:
  - Variant calls
  - QC – coverage etc
- Checked for integrity and to ensure illumina are upholding the agreement of the contract – certain number of bases at a certain quality
- Once happy GEL "accepts" the delivery

Data Storage

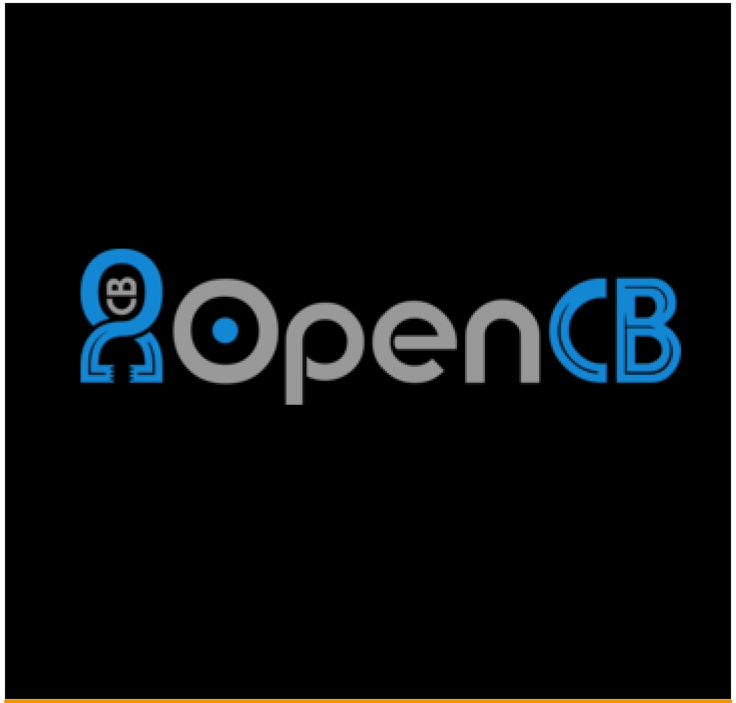
GEL Storage

Data Processed in 1 day = 20 Petabytes

- Rare Disease:
  - Each Genome: 100Gb
  - Trio is preferred so 300Gb per participant
- Cancer
  - Germline: 100Gb
  - Tumour: 200Gb
  - 300Gb per patient
- 10,000,000Gb = 10 Petabytes

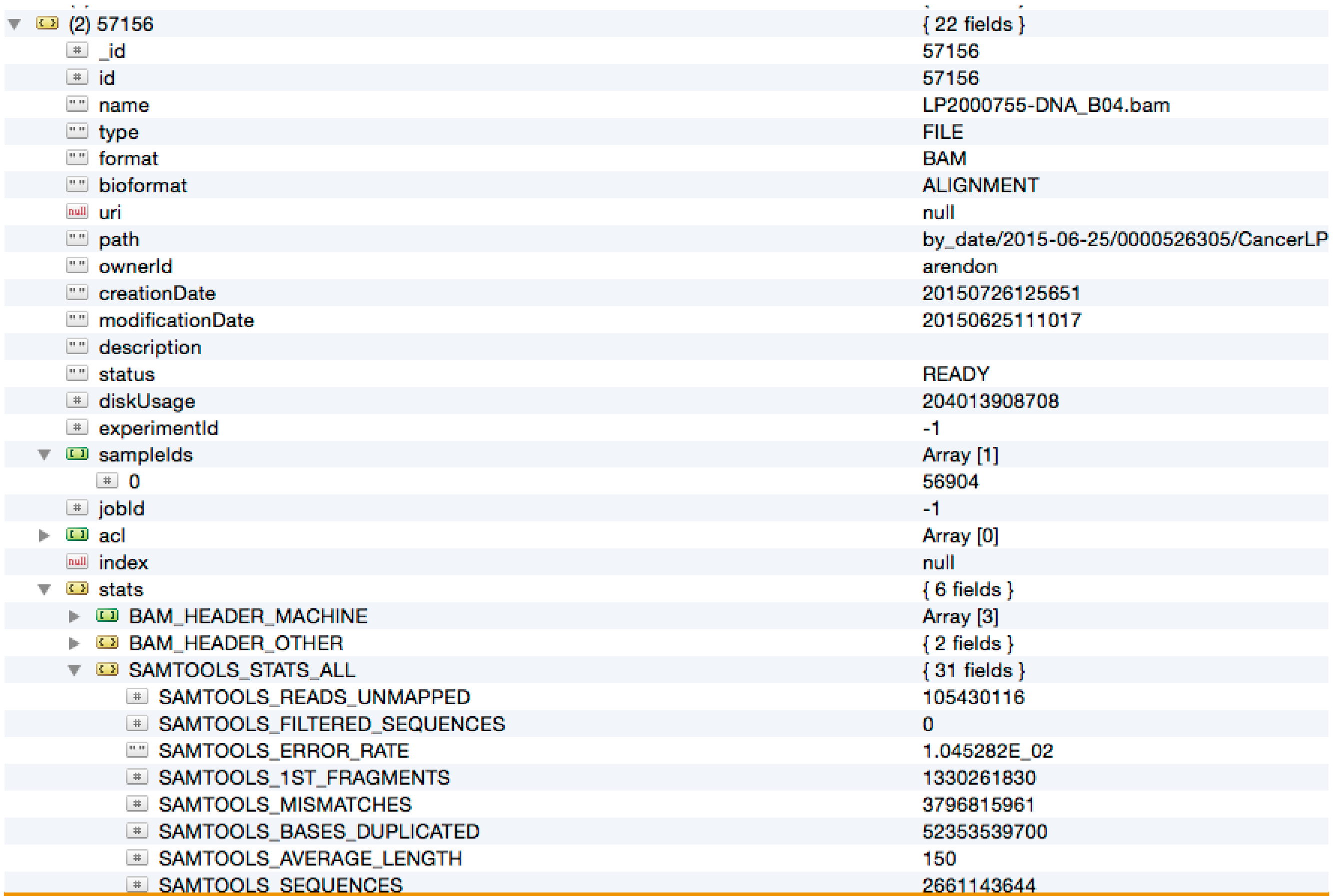
Databases

- Along with raw data files from illumina – variants are stored in a database called OpenCGA
- Allows annotation (adding information to variants like presence in public datasets)
- Allows fast internal frequency information
- Better for querying and producing reports



OpenCB – a suite of tools for genomic data

Catalog



```
▼ (2) 57156                                    { 22 fields }
      _id                                      57156
      id                                       57156
      name                                     LP2000755-DNA_B04.bam
      type                                     FILE
      format                                   BAM
      bioformat                                ALIGNMENT
      uri                                      null
      path                                     by_date/2015-06-25/0000526305/CancerLP
      ownerId                                  arendon
      creationDate                             20150726125651
      modificationDate                         20150625111017
      description
      status                                   READY
      diskUsage                                204013908708
      experimentId                             -1
    ▼ sampleIds                                Array [1]
        0                                      56904
      jobId                                    -1
    ▶ acl                                      Array [0]
      index                                    null
    ▼ stats                                    { 6 fields }
      ▶ BAM_HEADER_MACHINE                     Array [3]
      ▶ BAM_HEADER_OTHER                       { 2 fields }
      ▼ SAMTOOLS_STATS_ALL                     { 31 fields }
          SAMTOOLS_READS_UNMAPPED              105430116
          SAMTOOLS_FILTERED_SEQUENCES          0
          SAMTOOLS_ERROR_RATE                  1.045282E_02
          SAMTOOLS_1ST_FRAGMENTS               1330261830
          SAMTOOLS_MISMATCHES                  3796815961
          SAMTOOLS_BASES_DUPLICATED            52353539700
          SAMTOOLS_AVERAGE_LENGTH              150
          SAMTOOLS_SEQUENCES                   2661143644
```

Information on quality is stored in the database along with where to find the files

Initial Quality Control

- Key to a good clinical genome and high fidelity result: **Quality Control**
  - Delivery integrity
  - BAM & VCF validity
  - Coverage
  - Number of bases
  - Sex
  - Mendelian Errors
  - Inbreeding estimates
  - IBD estimation
  - Ancestry

Clinical Quality Genomes

Prioritisation = Tiering

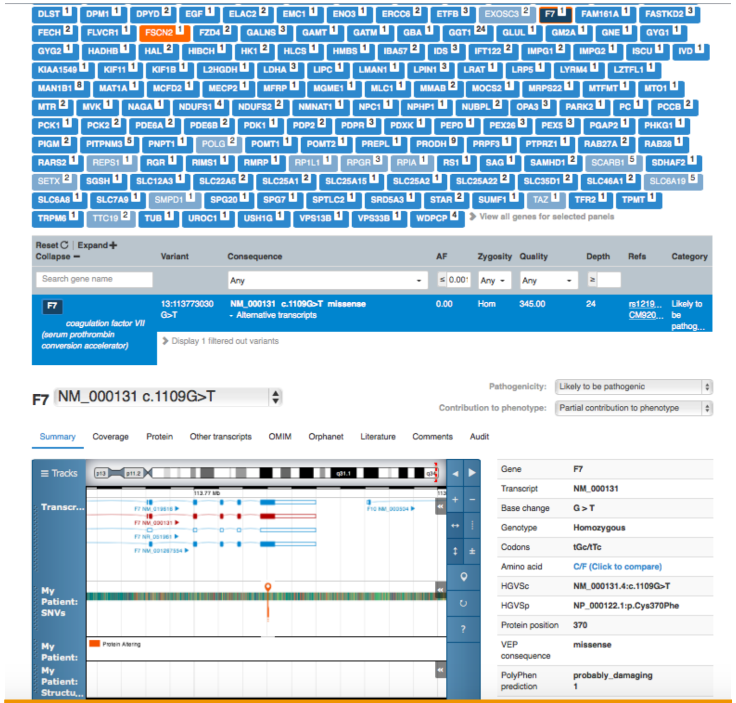
- NHS were worried about overloading clinical labs with too much validation work
- Prioritisation allows rules to be passed to GMCs on what they must validate, report etc
- Tiering based on mode of inheritance, phenotype, frequency in populations, frequency in disease groups etc

Tiering Ranks Variants

External Interpretation Partners

Sheffield Children's NHS
NHS Foundation Trust

- A number of external providers chosen after "bake-off"
- Provide interfaces so that GMCs can easily access and interpret the variants from the WGS data



DLST 1 DPM1 1 DPYD 2 EGF 1 ELAC2 2 EMC1 1 ENO3 1 ERCC6 2 ETFB 3 EXOSC3 2 F7 1 FAM161A 1 FASTKD2 3
FECH 2 FLVCR1 1 FSCN2 1 FZD4 2 GALNS 3 GAMT 1 GATM 1 GBA 1 GGT1 24 GLUL 1 GM2A 1 GNE 1 GYG1 1
GYG2 1 HADHB 1 HAL 2 HIBCH 1 HK1 2 HLCS 1 HMBS 1 IBA57 2 IDS 3 IFT122 2 IMPG1 2 IMPG2 1 ISCU 1 IVD 1
KIAA1549 1 KIF11 1 KIF1B 1 L2HGDH 1 LDHA 3 LIPC 1 LMAN1 1 LPIN1 3 LRAT 1 LRP5 1 LYRM4 1 LZTFL1 1
MAN1B1 8 MAT1A 1 MCFD2 1 MECP2 1 MFRP 1 MGME1 1 MLC1 1 MMAB 2 MOCS2 1 MRPS22 1 MTFMT 1 MTO1 1
MTR 2 MVK 1 NAGA 1 NDUFS1 4 NDUFS2 2 NMNAT1 1 NPC1 1 NPHP1 1 NUBPL 2 OPA3 3 PARK2 1 PC 1 PCCB 2
PCK1 1 PCK2 2 PDE6A 2 PDE6B 2 PDK1 1 PDP2 2 PDPR 3 PDXK 1 PEPD 1 PEX26 3 PEX5 3 PGAP2 1 PHKG1 1
PIGM 2 PITPNM3 5 PNPT1 1 POLG 2 POMT1 1 POMT2 1 PREPL 1 PRODH 9 PRPF3 1 PTPRZ1 1 RAB27A 2 RAB28 1
RARS2 1 REPS1 1 RGR 1 RIMS1 1 RMRP 1 RP1L1 1 RPGR 3 RPIA 1 RS1 1 SAG 1 SAMHD1 2 SCARB1 5 SDHAF2 1
SETX 2 SGSH 1 SLC12A3 1 SLC22A5 2 SLC25A1 2 SLC25A15 1 SLC25A2 1 SLC25A22 2 SLC35D1 2 SLC46A1 2 SLC6A19 5
SLC6A8 1 SLC7A9 1 SMPD1 1 SPG20 1 SPG7 1 SPTLC2 1 SRD5A3 1 STAR 2 SUMF1 1 TAZ 1 TFR2 1 TPMT 1
TRPM6 1 TTC19 2 TUB 1 UROC1 1 USH1G 1 VPS13B 1 VPS33B 1 WDPCP 4
View all genes for selected panels
Reset
Expand
Collapse
Tracks
p13
p11.2
q31.1
q34
113.77 Mb
Transcr...
F7 NM_019616
F7 NM_000131
F7 NR_051961
F7 NM_001267554
F10 NM_000504
My Patient: SNVs
My Patient:
Protein Altering
My Patient: Structu...

	Variant	Consequence	AF	Zygosity	Quality	Depth	Refs	Category
Search gene name		Any	≤ 0.001	Any	Any	≥		
F7 coagulation factor VII (serum prothrombin conversion accelerator)	13:113773030 G>T	NM_000131 c.1109G>T missense - Alternative transcripts	0.00	Hom	345.00	24	rs1219... CM920...	Likely to be pathog...

Display 1 filtered out variants

F7 NM_000131 c.1109G>T

Pathogenicity: Likely to be pathogenic

Contribution to phenotype: Partial contribution to phenotype

Summary Coverage Protein Other transcripts OMIM Orphanet Literature Comments Audit

Gene	F7
Transcript	NM_000131
Base change	G > T
Genotype	Homozygous
Codons	tGc/tTc
Amino acid	C/F (Click to compare)
HGVSc	NM_000131.4:c.1109G>T
HGVSp	NP_000122.1:p.Cys370Phe
Protein position	370
VEP consequence	missense
PolyPhen prediction	probably_damaging 1

Congenica Interpretation Interface

Clinical Report

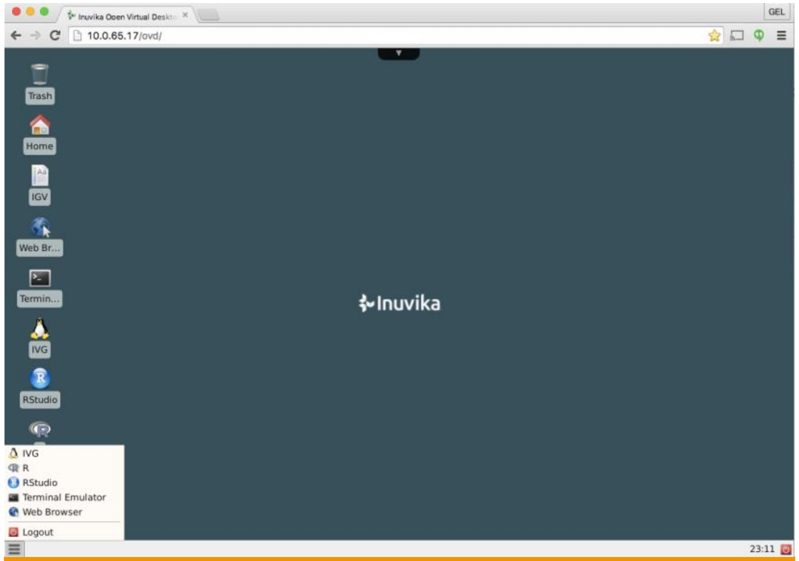
- Genomics England & Illumina not ISO accredited to deliver clinical results
- Findings must be validated by an orthogonal method
- Reports must be written by clinical scientist

GEL/Illumina Not ISO accredited

Data Access Policy

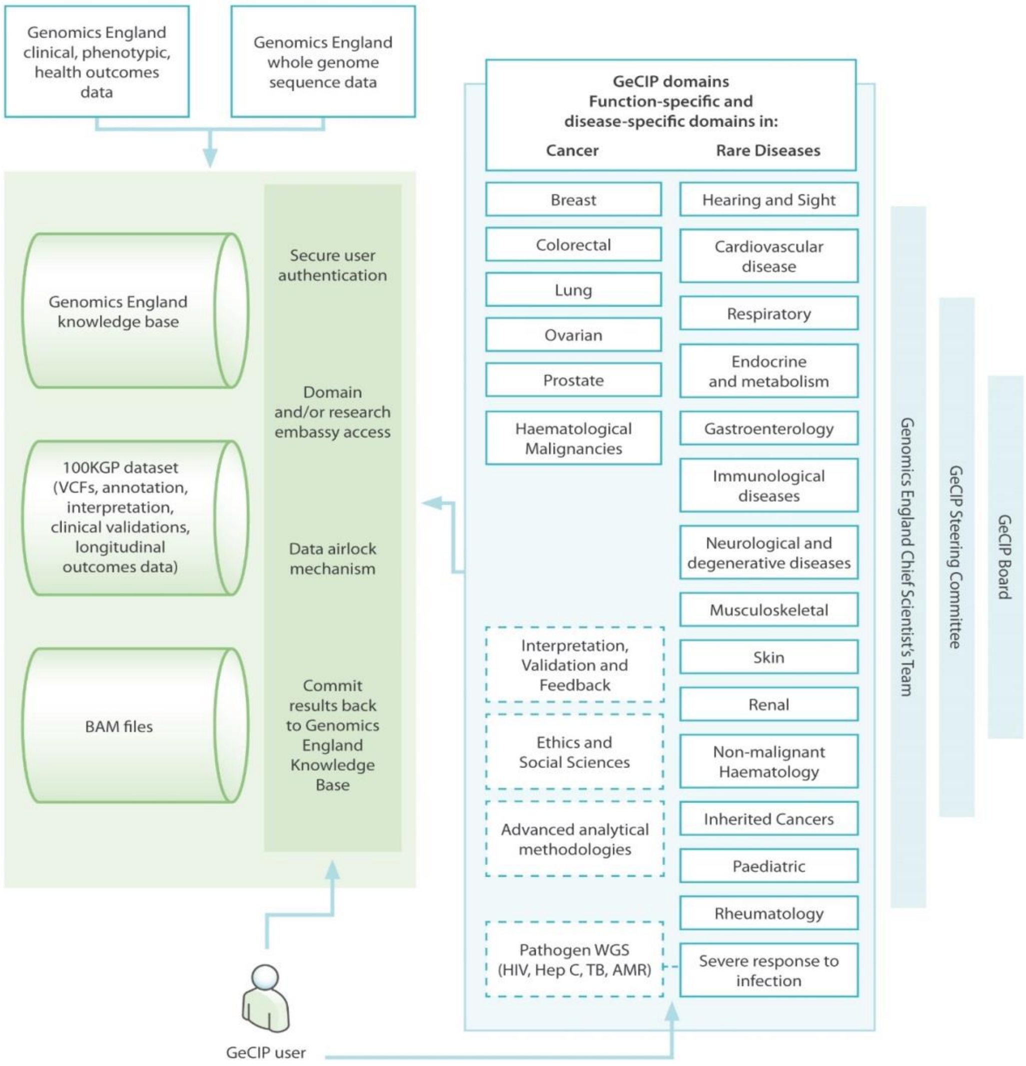
Sheffield Children's NHS
NHS Foundation Trust

- How can we access the data?
- Designed to protect patients
- Names & other identifiers not stored with WGS data
- All activity monitored & recorded
- **Diagnostics**
  - GMCs returning results to patients have full access - identifieable
  - Data kept in "reading library"
  - Electronic data access agreement
- **Researchers**
  - Non-identifiable data within "reading library"
  - Access Review Committee
  - GeCIPs
  - Electronic data access agreement
- **GENE – Pharma Collaborators**
  - Non-identifiable data within "reading library"
  - Electronic data access agreement
  - Selected Pharma partners
  - To promote cures for rare diseases



Virtual Desktop Environment

GeCIPs



Genomics England clinical, phenotypic, health outcomes data

Genomics England whole genome sequence data

Genomics England knowledge base

100KGP dataset (VCFs, annotation, interpretation, clinical validations, longitudinal outcomes data)

BAM files

Secure user authentication

Domain and/or research embassy access

Data airlock mechanism

Commit results back to Genomics England Knowledge Base

GeCIP user

GeCIP domains
Function-specific and disease-specific domains in:

Cancer

- Breast
- Colorectal
- Lung
- Ovarian
- Prostate
- Haematological Malignancies

Rare Diseases

- Hearing and Sight
- Cardiovascular disease
- Respiratory
- Endocrine and metabolism
- Gastroenterology
- Immunological diseases
- Neurological and degenerative diseases
- Musculoskeletal
- Skin
- Renal
- Non-malignant Haematology
- Inherited Cancers
- Paediatric
- Rheumatology
- Severe response to infection

Interpretation, Validation and Feedback

Ethics and Social Sciences

Advanced analytical methodologies

Pathogen WGS (HIV, Hep C, TB, AMR)

Genomics England Chief Scientist's Team

GeCIP Steering Committee

GeCIP Board

Figure 1: The structure of GeCIP

Group Participation

- Ethical/Governance Issues with WGS?
- Logistical Issues?

Take Home Messages

- Data is generated as a result of a long co-ordinated pipeline combining disparate organisations
- WGS + Health & Phenotype Data a valuable resource
- Data is stored at GEL and is only accessible through a virtual desktop
- But…. Important to remember WGS is not a magic bullet