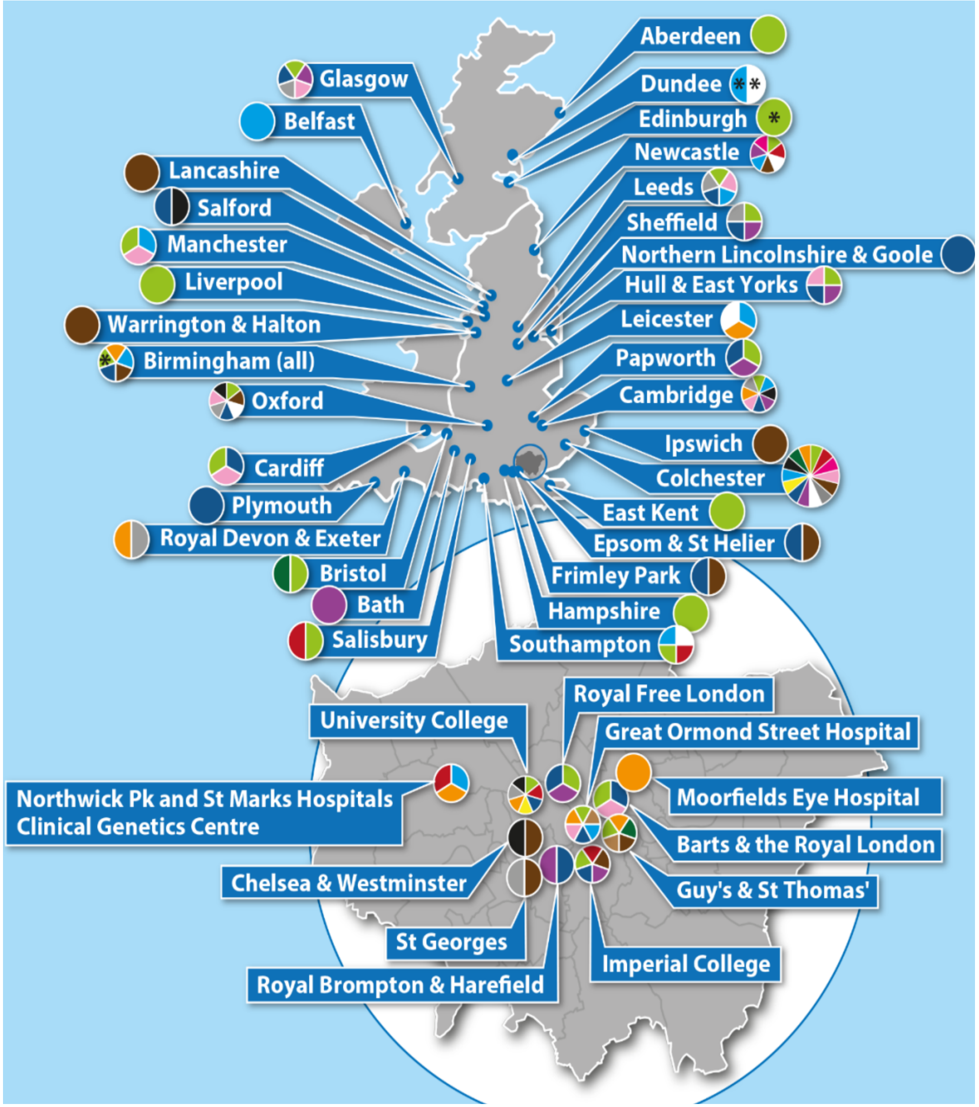
10K genomes from rare disease cohorts

Chris Penkett
University of Cambridge

© AP

The Rare Diseases Pilot

started in 2013



- **50** NHS Hospitals, **300** Clinical Care Teams
- PCR-free Whole Genome Sequencing in a **Clinically Accredited** Laboratory
- Clinical **Feedback** + Research

Main Rare Disease Cohorts

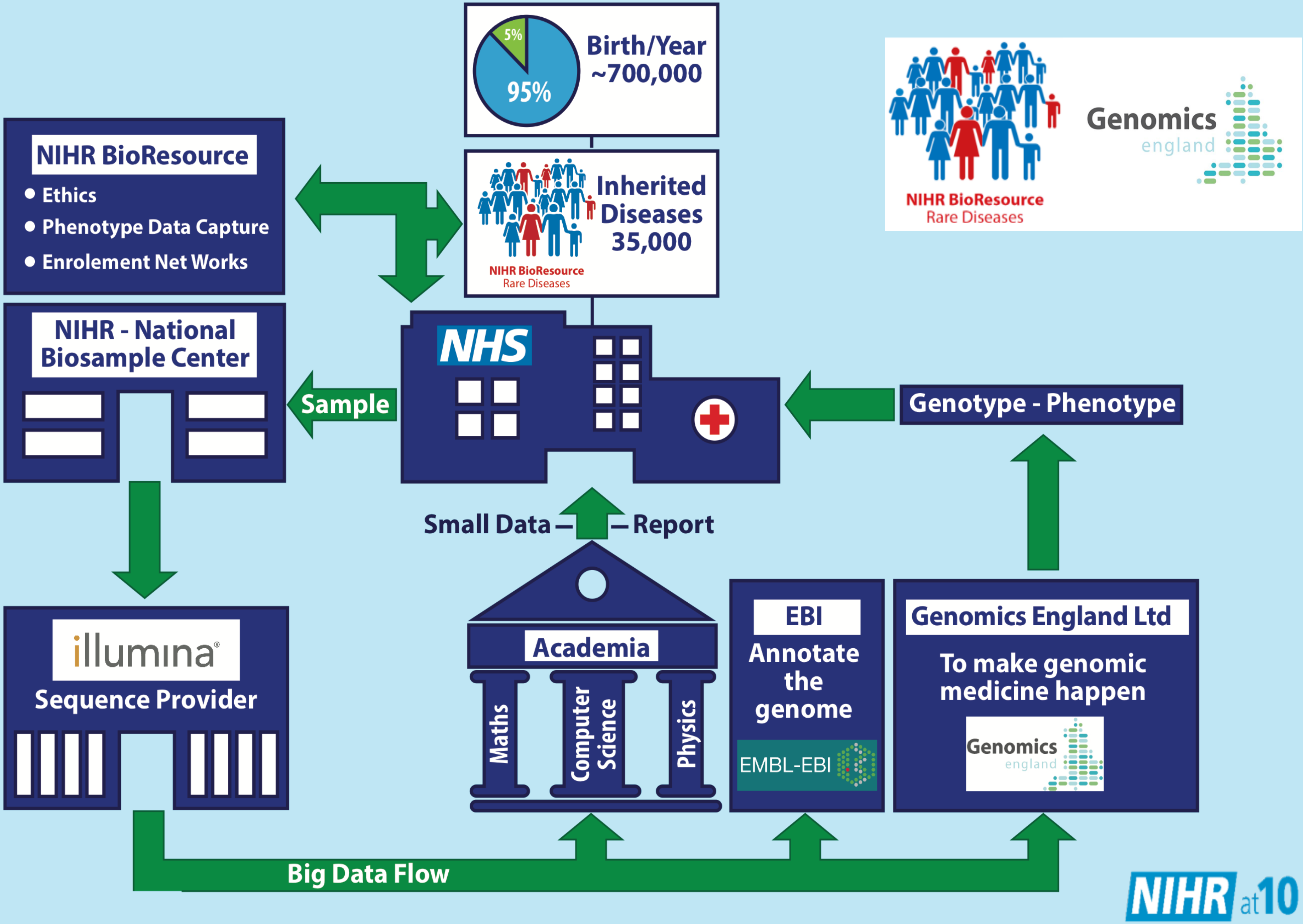
NIHR BioResource
Rare Diseases

NHS
National Institute for
Health Research

- 15 projects in total covering a variety of **Rare Diseases**
- **>10,000** Patients & Relatives
- Each project has a PI and (usually) a Chief Analyst to identify causal variants

	Projects	**Targeted organ/state**
BPD	Bleeding and platelets	Blood, coagulation etc.
PAH	Pulmonary Arterial Hypertension	Lung blood vessel
PID	Primary Immune Disorders	Immunity
SPEED	Retinal dystrophy + neurological disorders	Retinal, neurological

The NHS 100,000 Genome Project



Birth/Year ~700,000

Inherited Diseases 35,000

NIHR BioResource
Rare Diseases

Genomics england

NIHR BioResource

- Ethics
- Phenotype Data Capture
- Enrolement Net Works

NIHR - National Biosample Center

Sample

NHS

Genotype - Phenotype

illumina®

Sequence Provider

Small Data — — Report

Academia

Maths

Computer Science

Physics

EBI

Annotate the genome

EMBL-EBI

Genomics England Ltd

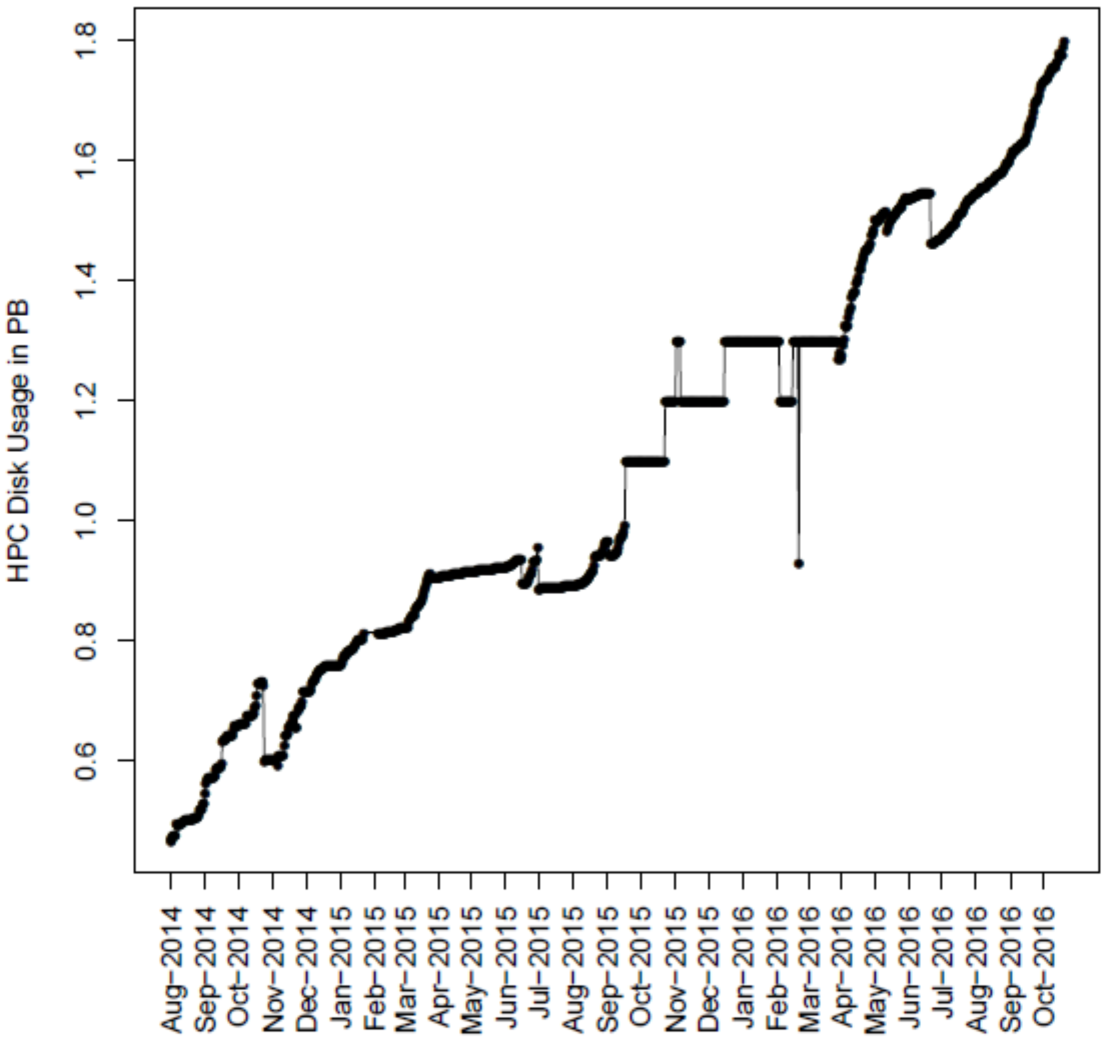
To make genomic medicine happen

Genomics england

Big Data Flow

NIHR at10

NIHR BioResource – The Data



Big Data

Now over 2 PB in > 3 years

~70GB raw data per genome

~0.5GB “just the variants” per genome

WGS Data QC

Data is checked/annotated for:

- Errors introduced during data transfer (md5 checksum)
- Create smaller cram files to archive at EGA (EBI) without raw data loss
- WGS quality metrics
- Genetic gender vs manifest gender
- Relatedness

  Check family structure as reported

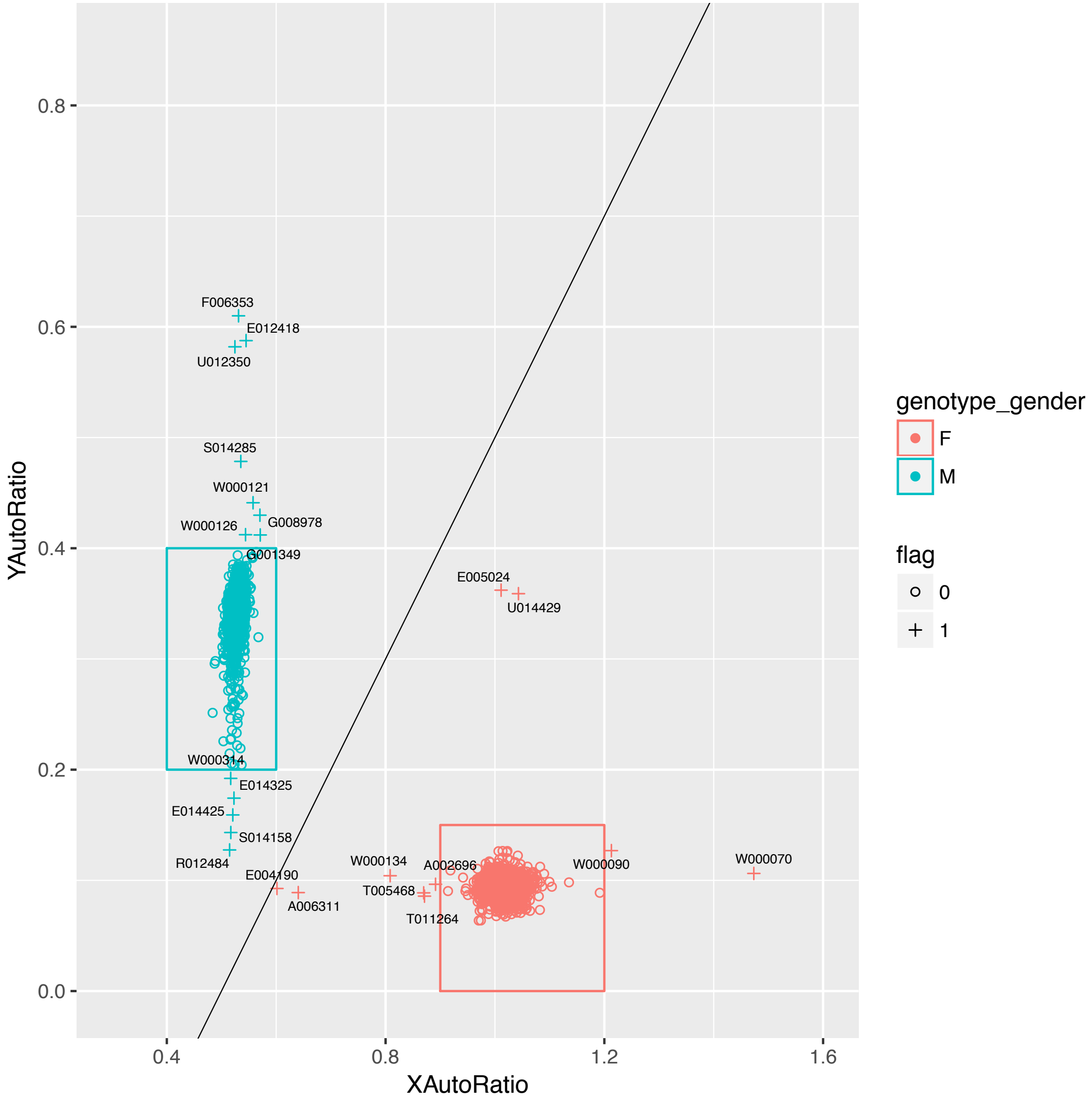
  Identity unknown family recruitments

  Detect duplicates and/or identical twins

  Identify subset of unrelated individuals for unbiased allele frequency calculation
- Ethnicity

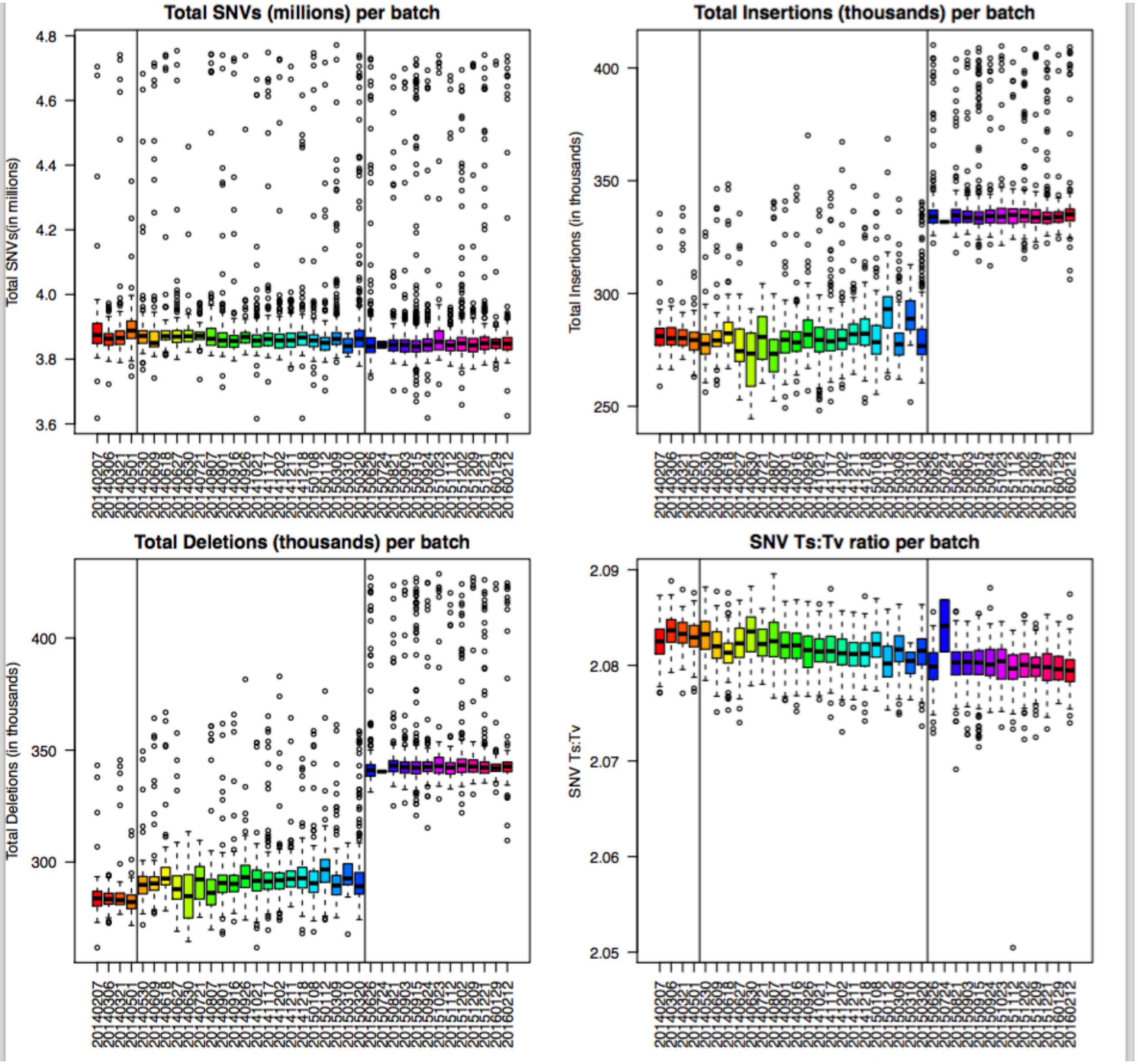
NIHR BioResource
Rare Diseases

Genomic Gender from X/Y vs Autosome Coverage

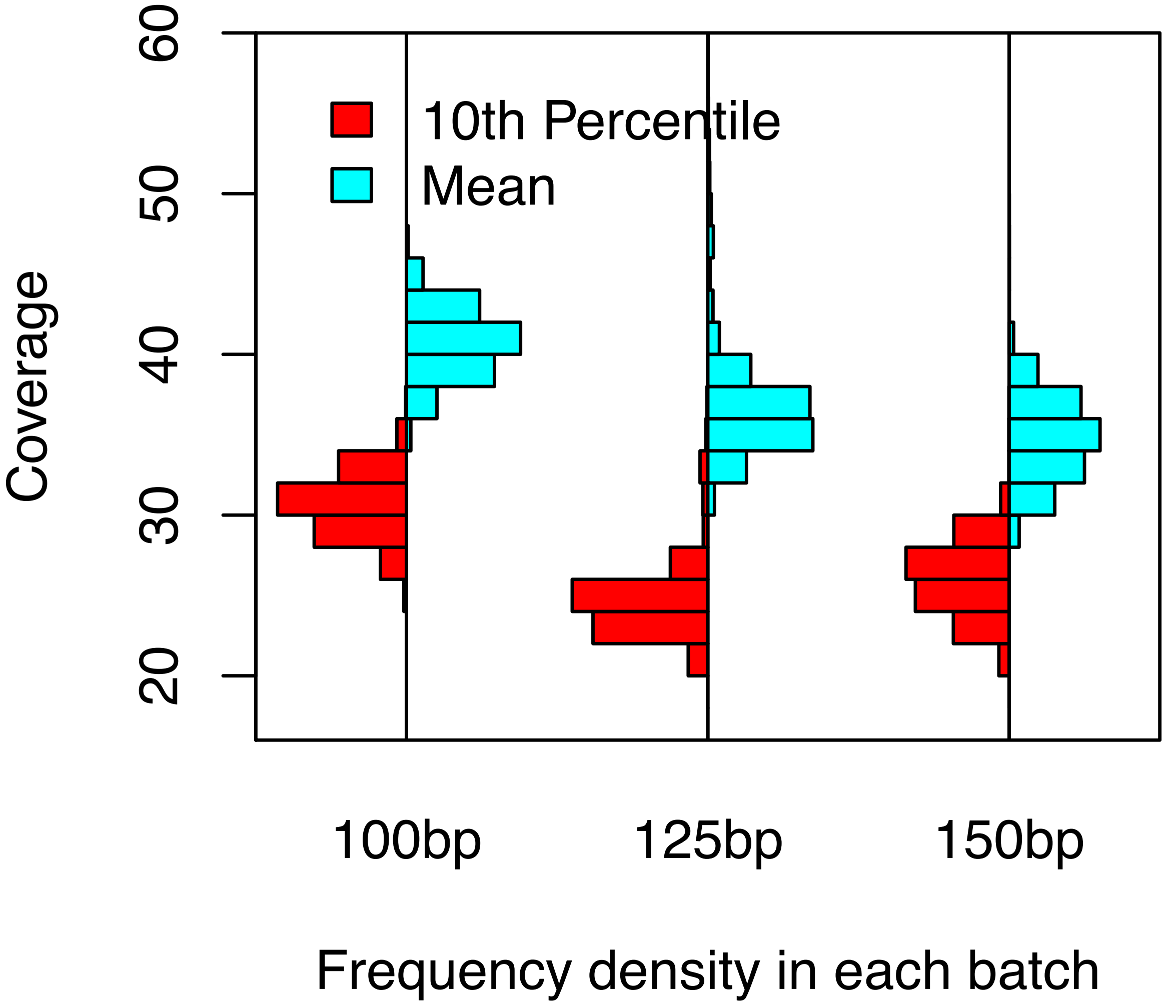


Summary of Data Parameters for 3 Read Lengths

NIHR BioResource
Rare Diseases



Overall Genome Coverage



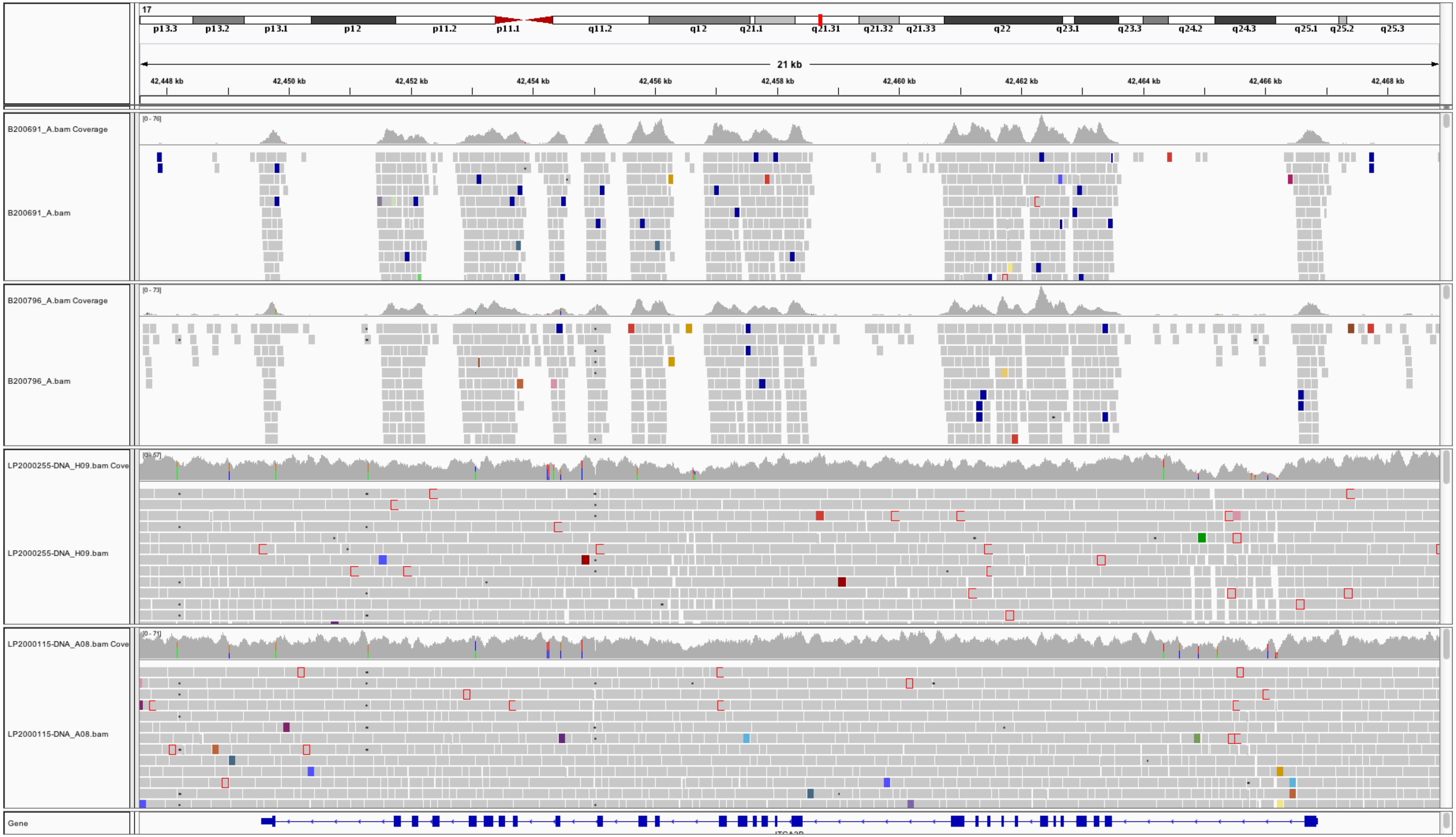
Coverage is the average number of reads representing a given nucleotide in the reconstructed sequence

For detection of SNPs and rearrangements, publications recommend from 10× to 30× depth of coverage

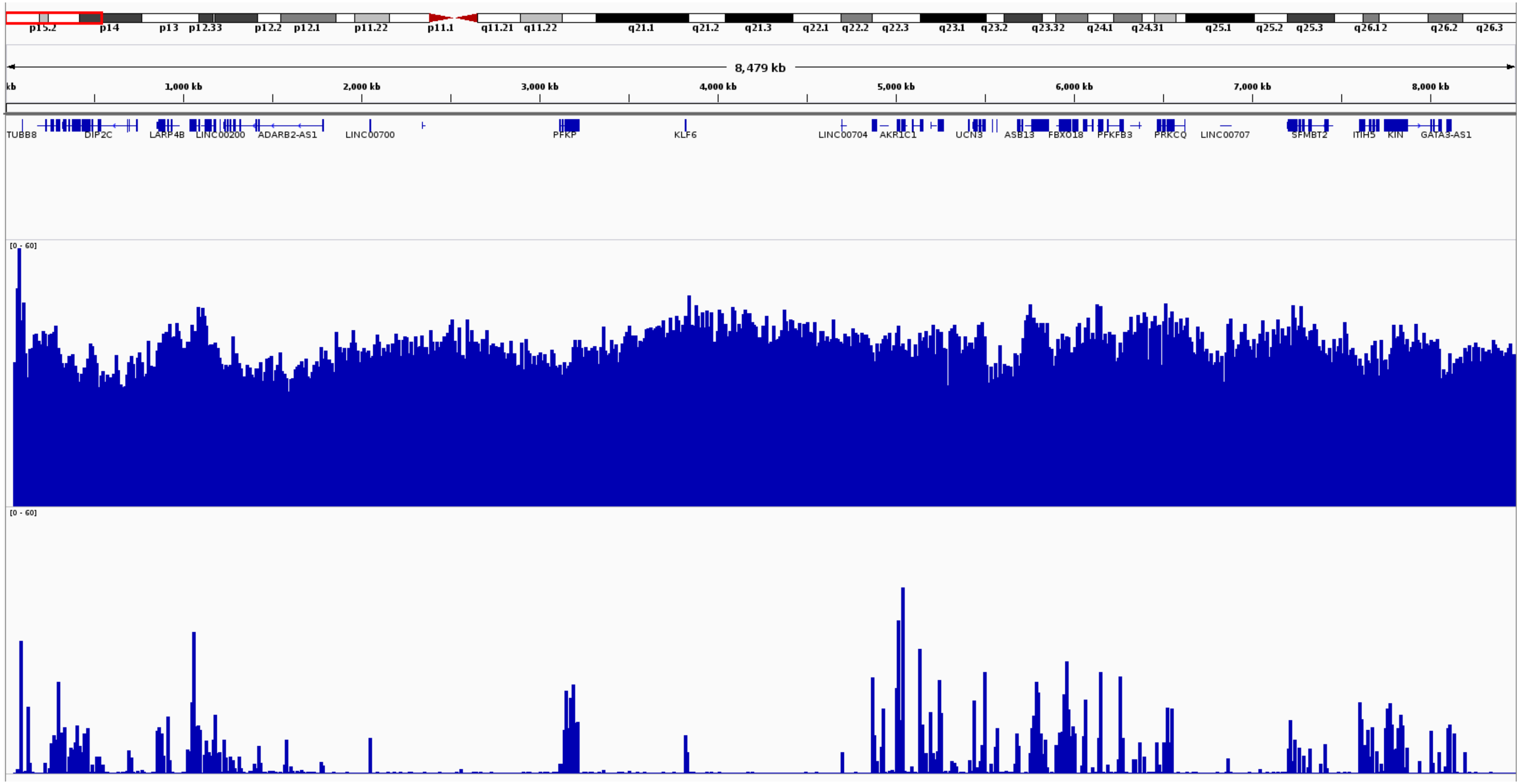
More than 90% of the genome has 25x coverage

NIHR BioResource
Rare Diseases

Genomes versus Exomes

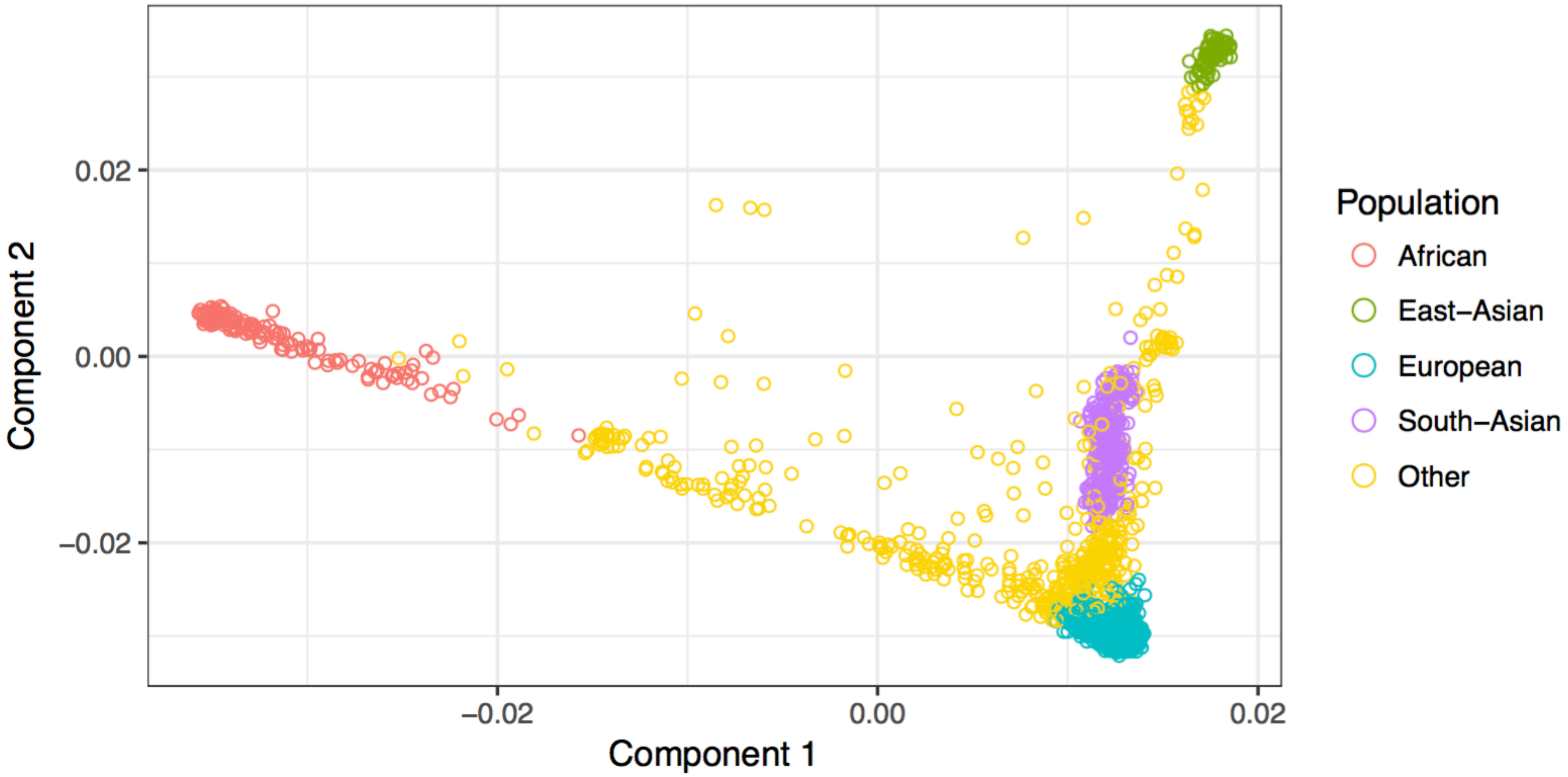


Example WGS10K Coverage vs WES



Genomic Ethnicity Determination from Principal Components Analysis

NIHR BioResource
Rare Diseases

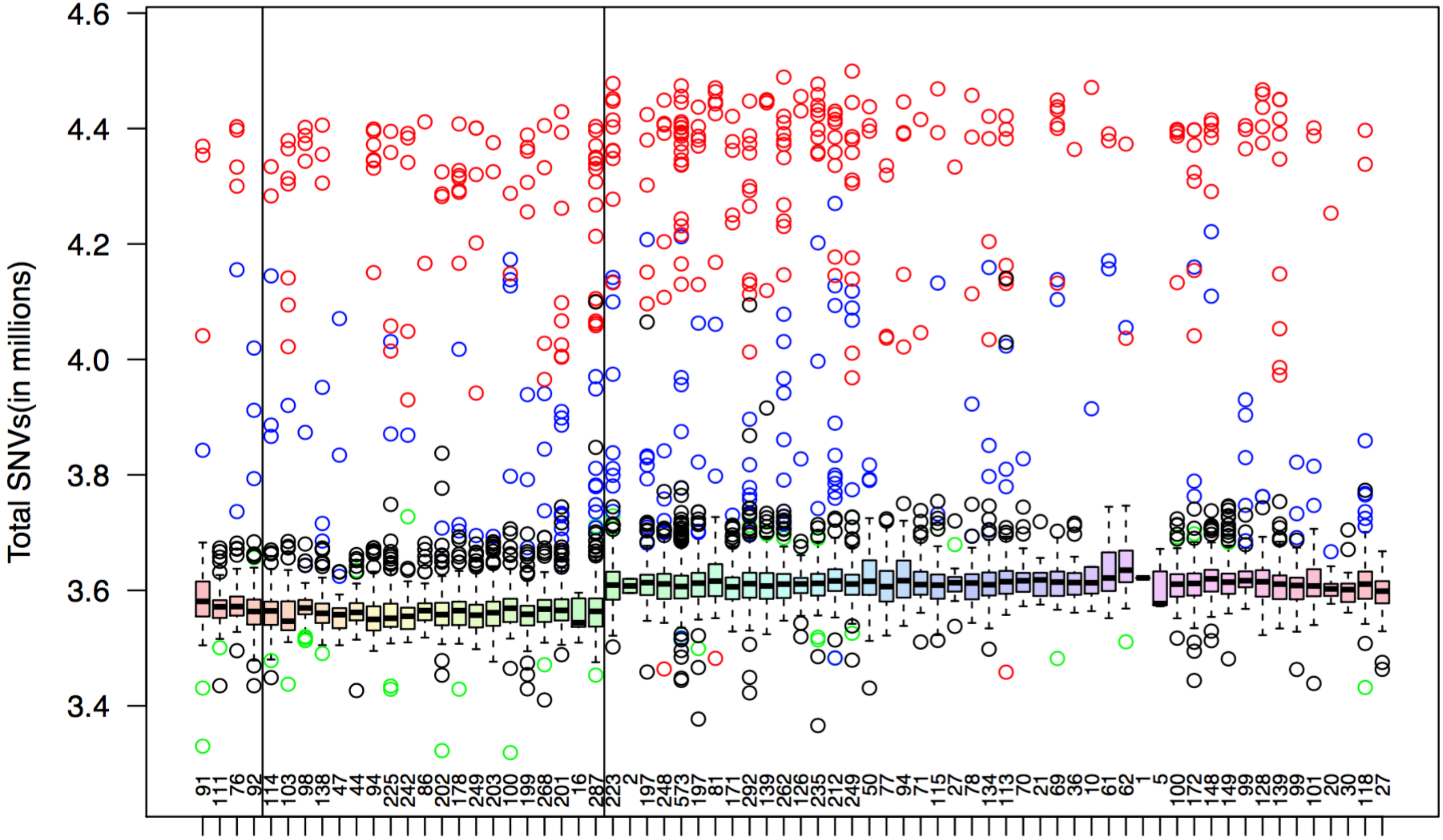


Batch Data, SNVs per Person and Ethnicity

NIHR BioResource
Rare Diseases

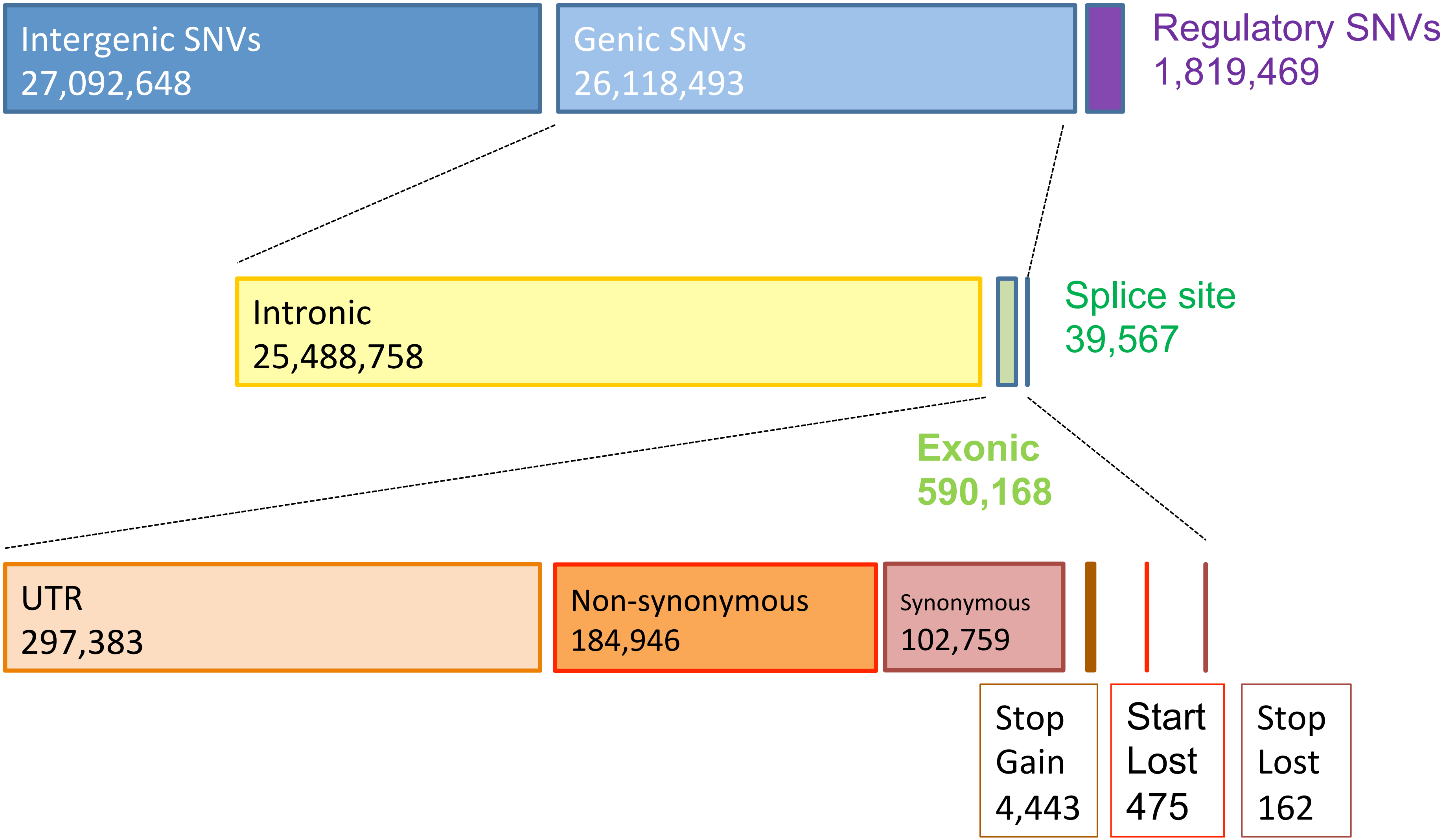
NHS
National Institute for
Health Research

Average 3.9 Million SNVs per person
Average 3.6 Million post QC filtering



55M Unique Variants in 7,204 Genomes

most are in the non-coding space regulating transcription and translation



Intergenic SNVs
27,092,648

Genic SNVs
26,118,493

Regulatory SNVs
1,819,469

Intronic
25,488,758

Splice site
39,567

Exonic
590,168

UTR
297,383

Non-synonymous
184,946

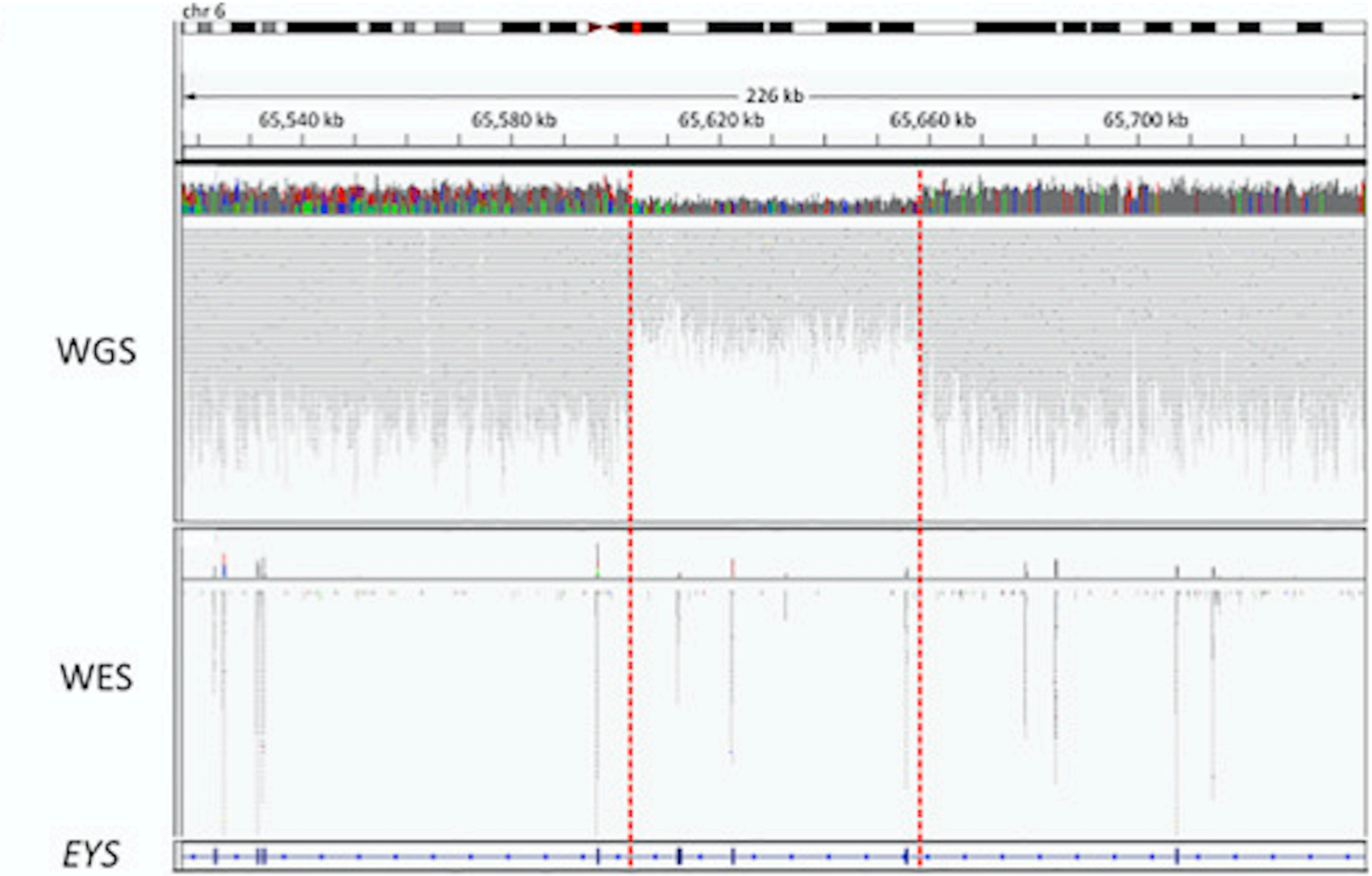
Synonymous
102,759

Stop Gain
4,443

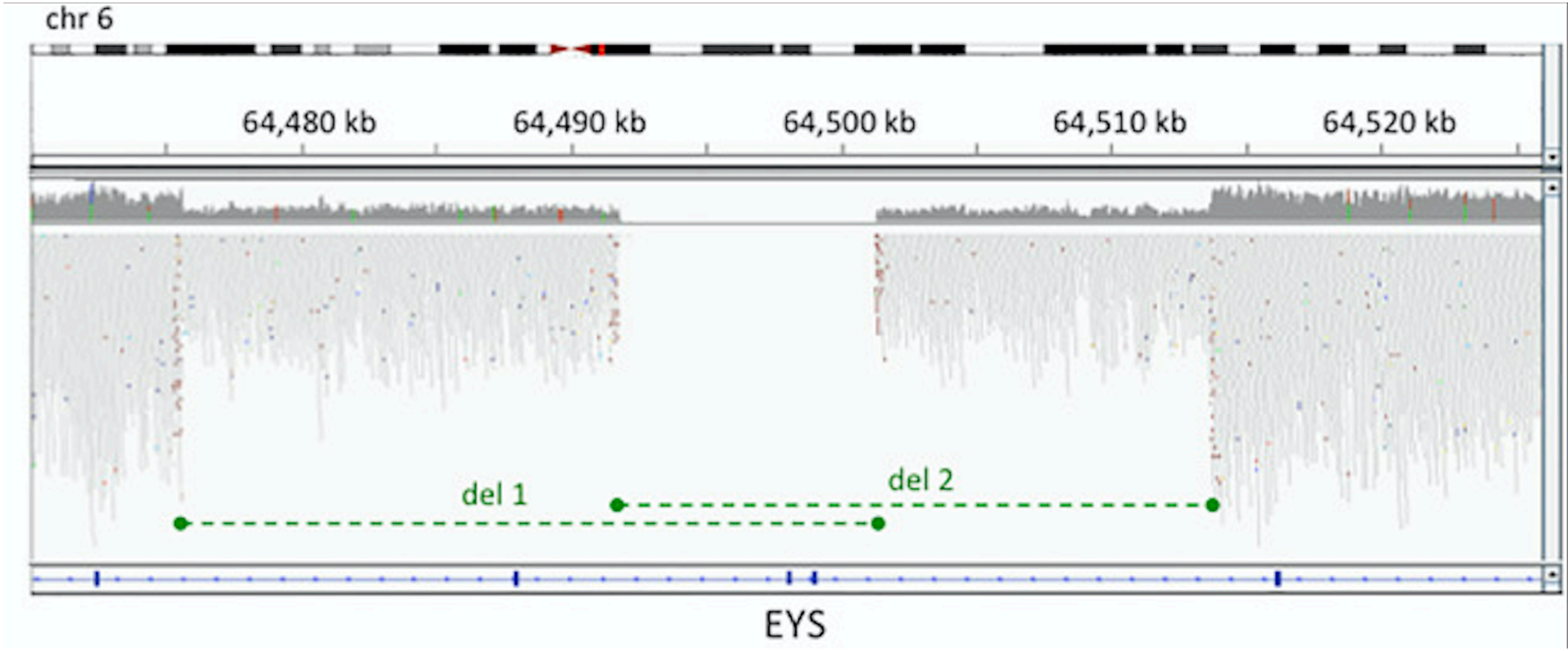
Start Lost
475

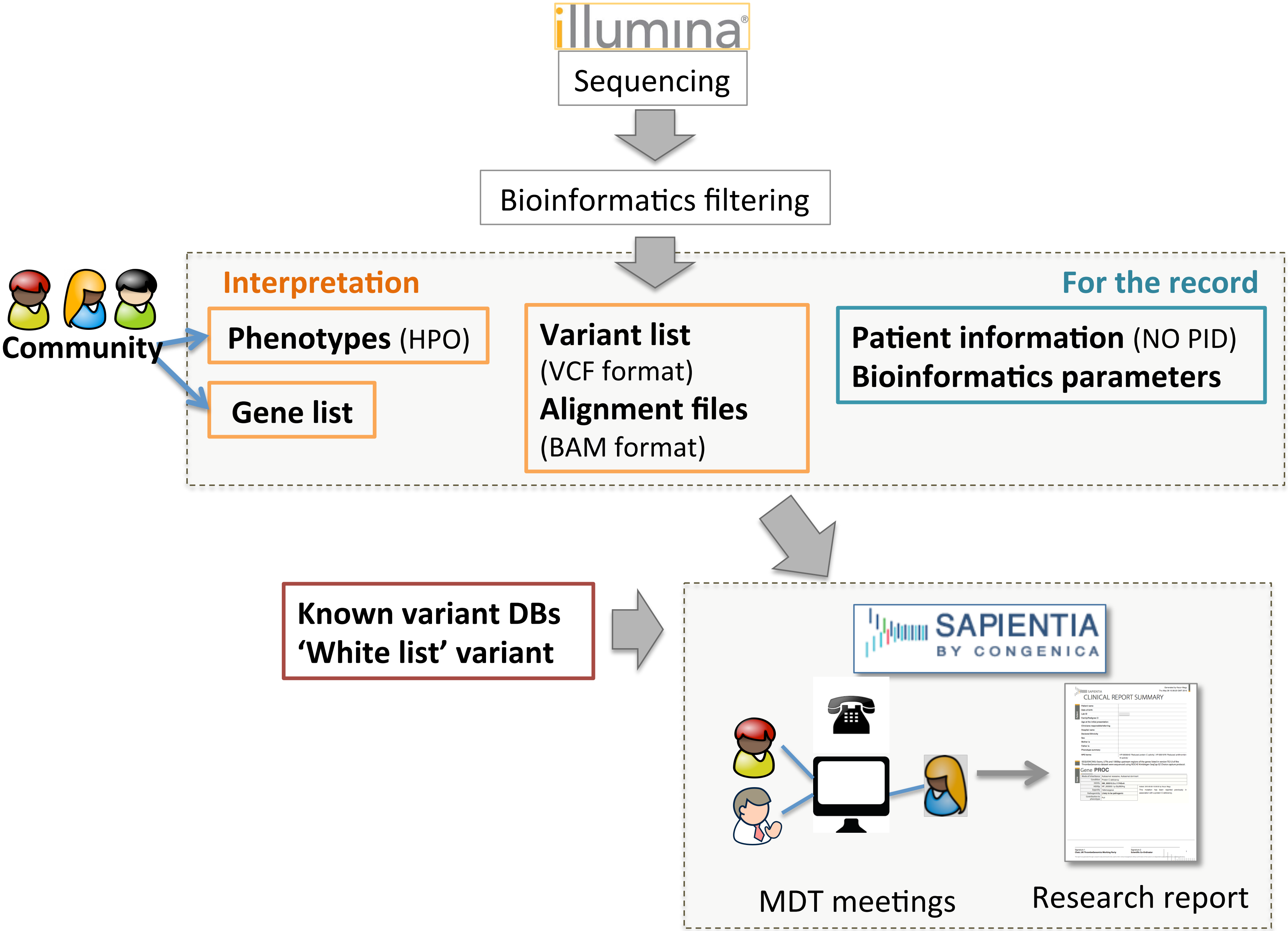
Stop Lost
162

Example of CNV Detection for Genome and Exome Data



Example of Two Large Overlapping Deletions in WGS Data





Sequencing

Bioinformatics filtering

Interpretation

For the record

Community

Phenotypes (HPO)

Gene list

Variant list
(VCF format)
Alignment files
(BAM format)

Patient information (NO PID)
Bioinformatics parameters

Known variant DBs
'White list' variant

MDT meetings

Research report

MDTs

- Multi-Disciplinary Team meetings
  - 2 clinician/clinical geneticist, 1 chairing
  - Project coordinator
  - Pertinent Finding team member
  - Analyst

☎ 🖥

Reporting Across the Cohorts

Projects	Pre-screening	Cases reviewed at MDT	Diagnostic yield %
BPD – Bleeding and Platelet Disorders	Extensive	1162	10 %
PAH – Pulmonary Arterial Hypertension	Moderate	781	18 %
PID – Primary Immunodeficiency	Extensive	720	9 %
SPEED RD – Retinal Dystrophy	Moderate	284	63 %
SPEED neuro – Paediatric Neurodevelopmental	Limited	243	32 %

Novel variants: ~50%

The Teams

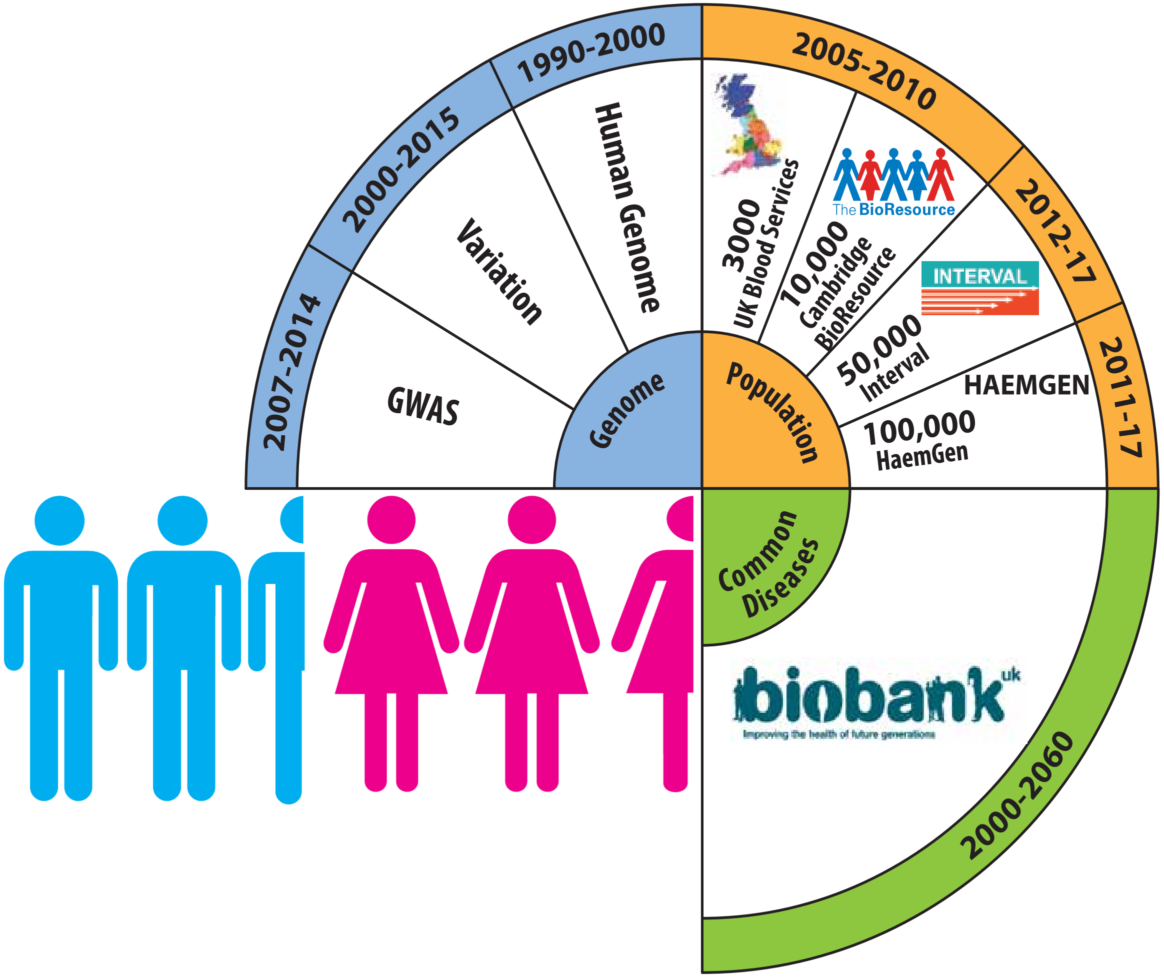
Sri Deevi
Salih Tuna
Olga Shamardina
Fengyuan Hu
Kathy Stirrups
Stuart Meacham
Tony Attwood
Stefan Gräf
Matthias Haimel
Marta Bleda
Ernest Turro
Daniel Greene
Keren Carss
Alba Sanchis-Juan
Hana Lango Allen
Karyn Megy
Louise Daugherty
Tim Young
Roger James
Catherine Titterton
Lucy Raymond
Willem Ouwehand

- CATGO
- NIHR-RD Enrolment team
  - Sofie Ashford
  - Sofia Papadia
- HPC
  - Stuart Rankin
  - Wojciech Turek
  - Paul Calleja
  - Nacho (Ignacio) Medina (GEL/EBI)
  - Jacobo Coll (GEL)
- BLUEPRINT team
- Admin team
- Illumina sequencing/bioinformatics teams
  - Russell Grocock
  - John Peden
  - Christian Bourne
  - Sean Humphray
  - Terry Gerighty
- GEL bioinformatics team
  - Augusto Rendon
  - Katherine Smith
- EGA team at EBI
  - Jeff Almeida-King

NIHR BioResource - Rare Diseases

Rare Disease/Condition	Acronym/ Approved in month/year	Lead Investigator	Capacity allocated	Samples sent for sequencing	WGS10K Samples received at the HPC
Bleeding and Platelet Disorders	BPD 12/12	Prof Willem Ouwehand	1250	839 (127 WES)	759
Cerebral small vessel	CSVD 04/14	Prof Hugh Markus	250	125	110
Ehler-Danlos Syndrome	EDS 07/13	Prof Tim Aitman	400	90* WES	0
Genomics England pilot	GEL Pilot 11/13	Prof Mark Caulfield	2000 (+ 3000)	4092 (10 dups)	2000
Hypertrophic Cardiomyopathy	HCM 05/12	Prof Hugh Watkins	300	146	134
Intrahepatic Cholestasis of Pregnancy	ICP 01/14	Prof Catherine Williams	270	90	76
Multiple Primary Malignant Tumours	MPMT 11/13	Prof Eamonn Maher	700	297	263
Neuropathic Pain Disorders	NPD 10/14	Prof Geoff Woods	250	41	39
Primary Immune Disorders	PID 12/12	Prof Ken Smith	1250	1119 (26 WES)	1080
Primary Membranoproliferative Glomerulonephritis	PMG 05/13	Dr Danny Gale	213	97	97
Pulmonary Arterial Hypertension	PAH 12/12	Prof Nick Morrell	1250	789	751
Specialist Pathology	SPEED 12/12	Prof Lucy Raymond	1250	1065 (188 WES)	1043
Stem Cell and Myeloid Disorders	SMD 10/14	Prof Irene Roberts	600	94	63
Steroid Resistant Nephrotic Syndrome	SRNS 03/13	Dr Ania Koziell	250	93	84
Leber Resistant Nephrotic Syndrome	LHON	Prof Patrick Chinnery	70	0	0

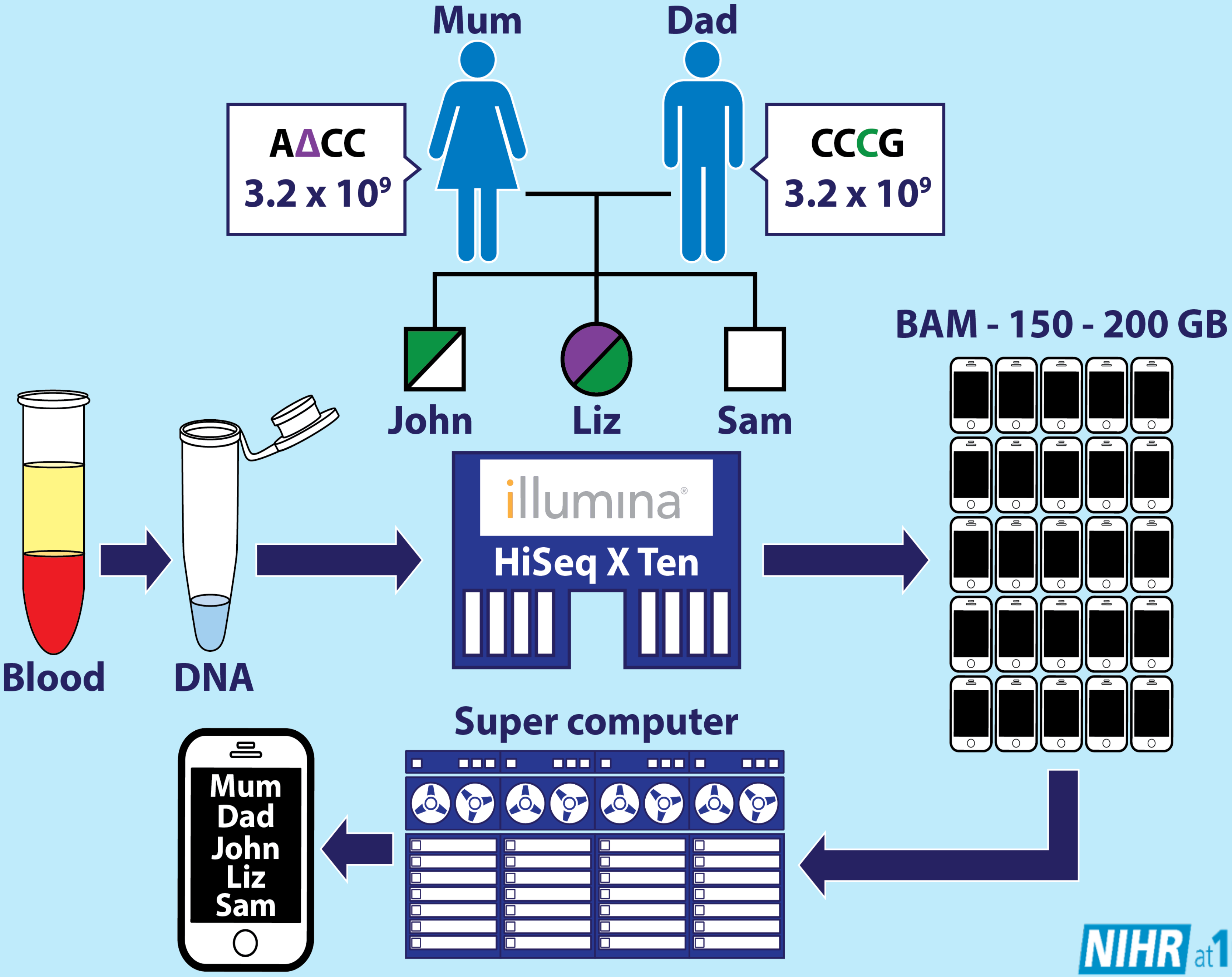
UK Biobank – 0.5 Million Genotyped Volunteers



= 100,000 males

= 100,000 females

Enrolment across the UK from 2006-10
Age at enrolment between 40-69 yrs
Linkage to NHS GP and Hospital records



Mum

Dad

AΔCC
3.2×10^9

CCCG
3.2×10^9

John

Liz

Sam

Blood

DNA

illumina

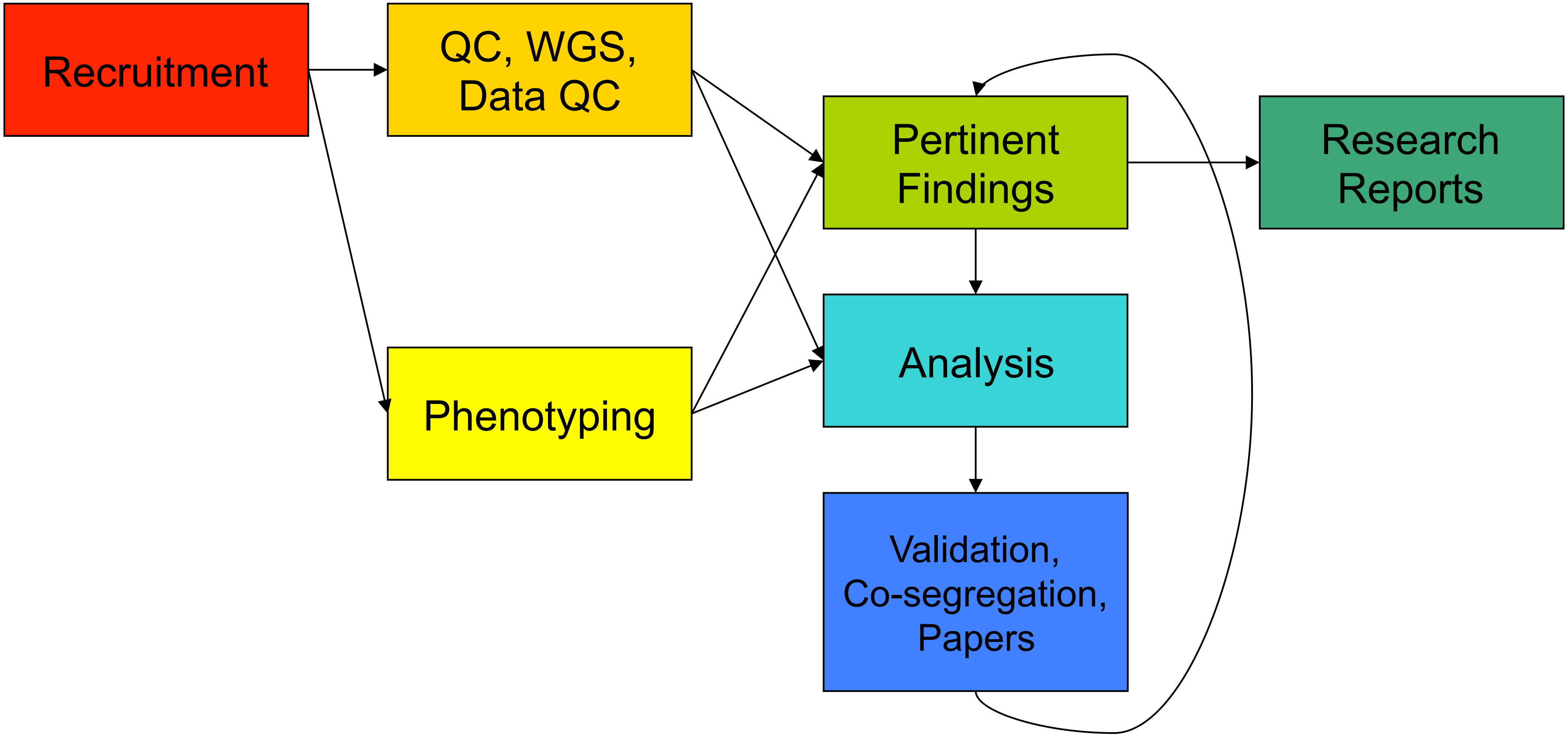
HiSeq X Ten

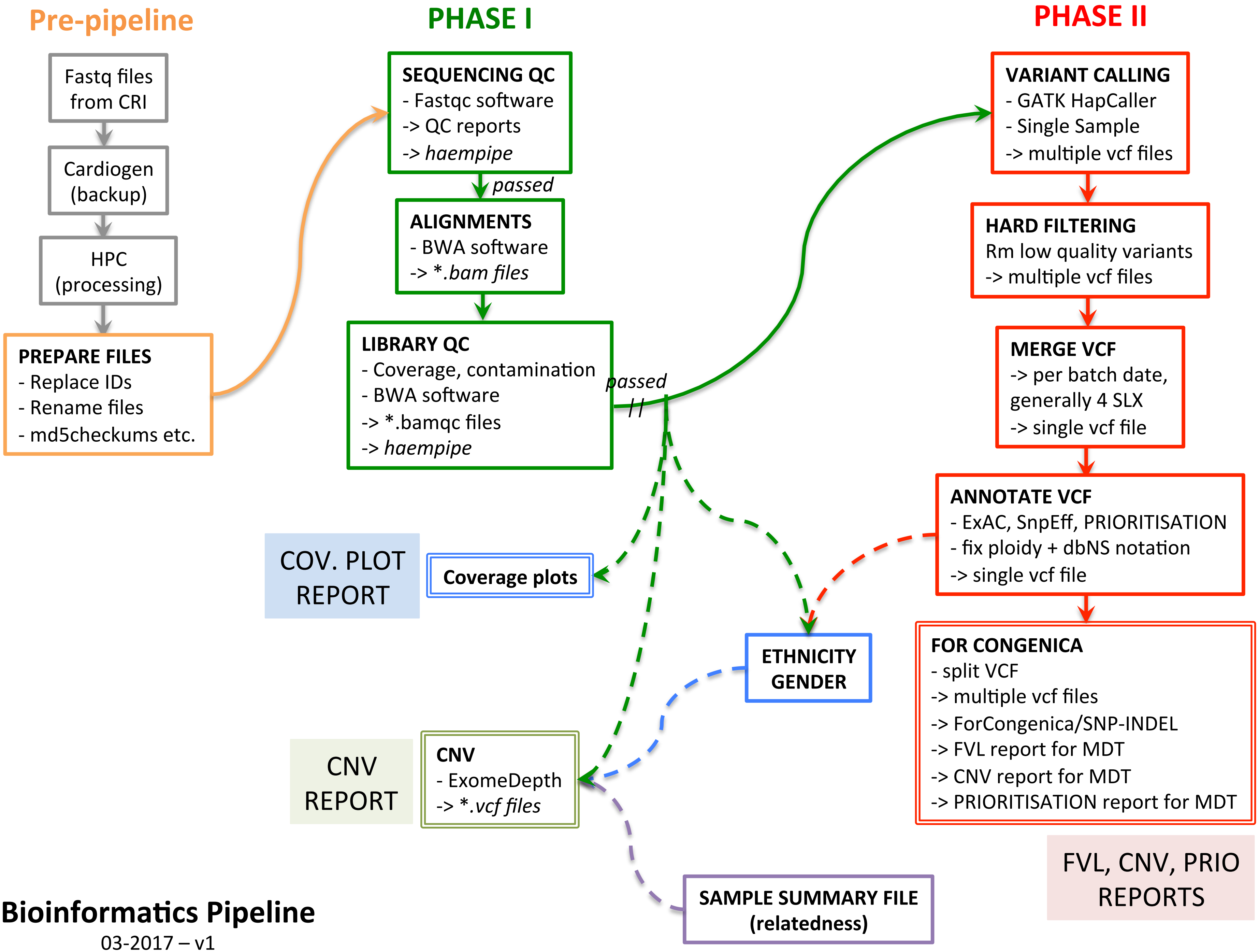
BAM - 150 - 200 GB

Super computer

Mum
Dad
John
Liz
Sam

NIHR BioResource – The Process





Pre-pipeline

Fastq files from CRI

Cardiogen (backup)

HPC (processing)

PREPARE FILES
- Replace IDs
- Rename files
- md5checkums etc.

PHASE I

SEQUENCING QC
- Fastqc software
-> QC reports
-> *haempipe*

passed

ALIGNMENTS
- BWA software
-> **.bam files*

LIBRARY QC
- Coverage, contamination
- BWA software
-> *.bamqc files
-> *haempipe*

passed

//

COV. PLOT REPORT

Coverage plots

ETHNICITY GENDER

CNV REPORT

CNV
- ExomeDepth
-> **.vcf files*

SAMPLE SUMMARY FILE (relatedness)

PHASE II

VARIANT CALLING
- GATK HapCaller
- Single Sample
-> multiple vcf files

HARD FILTERING
Rm low quality variants
-> multiple vcf files

MERGE VCF
-> per batch date, generally 4 SLX
-> single vcf file

ANNOTATE VCF
- ExAC, SnpEff, PRIORITISATION
- fix ploidy + dbNS notation
-> single vcf file

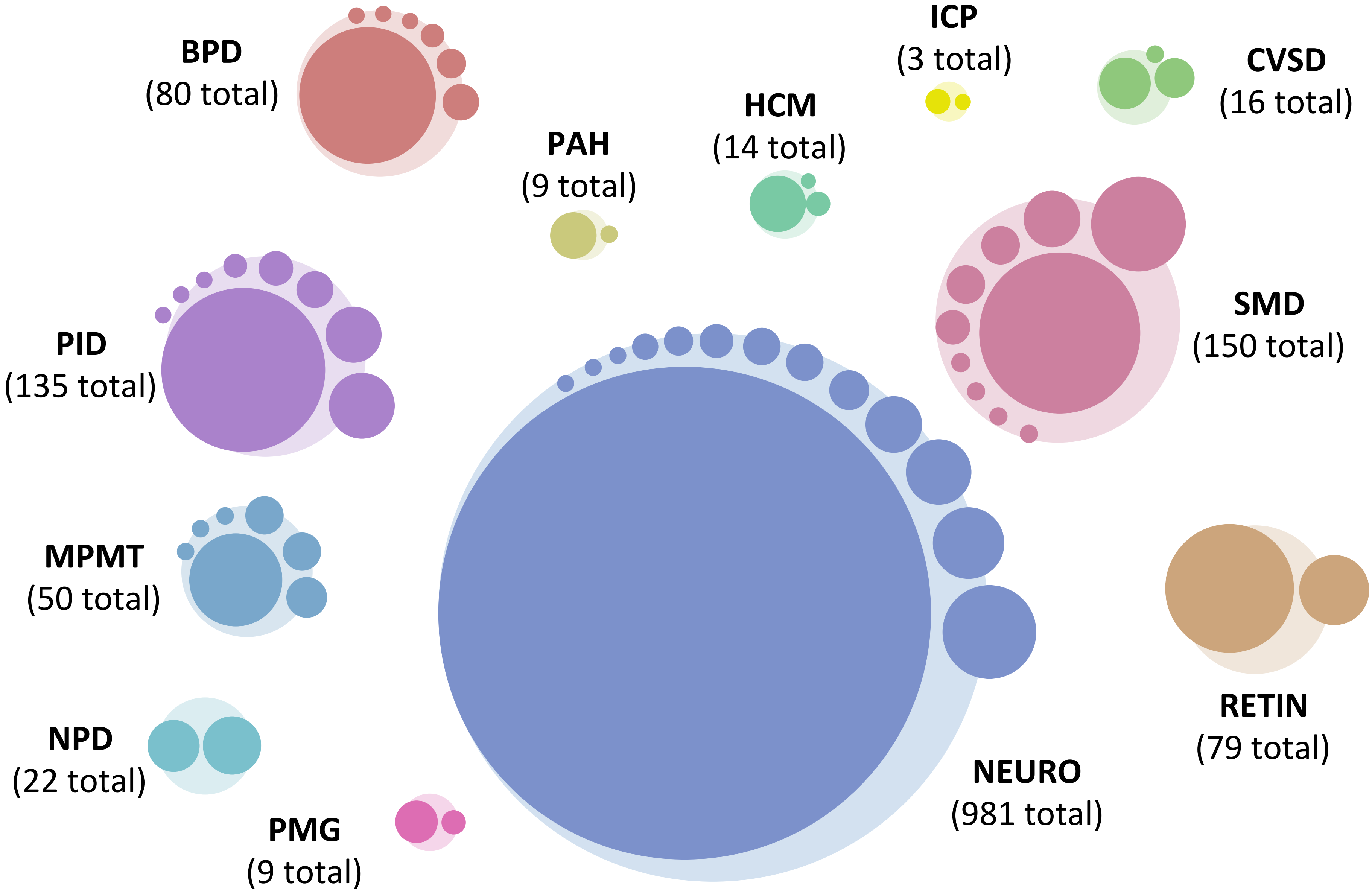
FOR CONGENICA
- split VCF
-> multiple vcf files
-> ForCongenica/SNP-INDEL
-> FVL report for MDT
-> CNV report for MDT
-> PRIORITISATION report for MDT

FVL, CNV, PRIO REPORTS

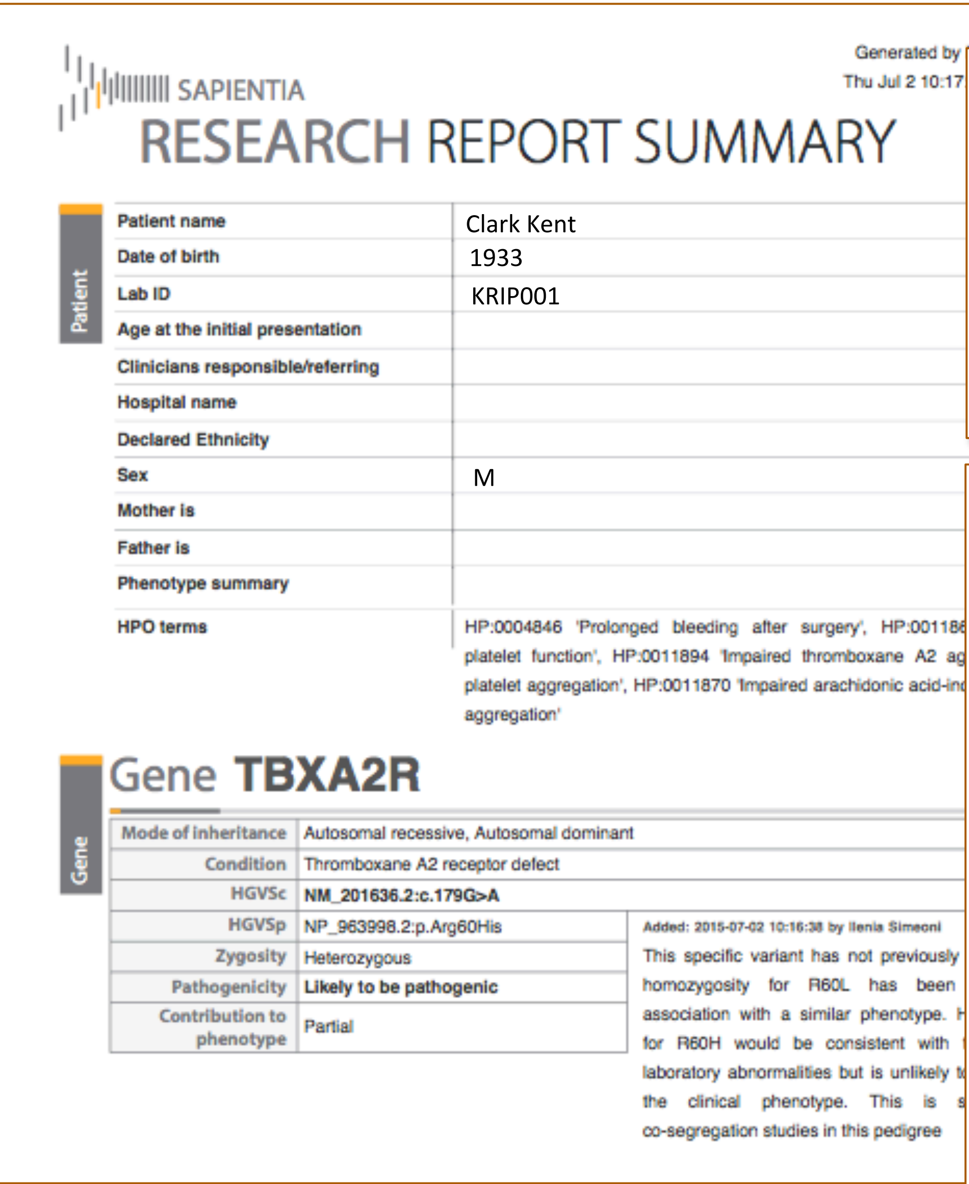
Bioinformatics Pipeline

03-2017 – v1

1,400 genes for clinical reporting



Research report



Generated by I
Thu Jul 2 10:17

SAPIENTIA

RESEARCH REPORT SUMMARY

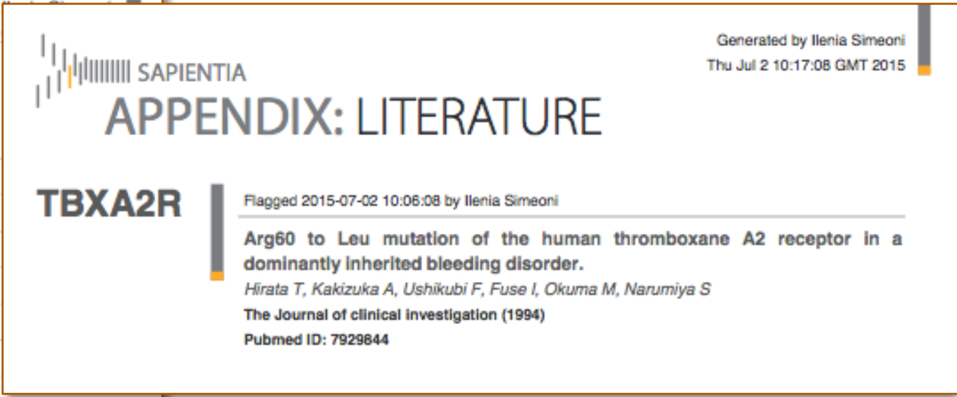
Patient name	Clark Kent
Date of birth	1933
Lab ID	KRIP001
Age at the initial presentation	
Clinicians responsible/referring	
Hospital name	
Declared Ethnicity	
Sex	M
Mother is	
Father is	
Phenotype summary	
HPO terms	HP:0004846 'Prolonged bleeding after surgery', HP:00118 platelet function', HP:0011894 'Impaired thromboxane A2 ag platelet aggregation', HP:0011870 'Impaired arachidonic acid-in aggregation'

Gene TBXA2R

Mode of inheritance	Autosomal recessive, Autosomal dominant
Condition	Thromboxane A2 receptor defect
HGVSc	NM_201636.2:c.179G>A
HGVSp	NP_963998.2:p.Arg60His
Zygosity	Heterozygous
Pathogenicity	**Likely to be pathogenic**
Contribution to phenotype	Partial

Added: 2015-07-02 10:16:38 by Ilenia Simeoni

This specific variant has not previously homozygosity for R60L has been association with a similar phenotype. H for R60H would be consistent with t laboratory abnormalities but is unlikely t the clinical phenotype. This is s co-segregation studies in this pedigree



Generated by Ilenia Simeoni
Thu Jul 2 10:17:08 GMT 2015

SAPIENTIA

APPENDIX: LITERATURE

TBXA2R

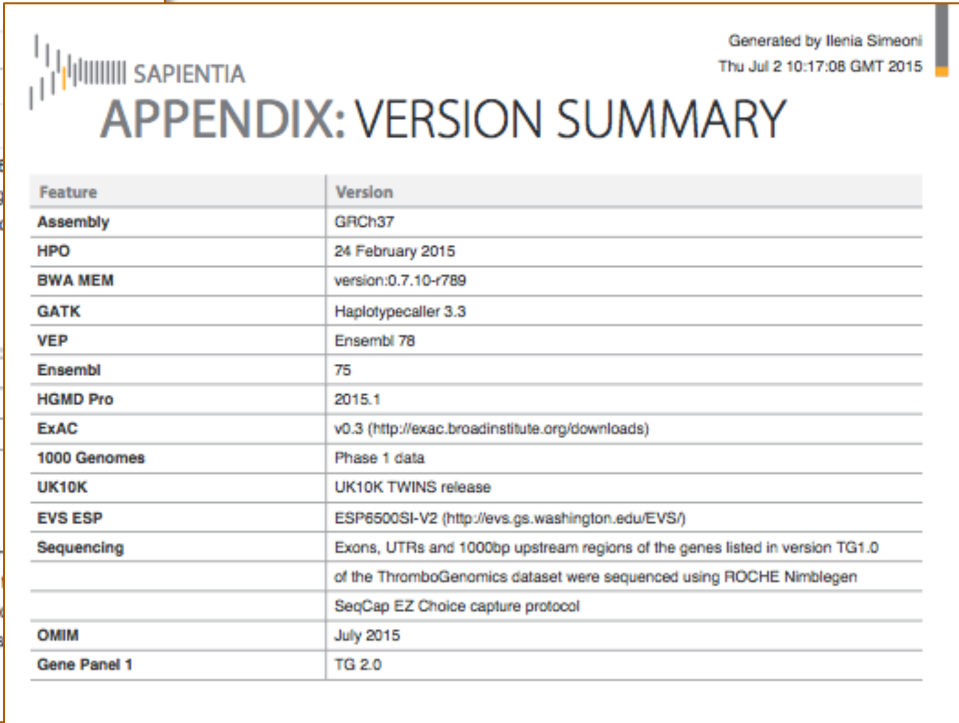
Flagged 2015-07-02 10:06:08 by Ilenia Simeoni

Arg60 to Leu mutation of the human thromboxane A2 receptor in a dominantly inherited bleeding disorder.

Hirata T, Kakizuka A, Ushikubi F, Fuse I, Okuma M, Narumiya S

The Journal of clinical investigation (1994)

Pubmed ID: 7929844



Generated by Ilenia Simeoni
Thu Jul 2 10:17:08 GMT 2015

SAPIENTIA

APPENDIX: VERSION SUMMARY

Feature	Version
Assembly	GRCh37
HPO	24 February 2015
BWA MEM	version:0.7.10-r789
GATK	Haplotypecaller 3.3
VEP	Ensembl 78
Ensembl	75
HGMD Pro	2015.1
ExAC	v0.3 (http://exac.broadinstitute.org/downloads)
1000 Genomes	Phase 1 data
UK10K	UK10K TWINS release
EVS ESP	ESP6500SI-V2 (http://evs.gs.washington.edu/EVS/)
Sequencing	Exons, UTRs and 1000bp upstream regions of the genes listed in version TG1.0
	of the ThromboGenomics dataset were sequenced using ROCHE Nimblegen
	SeqCap EZ Choice capture protocol
OMIM	July 2015
Gene Panel 1	TG 2.0