

A Targeted Capture Approach to Next Generation Sequencing

Christopher Watson

Yorkshire Regional Genetics Laboratory

St James's University Hospital

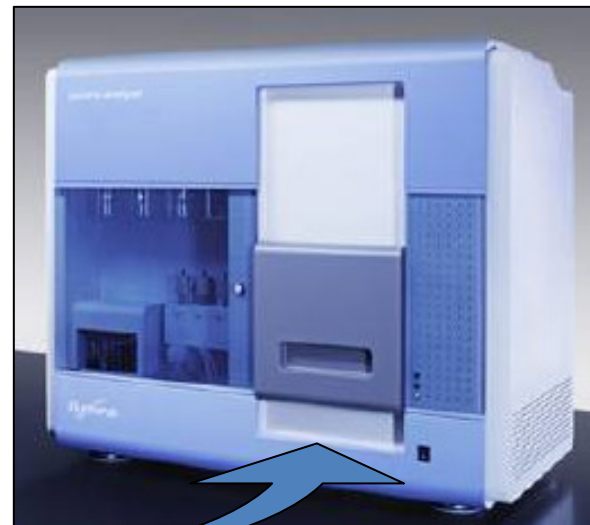
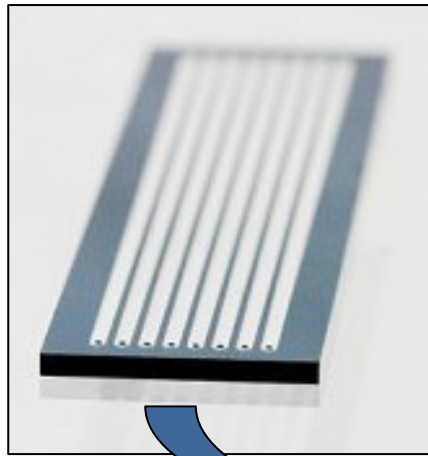
Next Generation Sequencing

ABI Solid, Roche 454, Illumina GAIIx

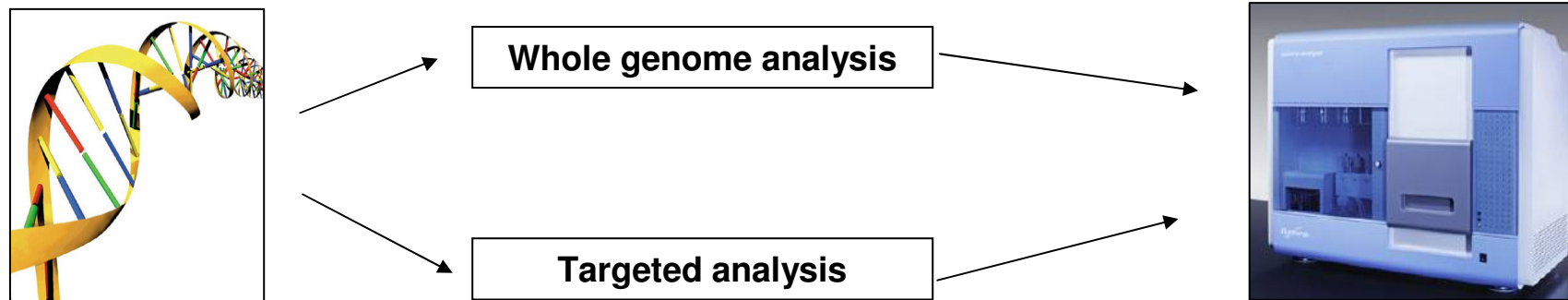
Short single reads

8x lanes (~2 gigabases per lane)

A significant increase in sequencing data output



Methods to obtain DNA for sequencing



1. Long range PCRs

- > *BRCA1, BRCA2, MLH1, MSH2, MSH6*
- > Multiple samples per run

2. Hybridisation enrichment of genomic DNA

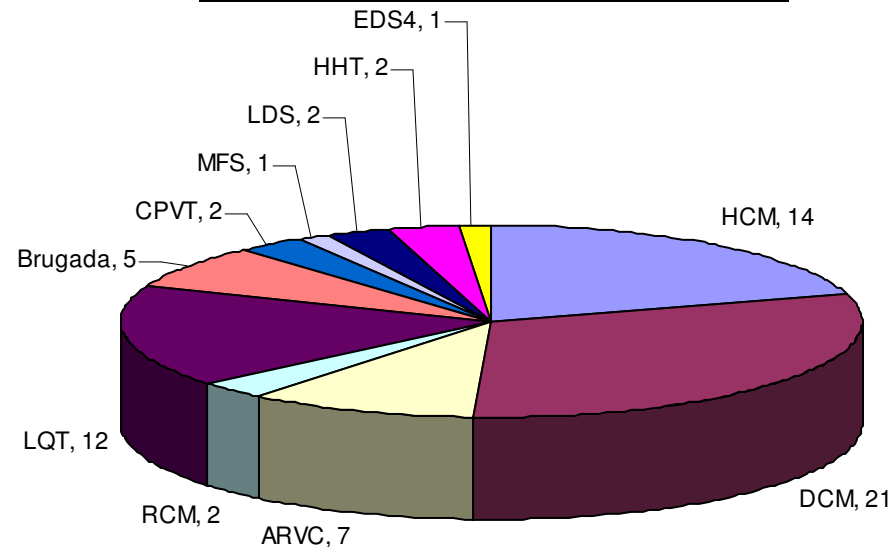
- > Suitable for multiple genes (exome sequencing)
- > Exclusion of wasteful data not routinely analysed (intronic sequences)

Cardiac disease as a model

Distinct phenotypic subtypes

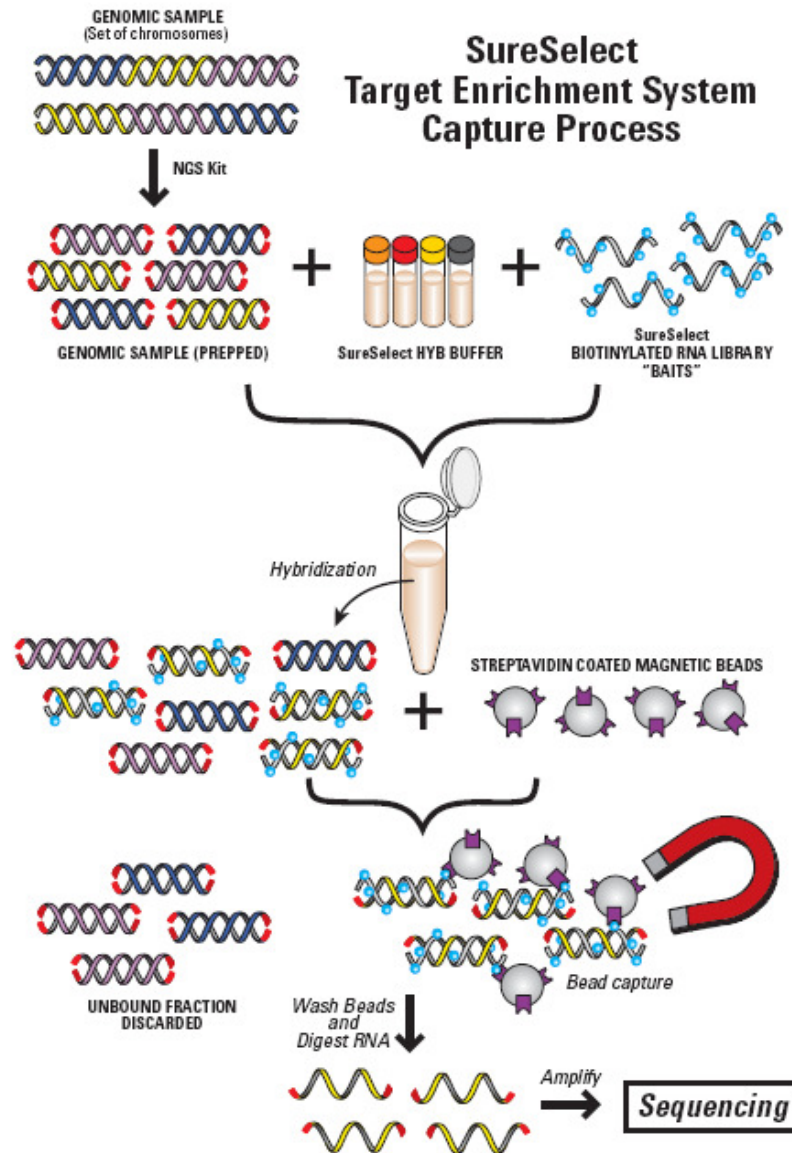
Genetically heterogeneous

The number of genes per cardiac subtype

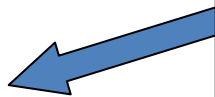


HCM = Hypertrophic cardiomyopathy, DCM = Dilated cardiomyopathy, ARVC = Arrhythmogenic right ventricular cardiomyopathy, RCM = Restrictive cardiomyopathy, LQT = Long QT syndrome, Brugada = Brugada syndrome, CPVT = Catecholaminergic Polymorphic Ventricular Tachycardia, MFS = Marfans syndrome, LDS = Loeys-Dietz syndrome, HHT = Hereditary hemorrhagic telangiectasia, EDS4 = Ehlers-Danlos syndrome type 4.

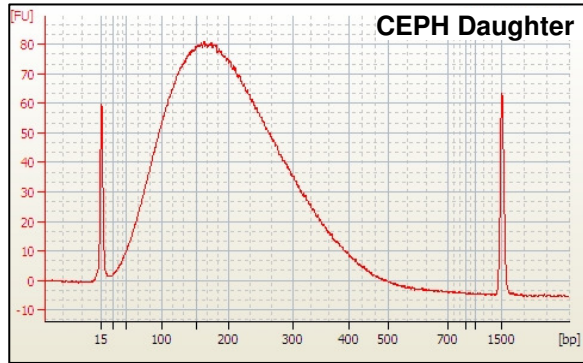
Agilent technologies – Pull down methodology



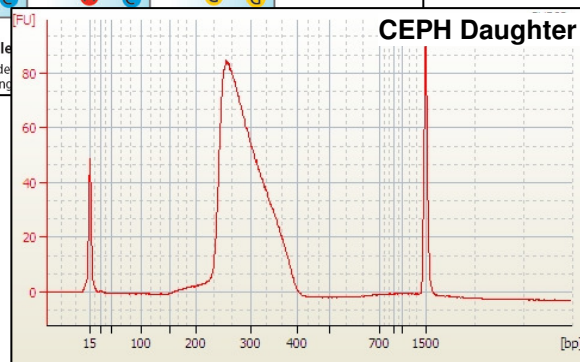
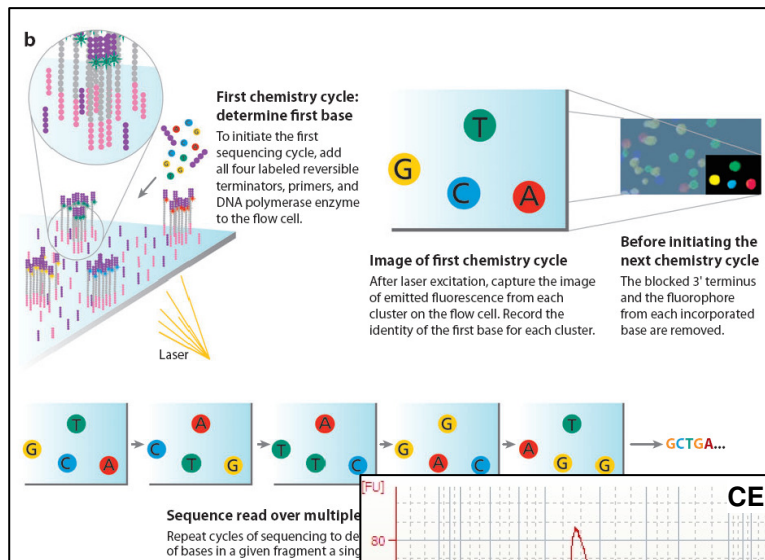
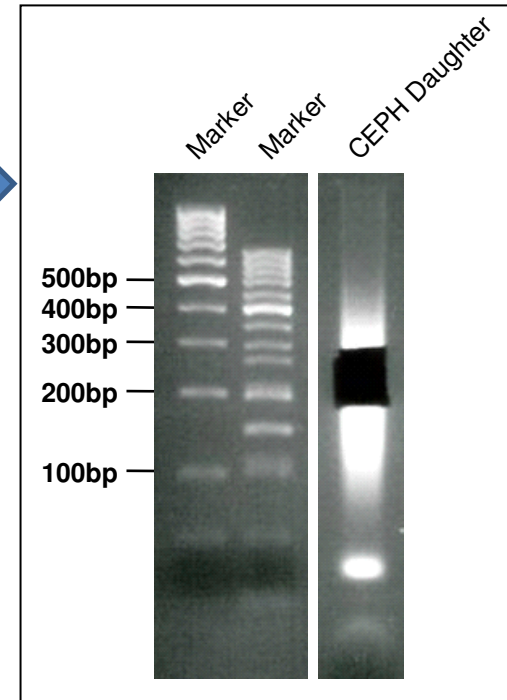
54 genes
1,297 exons
0.6 Mb



Sheared genomic DNA



End repair; A
addition; Adapter
ligation; Extract &
purify

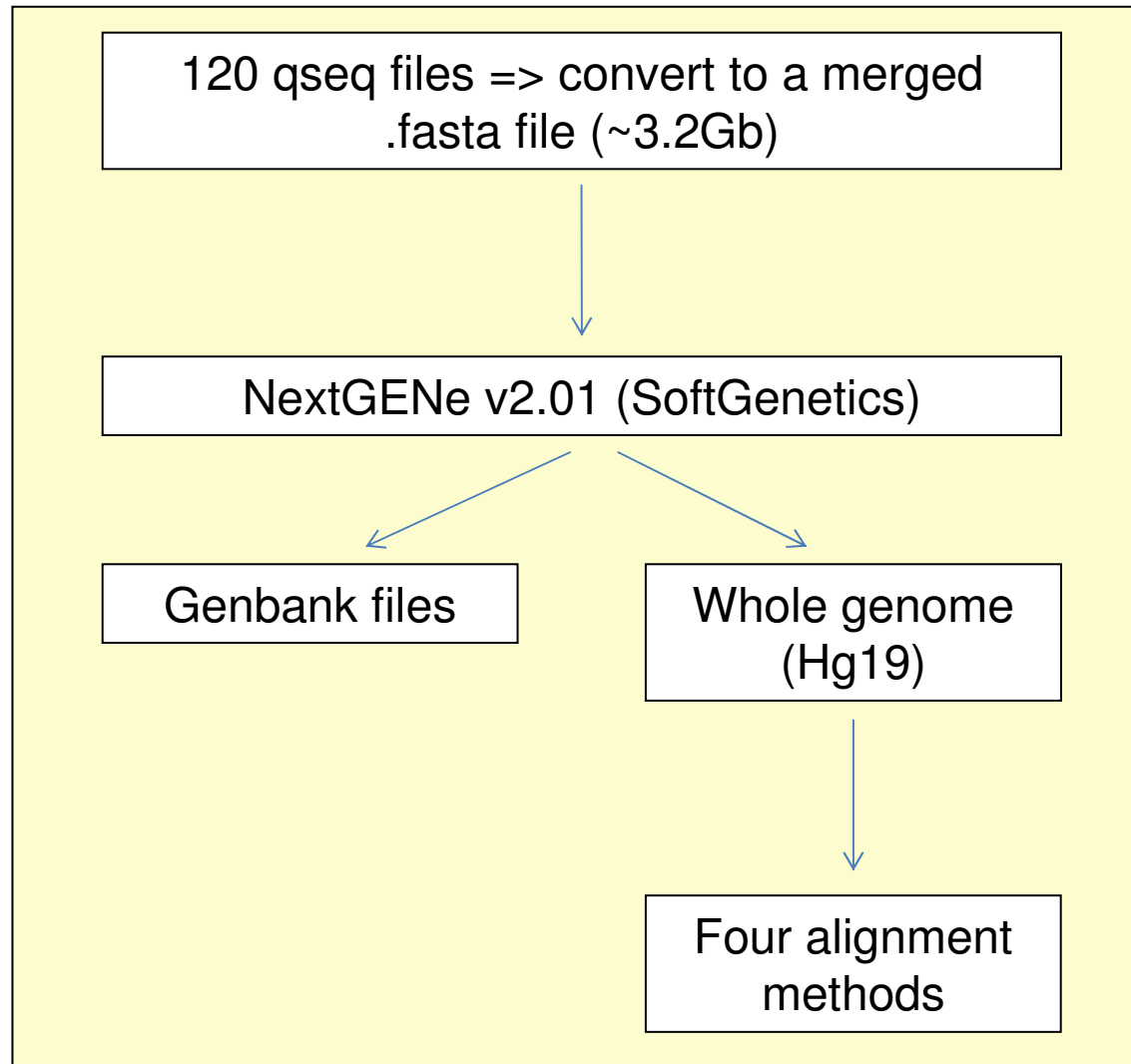


Hybridisation (24 hours @ 65°C)

Magnetic bead selection of
hybridised fragments

Post hybridisation PCR

Analysis



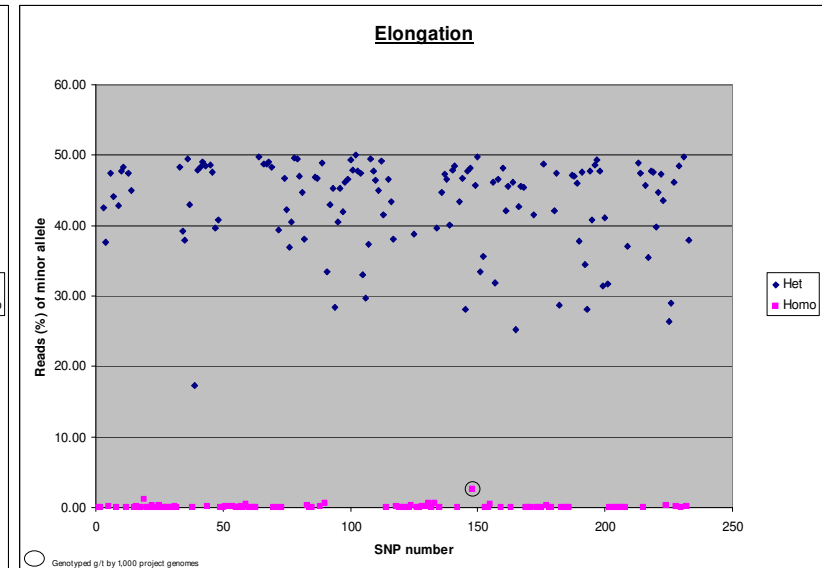
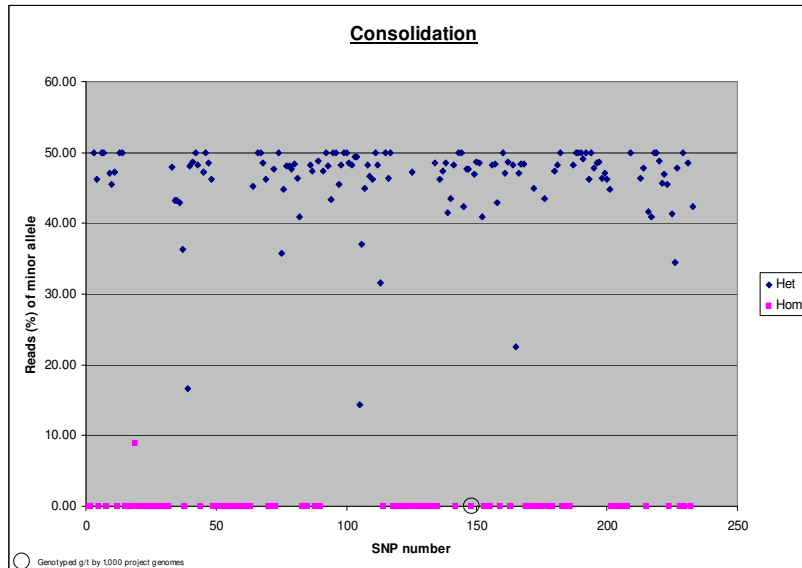
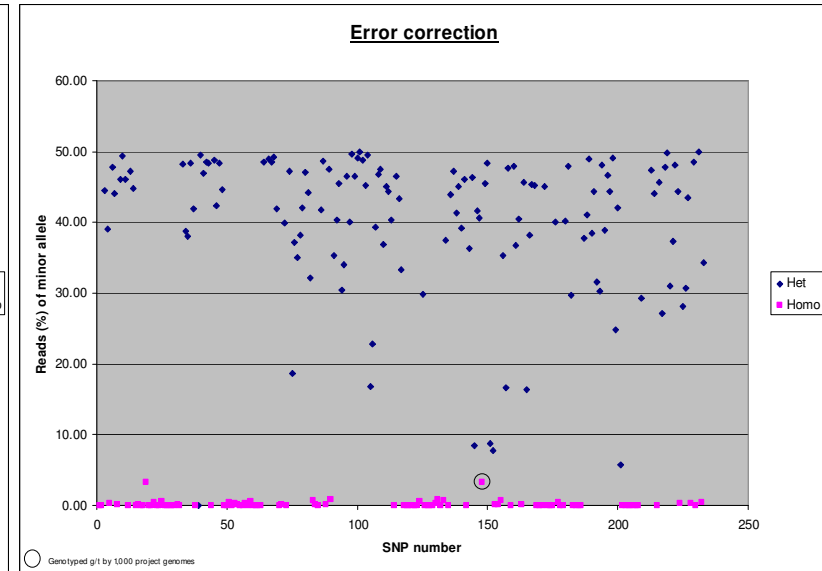
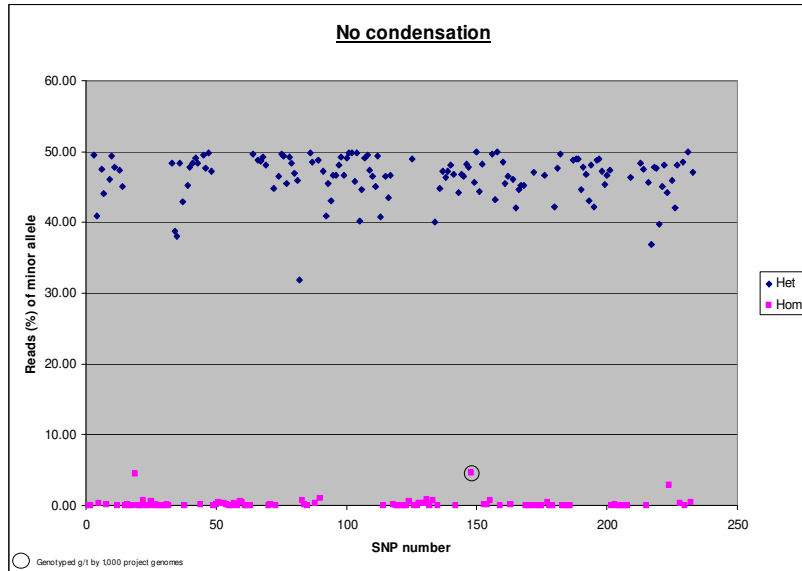
NextGene v2.01 condensation alignments

Removes base calling instrument error

Reads with the same 12bp anchor sequence grouped => flanking sequence processed

Alignment method	Principle	Run time
No condensation	Reads aligned no modifications	41 minutes
Consolidation	Overlapping sequences merged, consensus read used	3 hours 1 minute
Error correction	Overlapping sequences not merged, elongated read created	3 hours 29 minutes
Elongation	Low frequency errors corrected no extension / merger of reads	3 hours 37 minutes

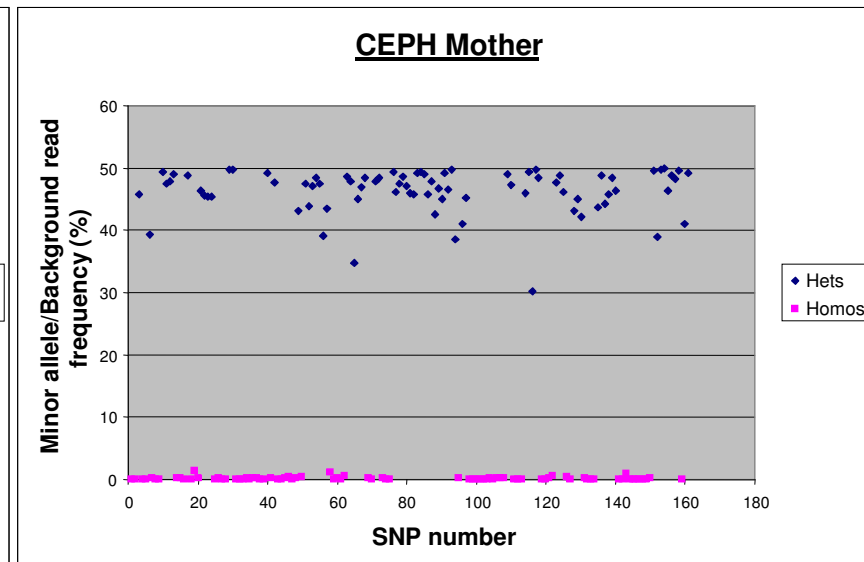
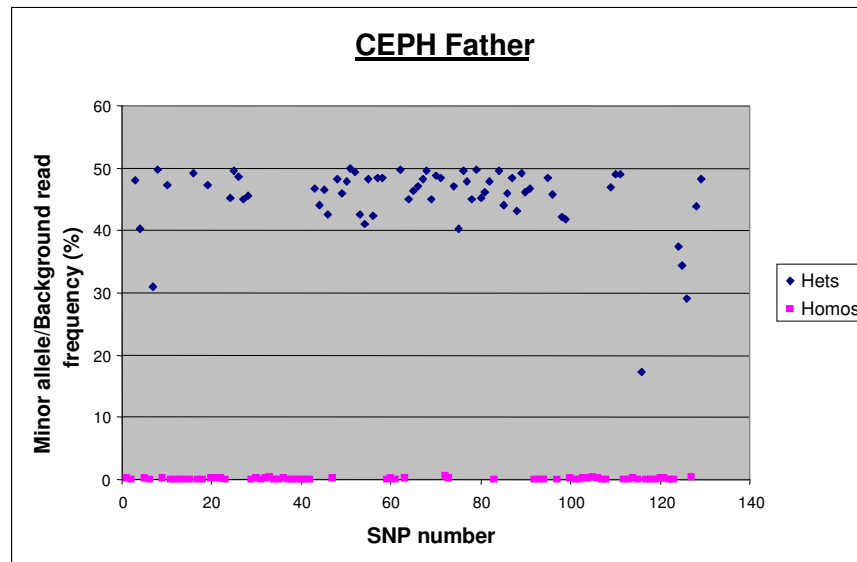
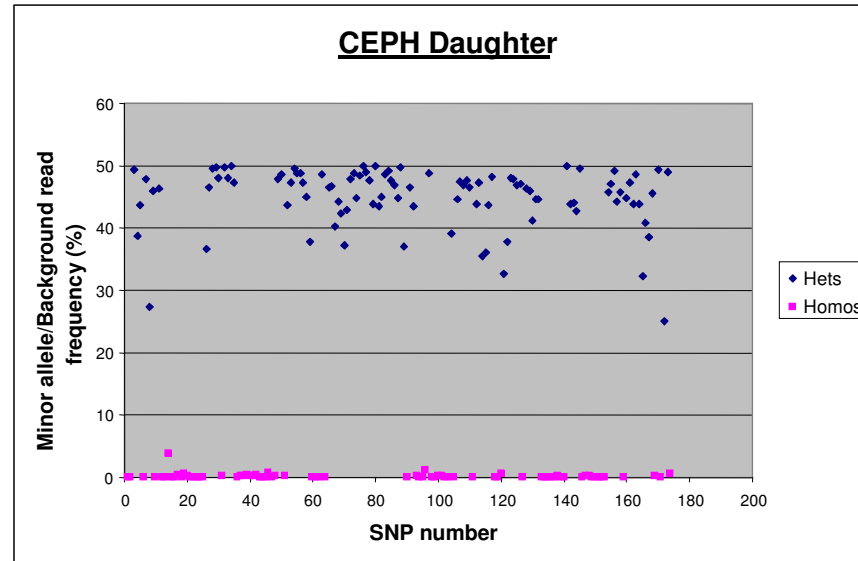
CEPH Daughter (known SNPs)



Run performance – no condensation alignments

Sample (ID)	Sequence reads		
	Mapped (%)	Unmapped (%)	Total
Daughter NA12878	20,552,590 (95.8)	899,296 (4.2)	21,451,886
Mother NA12892	27,250,191 (97.0)	846,837 (3.0)	28,097,028
Father NA12891	25,744,721 (97.2)	754,491 (2.8)	26,499,212
Average	24,515,834 (96.7)	833,541 (3.3)	25,349,375

Genotype separation of known SNPs



Concordance against 1,000 genomes data

7 discordant genotypes from 4 loci

Two good quality alignments - incorrect data?

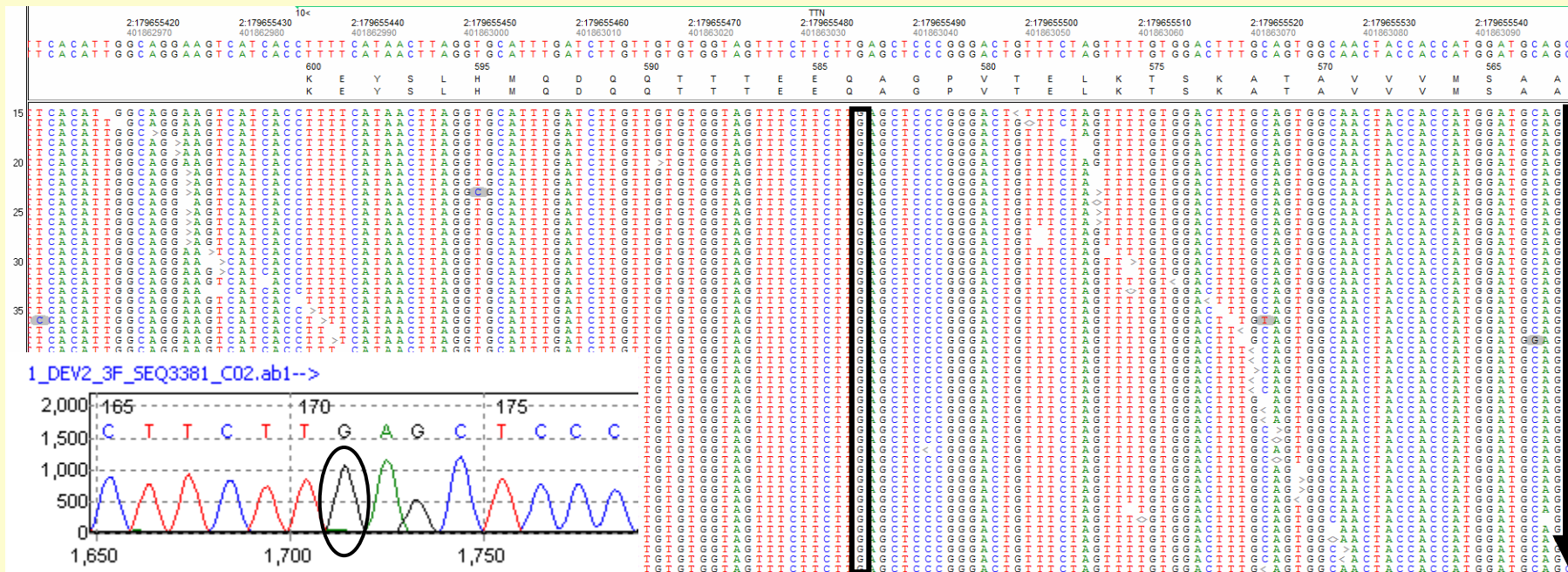
Two difficult alignments

Sample (ID)	Concordant SNP genotypes			Putative discordant genotypes
	Homozygous	Heterozygous	Total	
Daughter NA12878	74	100	174	3
Father NA12891	63	66	129	3
Mother NA12892	80	81	161	1
Total	217	247	464	7

Example 1 – Miscalled variant

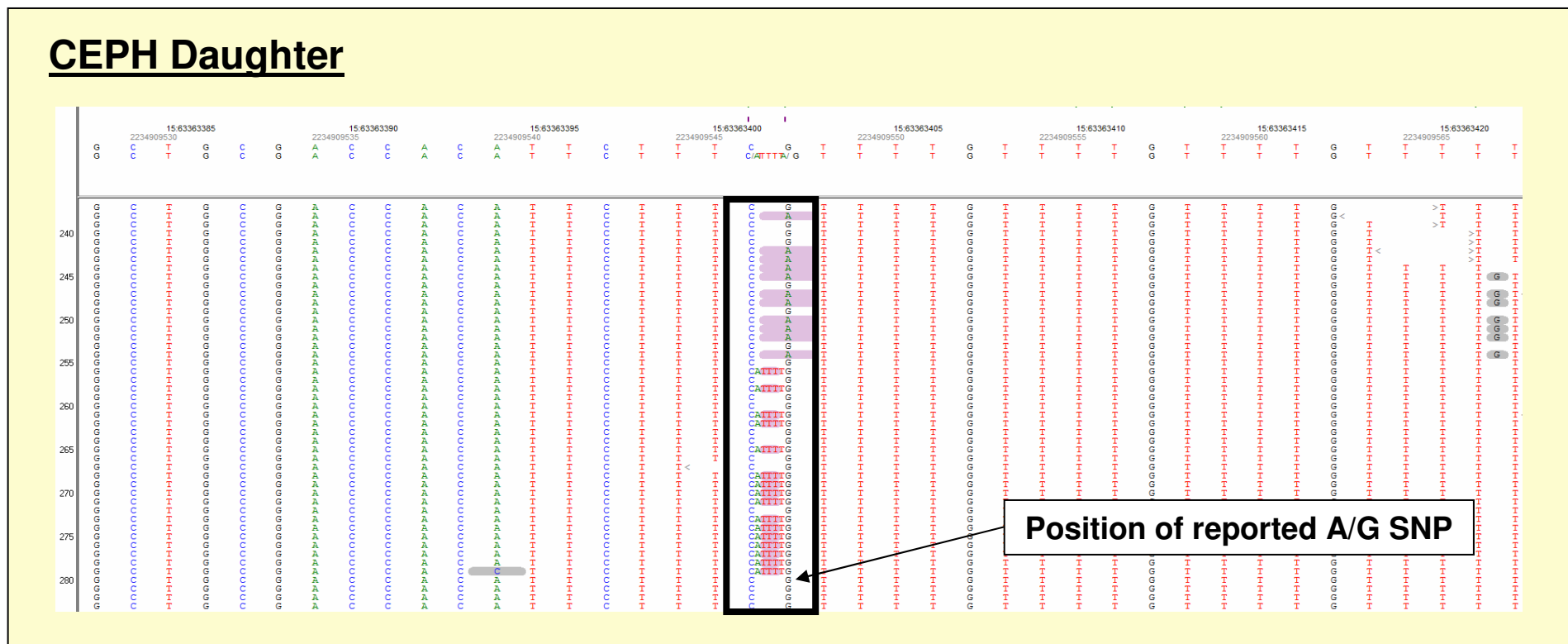
Chr	Position	Gene	dbSNP #	Daughter		Father		Mother		Sanger result
				1K	NGene	1K	NGene	1K	NGene	
2	179,655,485	TTN	rs72957309	G/T	G/G	G/T	G/G	G/T	G/G	Genotypes G/G homozygous

CEPH Father



Example 2 – Poor alignment

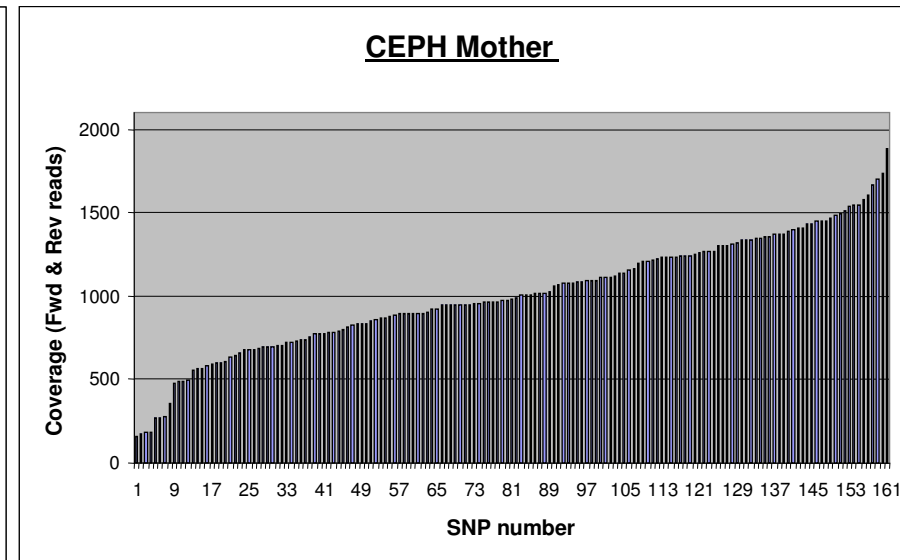
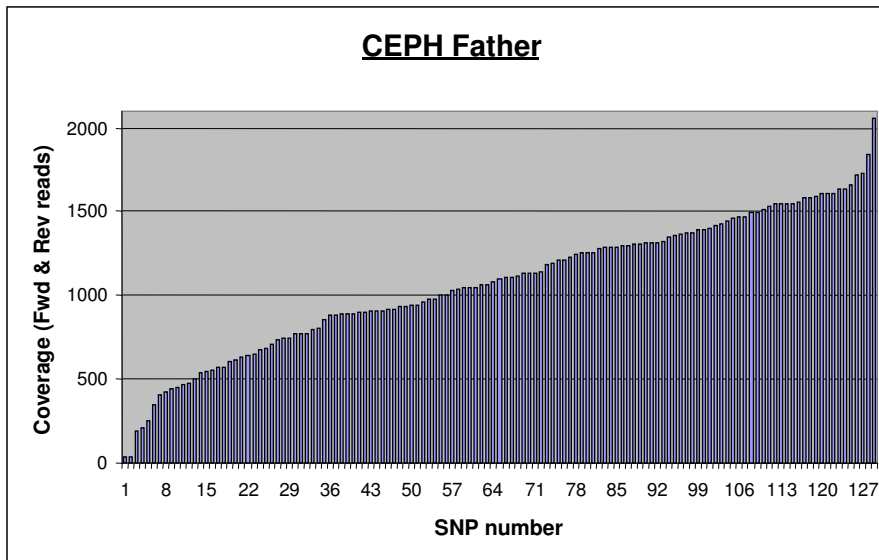
Chr	Position	Gene	dbSNP #	Daughter		Explanation
				1K	NGene	
15	63,363,402	TPM1	rs11558748	A/G	insATTTT heterozygous	Difficult alignment region. Het. insertion visible from reads



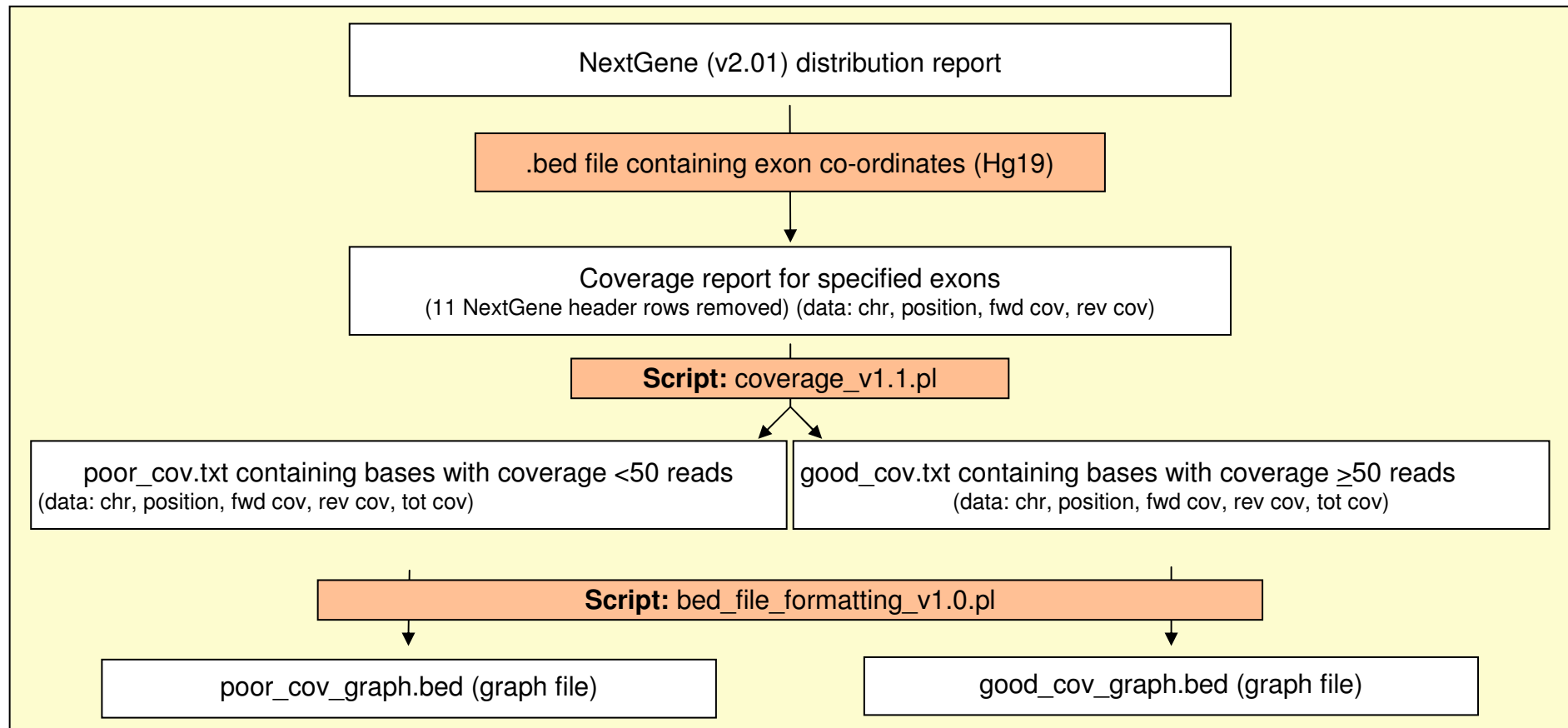
Explanation for all putatively discordant genotypes

Read depth of known SNPs

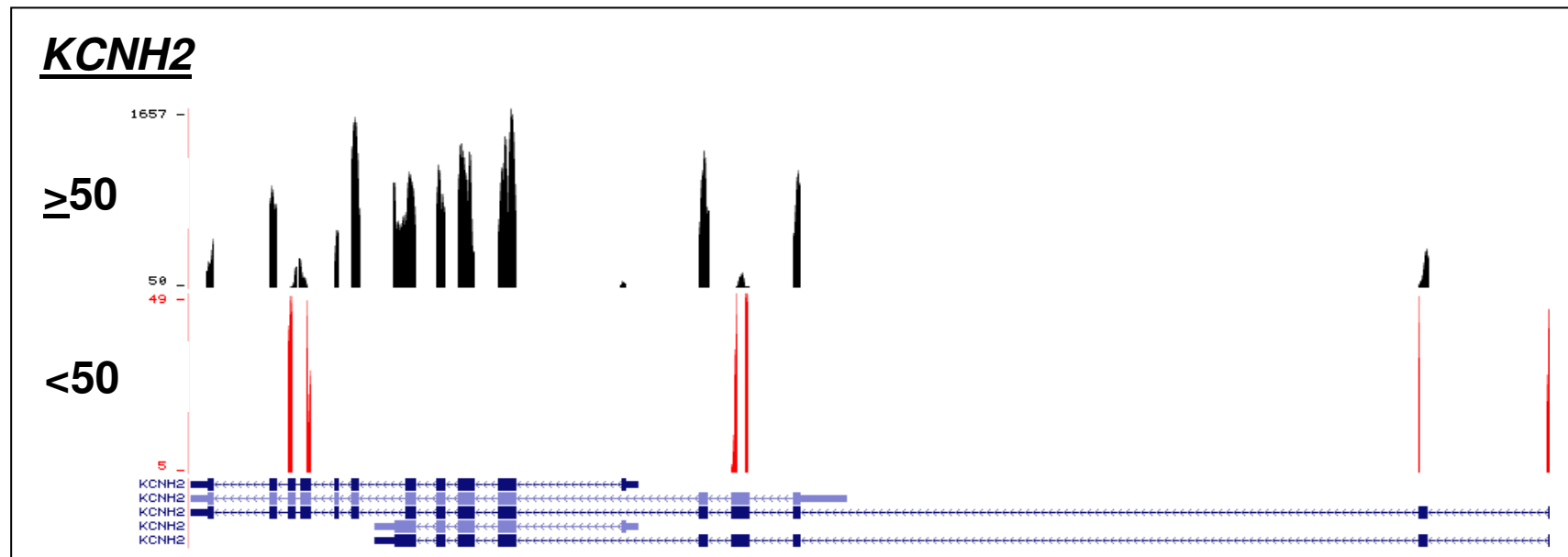
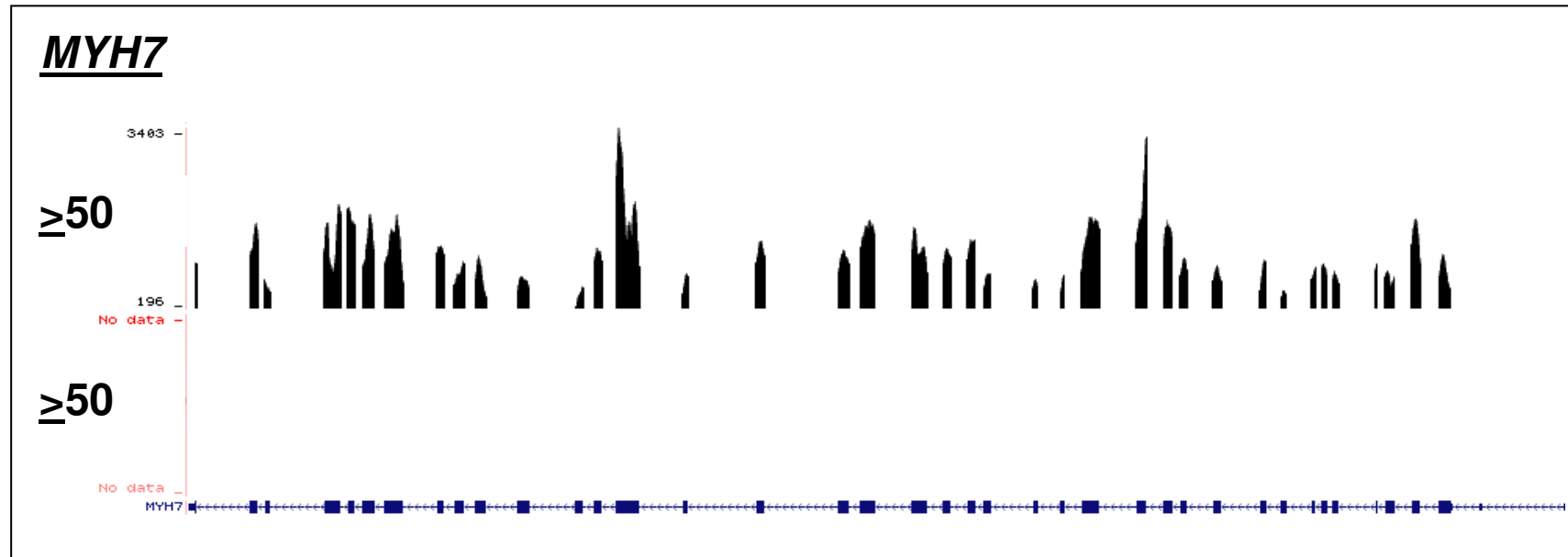
Fifteen missing genotypes: Daughter 5, Father 4, Mother 6
How to visualise the *'missing'* data?



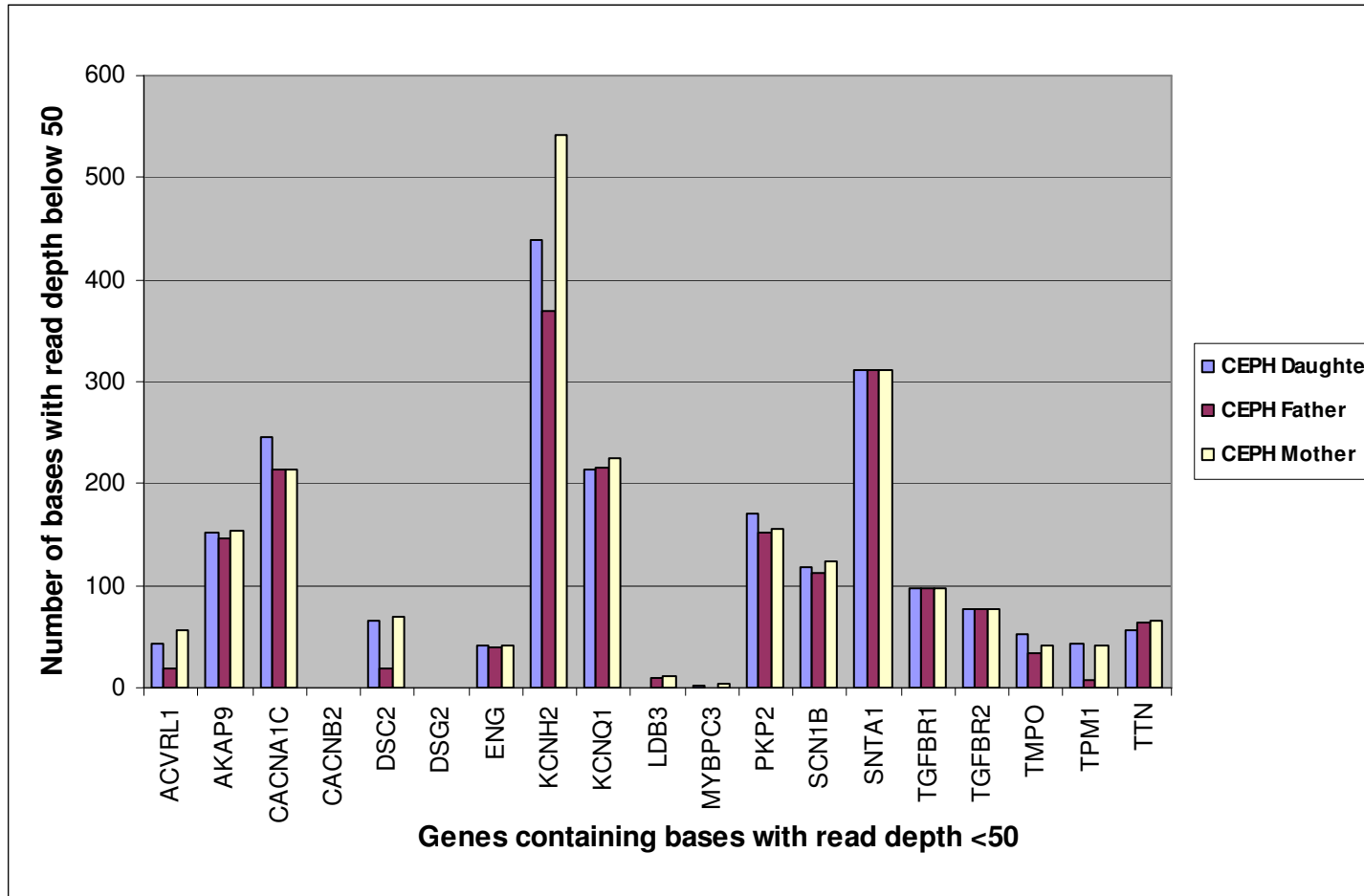
Read depth pipeline



Genome browser – CEPH Daughter



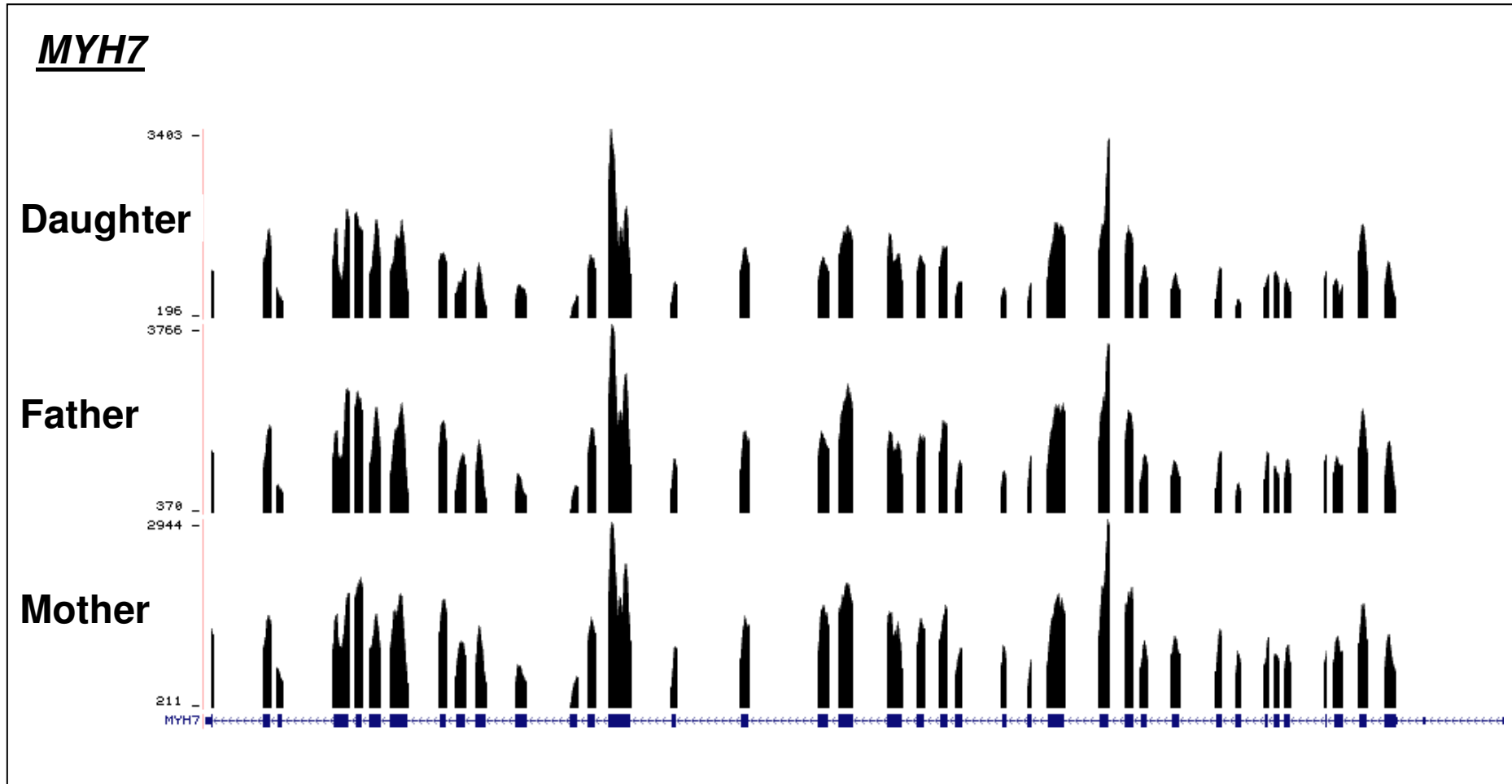
Genes with bases that have a read depth <50



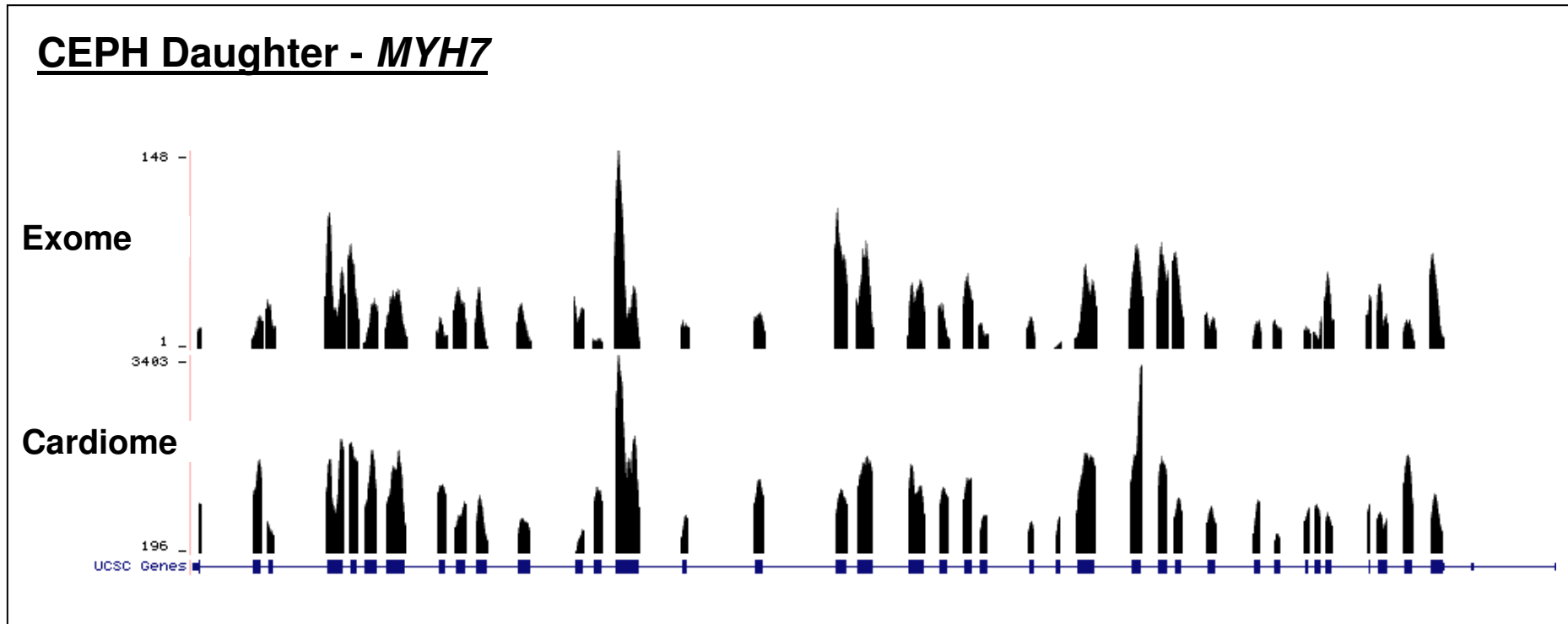
19 genes => 30/1,297 exons

97.7% of target exons / 99.2% of target CDS

Coverage replication between samples

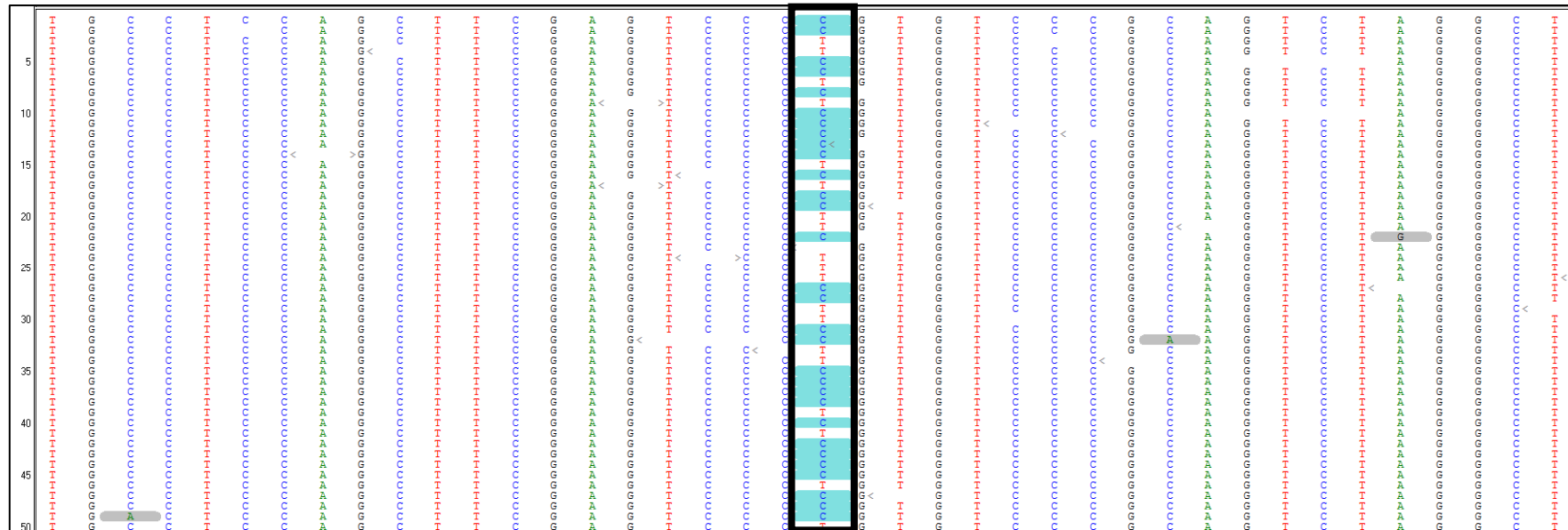


DNA sequence context specific



HCM Positive control

MYBPC3 c.927-2A>G splice site mutation



Chr	Chr Position	Gene	Ref Nucleotide	Coverage	C(%)	T(%)	A(%)	G(%)
11	47,367,923	MYBPC3	T	516	51.55	48.45	0	0

NB. MYBPC3 is encoded on the antisense strand, bases are therefore listed as T/C rather than A/G

Dilated cardiomyopathy case studies

Clinical details

Weakened and enlarged, unable to pump blood efficiently

Strong family histories, multiple affected family members

No previously identified familial mutation

Prior exclusion of *MYH7*, *TNNT2*, *TNNI3*, *LMNA* as the cause of DCM

	Case 1	Case 2
CDS +/- 20bp of 21 DCM genes	134	125
Exclusion of variants in dbSNP	15	17
Possibly causative	3	6

Conclusions

Successful design of cardiac capture probes

Significant number of concordant genotypes

Validation of positive HCM patient sample

Replication of coverage between runs/assay

Manipulation of large datasets

In-house expertise and knowledge of Next-Gen capture experiments

Future developments

Growing potential of multi gene analysis => informatic complexity of UV classification

Currently cost-prohibitive, investigating the possibility of sample multiplexing

Long range PCR service for 4 HCM genes and 5 LQT genes

Acknowledgements

Staff at the Yorkshire Regional Genetics Laboratory

Prof Graham Taylor – Sequencing

Prof Colin Johnson, Leeds – CEPH DNA