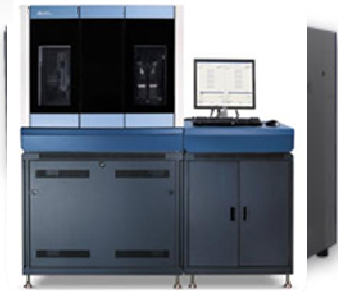


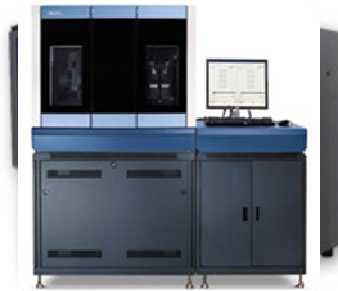
# Development and implementation of a clinical next generation sequencing service for patients with X-linked learning disability

Howard Martin

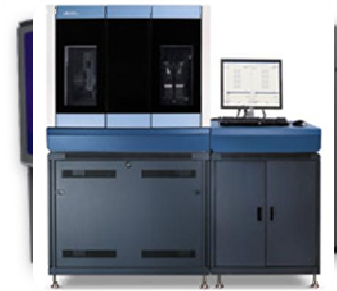
Cambridge Regional Molecular Genetics Laboratory  
Cambridge University Hospital NHS Foundation Trust



S5500x4  
1300 Gb



S5500x4  
1300 Gb



S5500x4  
1300 Gb



GAIIX  
100 Gb



GS FLX Titanium  
500 Mb



GS Junior  
35 Mb



PGM  
10-1000 Mb

# High Performance Computing Service



The Darwin cluster  
4000 cores, 8 TB of total memory



Utah Penguin Cloud  
Bioscope  
Di bayes  
Small intel



## Project background

- Investigation of a child with learning disability is one of the main referral reasons for paediatric, neurological and genetic services
- Common [1-2% of the population] ~5-10% of overall health care expenditure
- ~50% of cases with suspected genetic cause, underlying abnormality not identified
- ~10% of cases are estimated to be caused by single gene abnormality on the X
- Current approach is routine karyotype and FRAX testing.....
- ~100 genes now identified in association with syndromic and non-syndromic XLMR
- Local clinical and research expertise in identifying novel genes causing XLMR

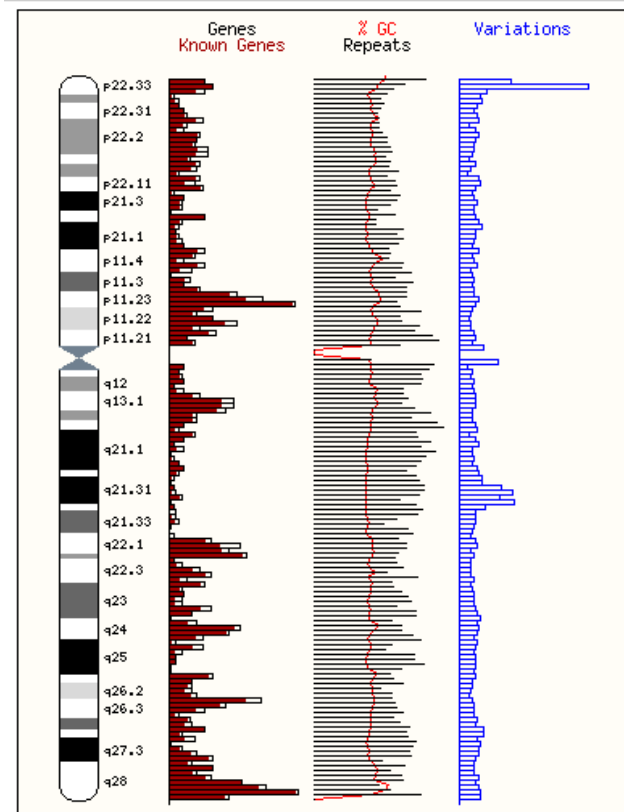
# Pilot proof of principle (PoP) project

- 10 patients (XLMR inheritance confirmed clinically)
- X exome previously sequenced by standard Sanger sequencing (7/10) 2009
- Approx 100 - 140 variants per patient (~880 variants proof of principle trial)
- Large data set of known recurrent variants on the X
- Small numbers of non recurrent variants

## Method

- Targeted enrichment and re-sequencing
- 3 enrichment platform approaches
  - Agilent SureSelect
  - *Febit HybSelect*
  - NimbleGen EZ
- 4 x 10 enrichment libraries generated
  - Agilent X Demo
  - Agilent X custom
  - Nimblegen solid phase custom
  - Nimblegen EZ liquid phase custom
- SOLiD sequencing platform

# X chromosome enrichment designs



861 known protein coding genes  
155,270,560 bp [GRCh37]

3 Mb capture region

- All platforms given the same design specifications
- Review of preliminary designs
  - ↓
- ID regions of poor probe placement (cause?)
  - ↓
- Augment designs to force additional probes
  - ↓
- Sign-off design, perform enrichments, sequence 3+4
  - ↓
- Analysis of results to ID 'most suitable' platform
  - ↓
- Re-designs to enhance regions of poor capture  
Agilent SureSelect custom library v4

## Enrichment libraries



Agilent SureSelect X Demo      xx,xxx baits      Liquid phase      Barcoded version



Agilent SureSelect X custom      xx,xxx baits      Liquid phase      Barcoded version  
Augmented custom version 4 in manufacture xx,xxx baits



NimbleGen custom      xxx,xxx baits      Solid phase      EZ Liquid phase  
Baylor and Beverly protocols required

# Analysis pipeline

## Primary analysis

Mapping

QC filtering

Optional

Indel realigning

Base score recalibration

## Secondary analysis

Indel detection

SNP detection

SNP filtering & clustering

SNP rescoring

## Summary enrichment performance

- **Agilent X\_exome genome mapped rates**

Average xx % Demo

Average xx % Custom

- **Agilent X\_exome X mapped reads [% on target]**

Average xx % mapped on target

- **Agilent X\_exome average depth of reads**

Average depth xx x [SOLiD3]

SOLiD3      Av xx M reads per Oct [SOLiD3]

Av xx M reads per Oct [SOLiD4]

# Typical SNP calling output for PoP study

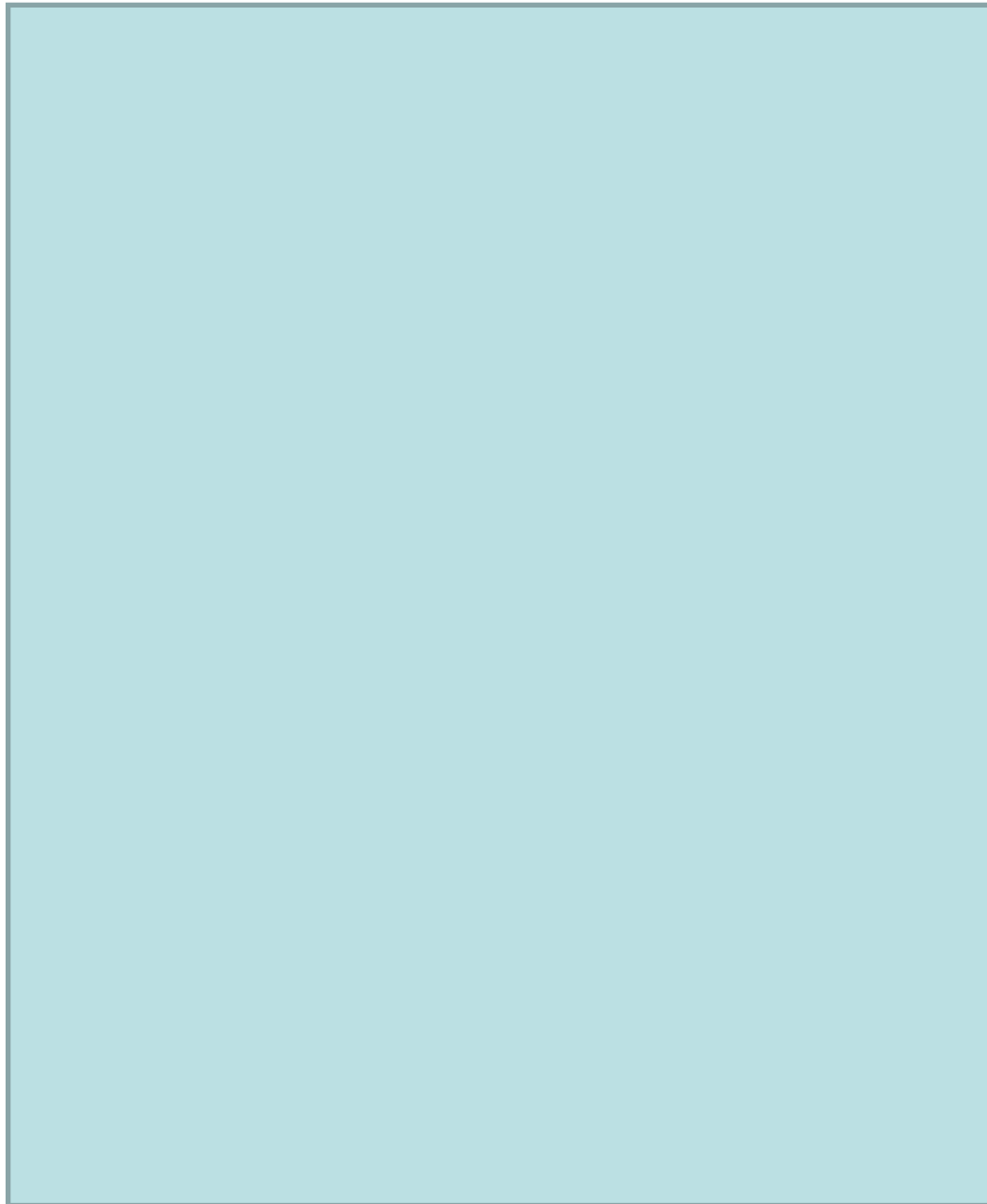
Sample ID	SNP ID	Reference	Sample	Quality
1	1	A	A	30
1	2	C	C	30
1	3	G	G	30
1	4	T	T	30
1	5	A	A	30
1	6	C	C	30
1	7	G	G	30
1	8	T	T	30
1	9	A	A	30
1	10	C	C	30
1	11	G	G	30
1	12	T	T	30
1	13	A	A	30
1	14	C	C	30
1	15	G	G	30
1	16	T	T	30
1	17	A	A	30
1	18	C	C	30
1	19	G	G	30
1	20	T	T	30
1	21	A	A	30
1	22	C	C	30
1	23	G	G	30
1	24	T	T	30
1	25	A	A	30
1	26	C	C	30
1	27	G	G	30
1	28	T	T	30
1	29	A	A	30
1	30	C	C	30
1	31	G	G	30
1	32	T	T	30
1	33	A	A	30
1	34	C	C	30
1	35	G	G	30
1	36	T	T	30
1	37	A	A	30
1	38	C	C	30
1	39	G	G	30
1	40	T	T	30
1	41	A	A	30
1	42	C	C	30
1	43	G	G	30
1	44	T	T	30
1	45	A	A	30
1	46	C	C	30
1	47	G	G	30
1	48	T	T	30
1	49	A	A	30
1	50	C	C	30
1	51	G	G	30
1	52	T	T	30
1	53	A	A	30
1	54	C	C	30
1	55	G	G	30
1	56	T	T	30
1	57	A	A	30
1	58	C	C	30
1	59	G	G	30
1	60	T	T	30
1	61	A	A	30
1	62	C	C	30
1	63	G	G	30
1	64	T	T	30
1	65	A	A	30
1	66	C	C	30
1	67	G	G	30
1	68	T	T	30
1	69	A	A	30
1	70	C	C	30
1	71	G	G	30
1	72	T	T	30
1	73	A	A	30
1	74	C	C	30
1	75	G	G	30
1	76	T	T	30
1	77	A	A	30
1	78	C	C	30
1	79	G	G	30
1	80	T	T	30
1	81	A	A	30
1	82	C	C	30
1	83	G	G	30
1	84	T	T	30
1	85	A	A	30
1	86	C	C	30
1	87	G	G	30
1	88	T	T	30
1	89	A	A	30
1	90	C	C	30
1	91	G	G	30
1	92	T	T	30
1	93	A	A	30
1	94	C	C	30
1	95	G	G	30
1	96	T	T	30
1	97	A	A	30
1	98	C	C	30
1	99	G	G	30
1	100	T	T	30

96% concordant SNP calls

## Diagnostic requirements

- LIMS
- SOPs and process controls
- Sample prep standardised reports
- Sample analysis log file reports
- QC reports of data quality
- Capture performance report
- No and low coverage report
- Indel report
- SNP report

# Sample analysis log file report



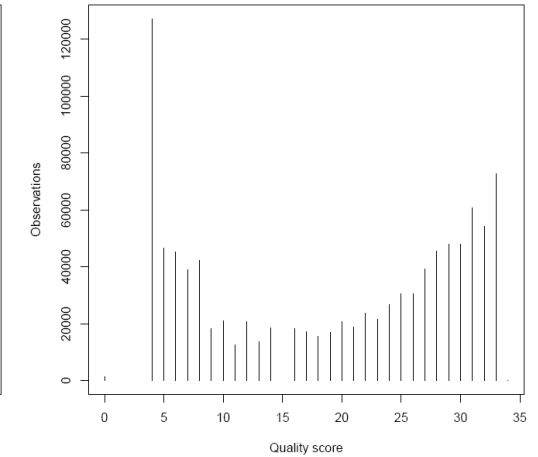
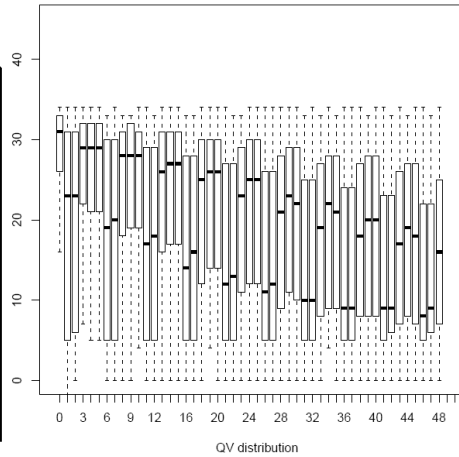
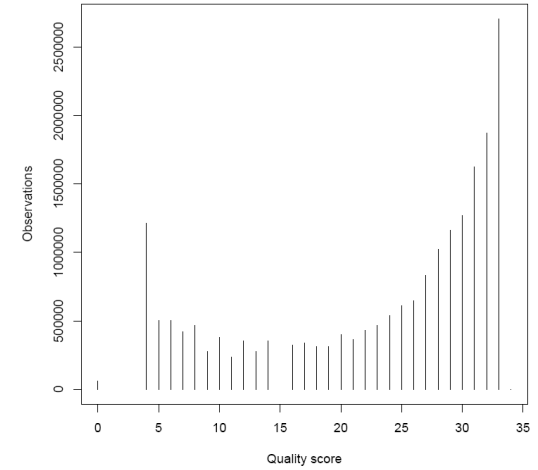
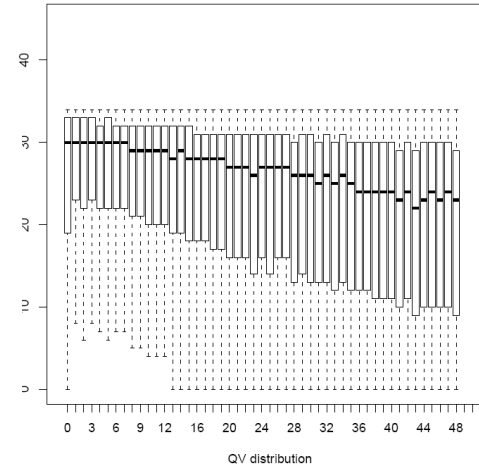
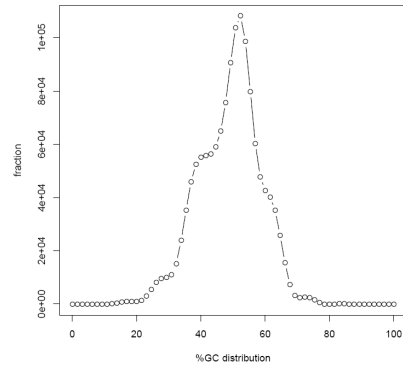
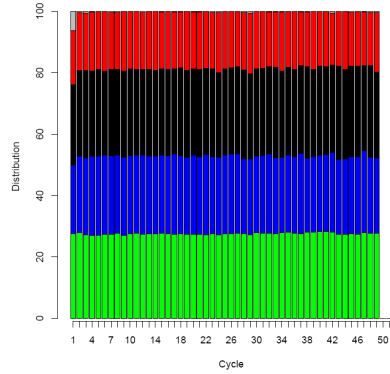
```
xx in total  
0 QC failure  
xx duplicates  
xx mapped (xx %)  
0 paired in sequencing  
0 read1  
0 read2  
0 properly paired (nan%)  
0 with itself and mate mapped  
0 singletons (nan%)  
0 with mate mapped to a different chr  
0 with mate mapped to a different chr (mapQ>=5)
```

version control enabled

# QC reports

## Summary

Input file
Sample size
Nr of reads
Bases $\geq$ Q30
Mappable prediction
Avg % AC



## Capture performance reports

base depth distribution	read #	%
0x depth		
1x depth		
2-10x depth		
10-20x depth		
20-30x depth		
>30x depth		
bait regions coverage	bait #	%
0 % of bait covered		
<70 % of bait covered		
70-80 % of bait covered		
80-90 % of bait covered		
90-99 % of bait covered		
100 % of bait covered		

Stats also generated for  
 BWA  
 Bowtie  
 BFAST  
 Bioscope

Agilent SureSelect

>30x depth                    xx%

100% bait coverage            xx%

NimbleGen Custom Solid phase (prelim)

>30x depth                    xx%

100% bait coverage            xx%

## No and Low Coverage reports

position	depth	gene	transcript	region	codon pos	protein pos
----------	-------	------	------------	--------	-----------	-------------

## indel reports

position	base(s)	evidence/depth	gene	transcript	region	codon pos	protein pos
----------	---------	----------------	------	------------	--------	-----------	-------------

## SNP reports

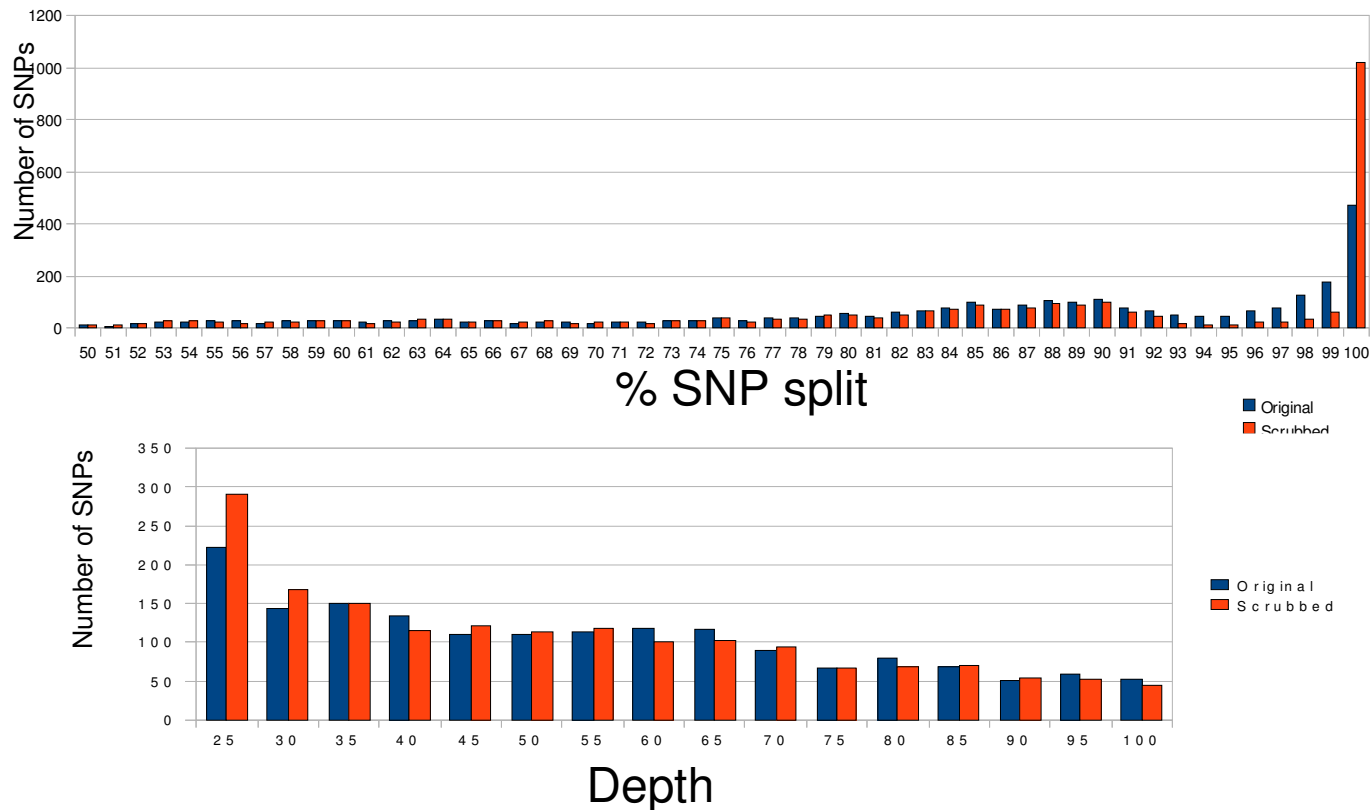
position	change	filter	score	depth	type	gene	transcript
----------	--------	--------	-------	-------	------	------	------------

region	codon pos	AA change	Grantham score	external ref	dbsnp freq	dbsnp flags	phylop	phast	pfam
--------	-----------	-----------	----------------	--------------	------------	-------------	--------	-------	------

And links to internal data bases [recurrence *etc*]

## Data scrubbing for quality improvement

- Observed a number of SNP calls with <100% base splits (X exome) or non 50:50 ratios in autosomes
- Splits are due to colour space correction to the reference
- Devised a per SNP removal of corrected bases and re assessment of SNP
- Significant improvement in SNP calling stringency and confidence



## **XLMR project future**

- NimbleGen EZ enrichments and cross comparisons
- Final enrichment platform choice and design improvement
- Evaluate SOLiD 4 chemistry on 5500xl
- Continued multiplexing with barcoded targeted enrichment
- Continued software evaluation and comparison
- XLMR diagnostic service imminent

## Thanks to

*Chris Clee*

Anthony Rogers

Dominique McCormick

Kim Brugger

Ilenia Simeoni

Owen McCann

Louise Godfrey

Jo Whittaker

John Todd

Lucy Raymond

Annabel Whibley

Patrick Tarpey