Cambridge University Hospitals NHS
NHS Foundation Trust

Developing targeted enrichment and re-sequencing on the SOLiD platform

Howard Martin

Cambridge Regional Molecular Genetics Laboratory
Cambridge University Hospital NHS Foundation Trust

Eastern Sequence and Informatics Hub (EASIH)

3 initial MRC funded hubs *May 2009* (Edinburgh, Liverpool, Cambridge)

Oxford included *June 2009*

Cambridge hub has a strategic aim to apply NGS to routine medical diagnostic use

SOLiD 3 50Gb

SOLiD 4 100GB

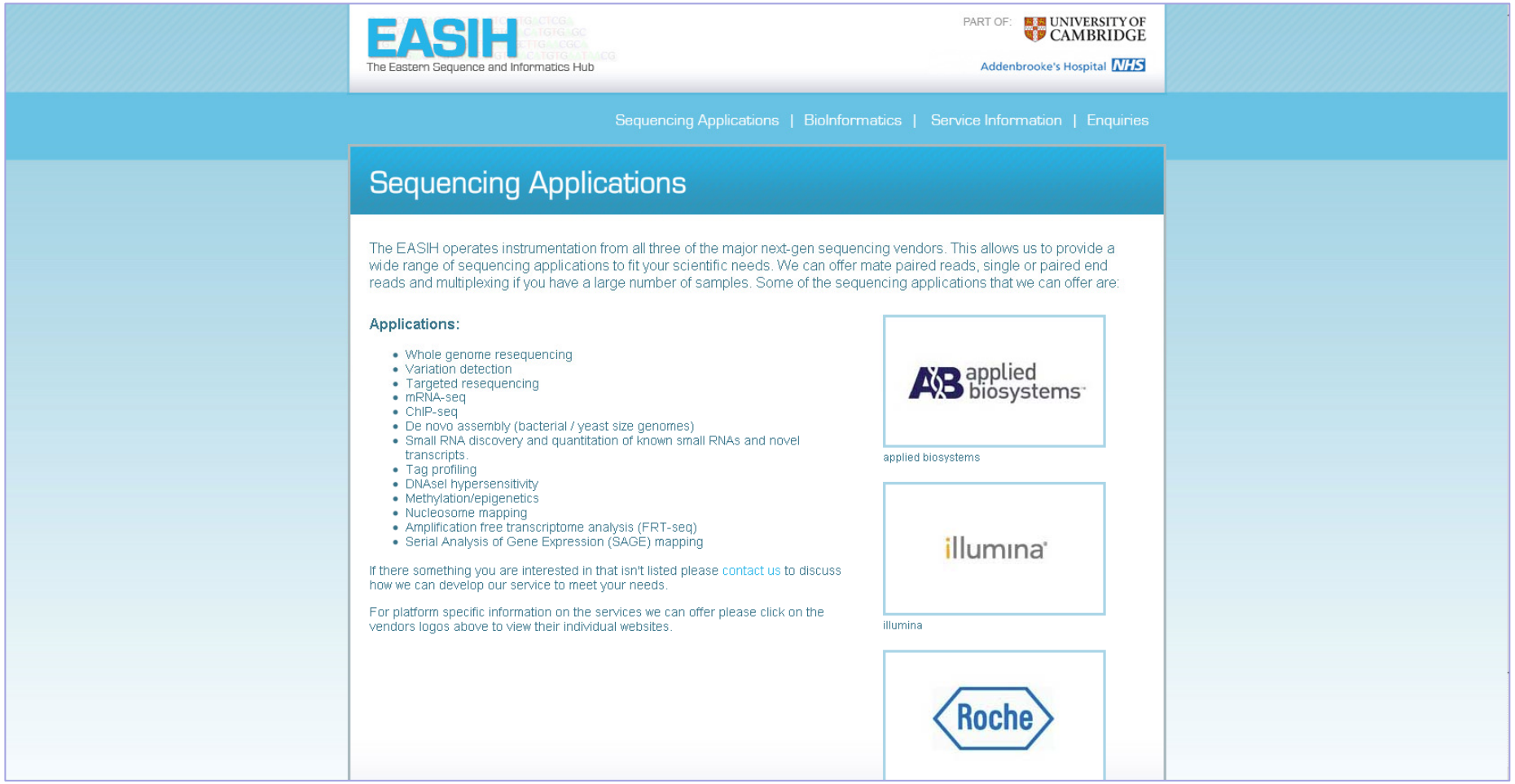
SOLiD 4 100Gb

GAIIx 50Gb

GS FLX Titanium 500Mb

www.easih.ac.uk

MRC Medical Research Council | UNIVERSITY OF CAMBRIDGE | NHS National Institute for Health Research | wellcome trust | life technologies | JDRF Juvenile Diabetes Research Foundation International — dedicated to finding a cure



Sequencing service provision

Offers a broad range of sequencing applications on all 3 main platforms

Sequence service
&
Bioinformatics support

Bioinformatics support

[sequence independent]

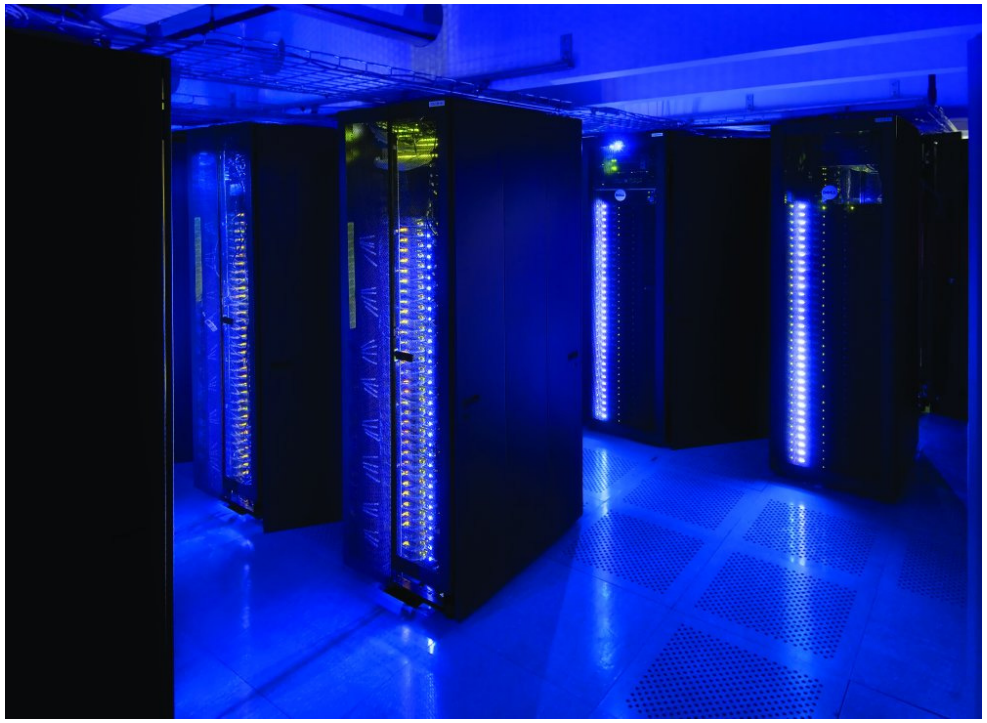
Diagnostic sequencing

ISO International Organization for Standardization

Cambridge University Hospitals NHS
NHS Foundation Trust

UNIVERSITY OF CAMBRIDGE 800 YEARS 1209~2009

High Performance Computing Service



The Darwin cluster

2340 x 3.0 GHz Intel Woodcrest cores, 4.6 TB of total memory

EMBL-EBI

European Bioinformatics Institute collaboration

Project aim ; develop a diagnostic service for X linked learning disability

- Investigation of a child with learning disability is on 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 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

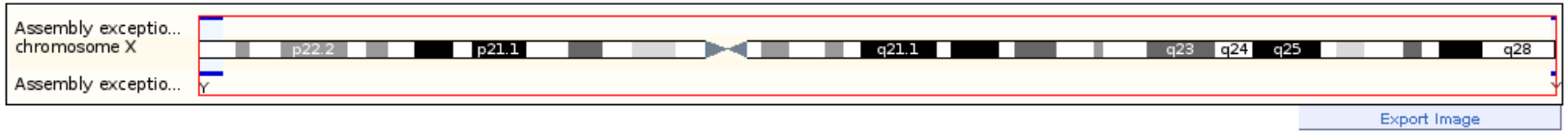
- Targeted enrichment and re-sequencing
- 3 enrichment platform approaches
  - Agilent SureSelect
  - *Febit HybSelect*
  - NimbleGen EZ
- *4* x 10 enrichment libraries generated
- SOLiD sequencing platform

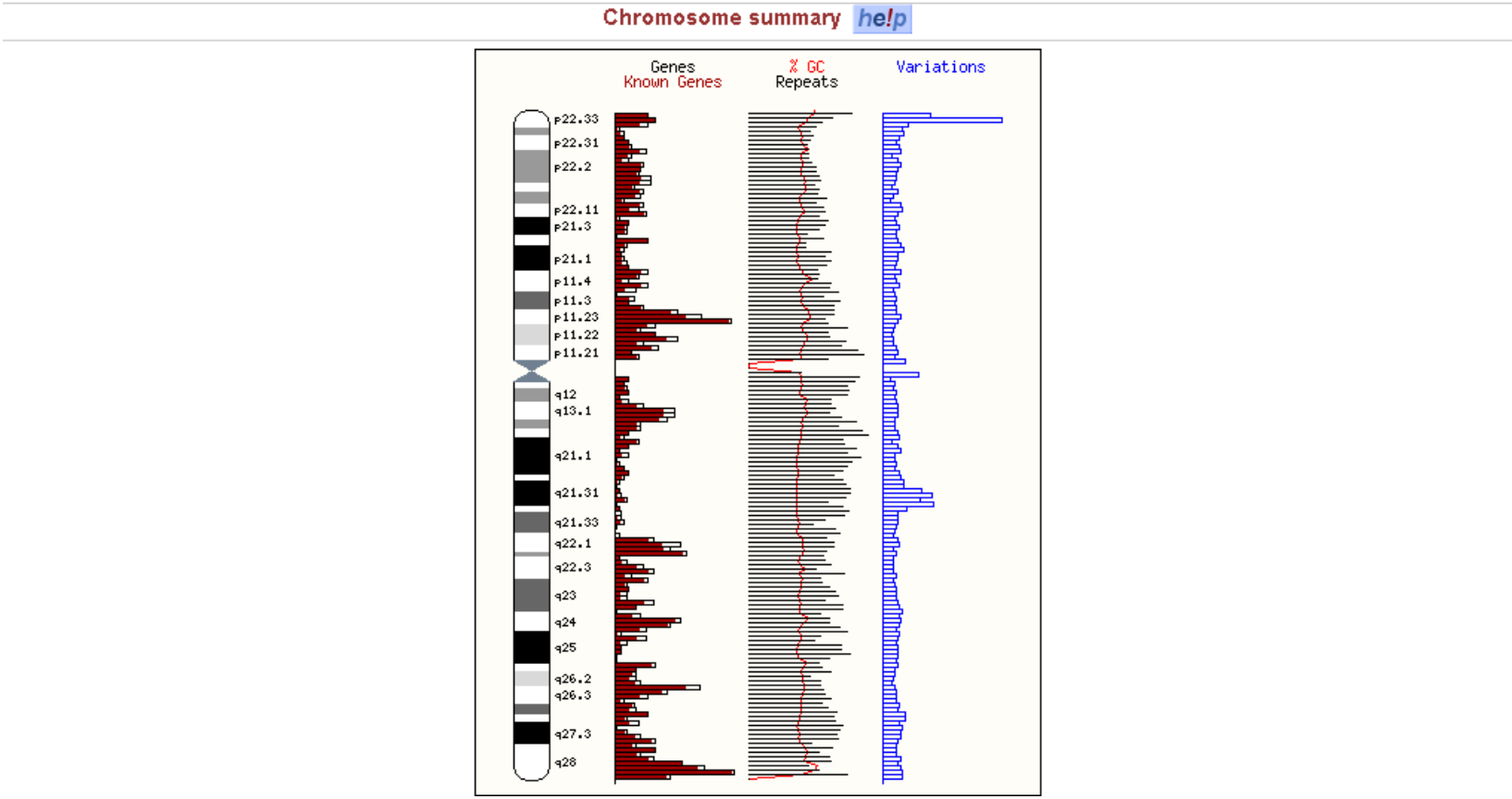
Trial each enrichment method

Compare

Hands on performance	•Ease of us •Suitability for automation
Customisation	•Design process including augmentation •Capture capacity of the designs
Performance	•Evenness of enrichment •Reproducibility of capture •Depth of coverage comparisons •SNP calling
Scalability	•Multiplexing capabilities

X chromosome





861 known protein coding genes

155,270,560 bp [GRCh37]

3 Mb capture region

•All platforms given the same design specifications [100 genes for Febit]

•Review of preliminary designs

•ID regions of poor probe placement (cause?)

•Augment designs to force additional probes

•Sign-off design, perform enrichments, sequence libraries (SOLiD 3 and 4)

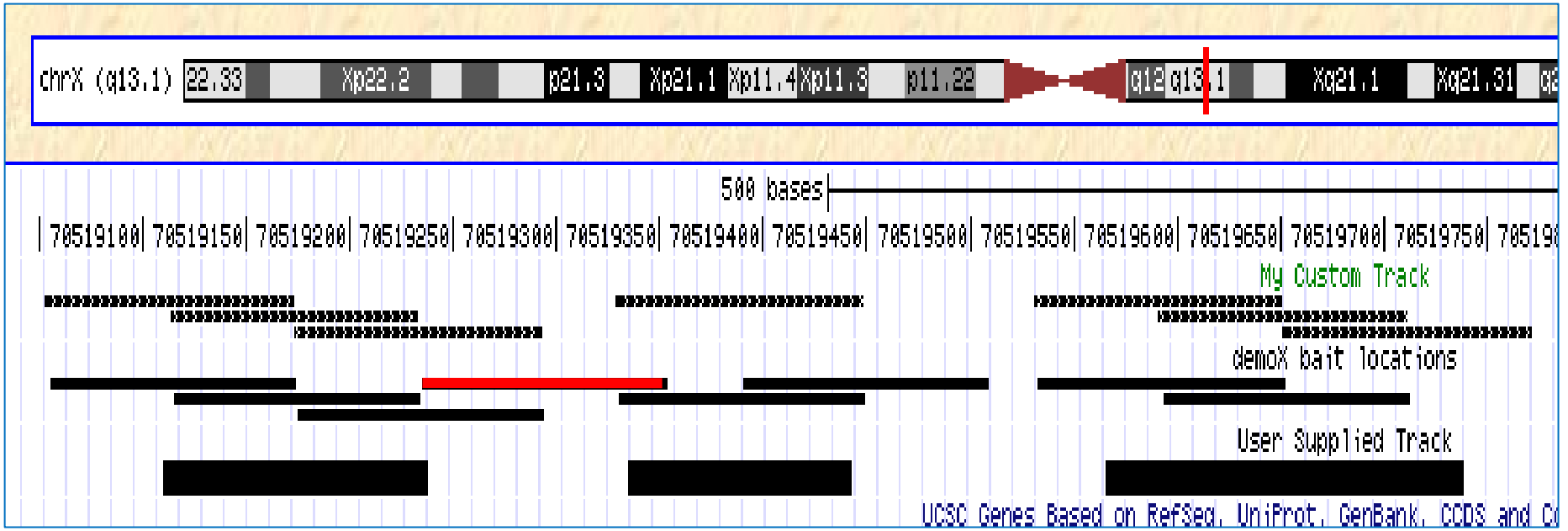
•Analysis of results to ID ‘most suitable’ platform

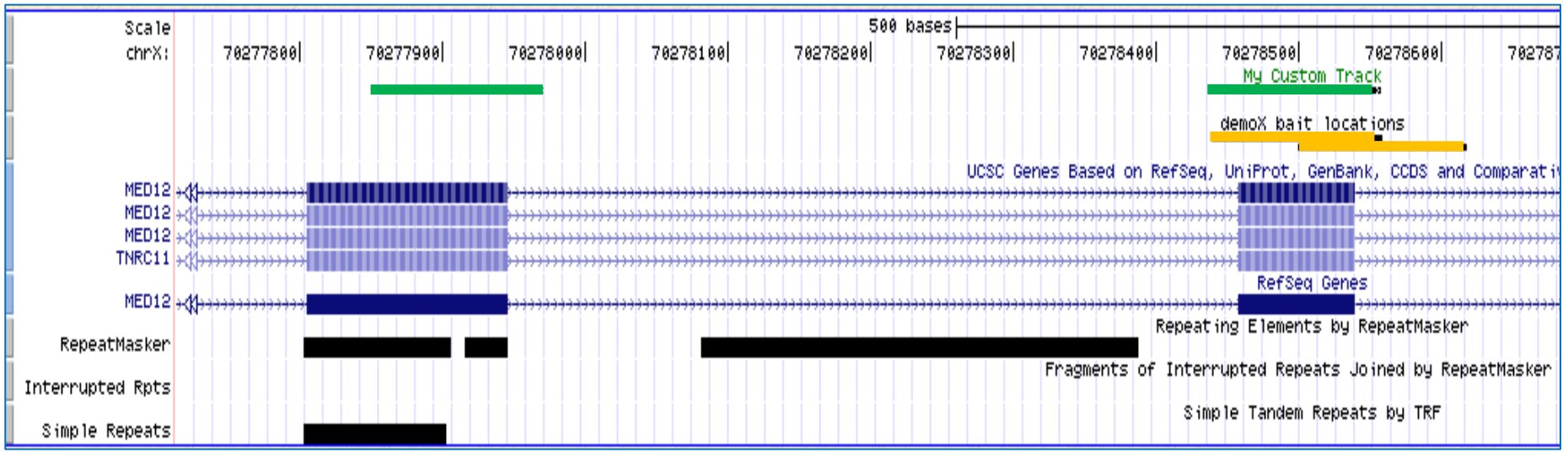
•Re-designs to enhance regions of poor capture

Enrichment libraries

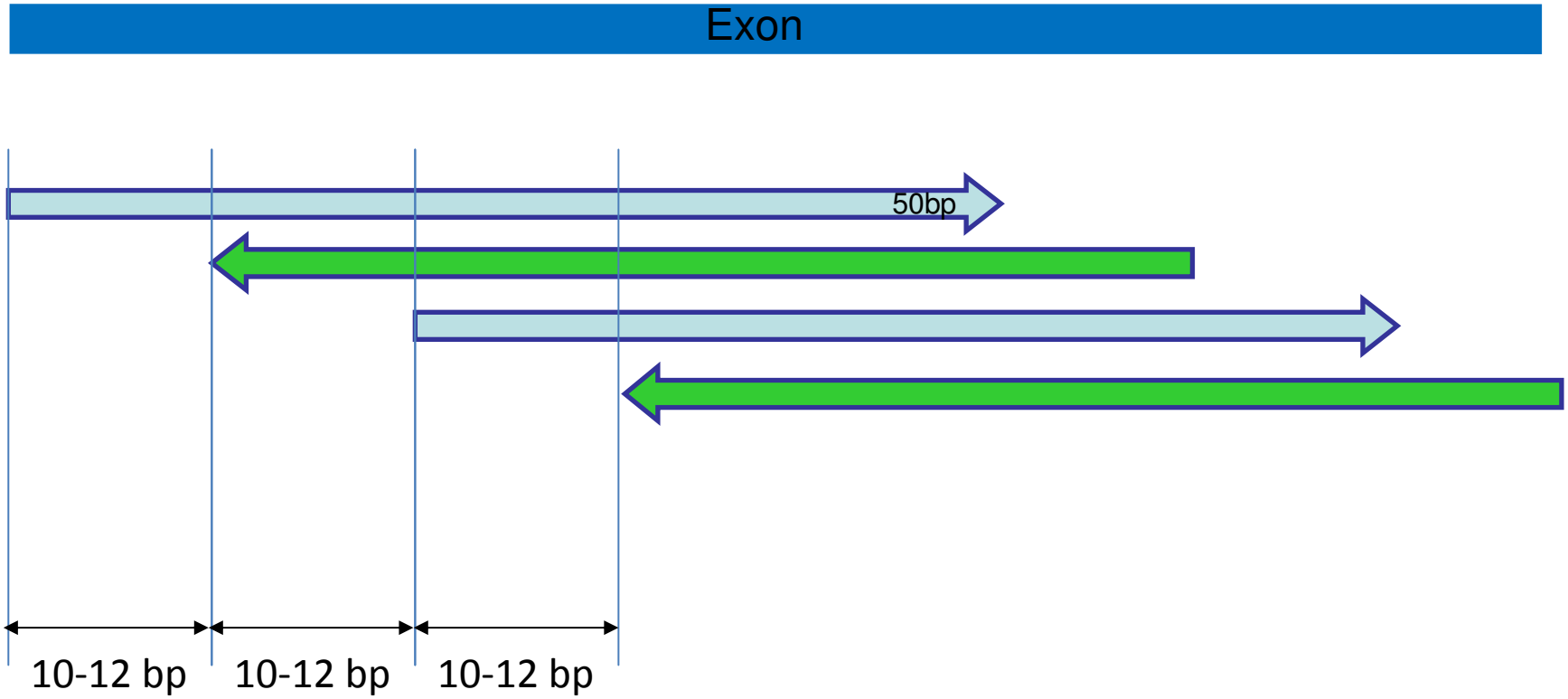
Agilent SureSelect X Demo	43,074 baits	Liquid phase	Barcoded version
Agilent SureSelect X custom	37,703 baits	Liquid phase	
Febit custom 100 XLMR genes	58,603 baits	Solid phase	
NimbleGen EZ custom	In progress	Solid phase	Liquid phase

Agi X_ex_demo vs Agi X_ex_custom

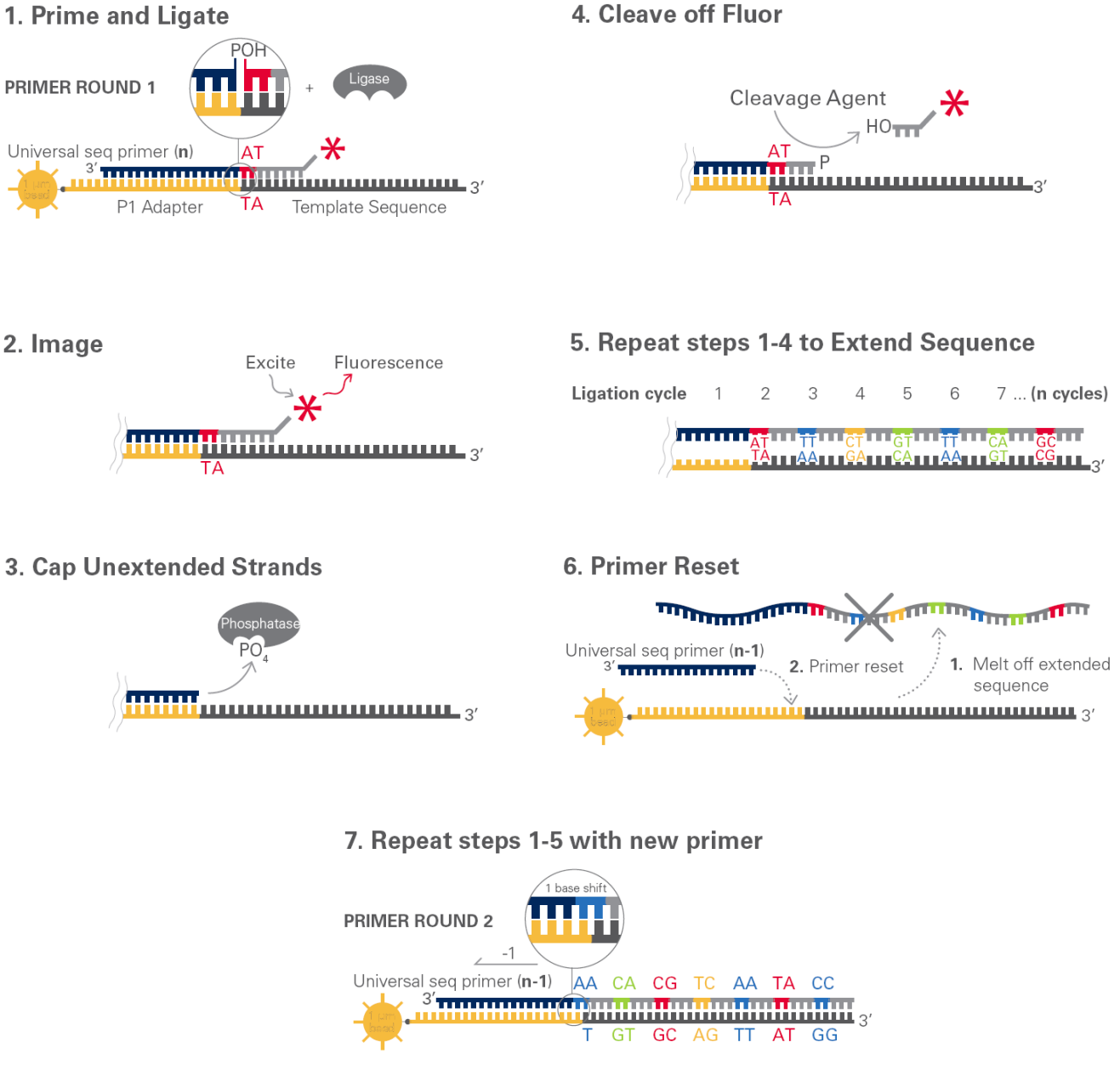


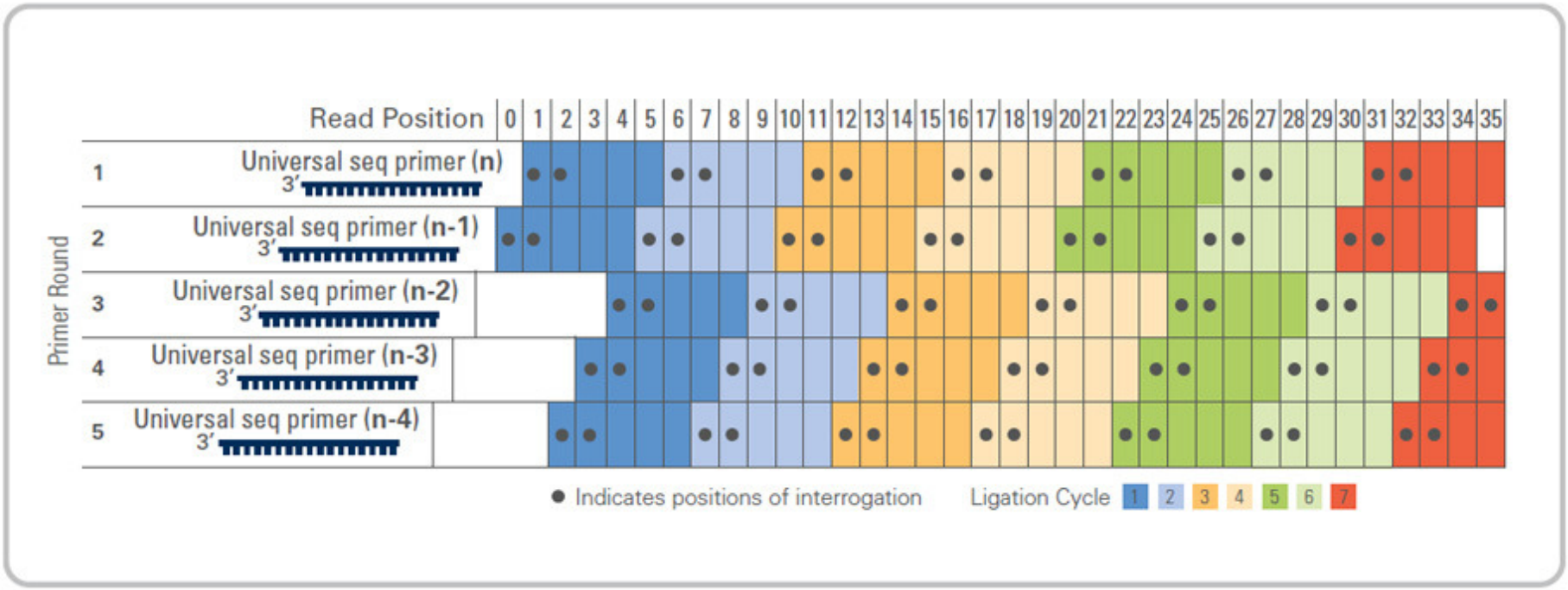


Febit design approach



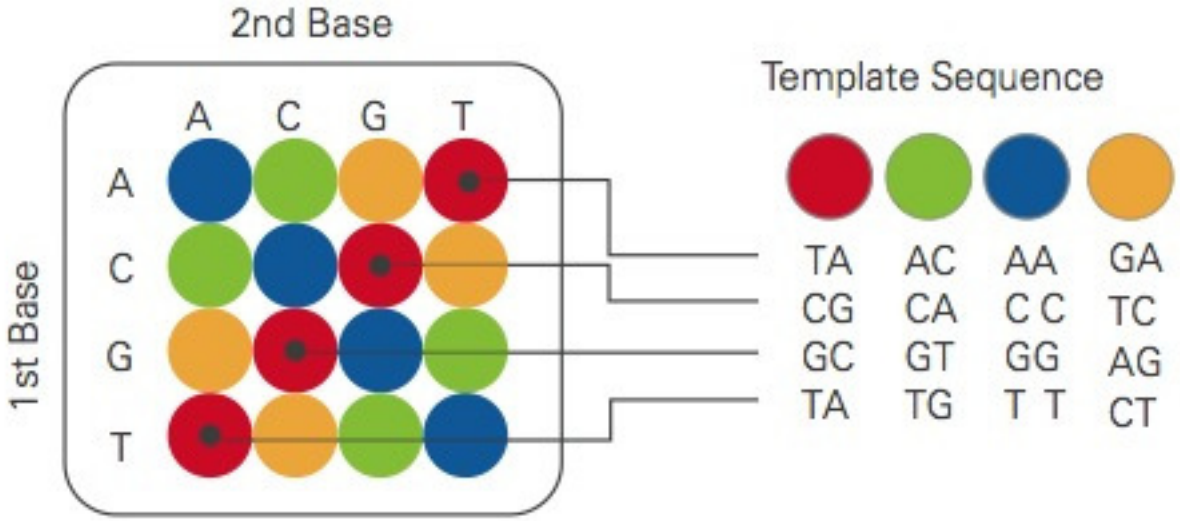
SOLiD chemistry




Read Position
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
Primer Round
1 Universal seq primer (n) 3′
2 Universal seq primer (n-1) 3′
3 Universal seq primer (n-2) 3′
4 Universal seq primer (n-3) 3′
5 Universal seq primer (n-4) 3′
● Indicates positions of interrogation
Ligation Cycle 1 2 3 4 5 6 7

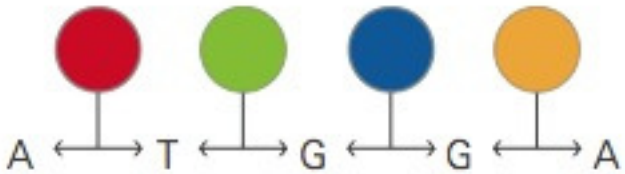
SOLiD™

Possible Dinucleotides Encoded By Each Color



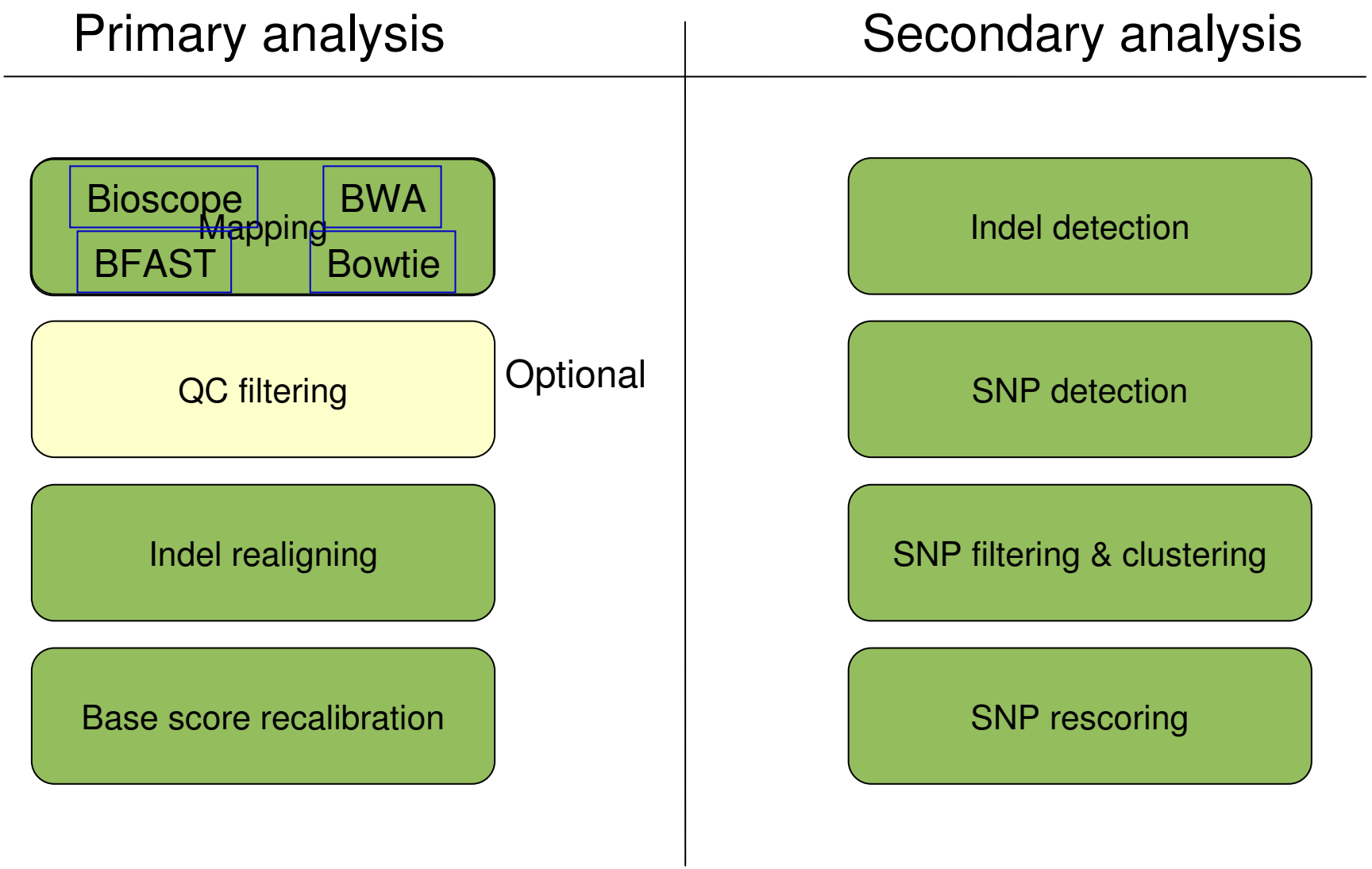
Double Interrogation

With 2 base encoding each base is defined twice

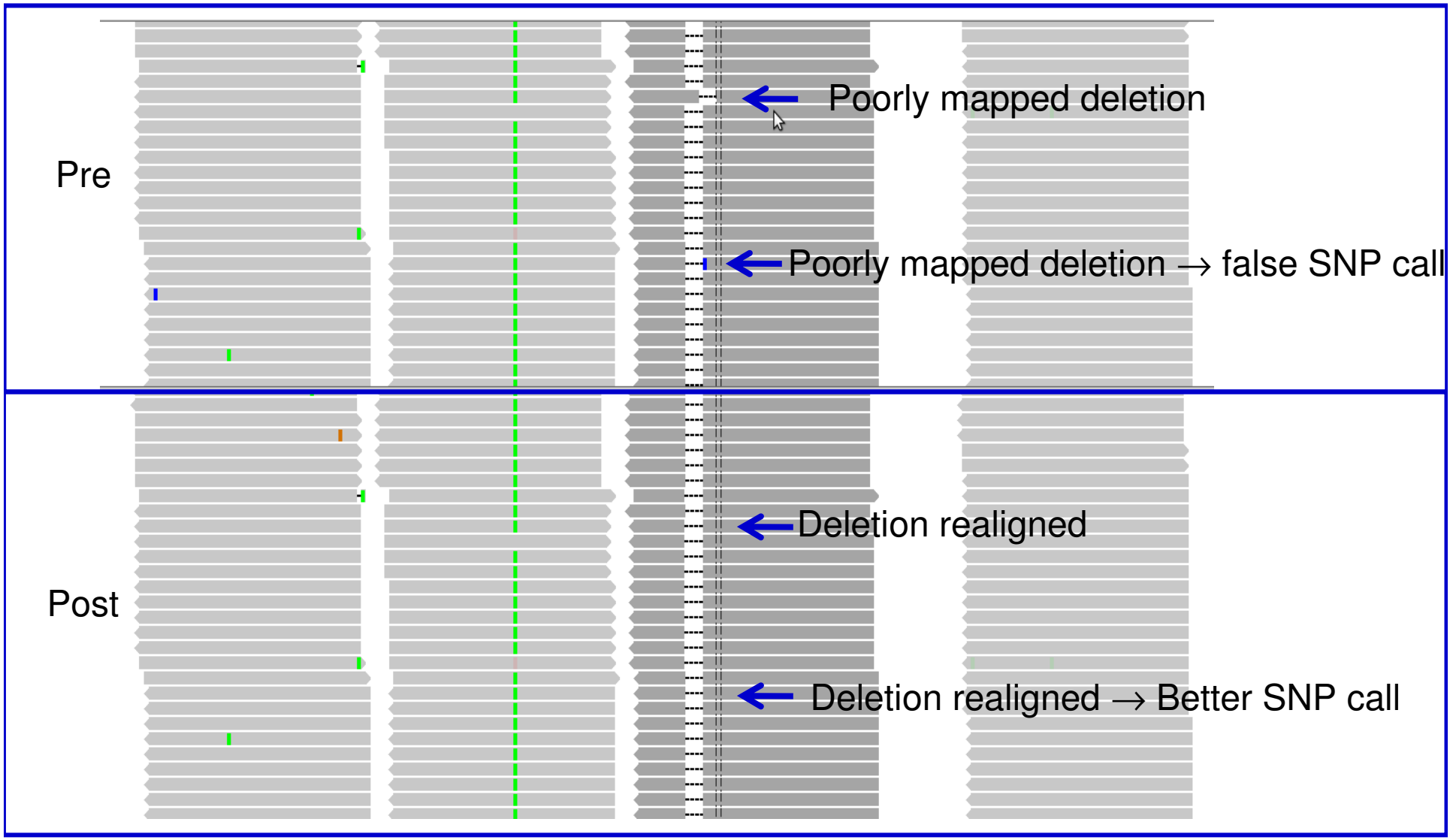


SOLiD™

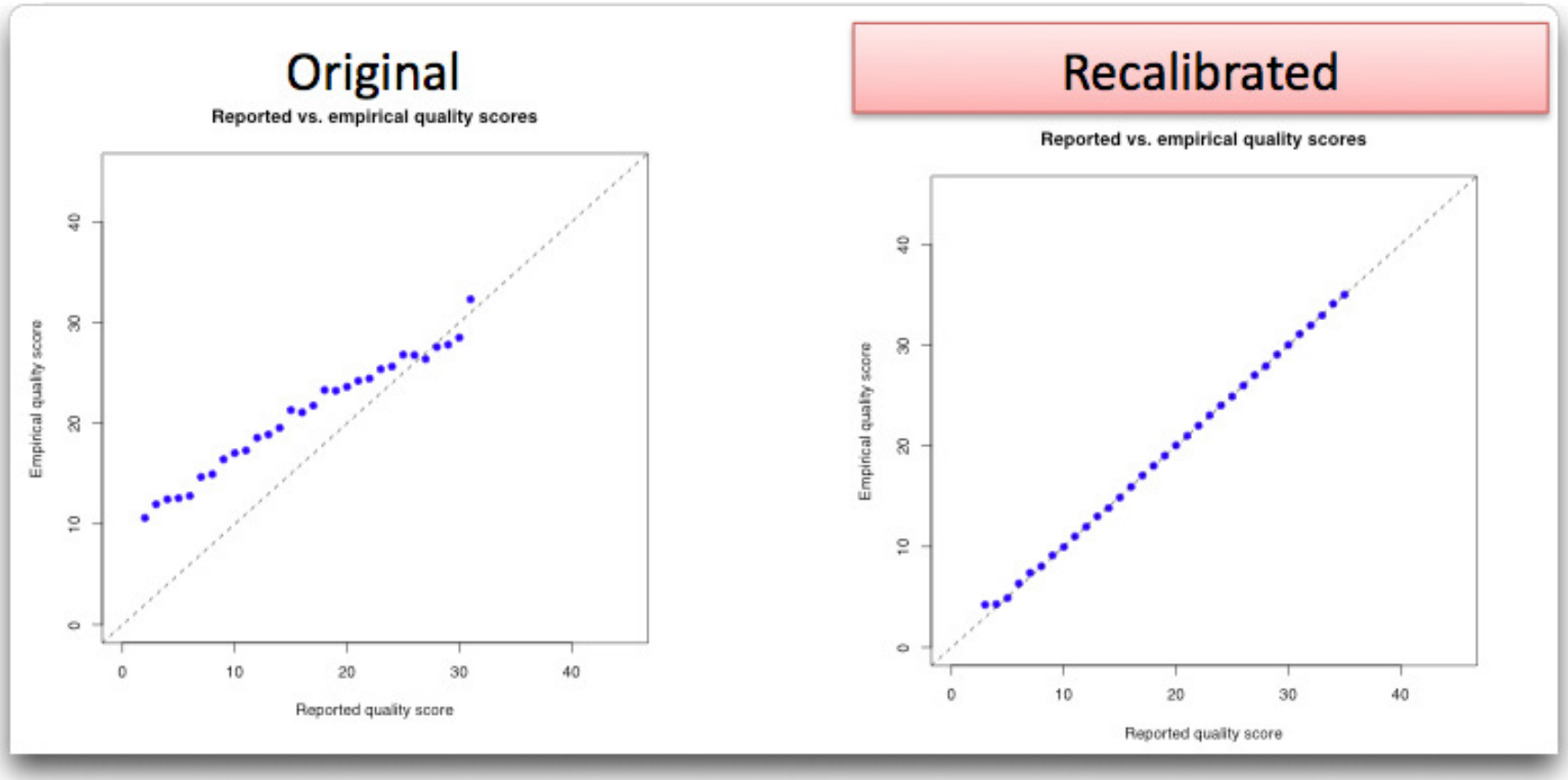
Analysis pipeline



Indel realigning



Base score recalibration [as for 1000 genomes project] for improved SNP calling etc



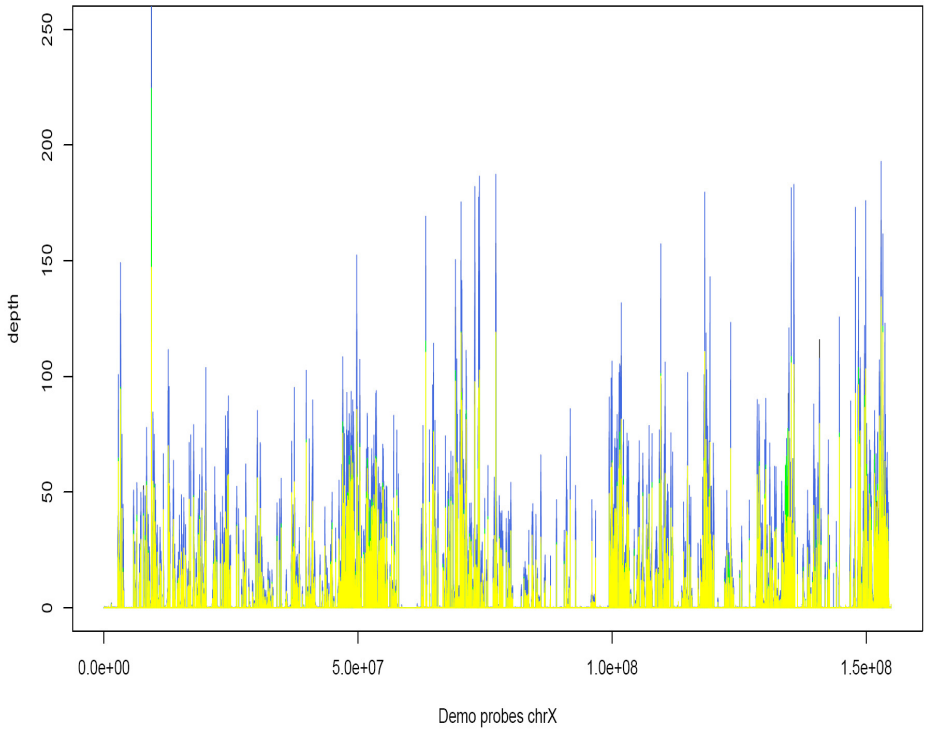
Empirical

reported

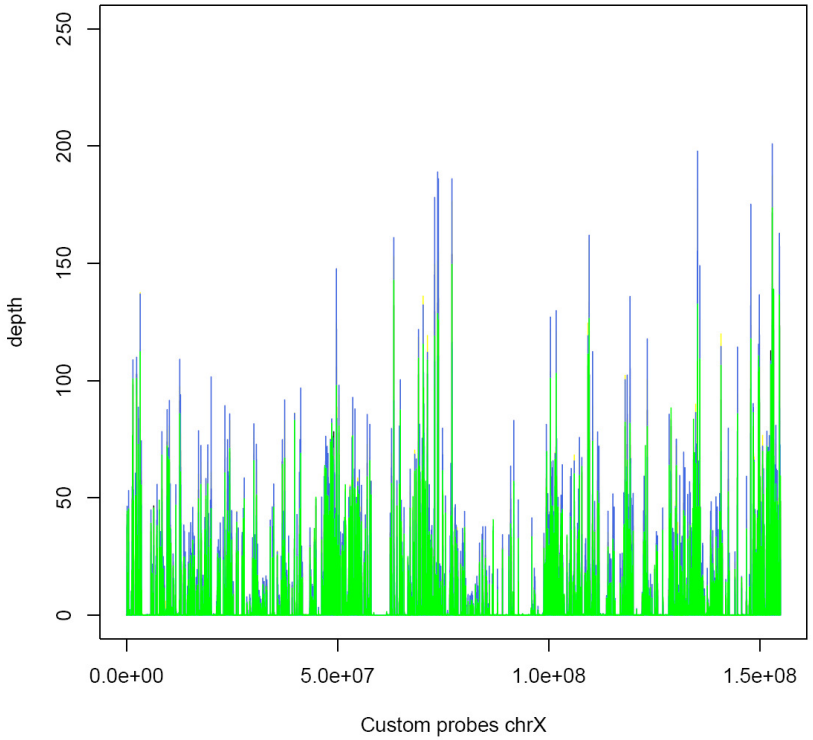
SOLiD under estimates QV for bases, Illumina over estimates

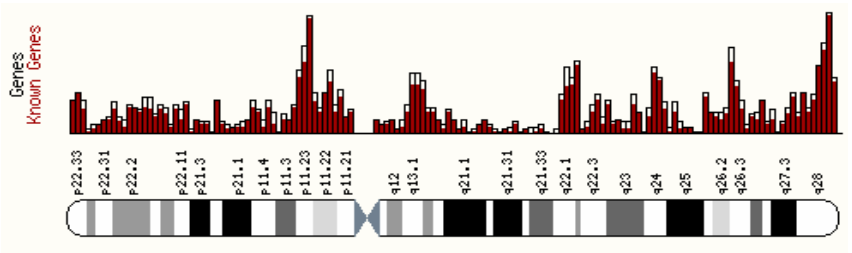
BROAD INSTITUTE

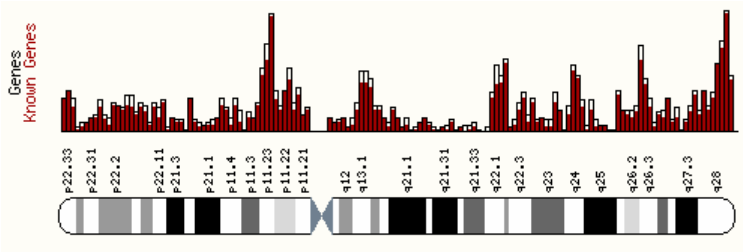
Agilent X exome demo



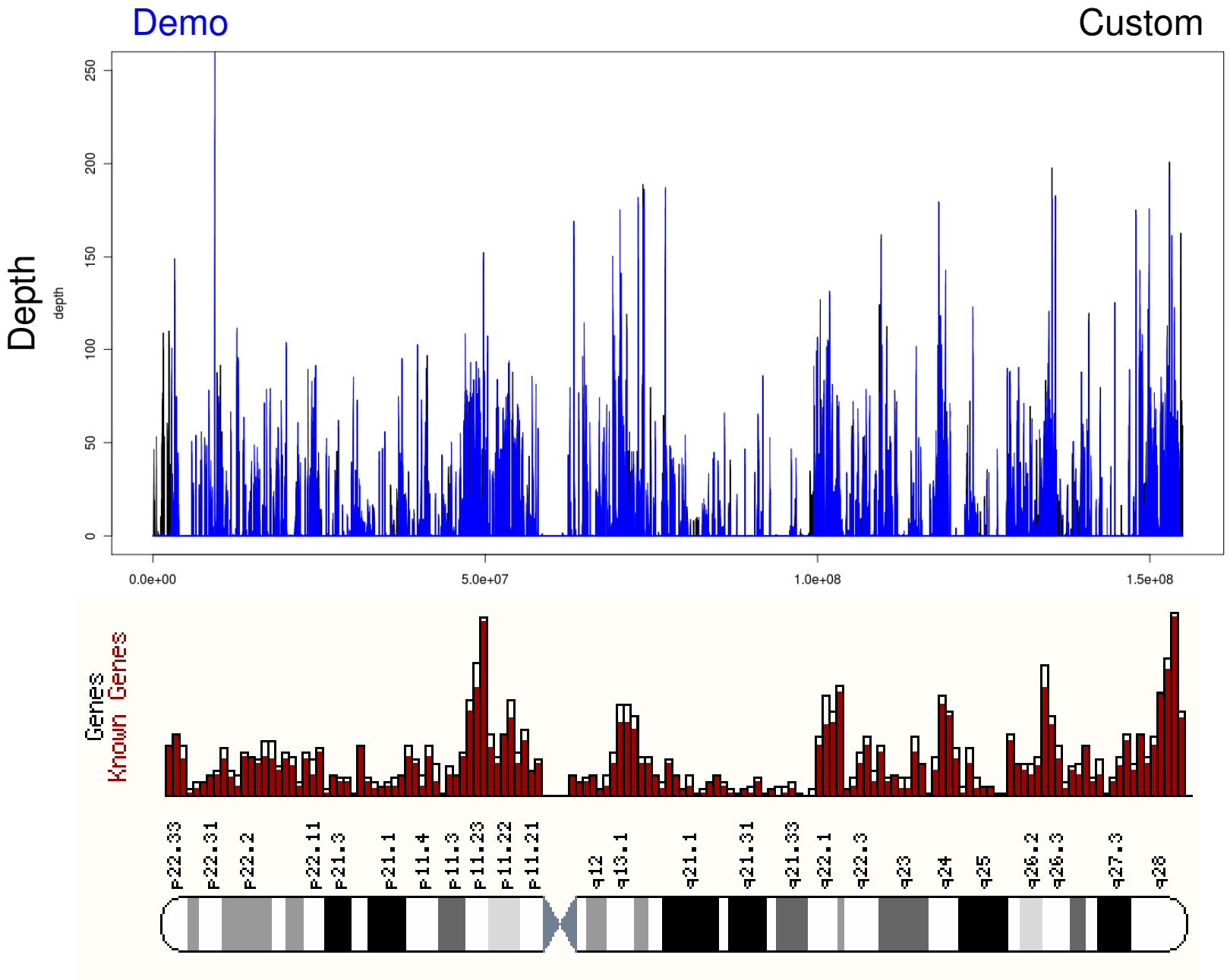
Agilent X exome custom





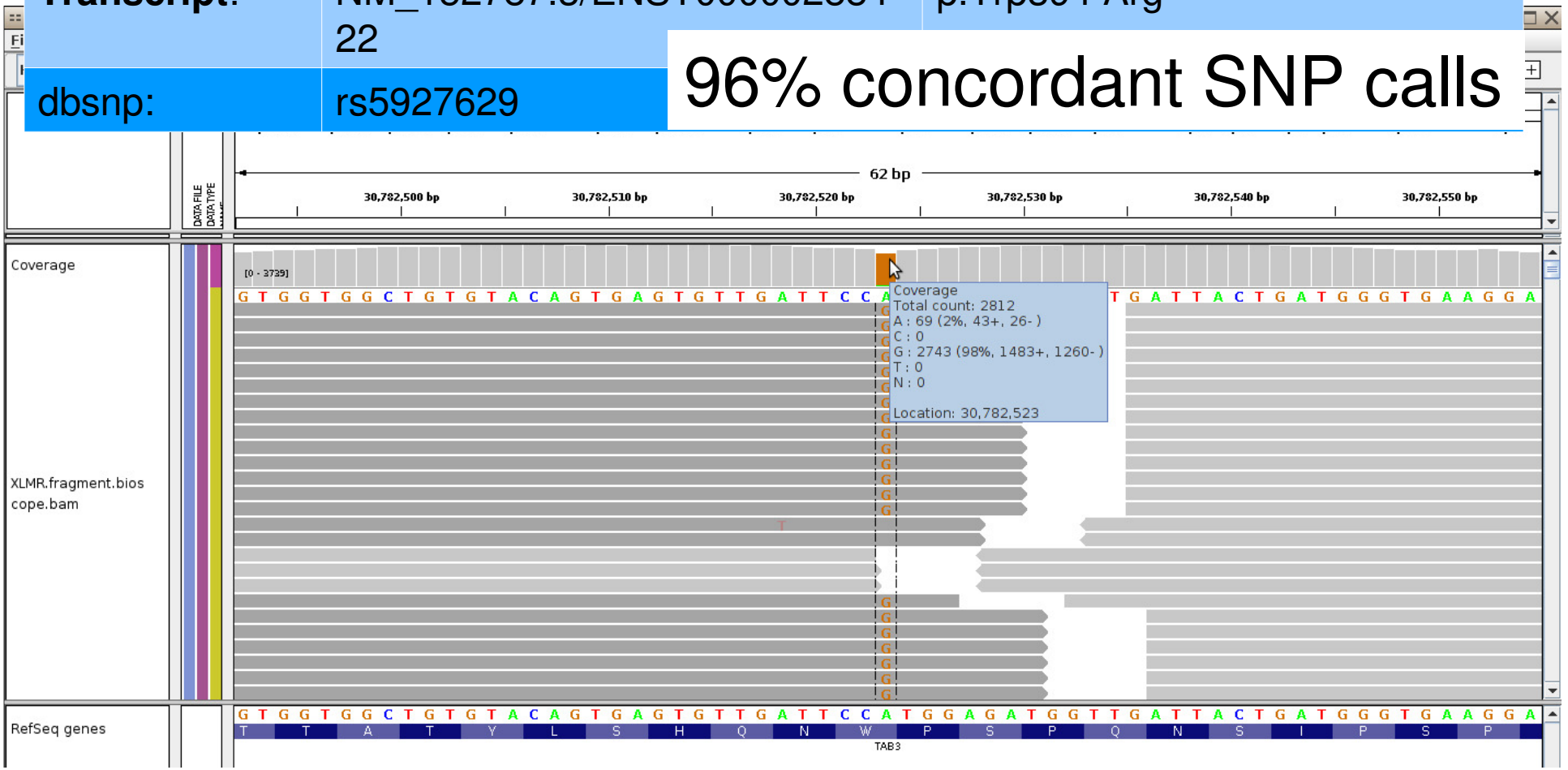


Reproducible enrichment spread

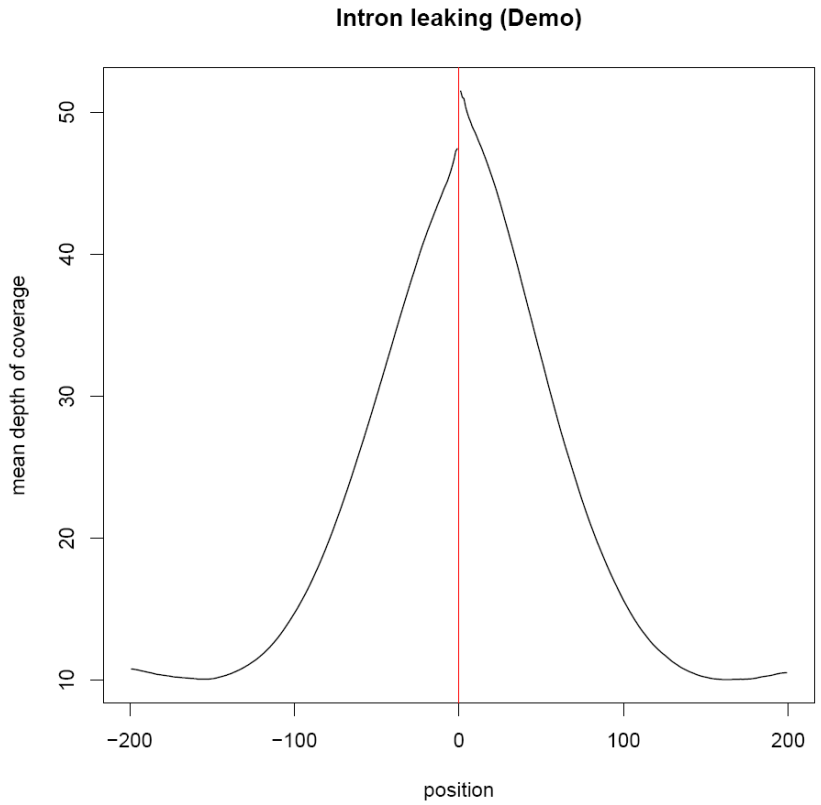

Demo
Custom
Depth
depth
0
50
100
150
200
250
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Genes
Known Genes
p22.33
p22.31
p22.2
p22.11
p21.3
p21.1
p11.4
p11.3
p11.23
p11.22
p11.21
q12
q13.1
q21.1
q21.31
q21.33
q22.1
q22.3
q23
q24
q25
q26.2
q26.3
q27.3
q28

chrX:30782523	A>G	score:104837	depth:2812	
Stats:	A: 69(2.45%)/6.57	C: 0(0.00%)/0	G: 2743(97.55%)/38.22	T: 0(0.00%)/0
Type:	NON_SYNONYMOUS_CODING			
Gene:	MAP3K7IP3/ENSG00000157625		c.1727	
Transcript:	NM_152787.3/ENST00000288422		p.Trp394 Arg	
dbsnp:	rs5927629			

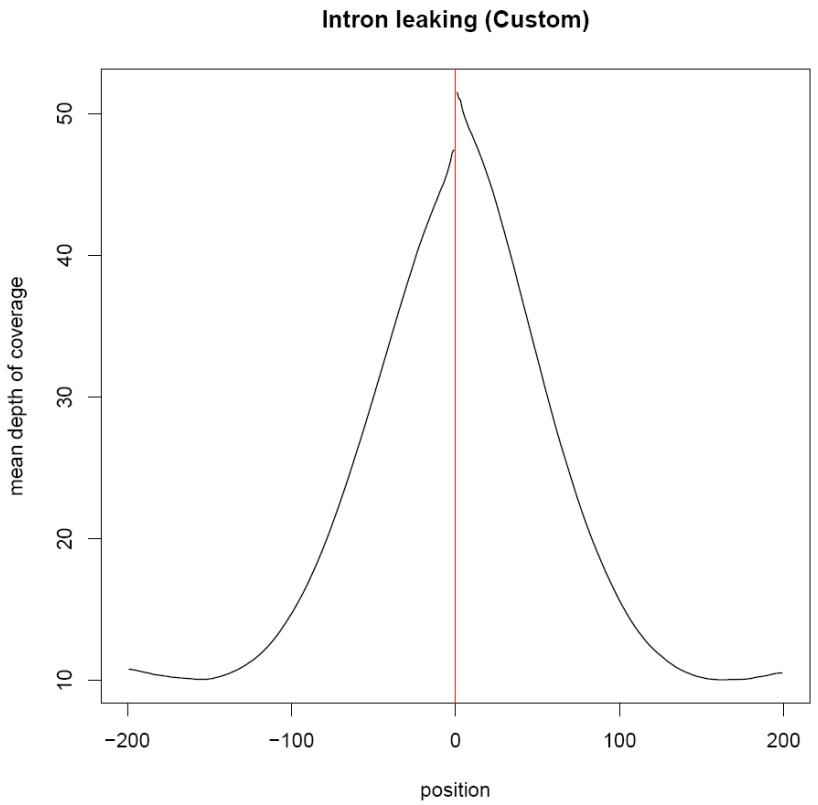
96% concordant SNP calls



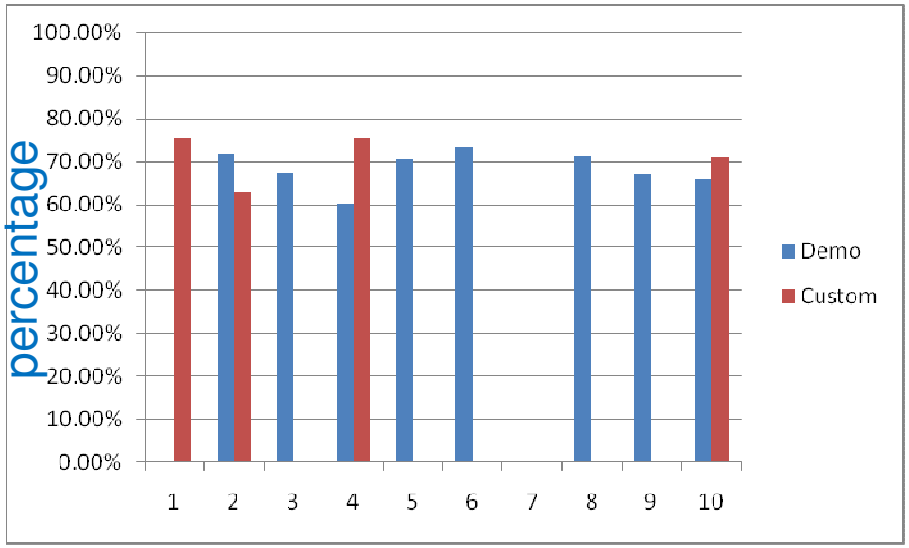
Depth of coverage at exonic boundaries



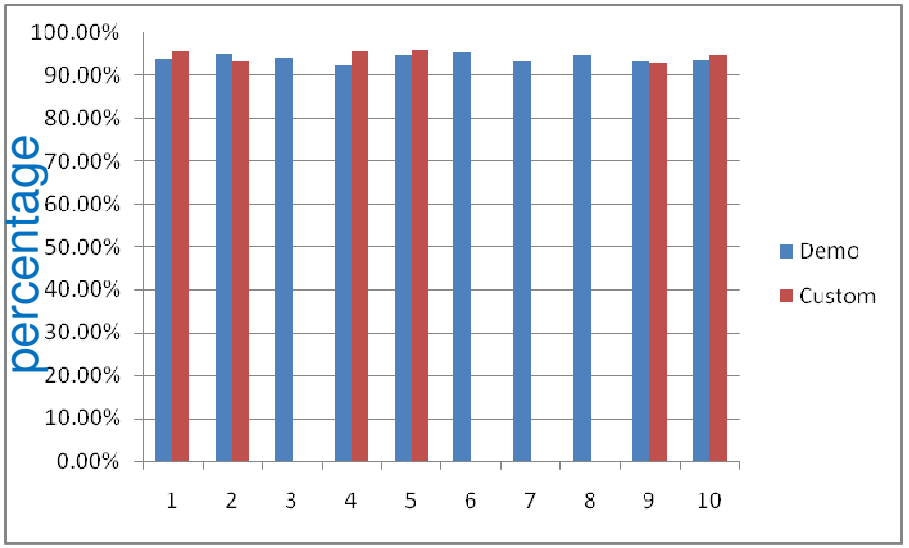
Agi_X_Ex_Demo



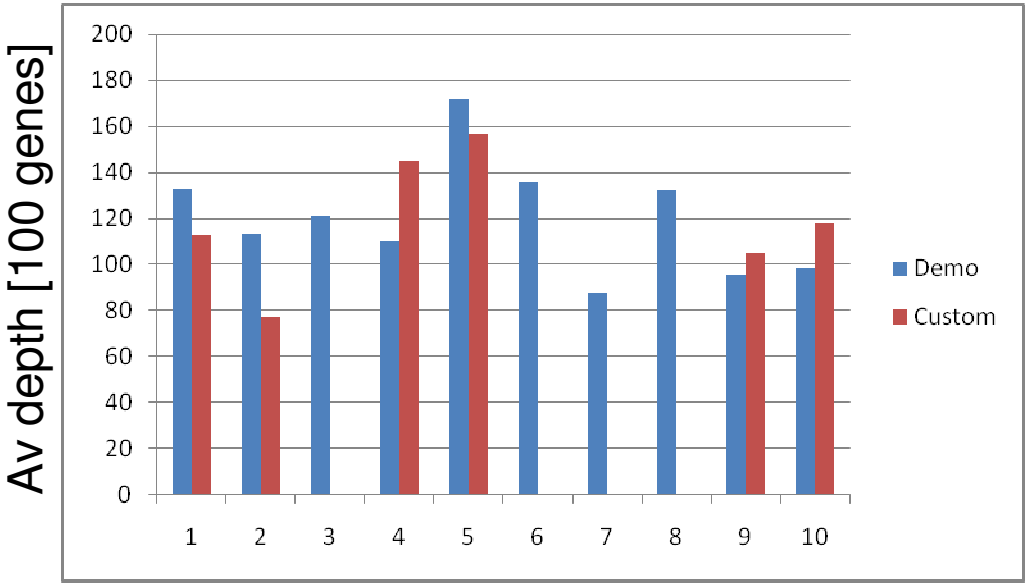
Agi_X_Ex_custom



Genome mapped reads on probe target
100 genes



X mapped reads on probe target
100 genes



Future designs

Sequencing strategy

Degree of multiplexing

Future

NimbleGen enrichments and cross comparisons

Full SNP calling comparisons

Enrichment platform choice and design improvement

Continued software evaluation and comparison

Evaluate SOLiD 4 chemistry [vs SOLiD 3] plus paired end sequence

Investigate barcoded targeted enrichment multiplexing

Thanks to

Chris Clee

Dominique McCormick

Ilena Simeoni

Jo Whittaker

John Todd

Anthony Rogers

Kim Brugger

Lucy Raymond

Annabel Whibley

Patrick Tarpey