

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)

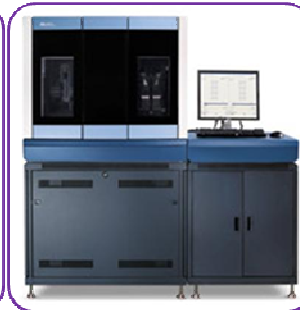
3 initial MRC funded hubs *May 2009* (Edinburgh, Liverpool, Cambridge)

Oxford included *June 2009*

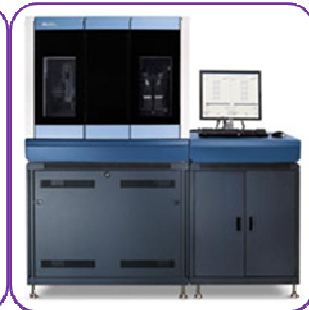
Cambridge hub has a strategic aim to apply NGS to routine medical diagnostic use



SOLiD 3
50Gb



SOLiD 4
100GB



SOLiD 4
100Gb



GAIx
50Gb



GS FLX Titanium
500Mb

EASIH Part of Cambridge University

For fast and cost-effective
production and analysis of large
scale sequence-based datasets



Subsequent to the completion of the Human Genome Project advances in genomic sequencing have led to a dramatic fall in time taken and cost to produce very large scale, high throughput sequence-based (HTS) datasets. Cambridge University and the cluster of closely-connected research Institutes in the region have a long and productive history in both the development and application of DNA sequencing technologies.

The creation of the Eastern Sequence and Informatics Hub (EASIH) based on major funding from the MRC, University of Cambridge and the NIHR Cambridge Biomedical Research Centre, will enable researchers working in our region to utilise a crucial 21st-century research approach in their studies. In collaboration with the nearby European Bioinformatics Institute, an international centre for sequence data analysis, EASIH will also provide researchers with access to the tools and protocols required to analyse these complicated and large-scale datasets, and in the design and implementation of projects.

EASIH, based at the Cambridge University Addenbrooke's Hospital, has a research development and strategic aim to apply HTS to routine medical diagnostic uses, in particular in HLA typing in transplantation and cord blood stem cells, and re-sequencing of disease genes in collaboration with NHS Blood and Transplant and NHS Regional Clinical Genetics Services.

EASIH
University of Cambridge

MAKE AN ENQUIRY

To get in touch or
make an enquiry
about your project
please use our fast
and easy form.



[ENQUIRY FORM](#)

TECHNOLOGIES USED:

Genome
Analyzer IIe



SOLID™ 4
System



FLX
System



Sequencing Applications

The EASIH operates instrumentation from all three of the major next-gen sequencing vendors. This allows us to provide a wide range of sequencing applications to fit your scientific needs. We can offer mate paired reads, single or paired end reads and multiplexing if you have a large number of samples. Some of the sequencing applications that we can offer are:

Applications:

- Whole genome resequencing
- Variation detection
- Targeted resequencing
- mRNA-seq
- ChIP-seq
- De novo assembly (bacterial / yeast size genomes)
- Small RNA discovery and quantitation of known small RNAs and novel transcripts.
- Tag profiling
- DNaseI hypersensitivity
- Methylation/epigenetics
- Nucleosome mapping
- Amplification free transcriptome analysis (FRT-seq)
- Serial Analysis of Gene Expression (SAGE) mapping

If there something you are interested in that isn't listed please [contact us](#) to discuss how we can develop our service to meet your needs.

For platform specific information on the services we can offer please click on the vendors logos above to view their individual websites.



applied biosystems



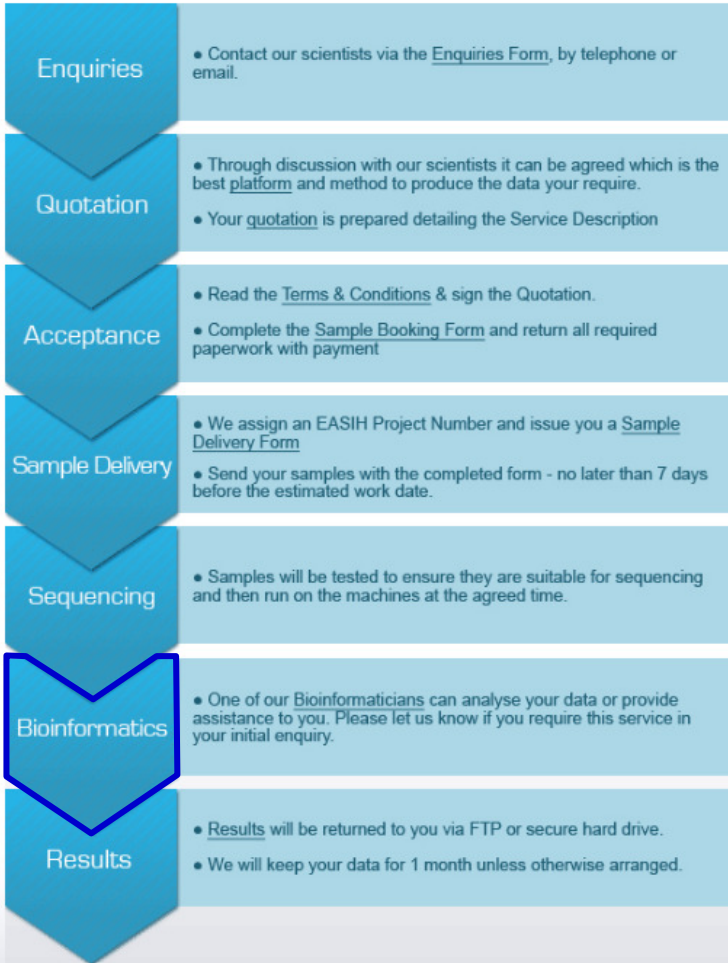
illumina



Sequencing service provision

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

Service Information



Service Information Documents

- [Enquiries Form](#)
- [Blank EASIH Quote](#)
- [Sample Booking Form](#)
- [Sample Delivery Form](#)
- [EASIH Terms and Conditions](#)
- [Sequencing Services Information Sheet](#)

Sequence service
&
Bioinformatics support

Bioinformatics support
[sequence independent]

Diagnostic sequencing



Cambridge University Hospitals **NHS**
NHS Foundation Trust

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 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 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

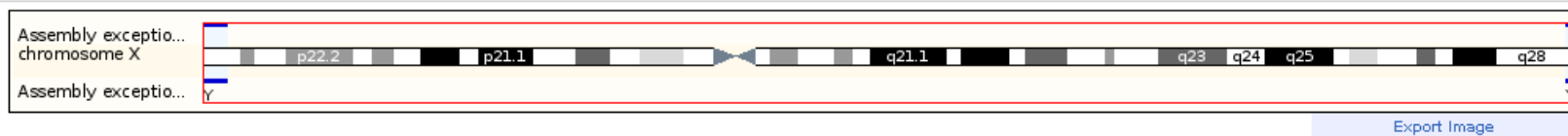
Performance

- Evenness of enrichment
- Reproducibility of capture
- Depth of coverage comparisons
- SNP calling

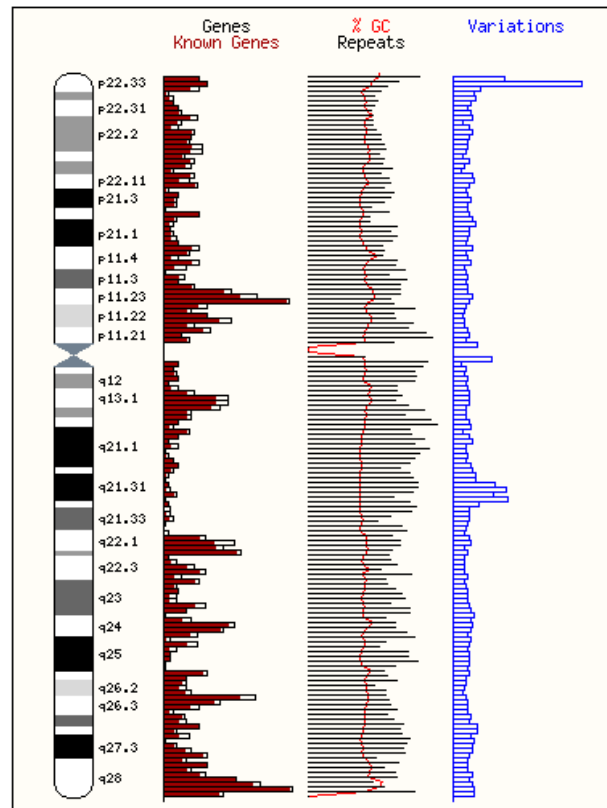
Scalability

- Multiplexing capabilities

X chromosome



Chromosome summary [help](#)



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 Technologies



Agilent SureSelect X Demo

43,074 baits

Liquid phase

Barcoded version



Agilent Technologies



Agilent SureSelect X custom

37,703 baits

Liquid phase

Febit custom 100 XLMR genes

58,603 baits

Solid phase

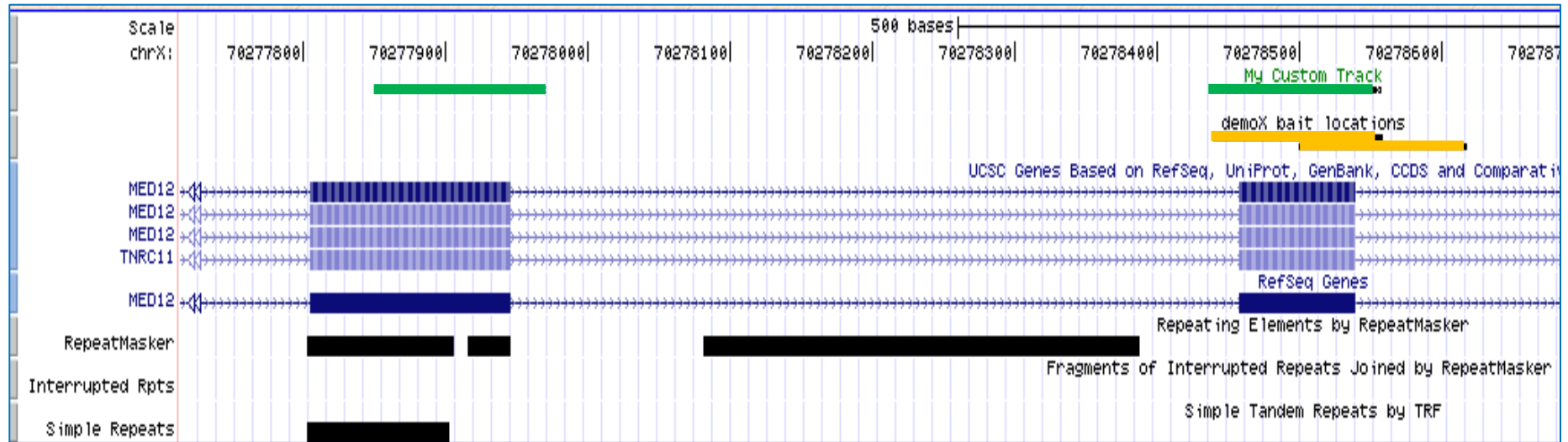
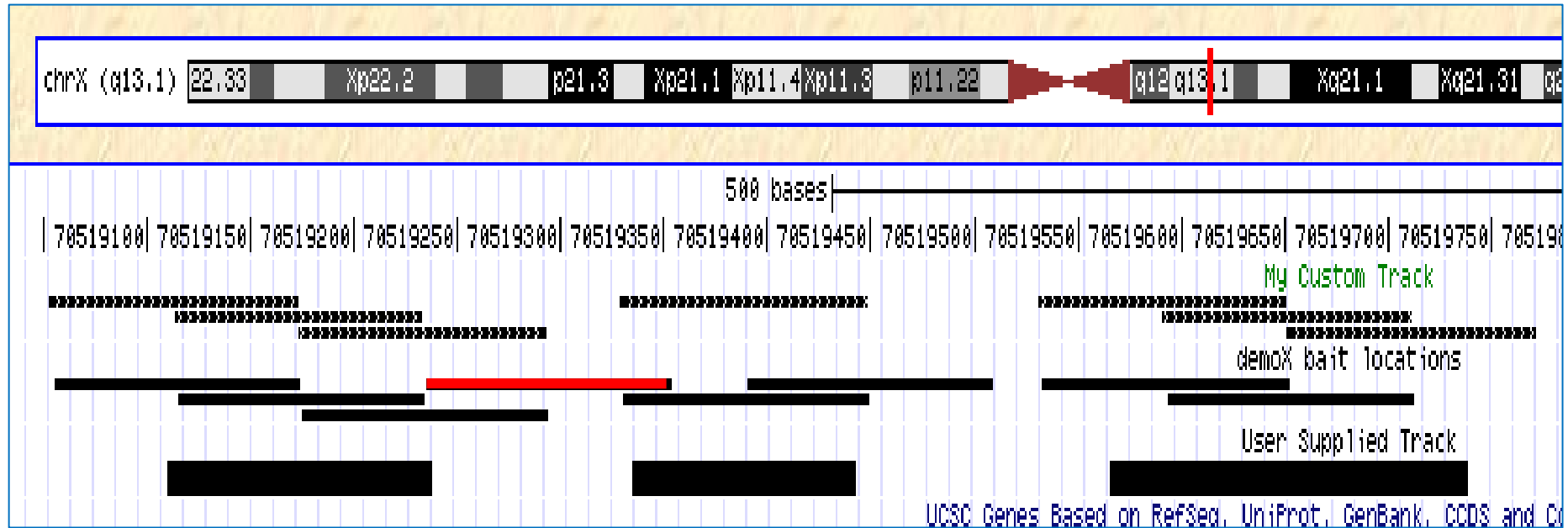
NimbleGen EZ custom

In progress

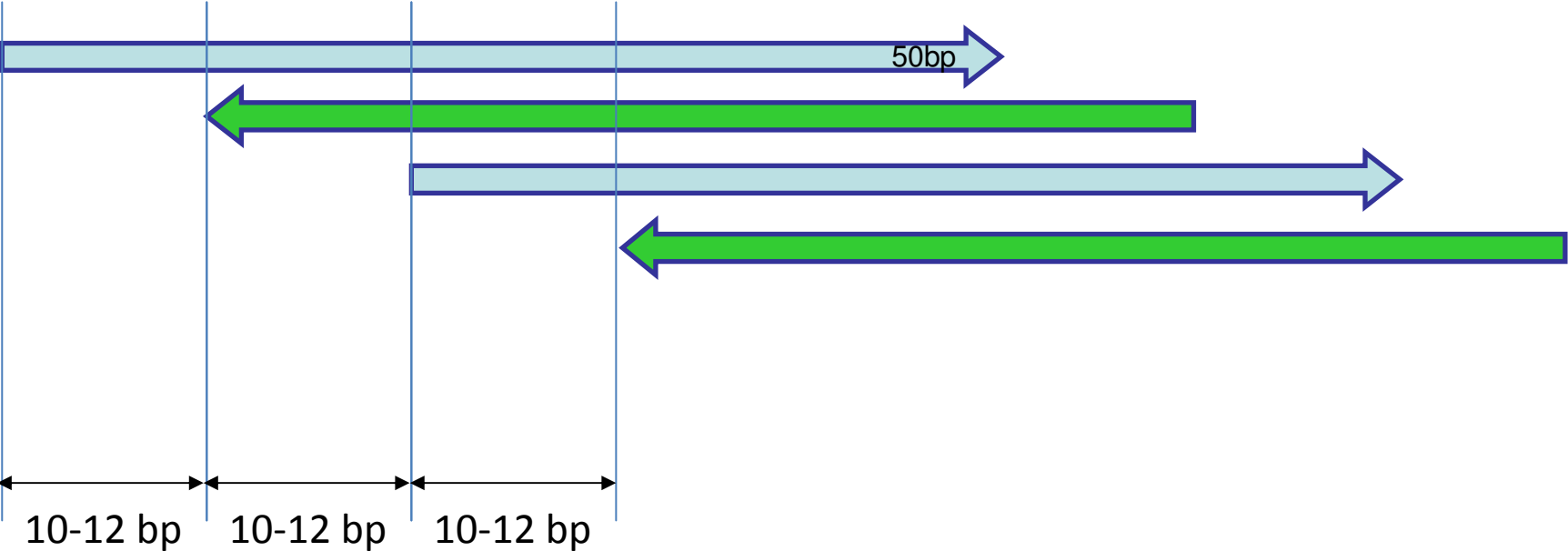
Solid phase

Liquid phase

Agi X_ex_demo vs Agi X_ex_custom

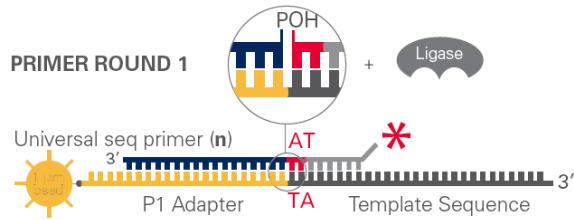


Febit design approach



SOLiD chemistry

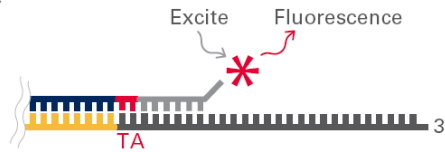
1. Prime and Ligate



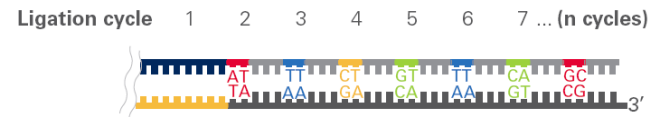
4. Cleave off Fluor



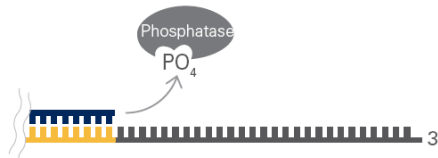
2. Image



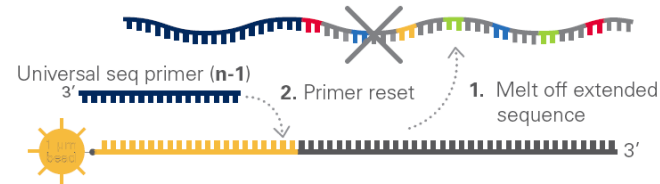
5. Repeat steps 1-4 to Extend Sequence



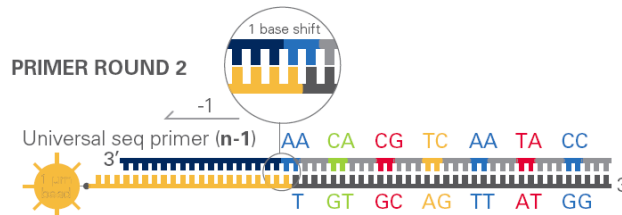
3. Cap Unextended Strands

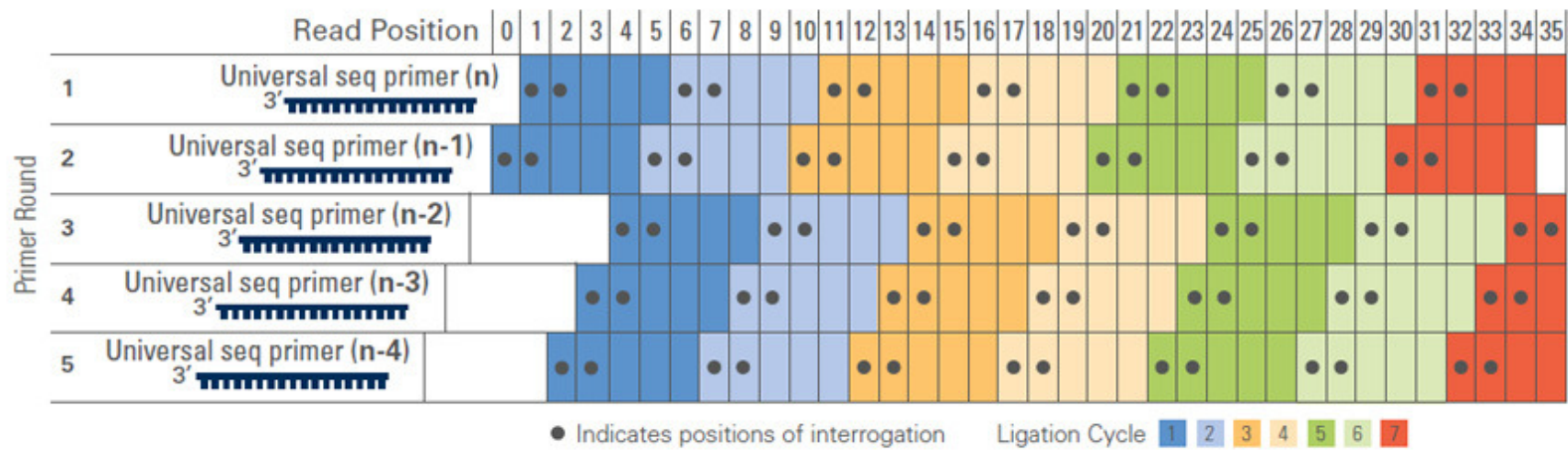


6. Primer Reset

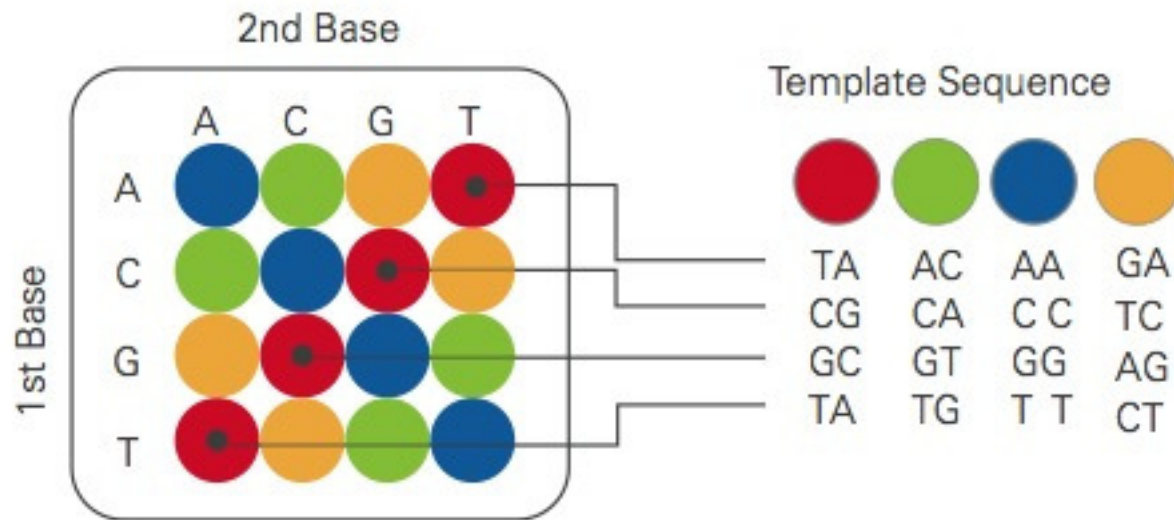


7. Repeat steps 1-5 with new primer



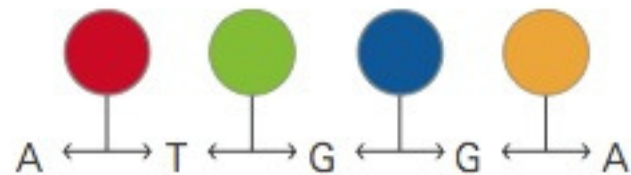


Possible Dinucleotides Encoded By Each Color



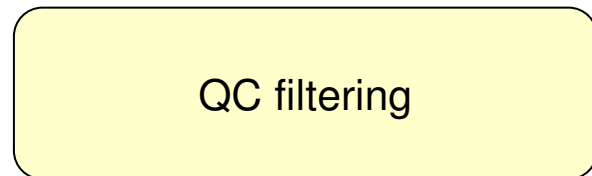
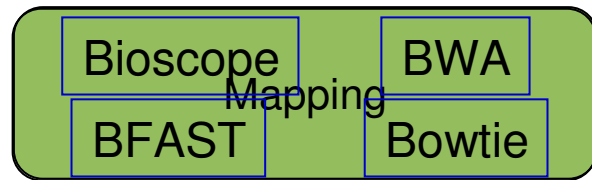
Double Interrogation

With 2 base encoding each base is defined twice

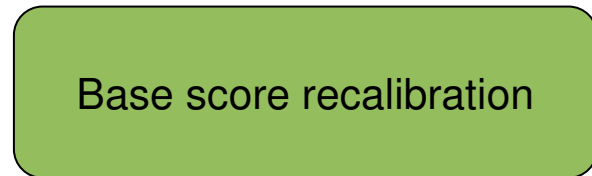
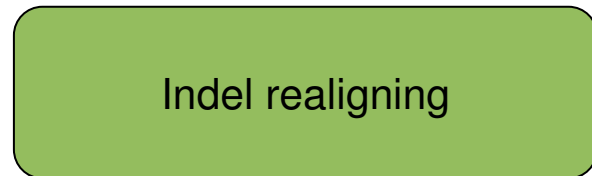


Analysis pipeline

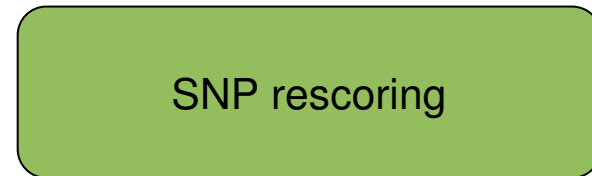
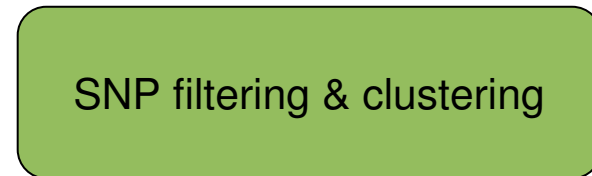
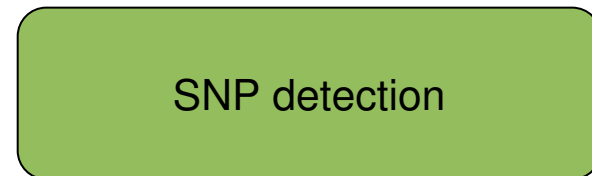
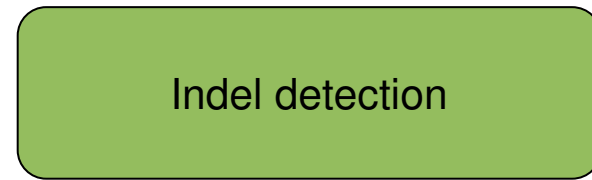
Primary analysis



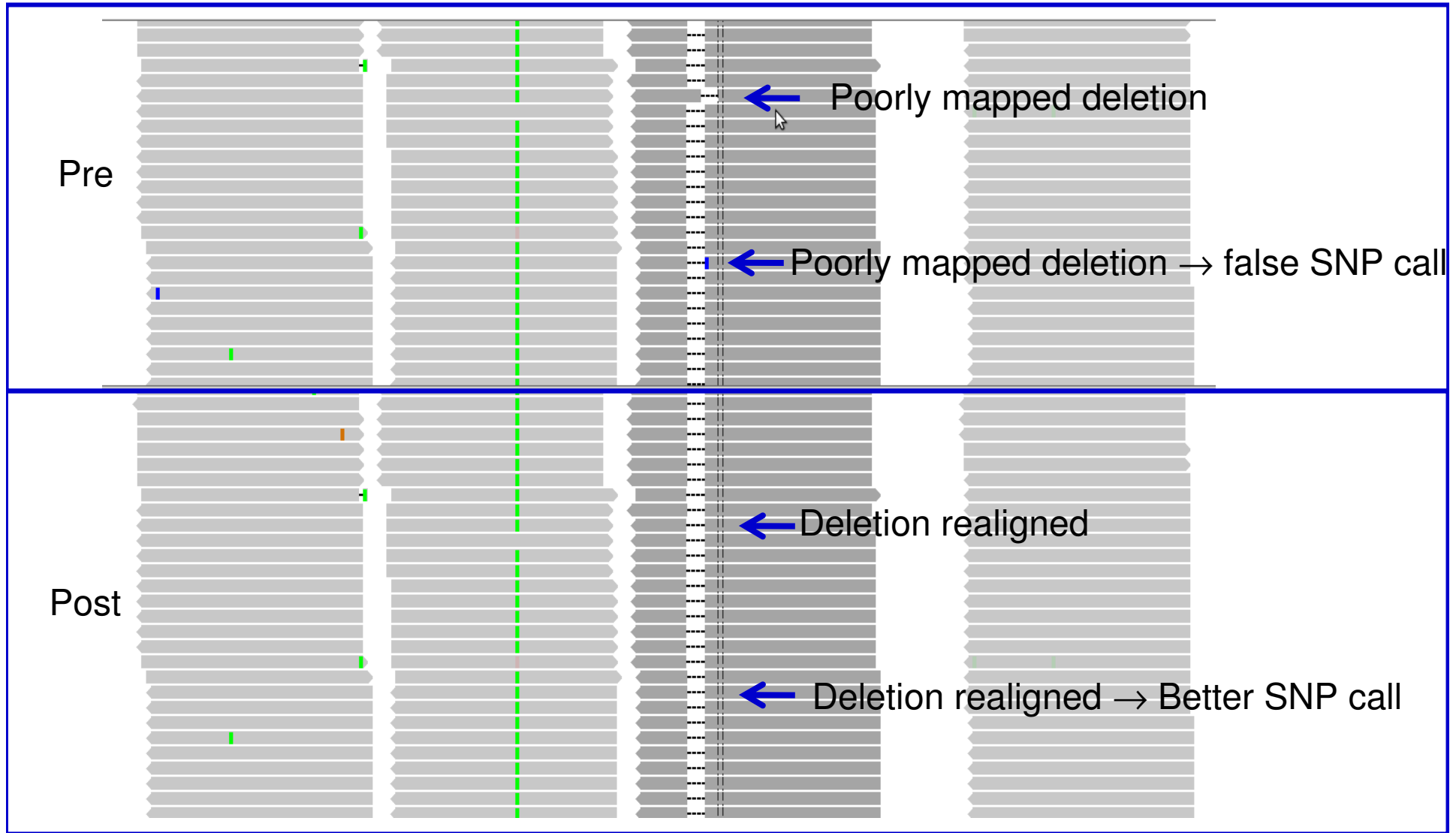
Optional



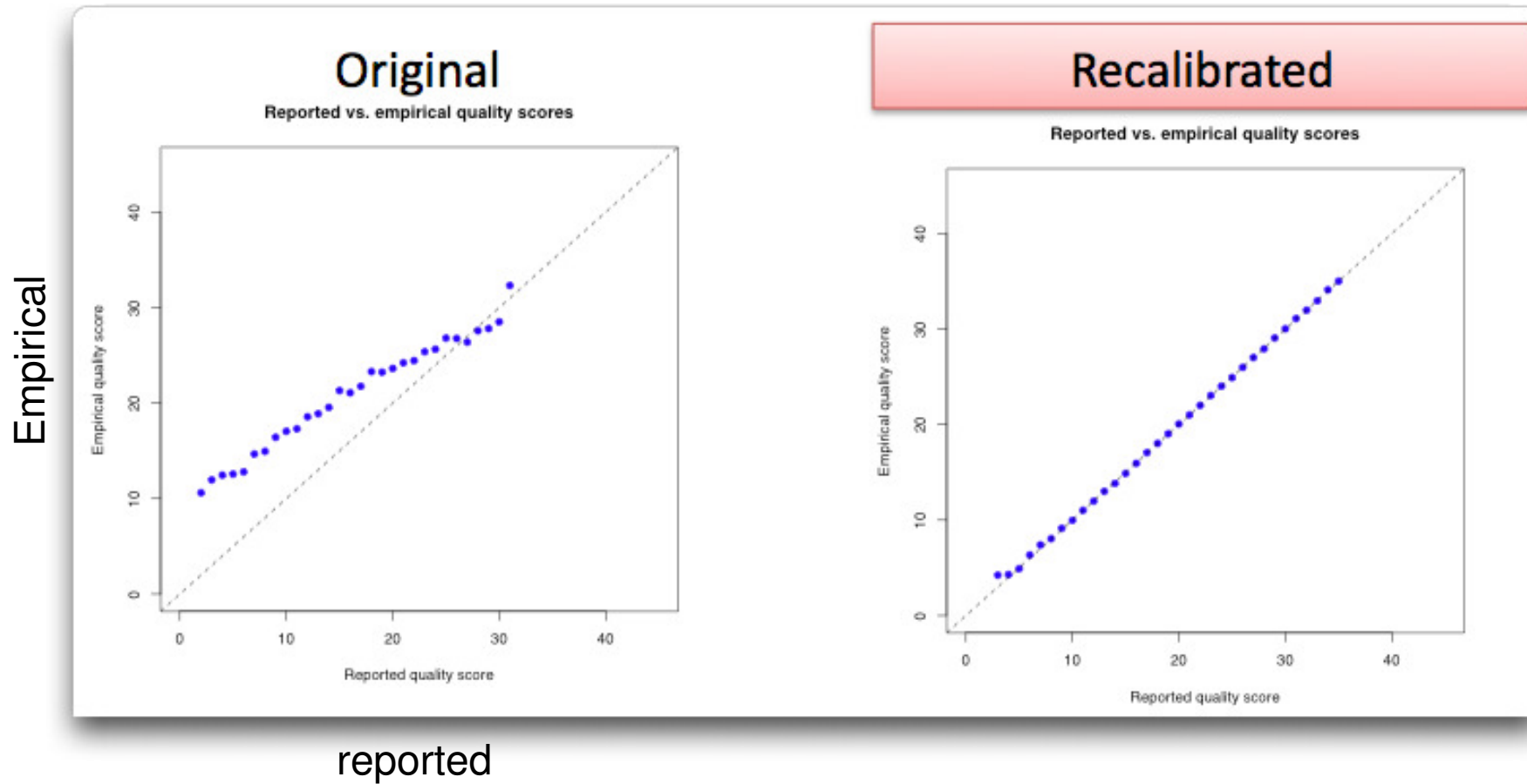
Secondary analysis



Indel realigning

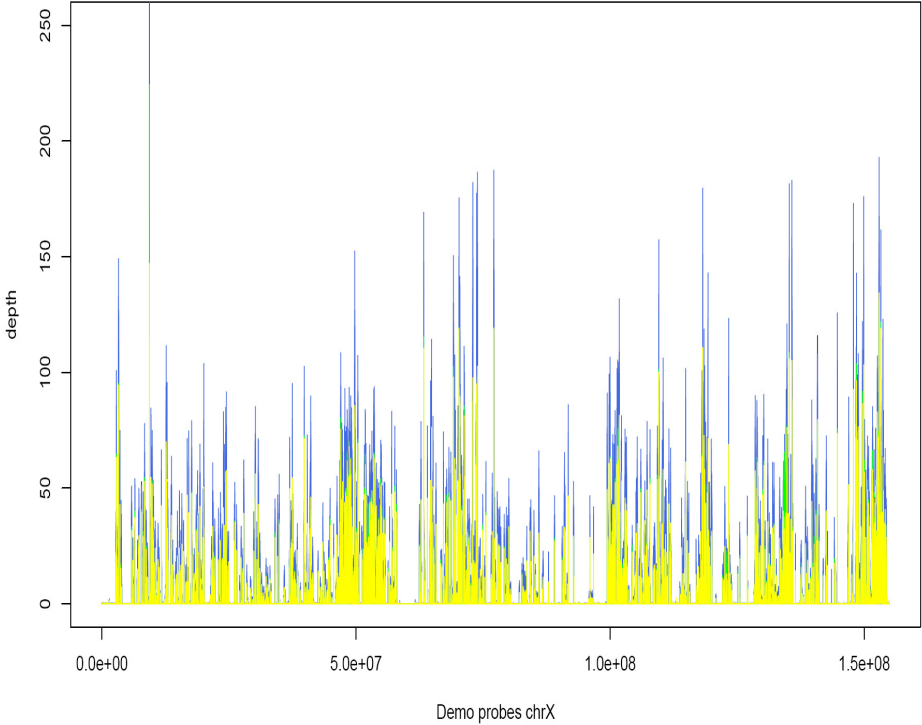


Base score recalibration [as for 1000 genomes project] for improved SNP calling etc

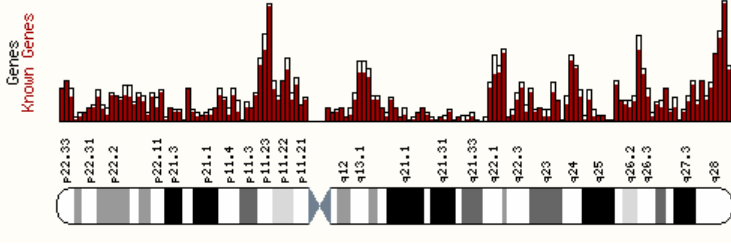
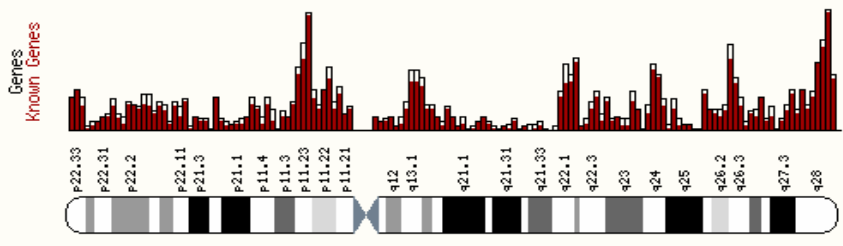
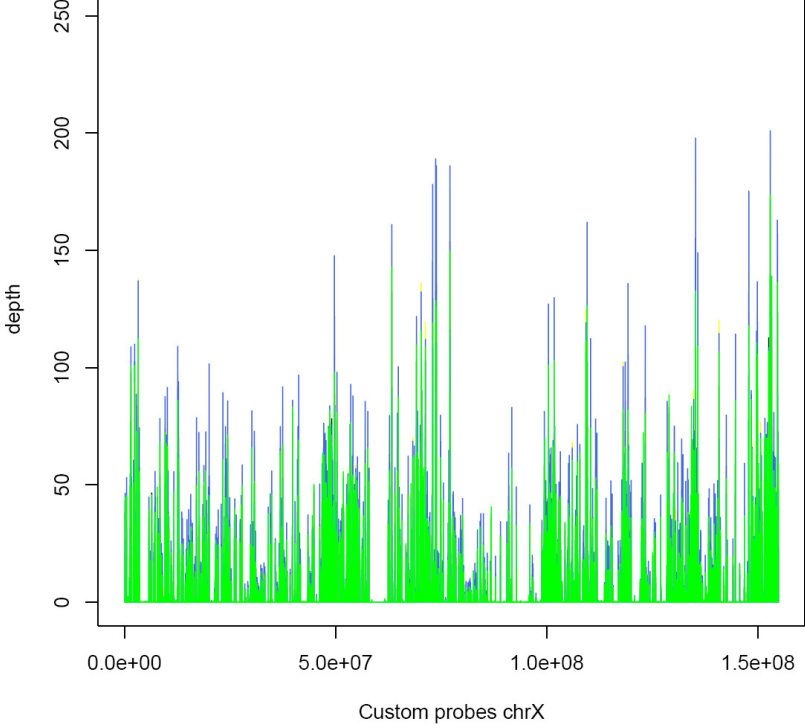


SOLiD under estimates QV for bases, Illumina over estimates

Agilent X exome demo



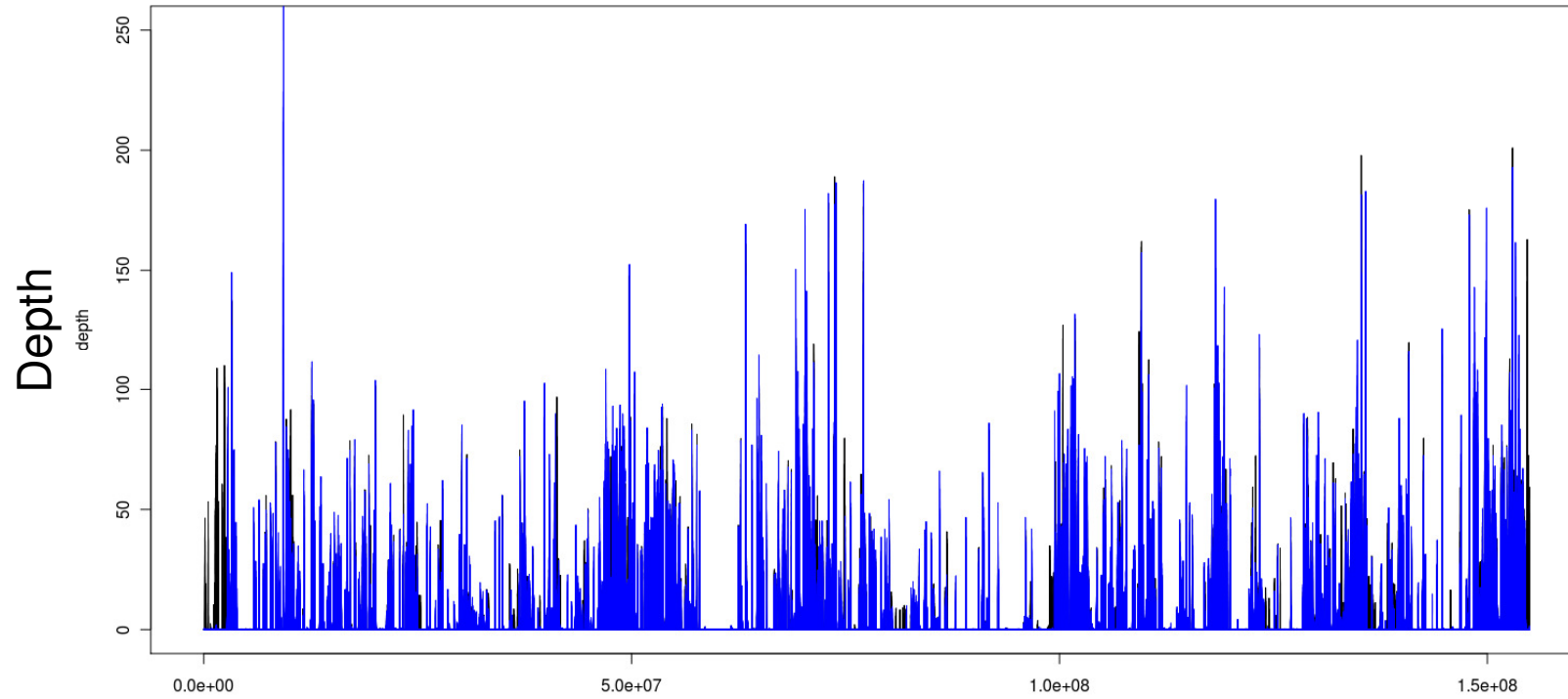
Agilent X exome custom



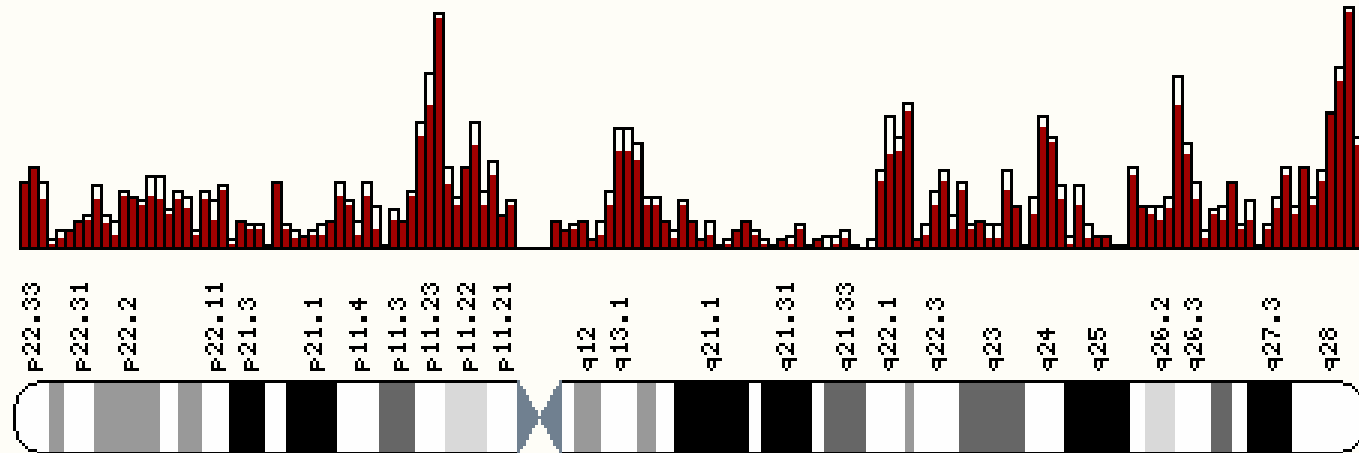
Reproducible enrichment spread

Demo

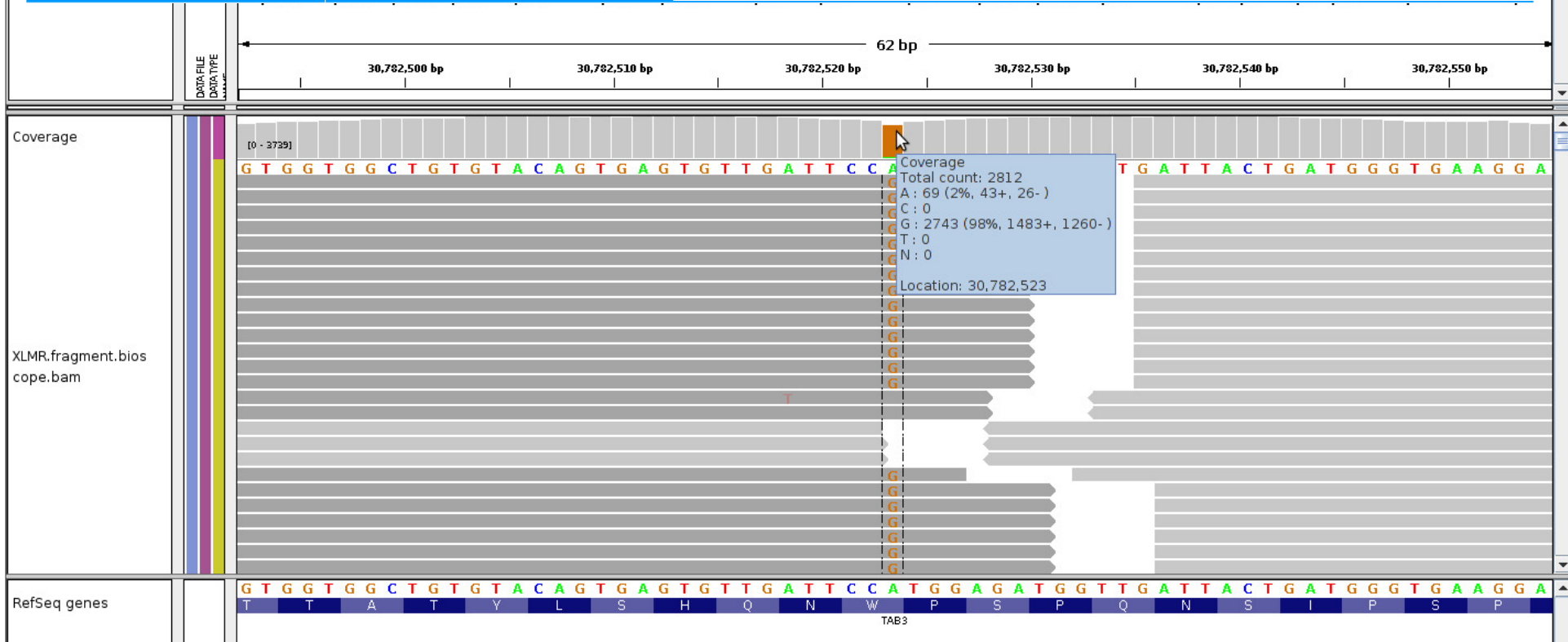
Custom



Genes
Known Genes

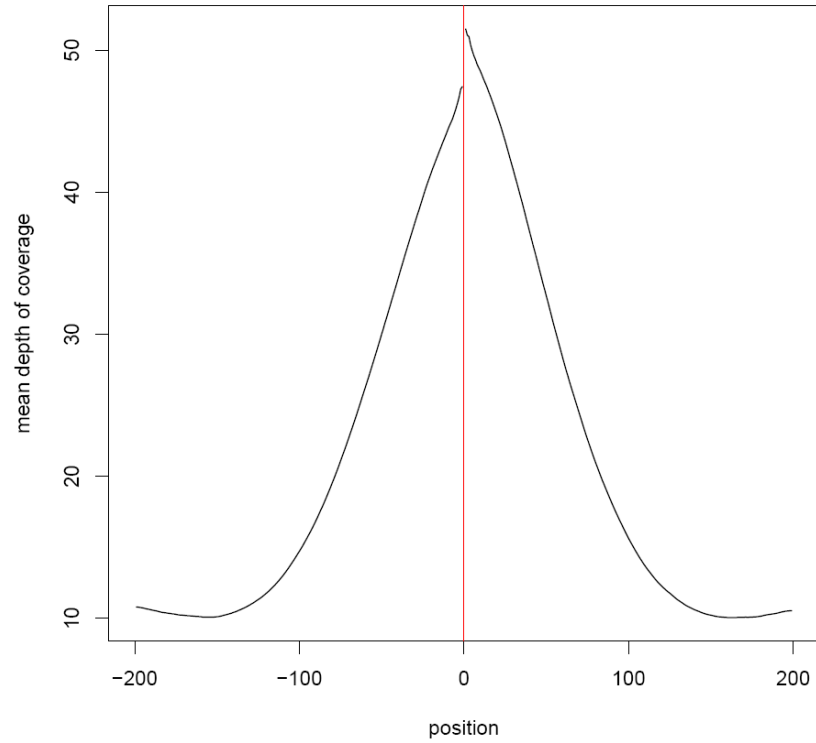


| | | | | |
|---------------------------|---------------------------------|---------------------------------|-----------------------|---------------|
| chrX:3078252 3 | A>G | score:104837 | depth:2812 | |
| Stats: | A: 69(2.45%)/6.57 | C: 0(0.00%)/0 | G: 2743(97.55%)/38.22 | T: 0(0.00%)/0 |
| Type: | NON_SYNONYMOUS_CODING | | | |
| Gene: | MAP3K7IP3/ENSG00000157625 | c.1727 | | |
| Transcript: | NM_152787.3/ENST000002884 22 | p.Trp394 Arg | | |
| dbsnp: | rs5927629 | 96% concordant SNP calls | | |



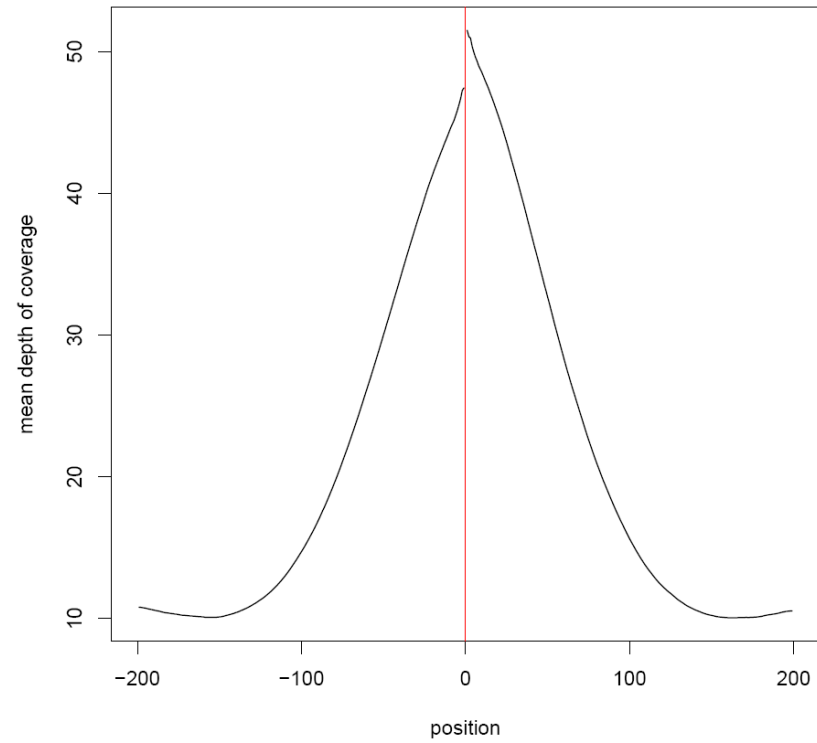
Depth of coverage at exonic boundaries

Intron leaking (Demo)

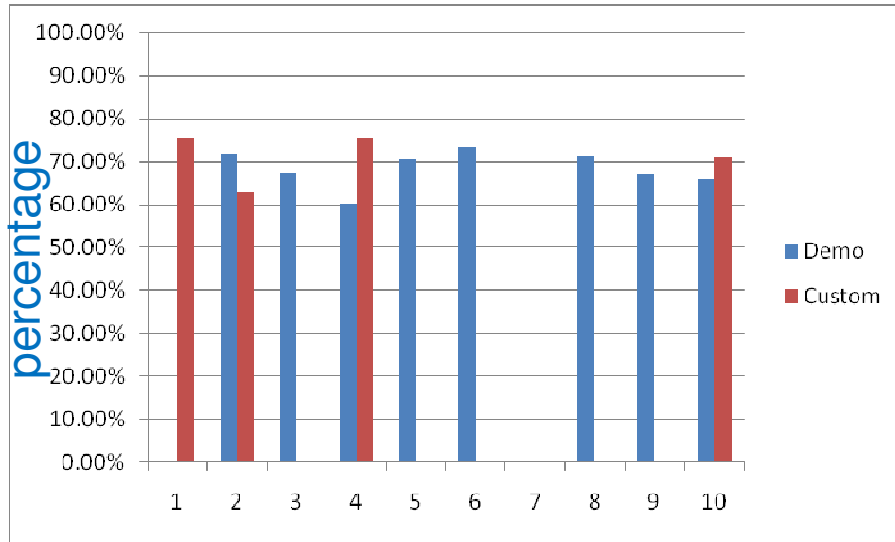


Agi_X_Ex_Demo

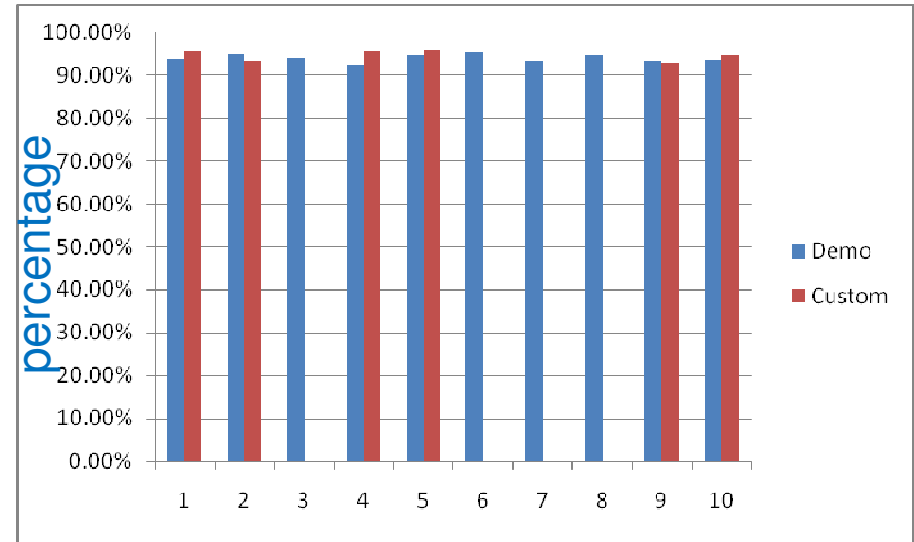
Intron leaking (Custom)



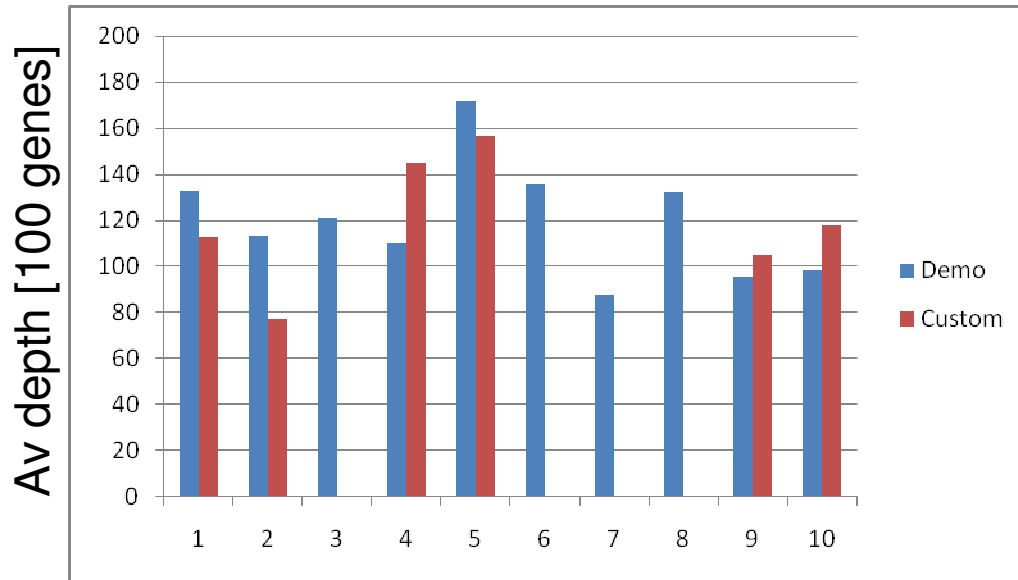
Agi_X_Ex_custom



Sample
Genome mapped reads on probe target
100 genes



Sample
X mapped reads on probe target
100 genes



Sample



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