UNIVERSITY OF CAMBRIDGE

BGC
Blood transfusion Genomics Consortium

NHS Blood and Transplant

Blood transfusion Genomics Consortium Fringe Meeting
24/06/2024

NGS to resolve complex structural variants

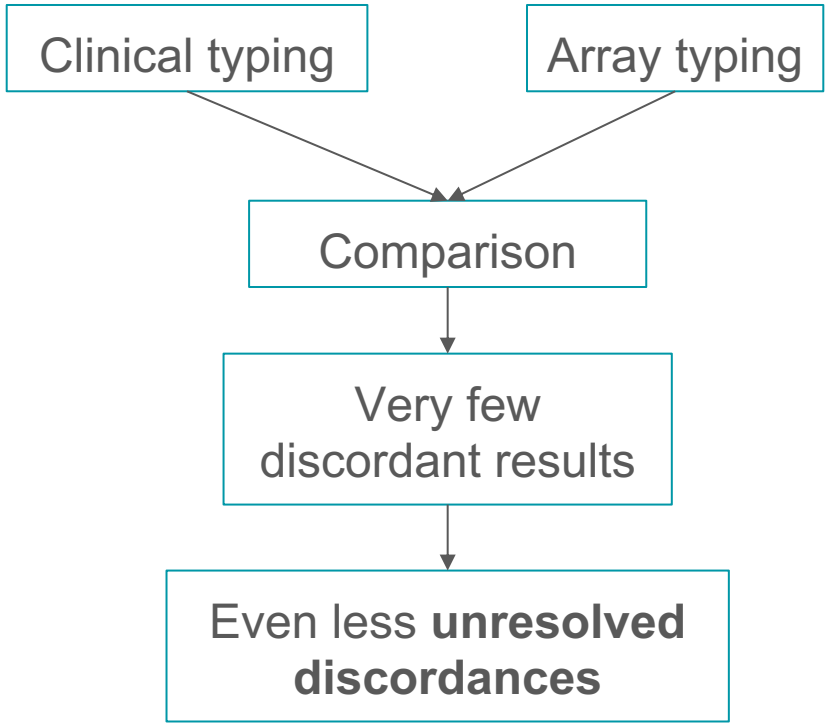
Olga Shamardina, PhD
os360@cam.ac.uk

Bioinformatician
Department of Haematology, University of Cambridge
(also, UCLH, Honorary Researcher and NHSBT, Honorary Bioinformatician)

Why do we need it

- **Patient safety:**
  - Resolve discordances
  - Resolve complex cases

- Potentially, discovery



Outline

- NGS technology
- Our panel
- Copy number estimations
- Some examples

NGS technology

NGS

Sample and library preparation

Target enrichment

Sequencing

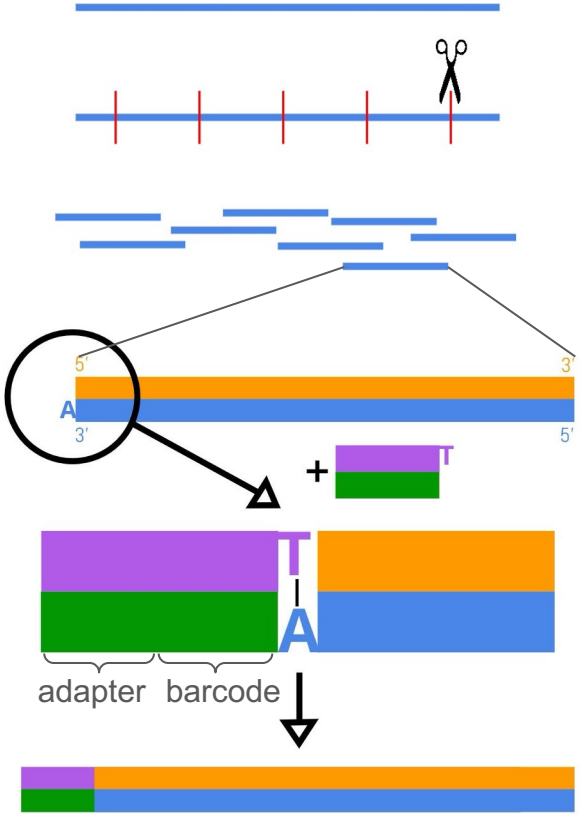
Analysis

NGS: Library preparation

- Enzymatic or mechanical fragmentation (~200 bp)
- Adapters are needed for the sequencing step

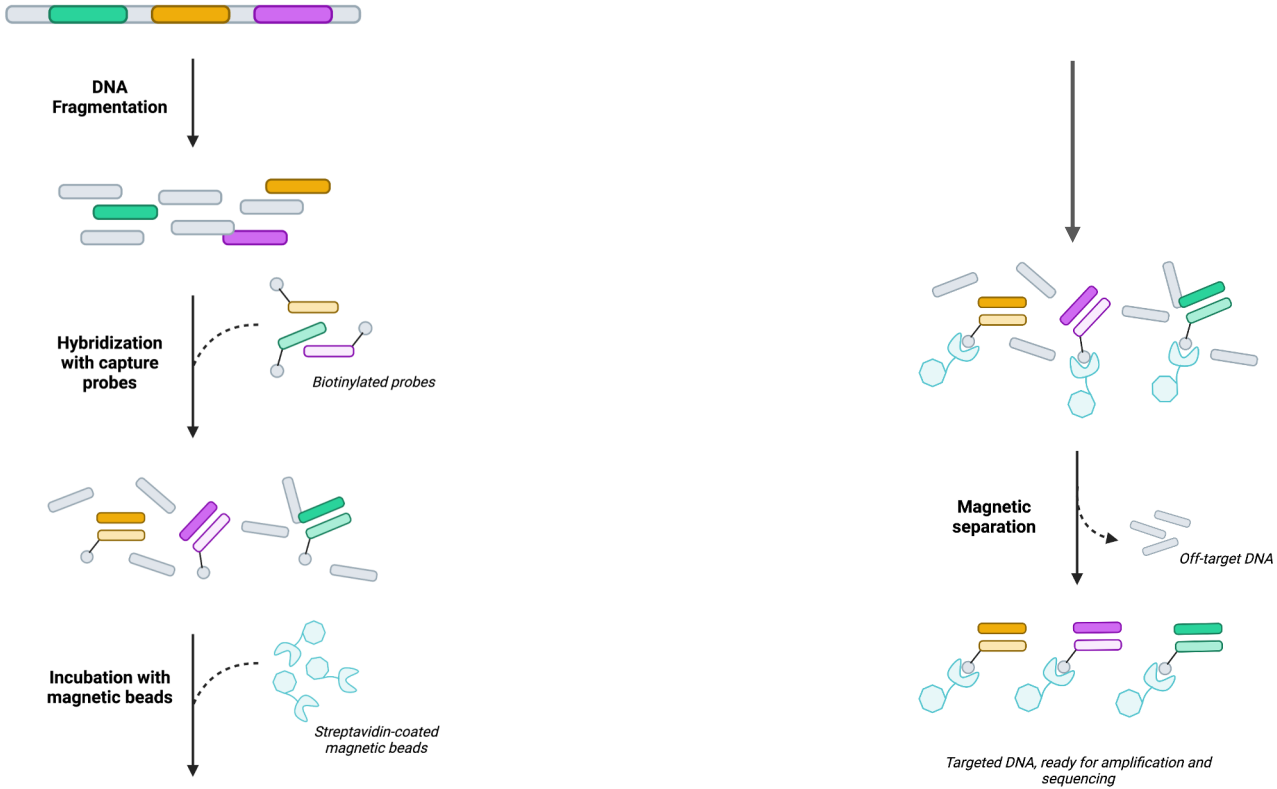
Steps:

- Blunting (in case of enzymatic fragmentation)
- A-tailing
- Adapter ligation (with barcodes for multiplexing)



https://www.twistbioscience.com/blog/science/exome-sequencing-101-part-1-library-preparation, modified

NGS: Target enrichment



https://www.twistbioscience.com/blog/science/Target-Capture-Hybridization-NGS, modified

NGS: Sequencing

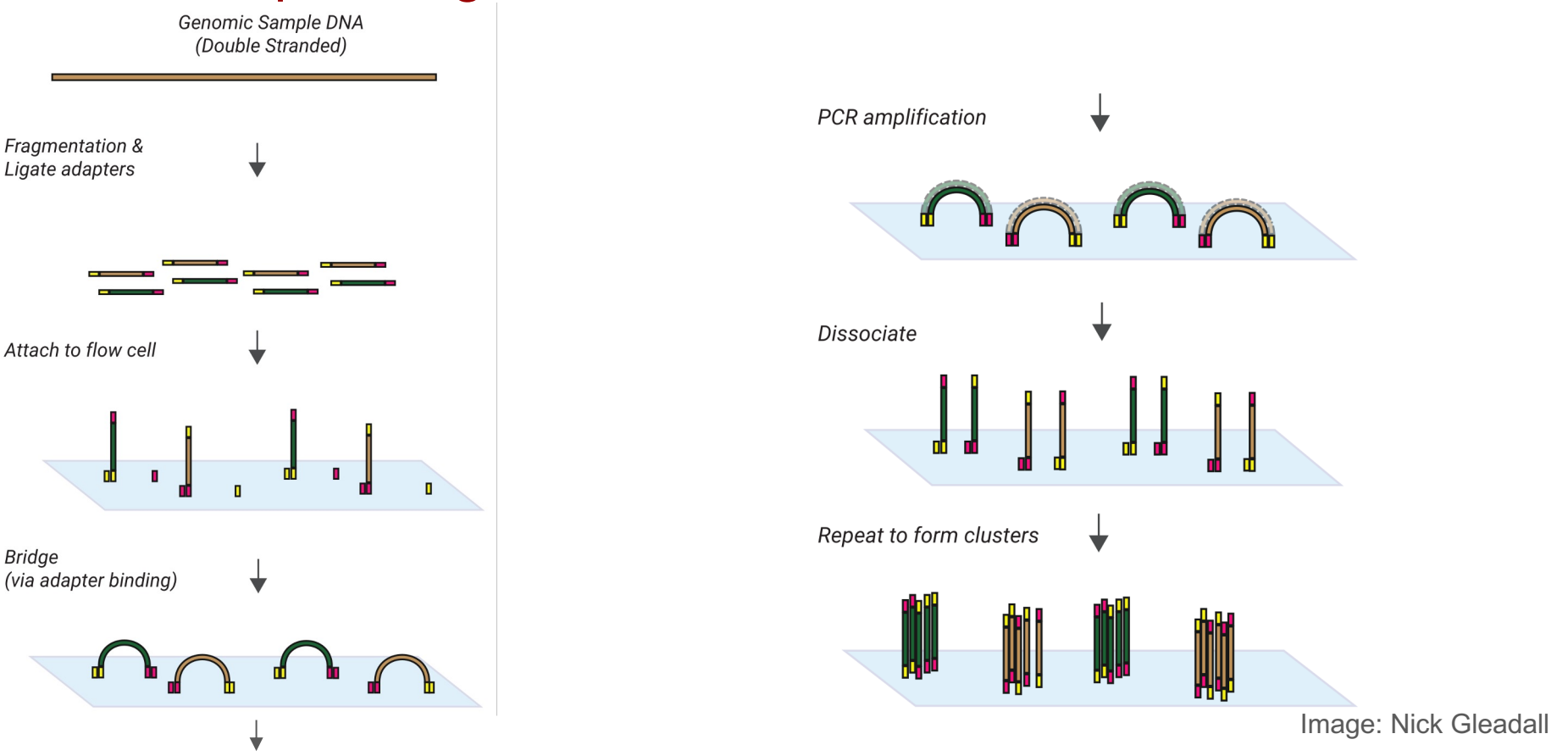
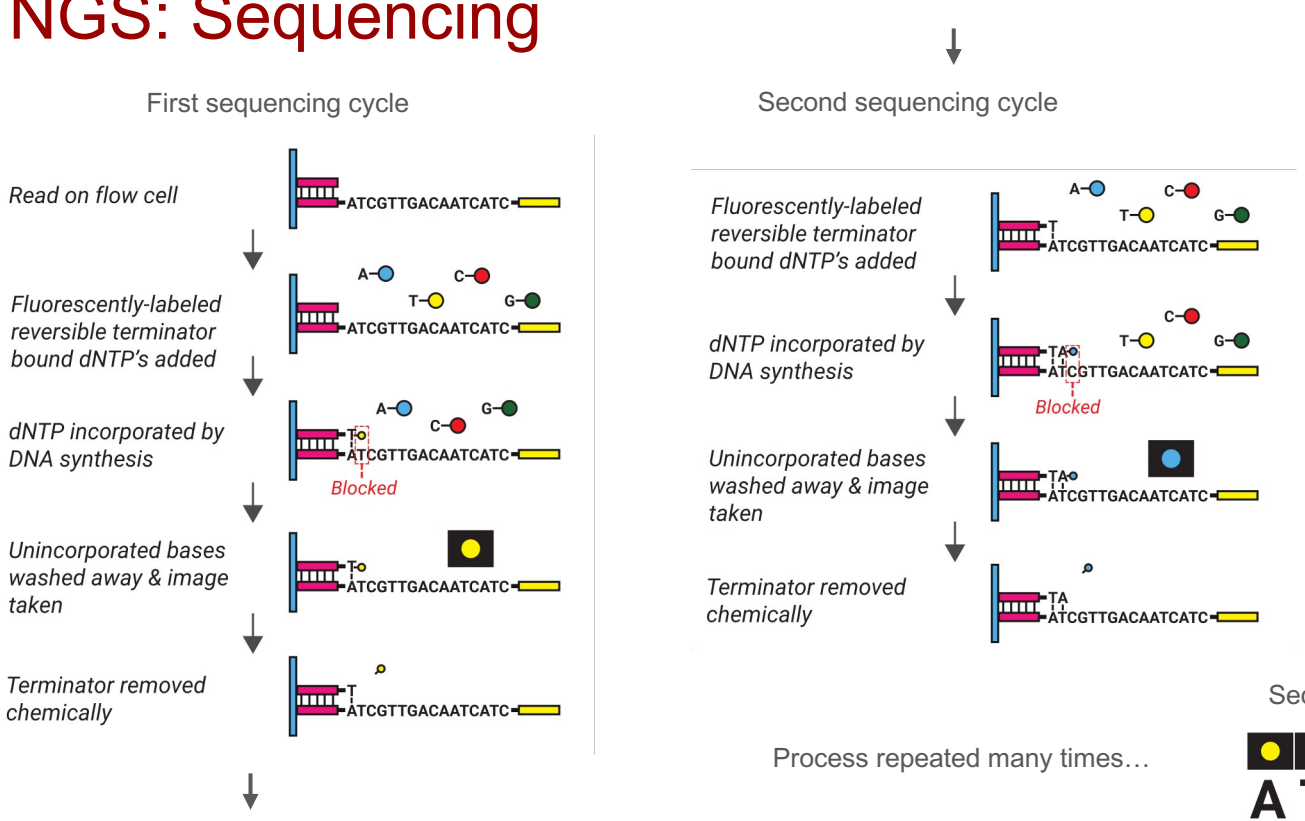


Image: Nick Gleadall

NGS: Sequencing



Sequence reconstructed from images:

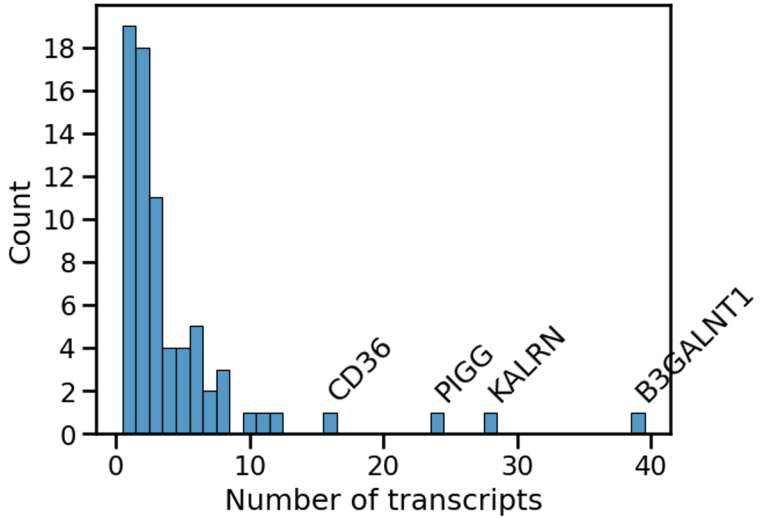
See also Illumina animation: https://youtu.be/fCd6B5HRaZ8
https://emea.illumina.com/science/technology/next-generation-sequencing/beginners.html

Images: Nick Gleadall

Our panel

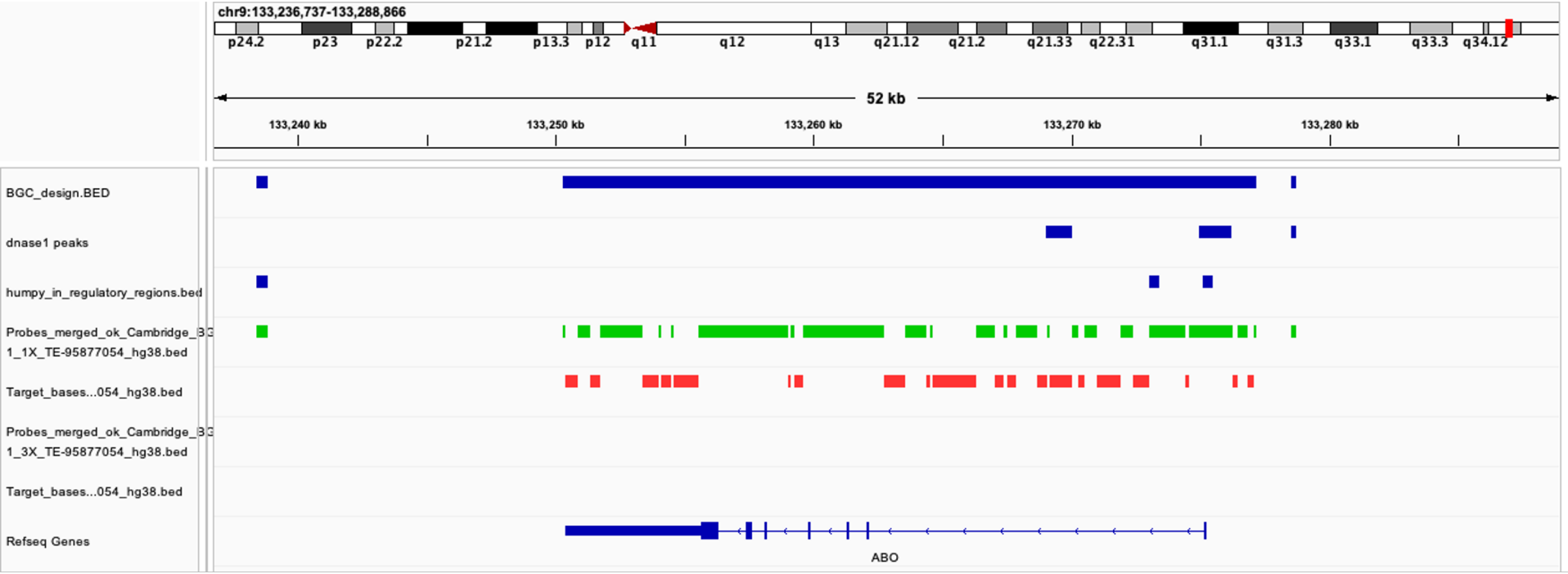
Gene panel

- GRCh38
- ISBT or MANE transcripts for each gene (+ all NM_/NR_ RefSeq transcripts)
- Also, regulatory regions identified from RNA-seq studies are included
- 1,794,208 base pairs



	exon only	complete
HEA	32	18
Blood group related	0	3
HPA	5	2
HLA	3	2
Other blood phenotypes	2	3
Sex calling	1	3

Target vs Capture



Not a problem in most cases, the excluded regions are mostly non-coding, but still we want to “rescue” some important genes

Anchors: RHD vs RHCE

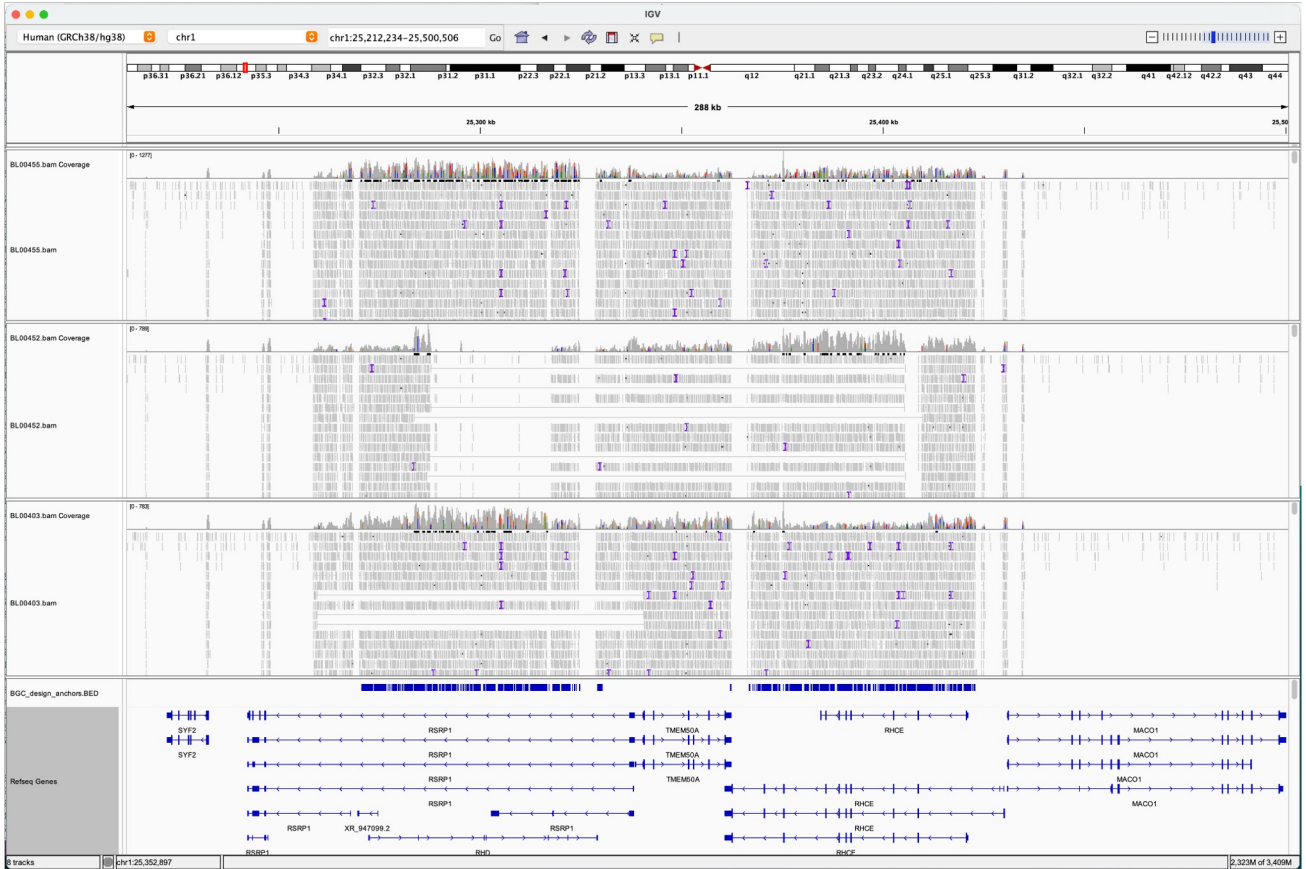
Bait size: 120bp; BLASTn line: 60bp

```
Query  13463  AGGGTCTGAGACCGGGAAAGGTGAGGGCTACCCAGGTGGCCCTGATGTTTTCTGCCAGCC  13522
              |||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  47346  AGGGTCTGAGACTGGGAAAGGTGAGGGCTACCCAGGTGGCCCTGATGTTTTCTGCCAGCC  47287

Query  13523  AGCTCACCAGGTCCCTCGCAGCAGGCGGCAAAGGGAGGGAGGTTTGCTGTGAAGATTATG  13582
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  47286  AGCTCACCAGGTCCCTCGCAGCAGGCGGCAAAGGGAGGGAGGTTTGCTGTGAAGATTATG  47227

Query  13583  TGGTTCCCAACAACAAGAGCGCTGGGCCTATCTCTGCCCTCTCTTTTCTGTGTGTCCTGG  13642
              |||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct  47226  TGGTTCCCAACAACAAGAGCACTGGGCCTATCTCTGCCCTCTCTTTTCTGTGTGTCCTGG  47167
```

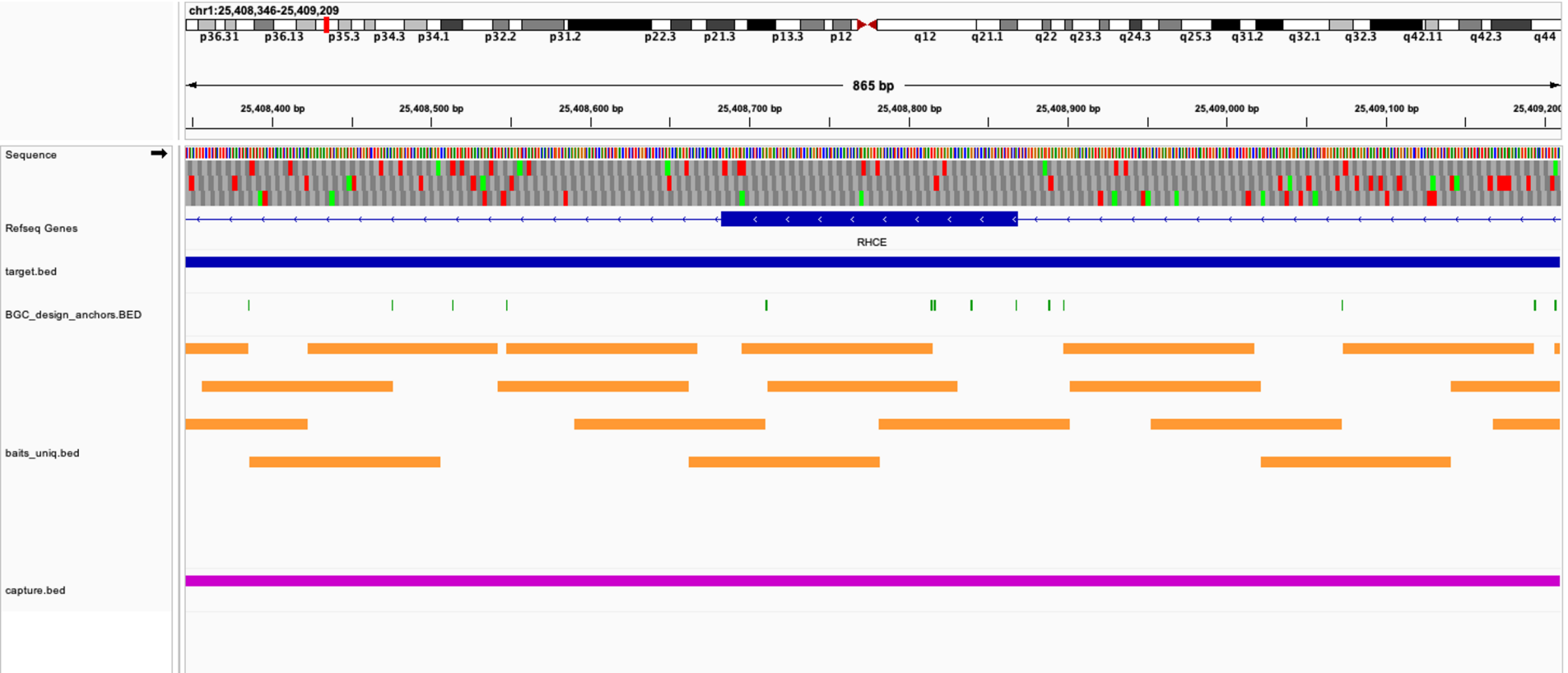
RH-locus coverage



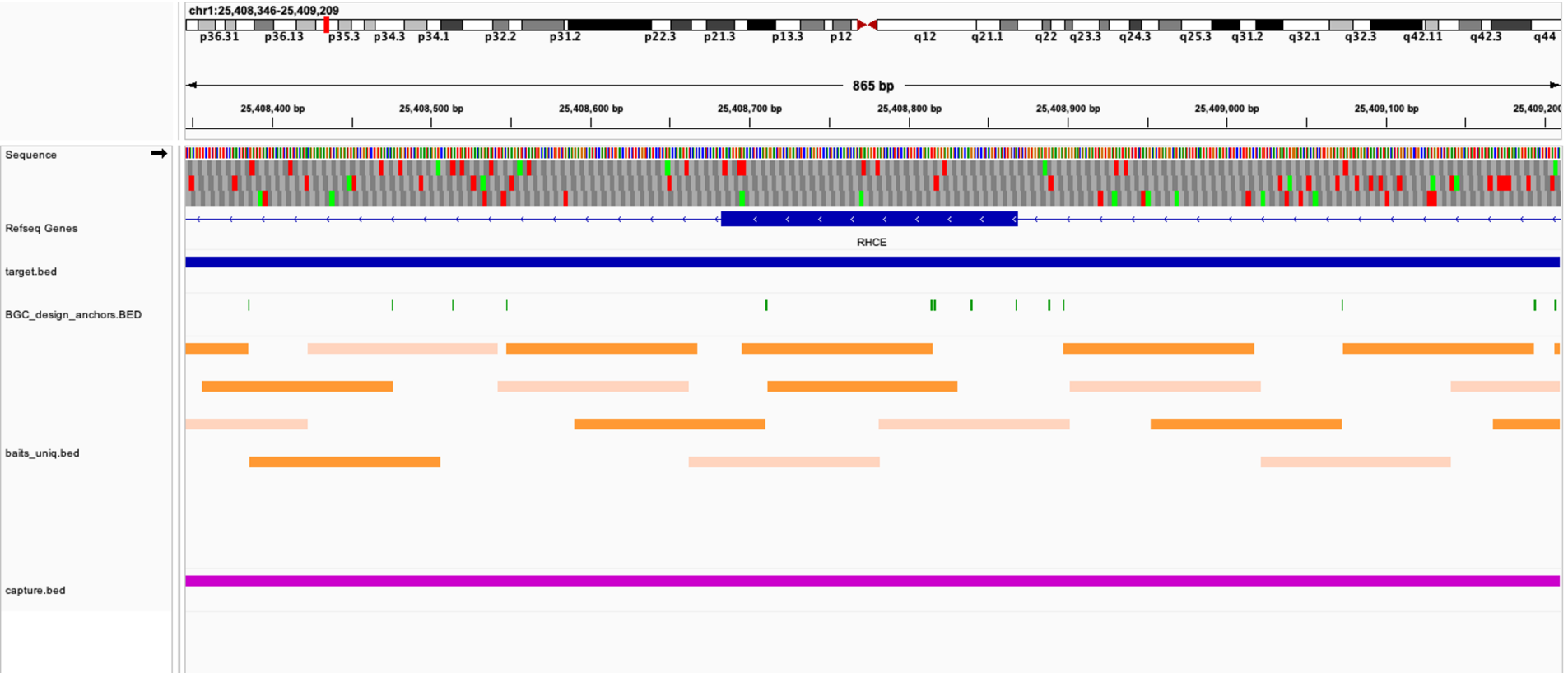
Difficult because of:

1. Complex rearrangements
2. Mapping artifacts

Baits for our panel



Baits for our panel



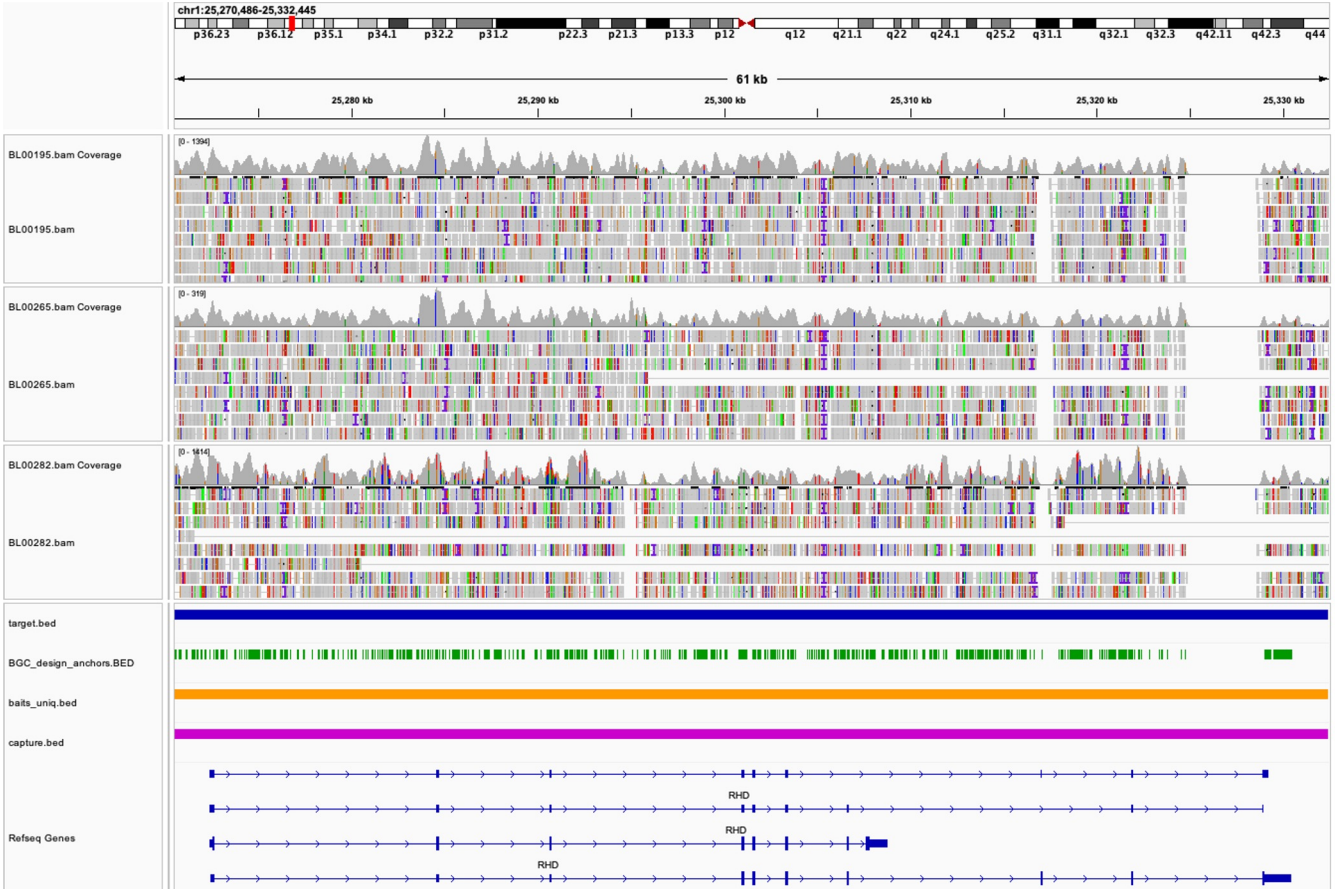
Processing sequencing data

- GATK Exome Germline Single Sample Pipeline
- GATK GenotypeGVCFs
- Custom scripts

Image: GATK

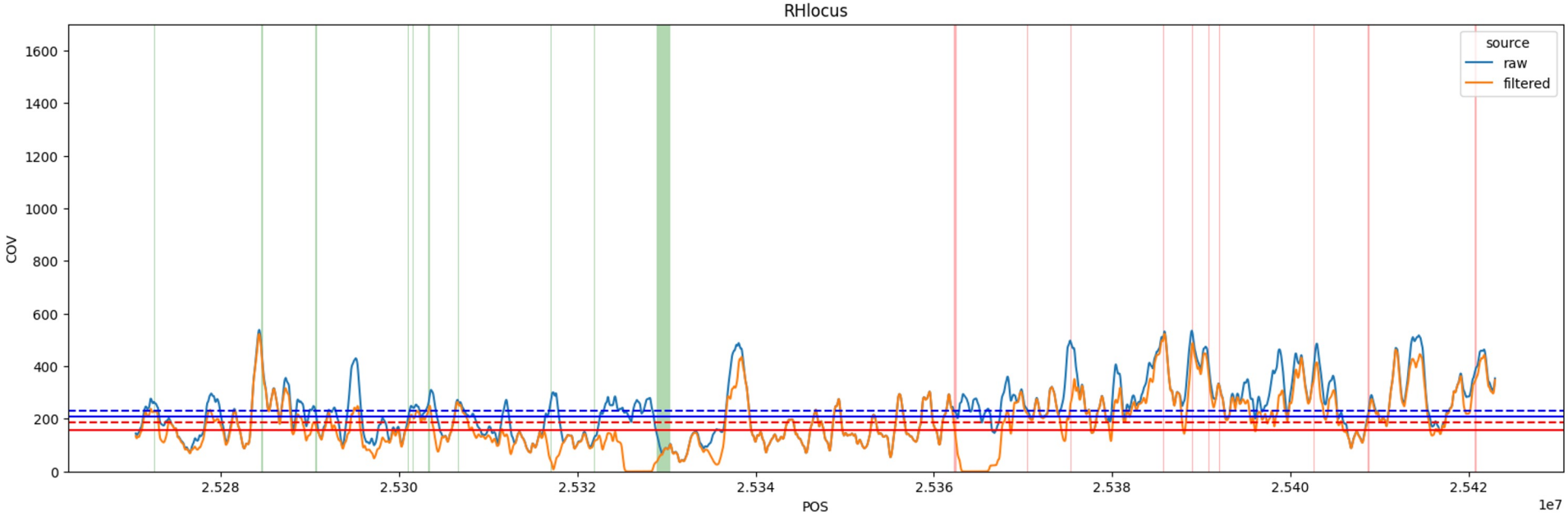
Copy number estimations

Copy number estimations: Coverage



Copy number estimations: Coverage

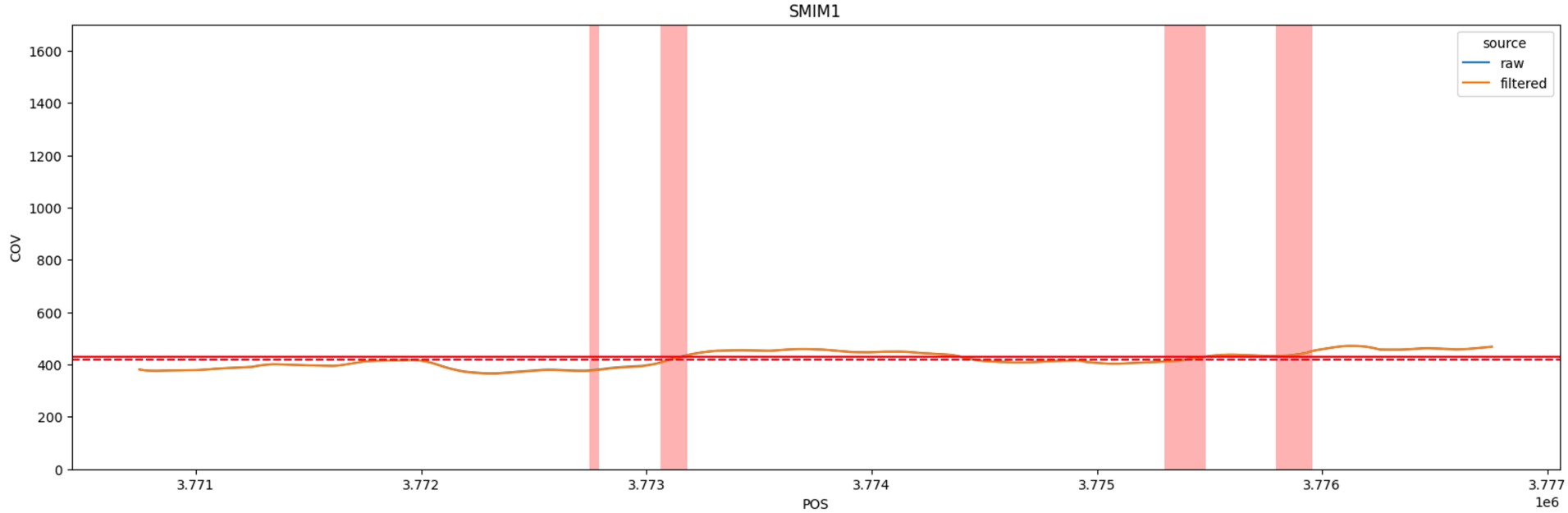
Mean value in window=1000 with step=1



raw => exclude UNMAP,QCFAIL,DUP
filtered => exclude UNMAP,SECONDARY,QCFAIL,DUP,SUPPLEMENTARY

Copy number estimations: Coverage

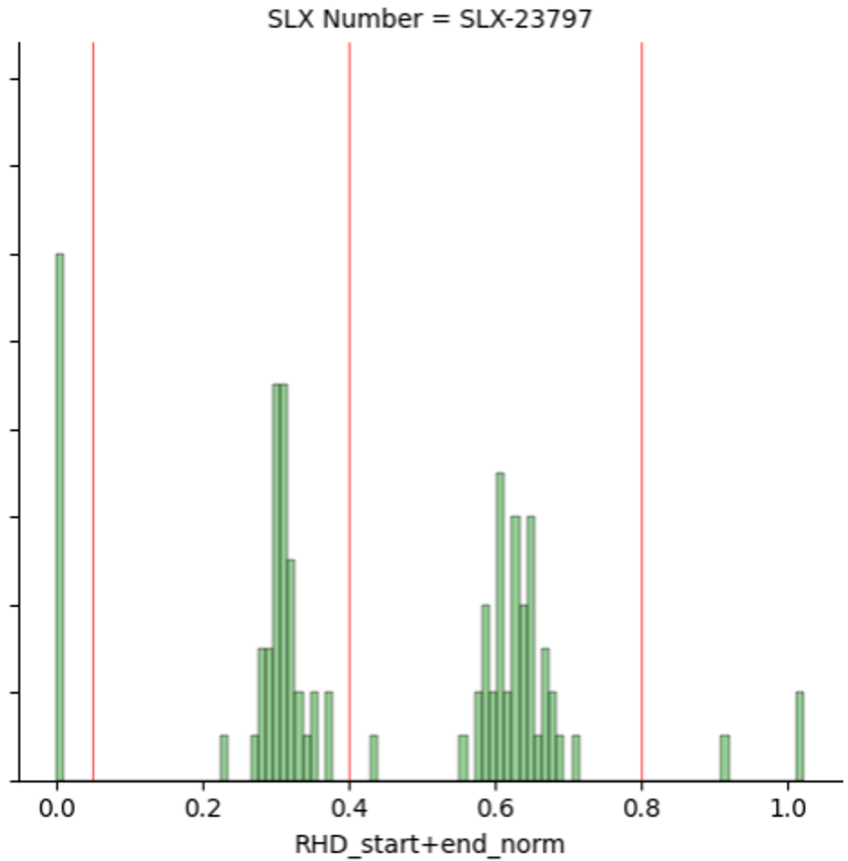
Mean value in window=1000 with step=1



raw => exclude UNMAP,QCFAIL,DUP
filtered => exclude UNMAP,SECONDARY,QCFAIL,DUP,SUPPLEMENTARY

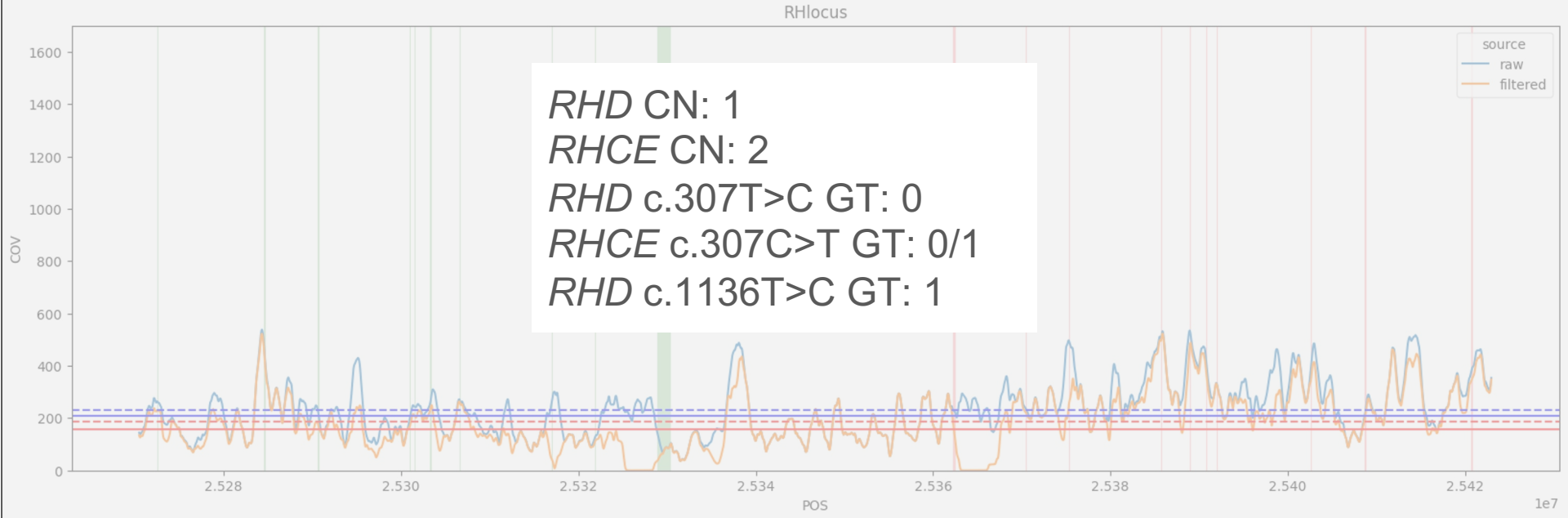
Copy number estimations: Coverage

- Filter
- Normalise (samples in NGS have **variable coverages**)
- Use **median** coverage per region
- Analyse by **batch**



Copy number estimations: Coverage

Mean value in window=1000 with step=1



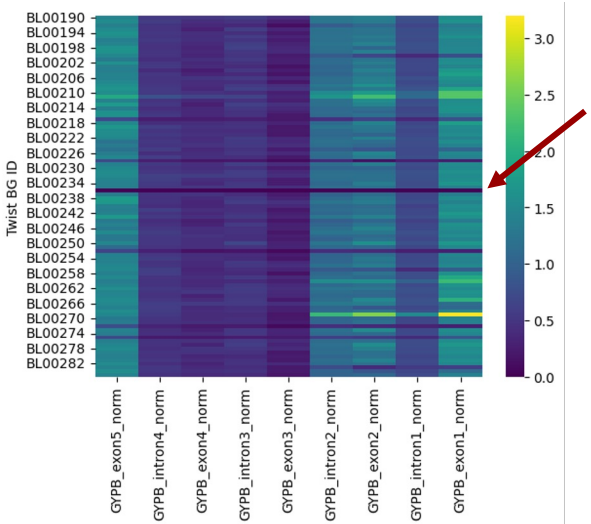
RHD CN: 1
RHCE CN: 2
RHD c.307T>C GT: 0
RHCE c.307C>T GT: 0/1
RHD c.1136T>C GT: 1

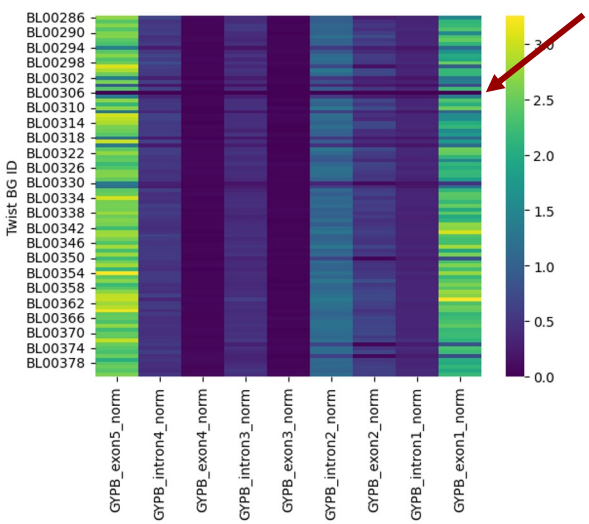
raw => exclude UNMAP,QCFAIL,DUP
filtered => exclude UNMAP,SECONDARY,QCFAIL,DUP,SUPPLEMENTARY

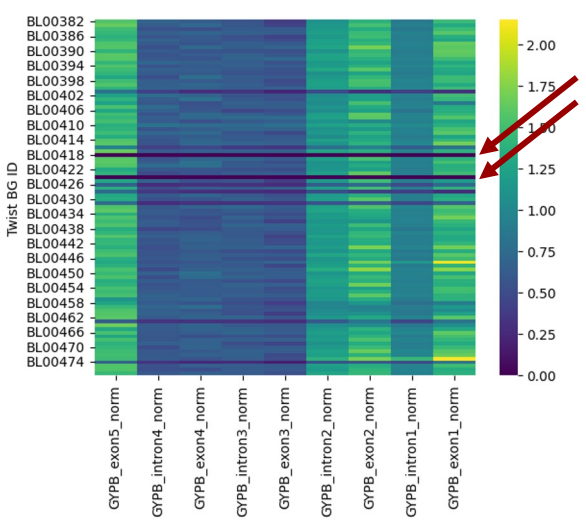
Some examples

MNS system: detect U- cases

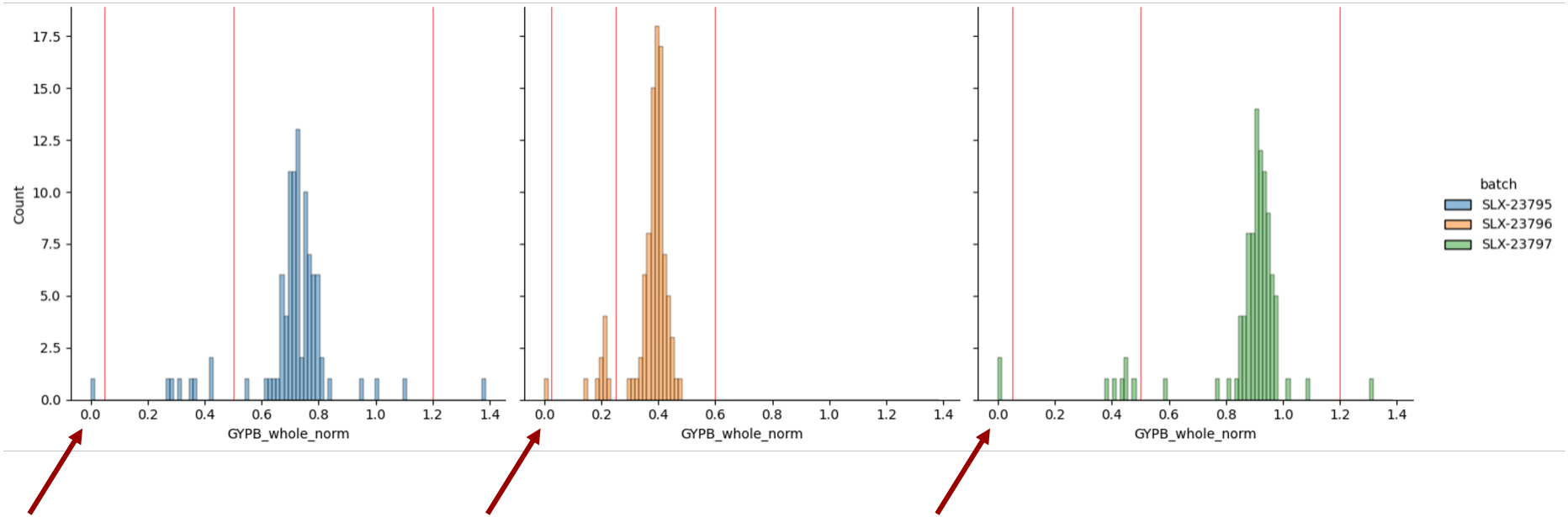
- For normalisation, using *GYPE* (*GYPA* doesn't work)
- Note variable coverage of regions:
  - For each sample
  - Between samples
  - Between batches







MNS system: detect U- cases



MNS system: MUR

The hybrid allele resulting from insertion of part of *GYPA* exon3 in *GYPB* pseudo exon3

1. Irregularities of coverage on a heatmap
   => "suspicious" samples
2. Visual assessment in IGV
   => **a common pattern of 6 variants**
3. Systematic check for
   TACCGGTTTCCTCTTCTGGAGGGTAAACAGTTCT
   in chr4:144000378-144000419
   => 3 “MUR” samples

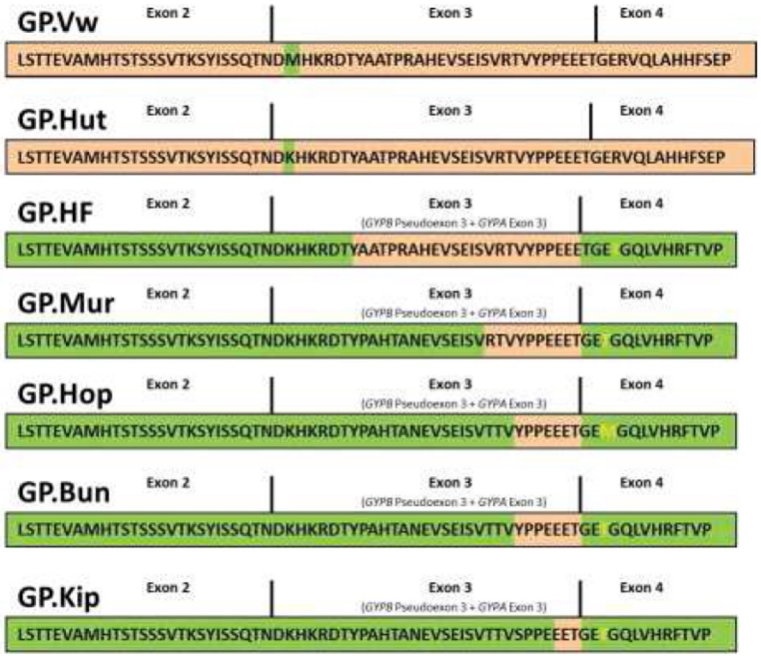
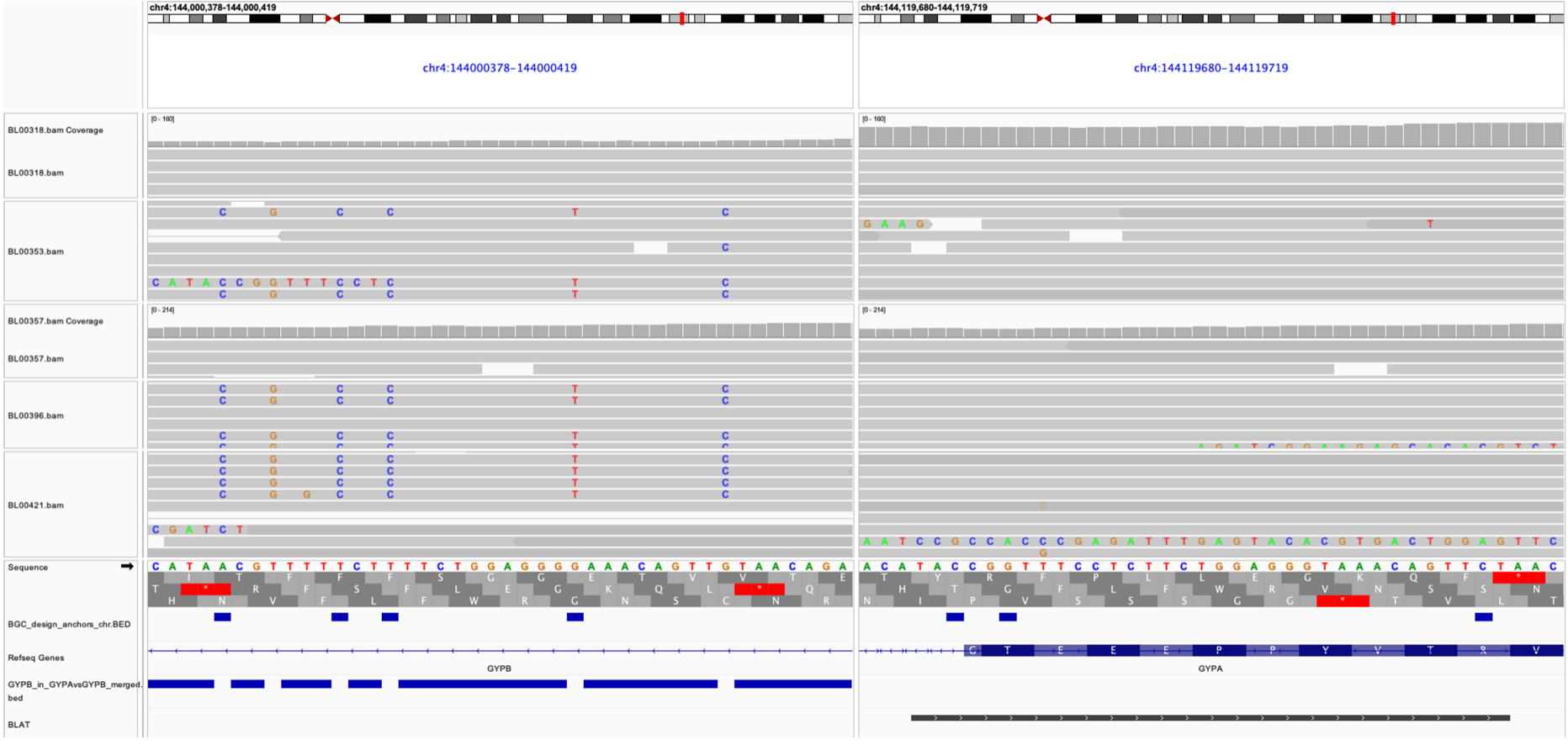


Image: Genghis Lopez

MNS system: MUR



RH system

- *SMIM1* (including introns) for normalisation
- *RHD*: exon1, beginning of intron1, exon9 with flanks, end of exon10
- *RHCE*: exon1, beginning of intron1, exon9 with flanks, exon10
- *RHD*/*RHCE* exon4-exon7
- *RHD*/*RHCE* exon2 (**5 “anchors”**)
- *RHD*/*RHCE* intron2 (**not workin**g)
- *RHD*/*RHCE* exon3
- *RHD*/*RHCE* intron3
- *RHD*/*RHCE* exon8 (**1 “anchor**”)

RH system

- Counts of variant contexts observed in the reads
- Filtering: exclude SECONDARY,QCFAIL,DUP,SUPPLEMENTARY
- Region: chr1:25272486-25420935

- c.307 counts: CCAGTTCCCT**T**CTGGGAAGGT|ACCTTCCCAG**A**AGGGAACTGG or ACCTTCCCAG**G**AGGGAACTGG|CCAGTTCCCT**C**CTGGGAAGGT
- ins109 full counts TGCAATGAGCTATGATTGTACCACTGGGAAGTGACAAAGGGCACCCTGGGGGATTTCAAATGGTGGTGGCCCTGG TTTGGTGTTGCTGCCAGGTGAGTCCTTAAGCTATA|TATAGCTTAAGGACTCACCTGGCAGCAACACCAAACCAG GGCCACCACCATTTGAAATCCCCCAGGGTGCCCTTTGTCACTTCCCAGTGGTACAATCATAGCTCATTGCA (not working)
- ins109 counts of halves TGCAATGAGCTATGATTGTACCACTGGGAAGTGACAAAGGGCACCCTGGGGGATT|TCAAATGGTGGTGGCCCTG GTTTGGTGTTGCTGCCAGGTGAGTCCTTAAGCTATA|TATAGCTTAAGGACTCACCTGGCAGCAACACCAAACCA GGGCCACCACCATTTGA|AATCCCCCAGGGTGCCCTTTGTCACTTCCCAGTGGTACAATCATAGCTCATTGCA
- c.1136 counts: ATAGCTCTCA**T**GTCTGGTCTC|GAGACCAGAC**A**TGAGAGCTAT or GAGACCAGAC**G**TGAGAGCTAT|ATAGCTCTCA**C**GTCTGGTCTC

RH-system: Steps

1. *RHD*/*RHCE* start plus end and exon4-exon7 copy number (first approximation)
2. Do we need to extend the size of the insertion of one gene into another?
3. Taking into account 1., 2., c.307 and 109bp insertion counts, infer genotypes for two c.307
4. Taking into account *RHD*/*RHCE* start plus end and exon8 copy number, c.1136 counts, infer genotype of c.1136T>C in *RHD*

RH-system: Steps

1. *RHD*/*RHCE* start plus end and exon4-exon7 copy number (first approximation)

Number of samples	RHD_cn	RHCE_cn	RHCE_exons_in_RHD	RHD_exons_in_RHCE
119	2	2	None	None
104	1	2	None	None
39	0	2	None	None
9	2	2	4-7	None
5	2	2	None	4-7
5	3	2	None	None
4	1	2	4-7	None
1	0	2 or 3	None	None
1	2	2	4-7,4-7	None
1	2	3	None	None

RH-system: Steps

1. *RHD*/*RHCE* start plus end and exon4-exon7 copy number (first approximation)

two copies of *RHD* and two copies of *RHCE*

Number of samples	RHD_cn	RHCE_cn	RHCE_exons_in_RHD	RHD_exons_in_RHCE
119	2	2	None	None
104	1	2	None	None
39	0	2	None	None
9	2	2	4-7	None
5	2	2	None	4-7
5	3	2	None	None
4	1	2	4-7	None
1	0	2 or 3	None	None
1	2	2	4-7,4-7	None
1	2	3	None	None

RH-system: Steps

1. *RHD*/*RHCE* start plus end and exon4-exon7 copy number (first approximation)

two copies of *RHD* and two copies of *RHCE*

CE exons 4-7 RHD*01N.07

Number of samples	RHD_cn	RHCE_cn	RHCE_exons_in_RHD	RHD_exons_in_RHCE
119	2	2	None	None
104	1	2	None	None
39	0	2	None	None
9	2	2	4-7	None
5	2	2	None	4-7
5	3	2	None	None
4	1	2	4-7	None
1	0	2 or 3	None	None
1	2	2	4-7,4-7	None
1	2	3	None	None

RH-system: Steps

2. Do we need to extend the size of the insertion of one gene into another? Consider **intron3**, **exon3**, **exon2** and if their copy number matches the insertion, extend but
    a. Take into account if 109bp insertion was detected (so whether to extend to exon 2 or if it is a case of mis-mapping)
    b. Check if it's a border case on the boundary of copy number bins

RH-system: Steps

Number of samples	RHD_cn	RHCE_cn	RHCE_exons_in_RHD	RHD_exons_in_RHCE
119	2	2	None	None
104	1	2	None	None
39	0	2	None	None
5	2	2	3i-7	None
5	2	2	None	2-7
5	3	2	None	None
4	2	2	4-7	None
2	1	2	2-7	None
1	0	2 or 3	None	None
1	1	2	3-7	None
1	1	2	4-7	None
1	2	2	3i-7,4-7	None
1	2	3	None	None

RH-system: Steps

two copies of *RHD* and two copies of *RHCE*

Number of samples	RHD_cn	RHCE_cn	RHCE_exons_in_RHD	RHD_exons_in_RHCE
119	2	2	None	None
104	1	2	None	None
39	0	2	None	None
5	2	2	3i-7	None
5	2	2	None	2-7
5	3	2	None	None
4	2	2	4-7	None
2	1	2	2-7	None
1	0	2 or 3	None	None
1	1	2	3-7	None
1	1	2	4-7	None
1	2	2	3i-7,4-7	None
1	2	3	None	None

RH-system: Steps

two copies of *RHD* and two copies of *RHCE*

CE exons 4-7 RHD*01N.07

Number of samples	RHD_cn	RHCE_cn	RHCE_exons_in_RHD	RHD_exons_in_RHCE
119	2	2	None	None
104	1	2	None	None
39	0	2	None	None
5	2	2	3i-7	None
5	2	2	None	2-7
5	3	2	None	None
4	2	2	4-7	None
2	1	2	2-7	None
1	0	2 or 3	None	None
1	1	2	3-7	None
1	1	2	4-7	None
1	2	2	3i-7,4-7	None
1	2	3	None	None

RH-system: Steps

two copies of *RHD* and two copies of *RHCE*

CE exons 2-7 RHD*01N.05

CE exons 4-7 RHD*01N.07

Number of samples	RHD_cn	RHCE_cn	RHCE_exons_in_RHD	RHD_exons_in_RHCE
119	2	2	None	None
104	1	2	None	None
39	0	2	None	None
5	2	2	3i-7	None
5	2	2	None	2-7
5	3	2	None	None
4	2	2	4-7	None
2	1	2	2-7	None
1	0	2 or 3	None	None
1	1	2	3-7	None
1	1	2	4-7	None
1	2	2	3i-7,4-7	None
1	2	3	None	None

Summary of results

- Panel performance: 450bp mean bait coverage, 2% uncovered bases, 99.2% genotype concordance in duplicate samples

- Review of the NGS results for array vs. serology discordances: RH (24), MNS, JK (5 each), FY, LU (4 each), DO and KEL (1 each) systems and 3 HPA antigen discordances => confirmed the results of one of the previous tests (serology, array, MLPA) including RHD-RHCE(3-7)-RHD and RHCE-RHD(2-7)-RHCE hybrids and MUR hybrid allele of the MNS system