USTAR Center for Genetic Discovery

rufus

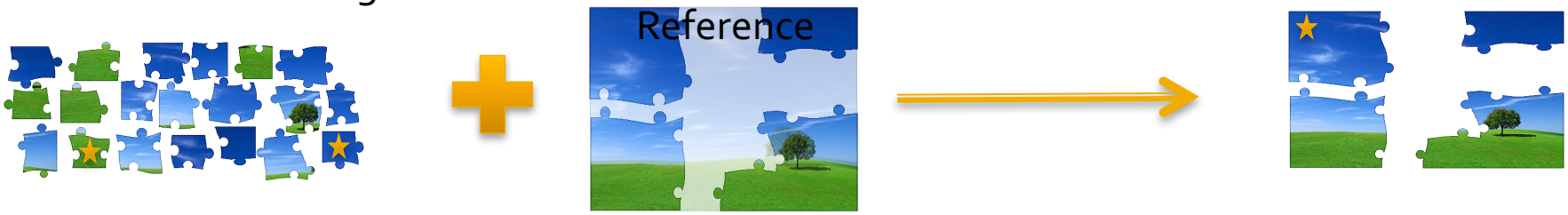
Reference independent variant detection

Andrew Farrell, PhD
Research Associate
The Marth Lab
University of Utah
Ustar Center for Genetic Discovery

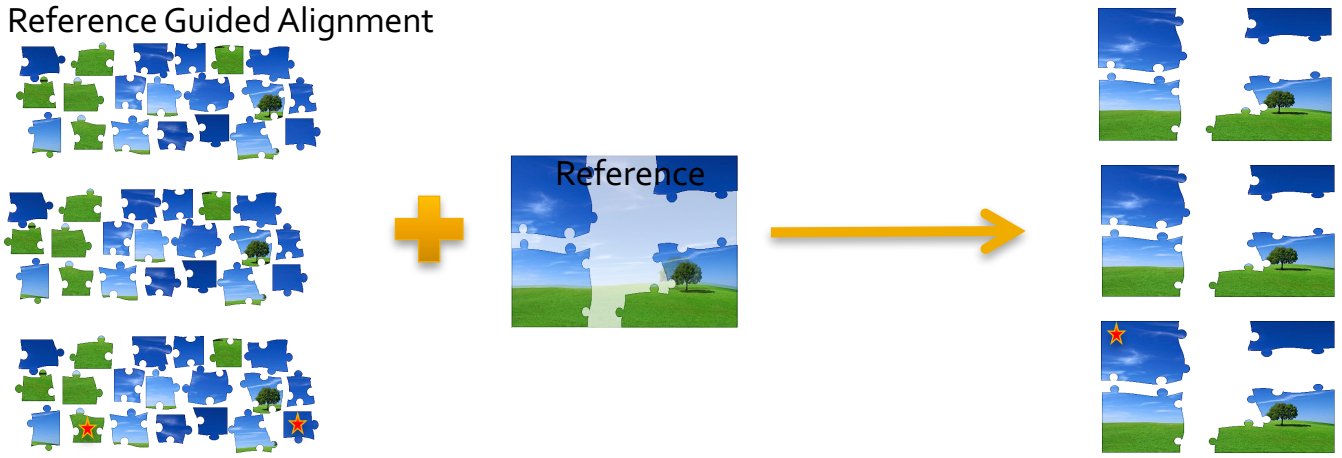
Whole Genome Analysis Methods

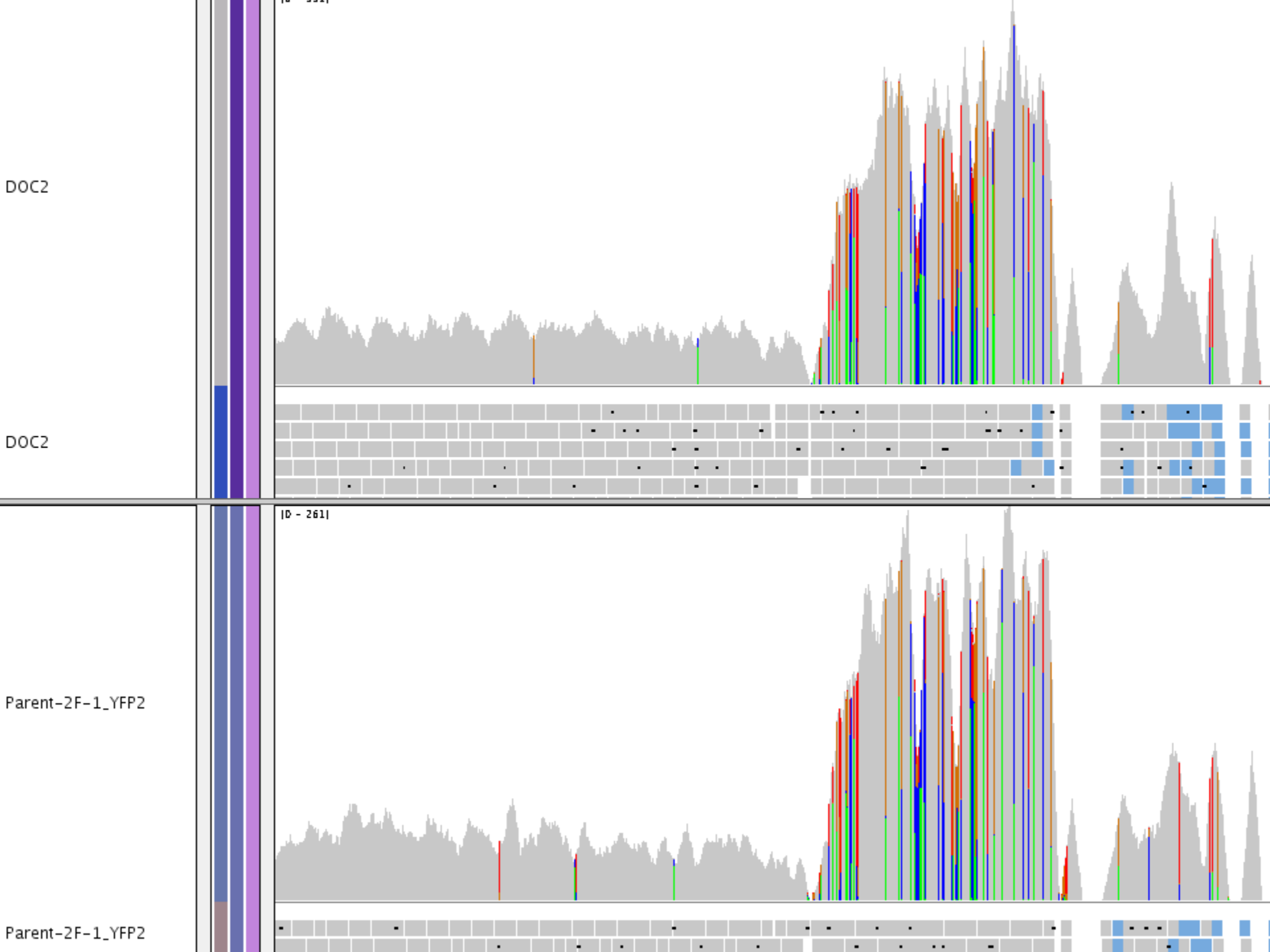
De Novo Assembly

Reference Guided Alignment



Whole Genome Analysis Methods




DOC2
DOC2
[0 - 261]
Parent-2F-1_YFP2
Parent-2F-1_YFP2

Reference Guided Alignment

- Limitations
  - Reference sequence is required
  - Reference errors
  - Reference bias
    - Bias against mutations
  - Reads must be reasonably unique - mappable

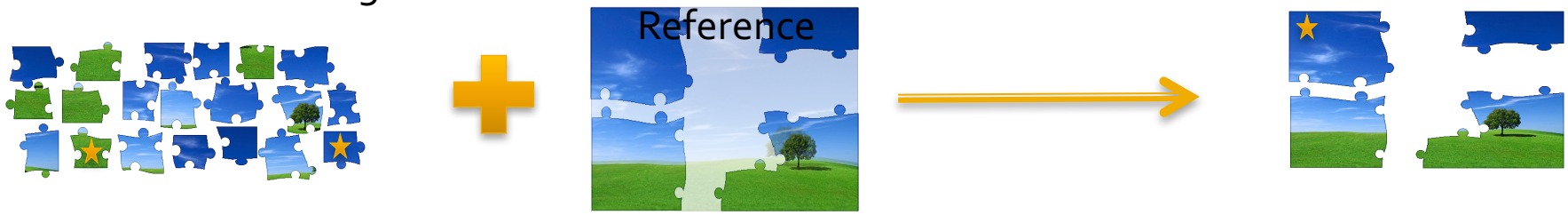
De Novo Assembly

- Limitations
  - Difficult to reconstruct repetitive regions
  - Multiple sequencing library preparations and addition segueing required
  - Computationally expensive
  - Rare variants tend to get lost

Whole Genome Analysis Methods

De Novo Assembly

Reference Guided Alignment

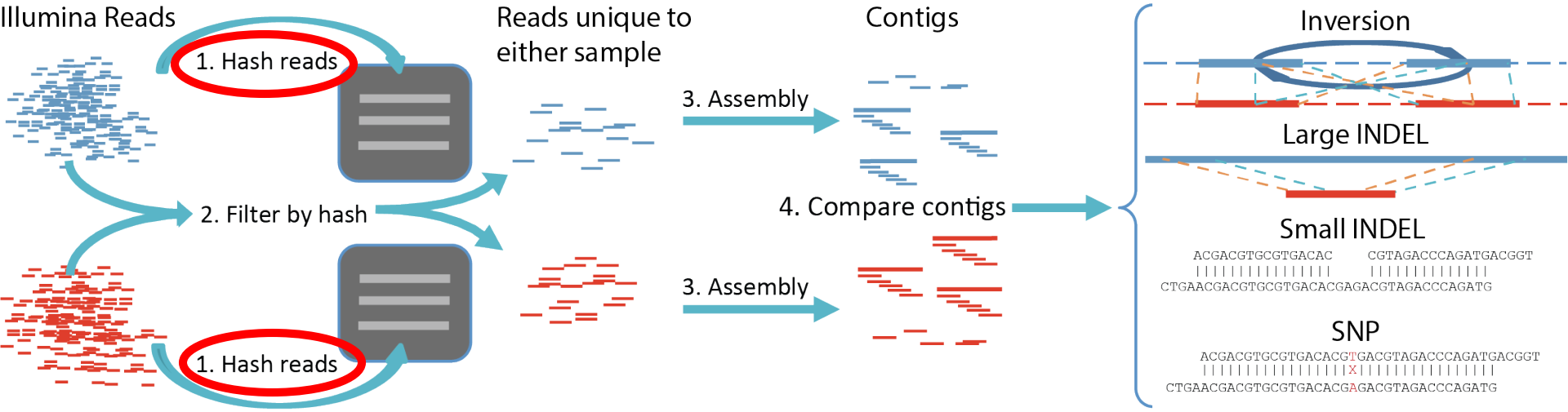


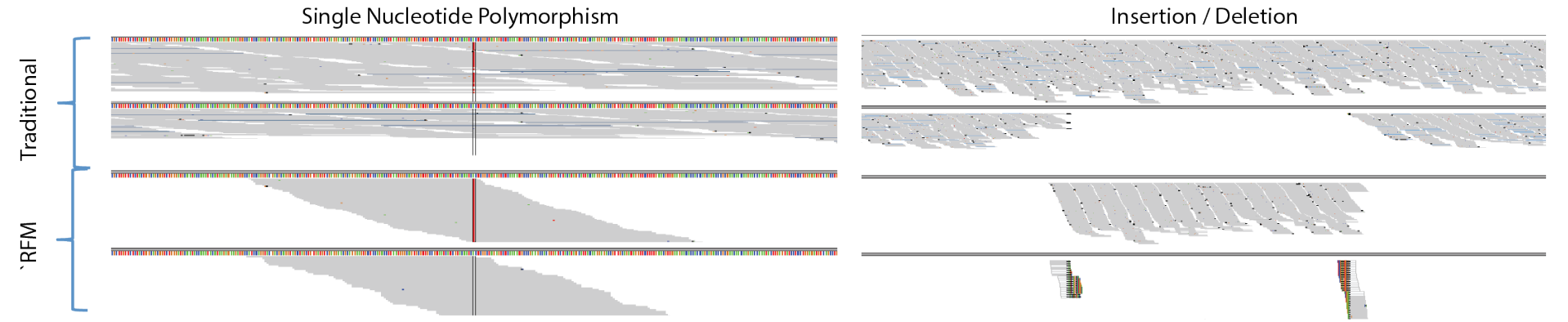
rufus

RUFUS: Reference Free Variant Detection

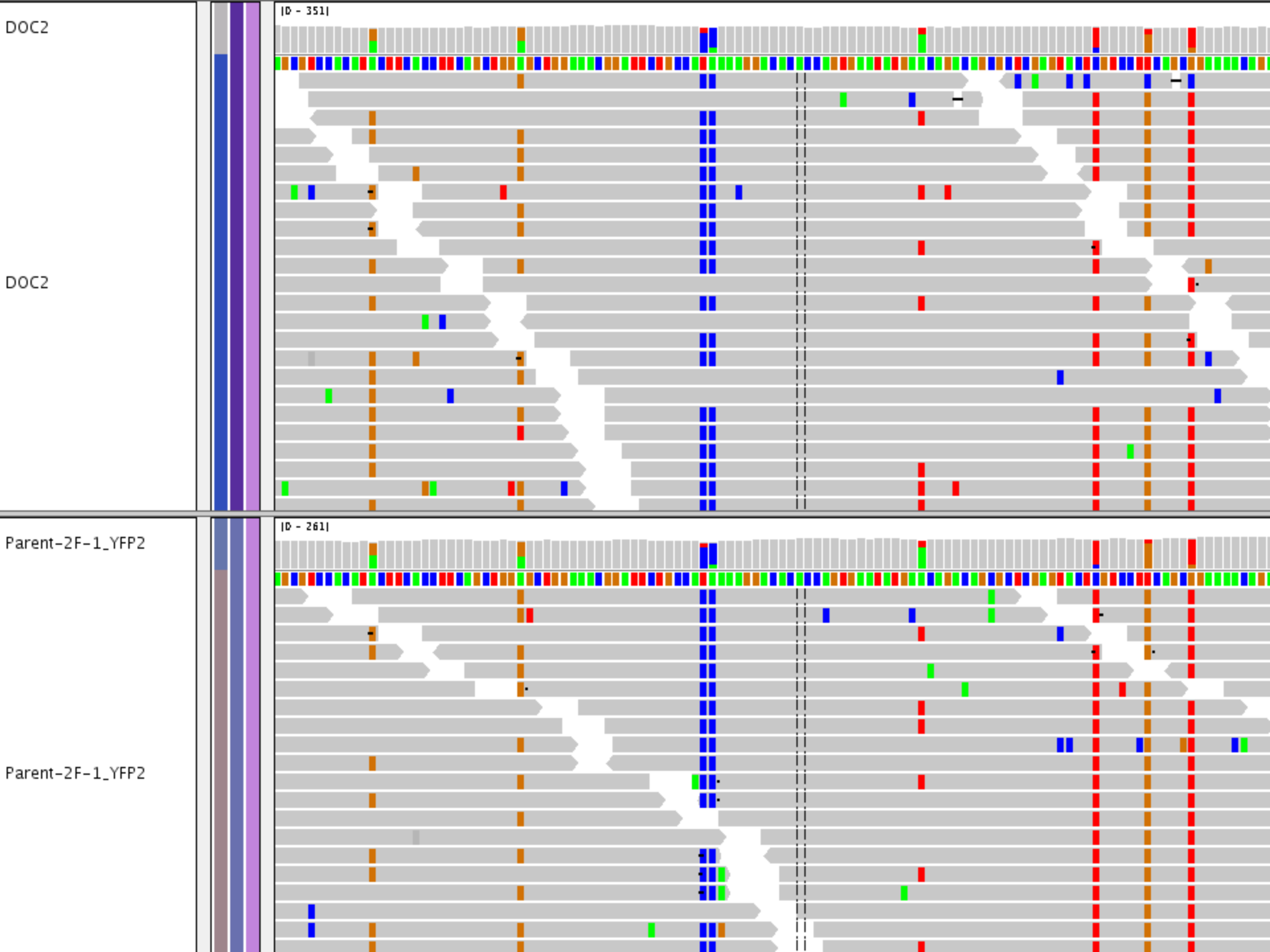
- Variants are identified by directly comparing raw sequence reads between samples
- Identify variant reads before determining context
  - Eliminates mapping errors
  - Not restricted to the "Mappable Genome"
  - No assembled reference needed: Works in any organism that can be sequenced
- Reads identified by dissimilarity

RUFUS: read driven analysis

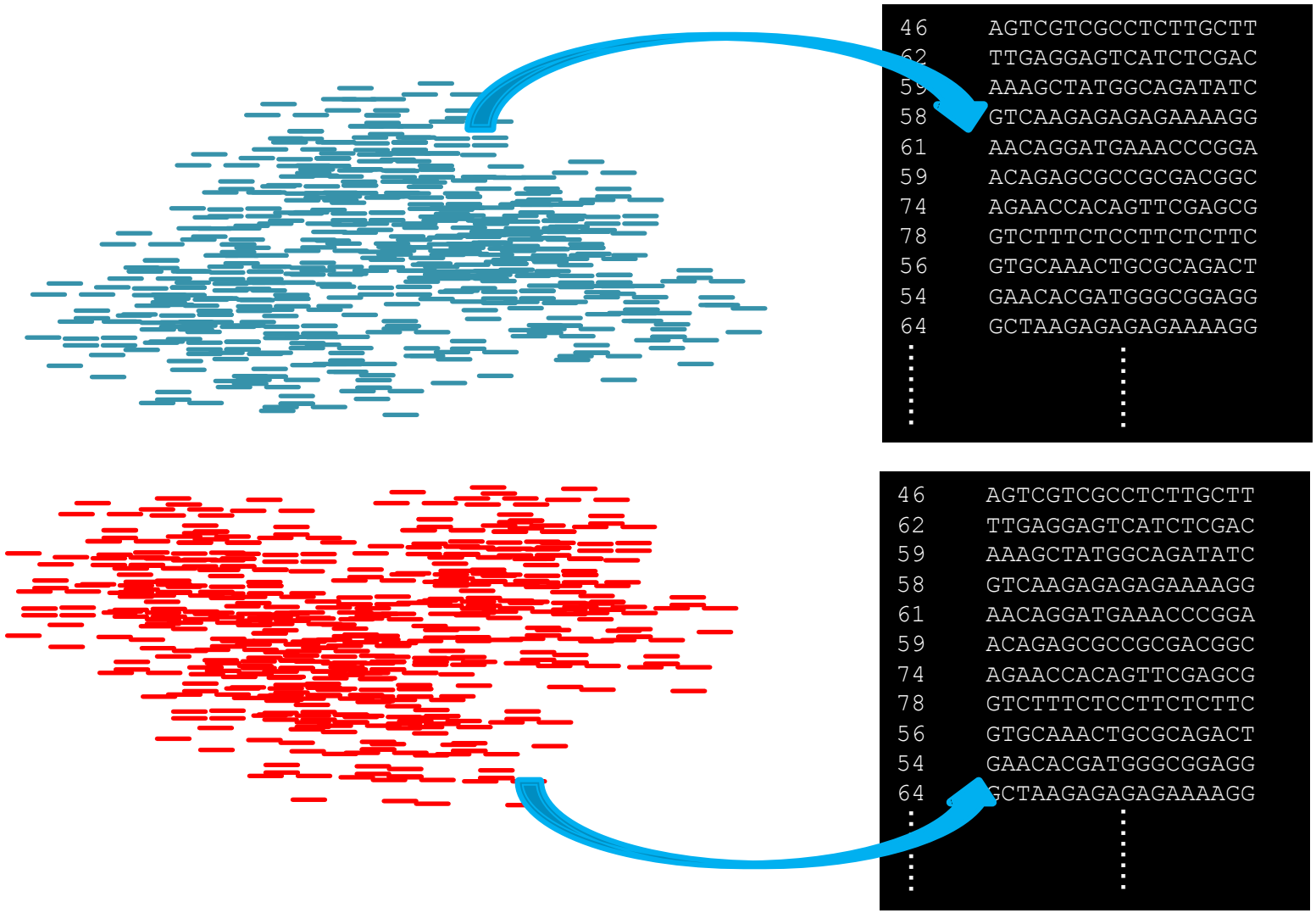




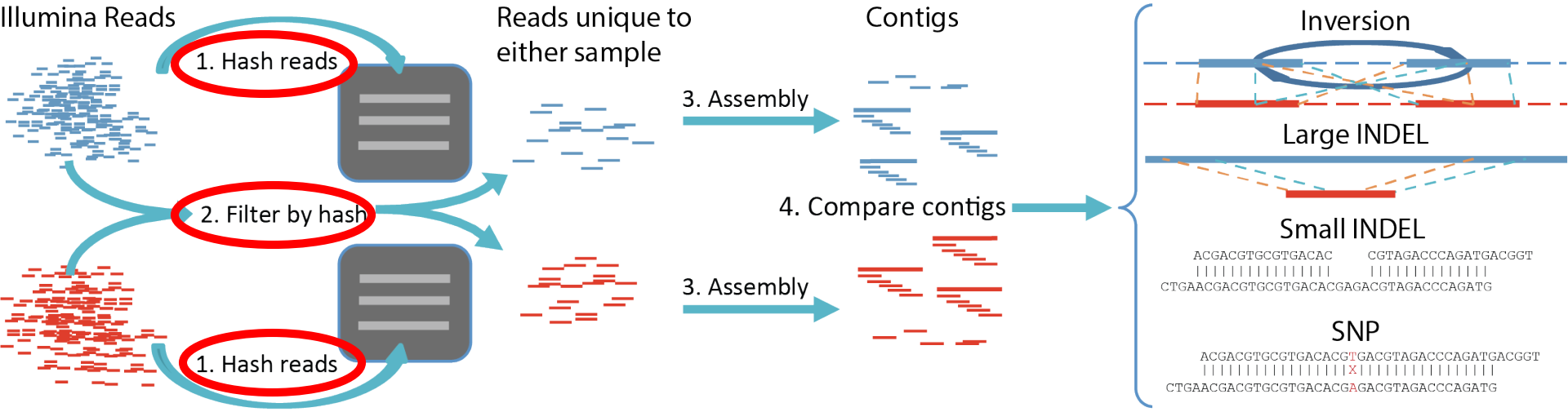

DOC2
DOC2
Parent-2F-1_YFP2
Parent-2F-1_YFP2

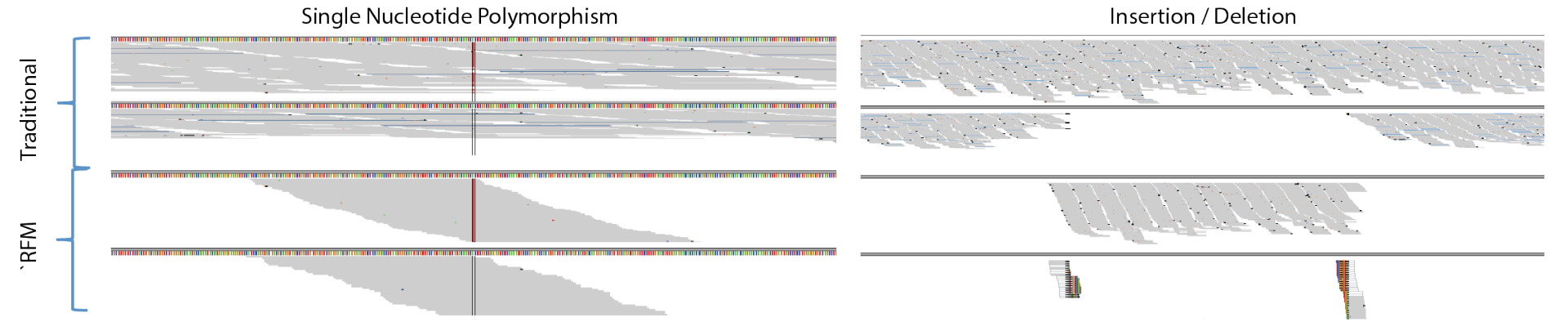

DOC2
DOC2
Parent-2F-1_YFP2
Parent-2F-1_YFP2

1: Use raw reads to create K-mer table



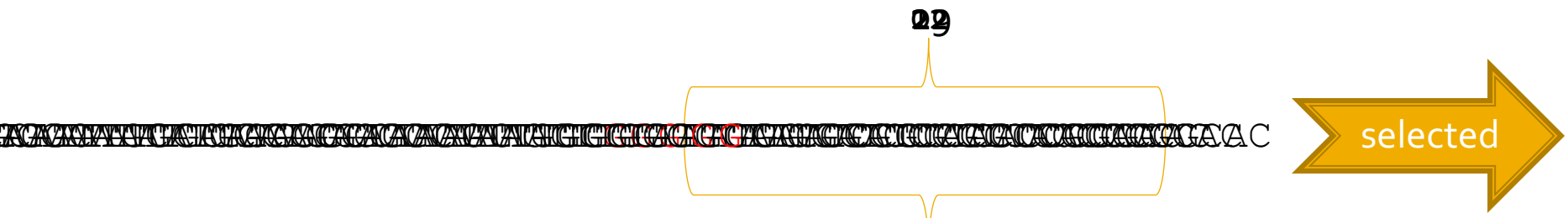
RUFUS: read driven analysis





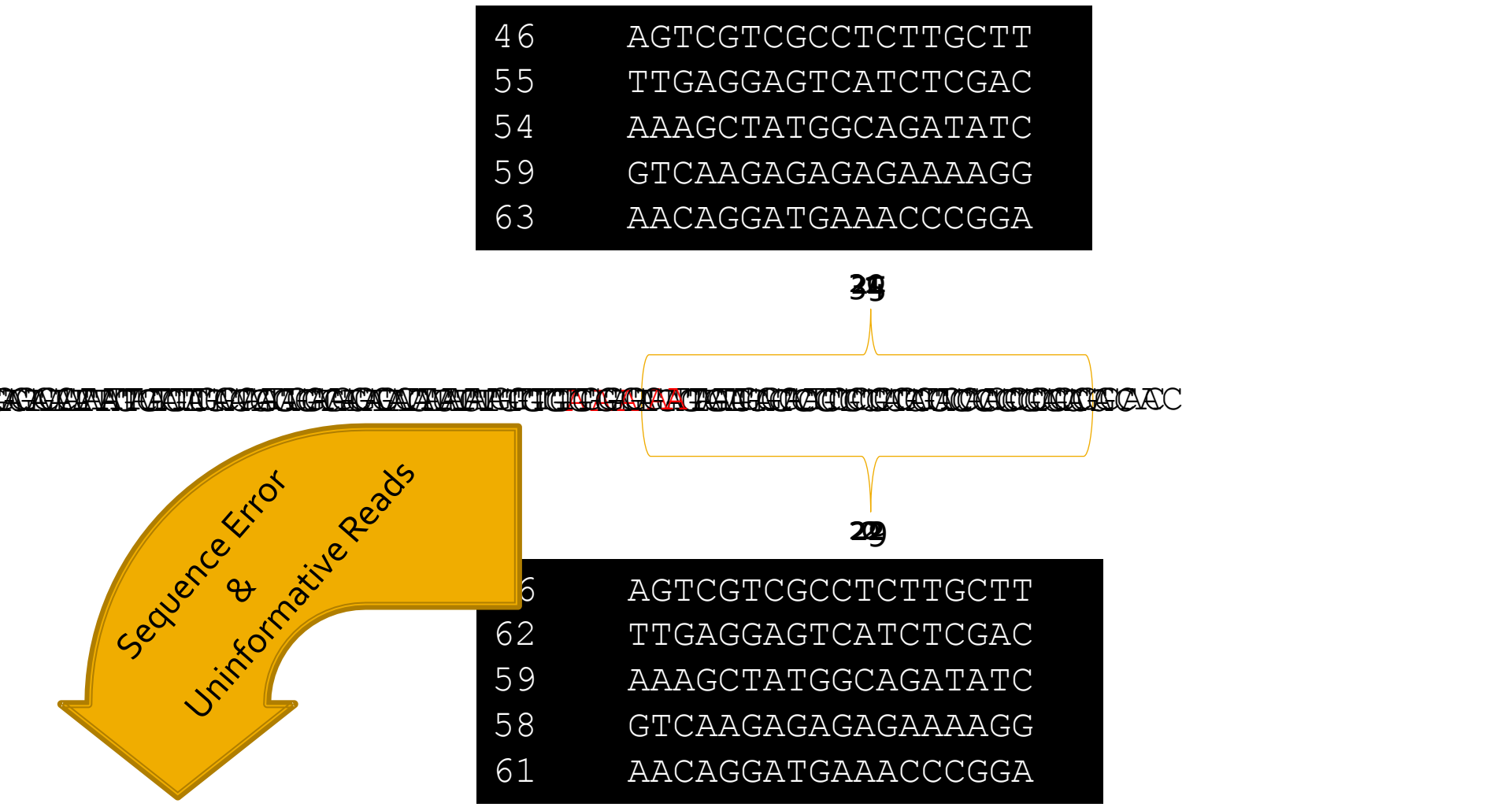
Step 2: Filtering

46	AGTCGTCGCCTCTTGCTT
55	TTGAGGAGTCATCTCGAC
54	AAAGCTATGGCAGATATC
59	GTCAAGAGAGAGAAAAGG
63	AACAGGATGAAACCCGGA

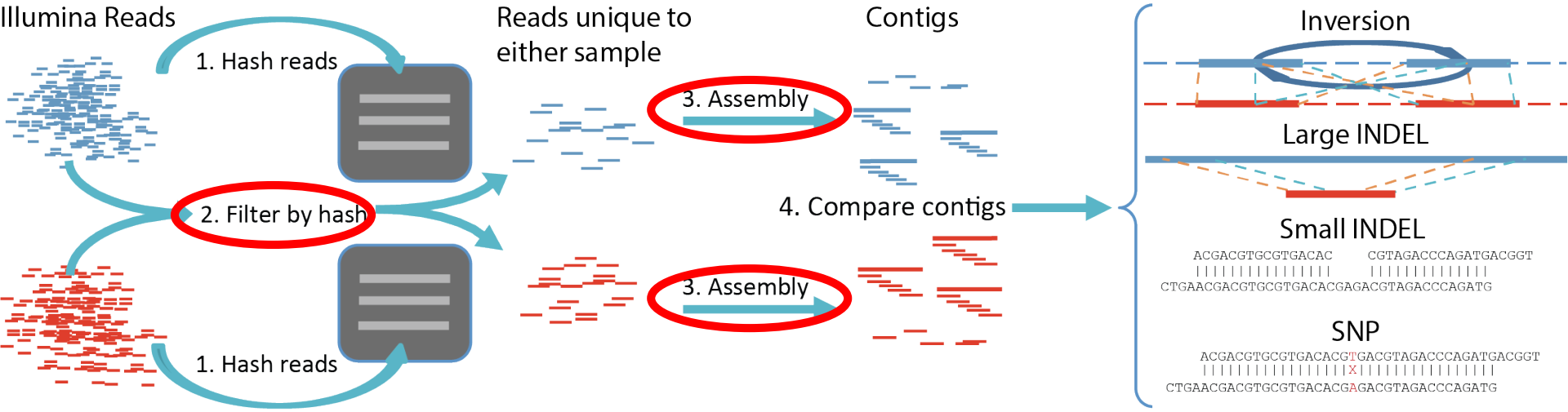


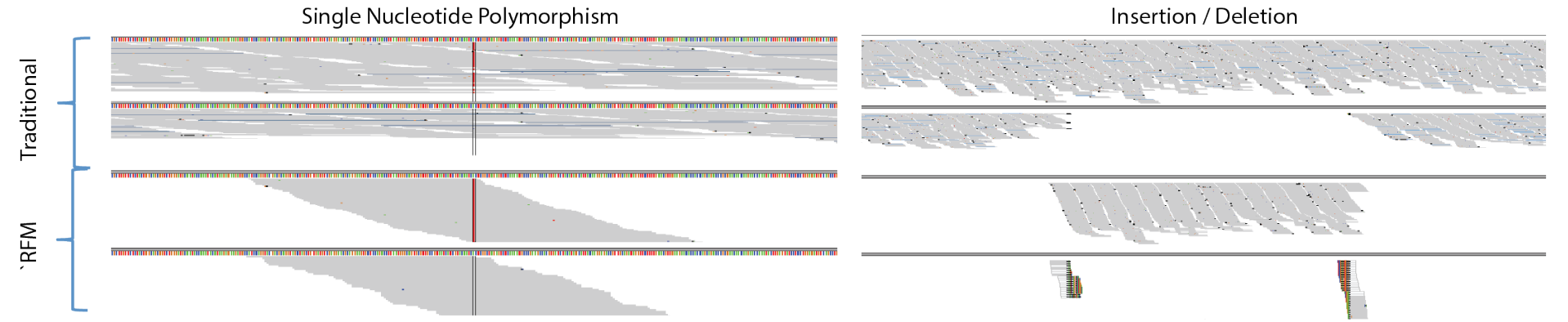
46	AGTCGTCGCCTCTTGCTT
62	TTGAGGAGTCATCTCGAC
59	AAAGCTATGGCAGATATC
58	GTCAAGAGAGAGAAAAGG
61	AACAGGATGAAACCCGGA

Step 2: Filtering Sequence Error

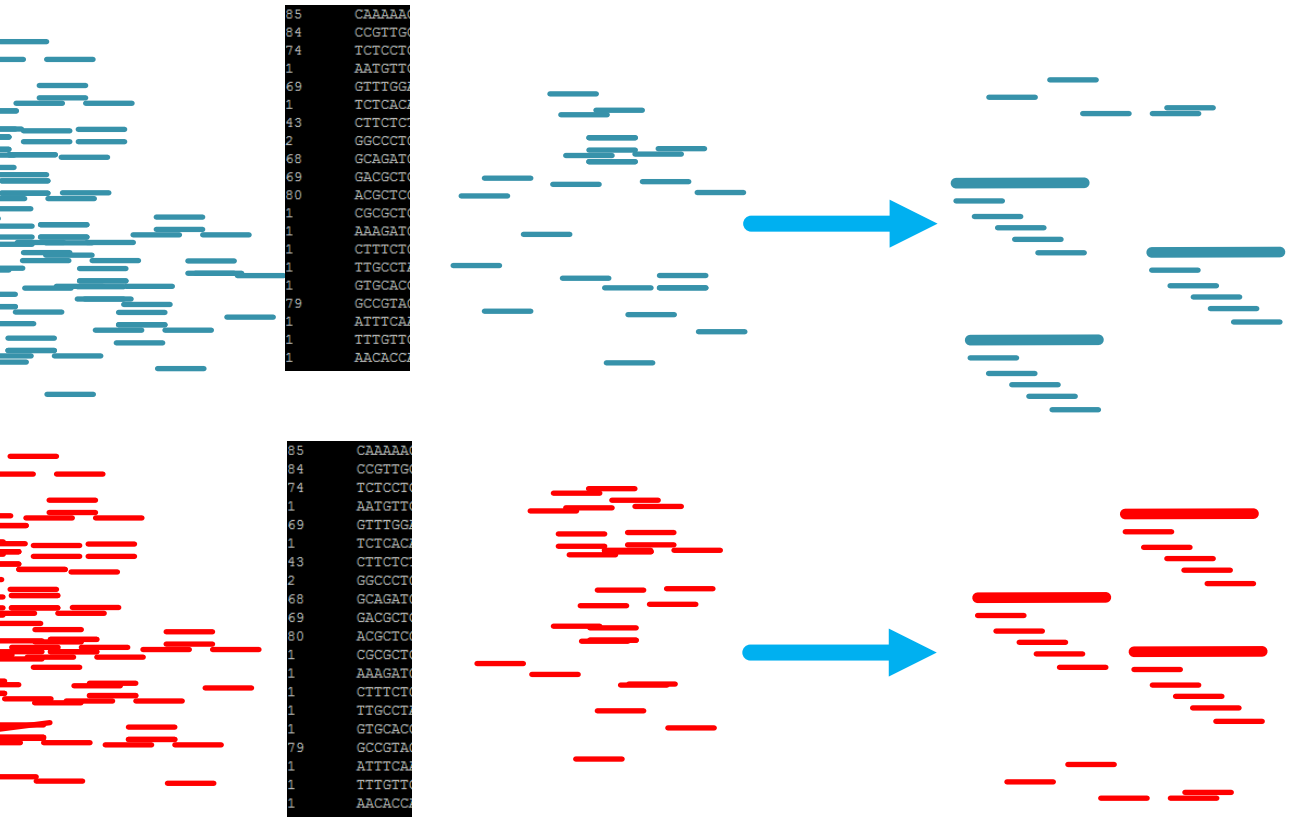


RUFUS: read driven analysis



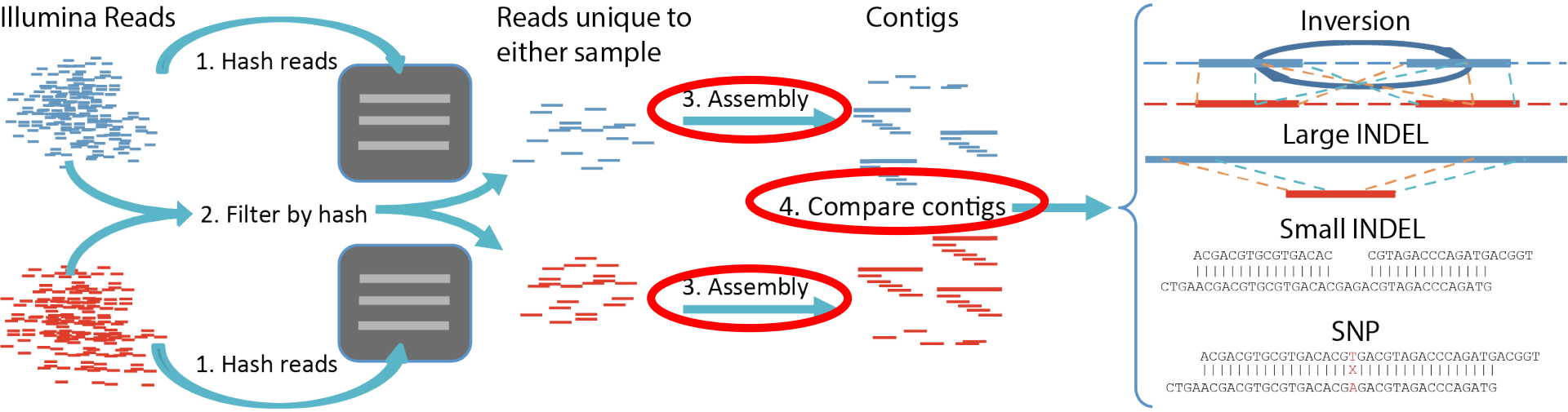


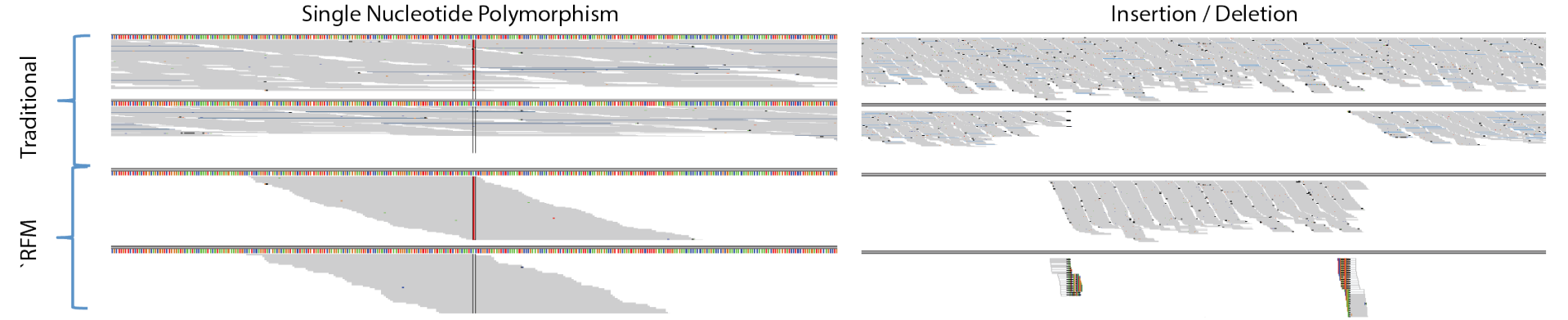
3: Assembly



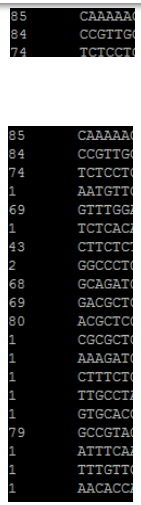
- Greatly reduced library
  - Over 99.5% reduction
  - Allows very sensitive computationally expensive methods
- Removed repetitive elements

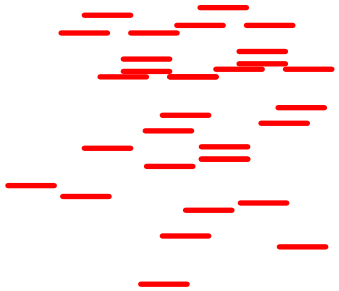
RUFUS: read driven analysis

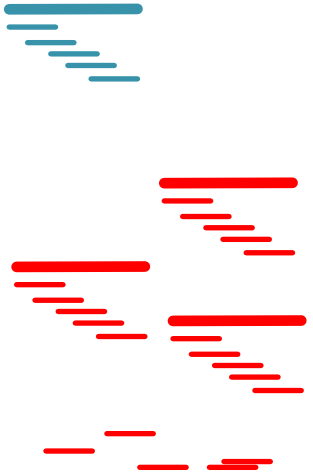




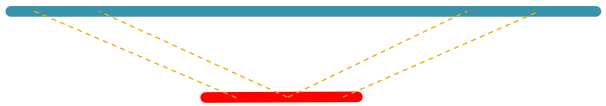
4: Compare contigs







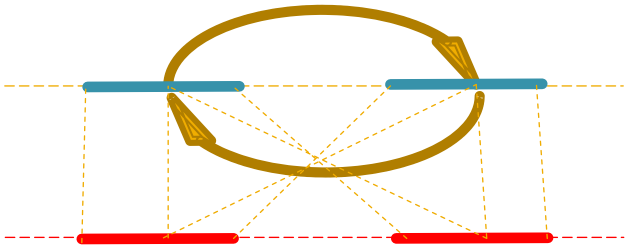
Large INDEL



SNP

```
ACGACGTGCGTGACACGTGACGTAGACCCAGATGACGGT
|||||||||||||||||| ||||||||||||||||||
CTGAACGACGTGCGTGACACGAGACGTAGACCCAGATG
```

Inversion

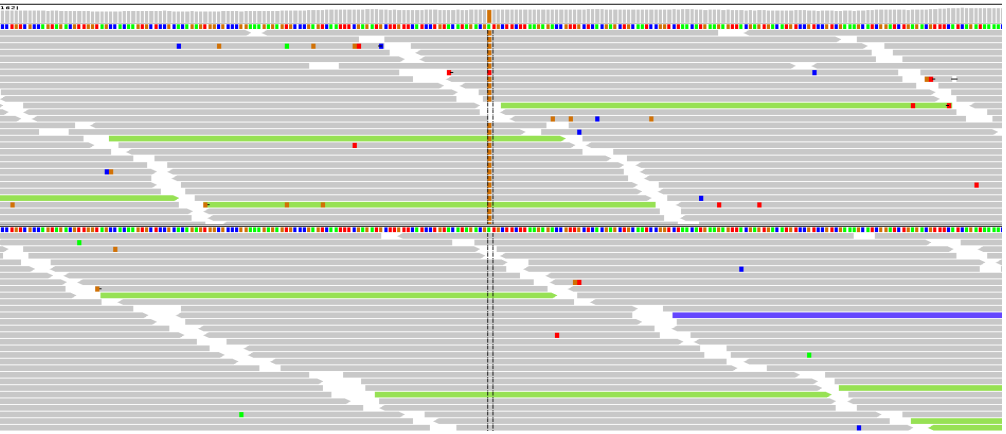


Small INDEL

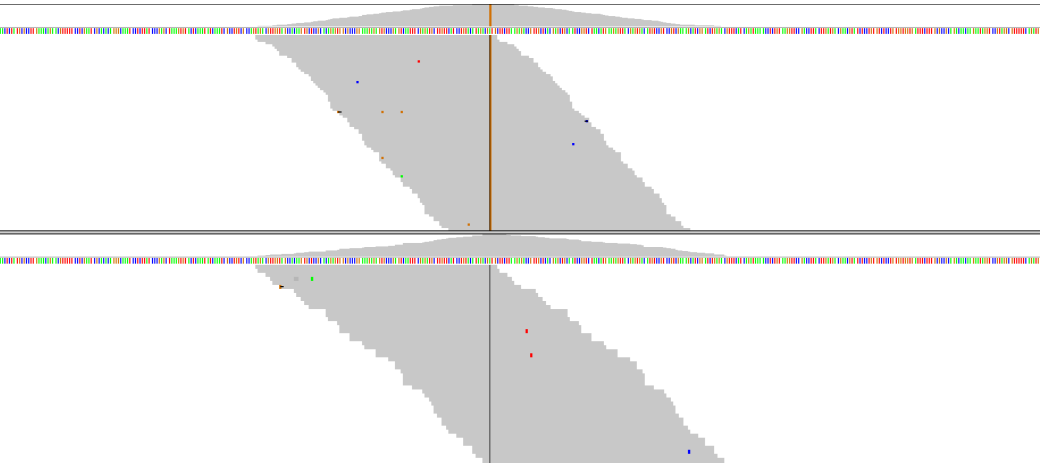
```
ACGACGTGCGTGACAC     CGTAGACCCAGATGACGGT
|||||||||||||||| ||||||||||||||||
CTGAACGACGTGCGTGACACGAGACGTAGACCCAGATG
```

RESULTS: SNP example

- Traditional Mapping



- RUFUS Filtered Reads

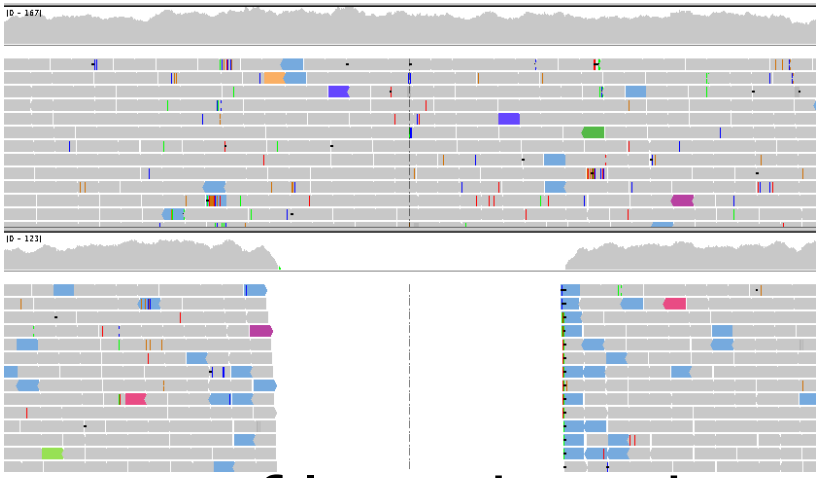


- RUFUS SNP call

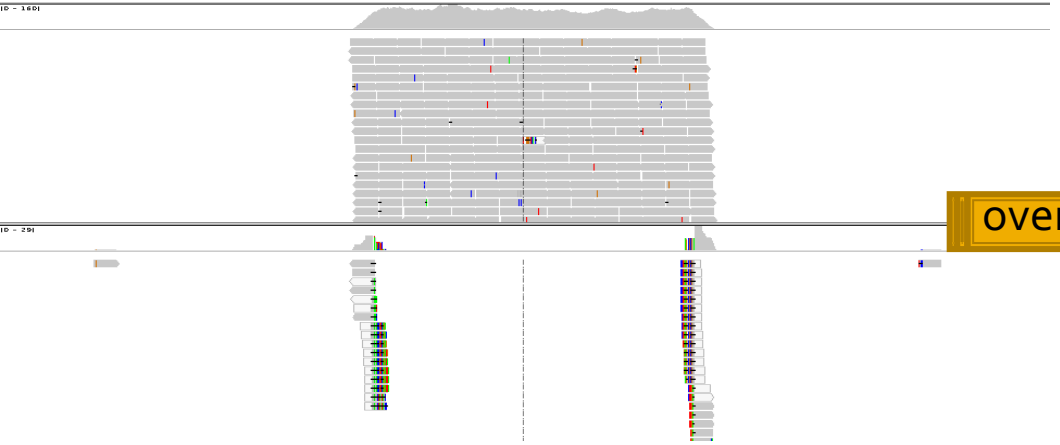
```
Query= @NODE_177_193

Length=193

                                                                      Score     E
Sequences producing significant alignments:                           (Bits)  Value

  @NODE_869_188                                                        340     3e-95


> @NODE_869_188
Length=188

 Score =  340 bits (184),  Expect = 3e-95
 Identities = 186/187 (99%), Gaps = 0/187 (0%)
 Strand=Plus/Minus

Query  1    CTCAGATCCGATGCTTTCGACGGACGGACATGTCACTCGTAACTCTTCATGTTAGACCGG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  187  CTCAGATCCGATGCTTTCGACGGACGGACATGTCACTCGTAACTCTTCATGTTAGACCGG  128

Query  61   ATTGCAGCTCGTGAGCAACGGTCTCTTAAGAAGCACCGTCTATGTACAGGATGAACAAGC  120
            |||||||||||||||||||||||||||||||| ||| ||||||||||||||||||||||
Sbjct  127  ATTGCAGCTCGTGAGCAACGGTCTCTTAAGAAGCATCGTCTATGTACAGGATGAACAAGC  68

Query  121  ATCTCGAAATTGACTCGGGCACTCTCTTTCCGGGCGCCACTCTGTGCGGAGCACTTGTGG  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  67   ATCTCGAAATTGACTCGGGCACTCTCTTTCCGGGCGCCACTCTGTGCGGAGCACTTGTGG  8

Query  181  CTACCAA  187
            |||||||
Sbjct  7    CTACCAA  1
```

Insertion/Deletion Detection

- Traditional mapping

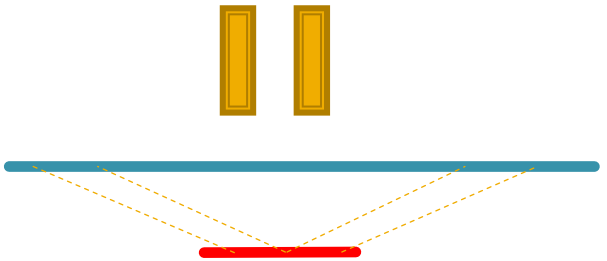


- RUFUS filtered reads

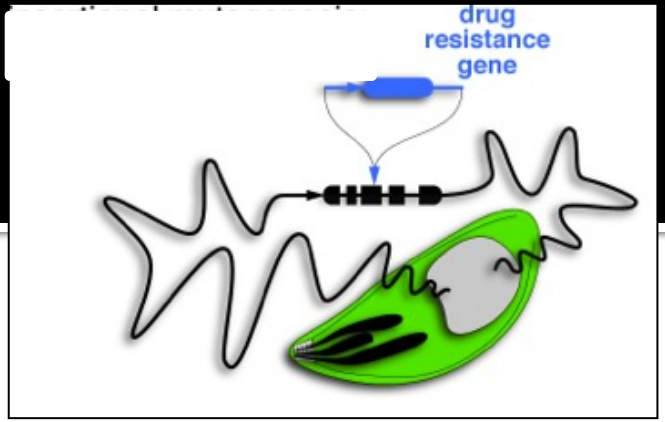


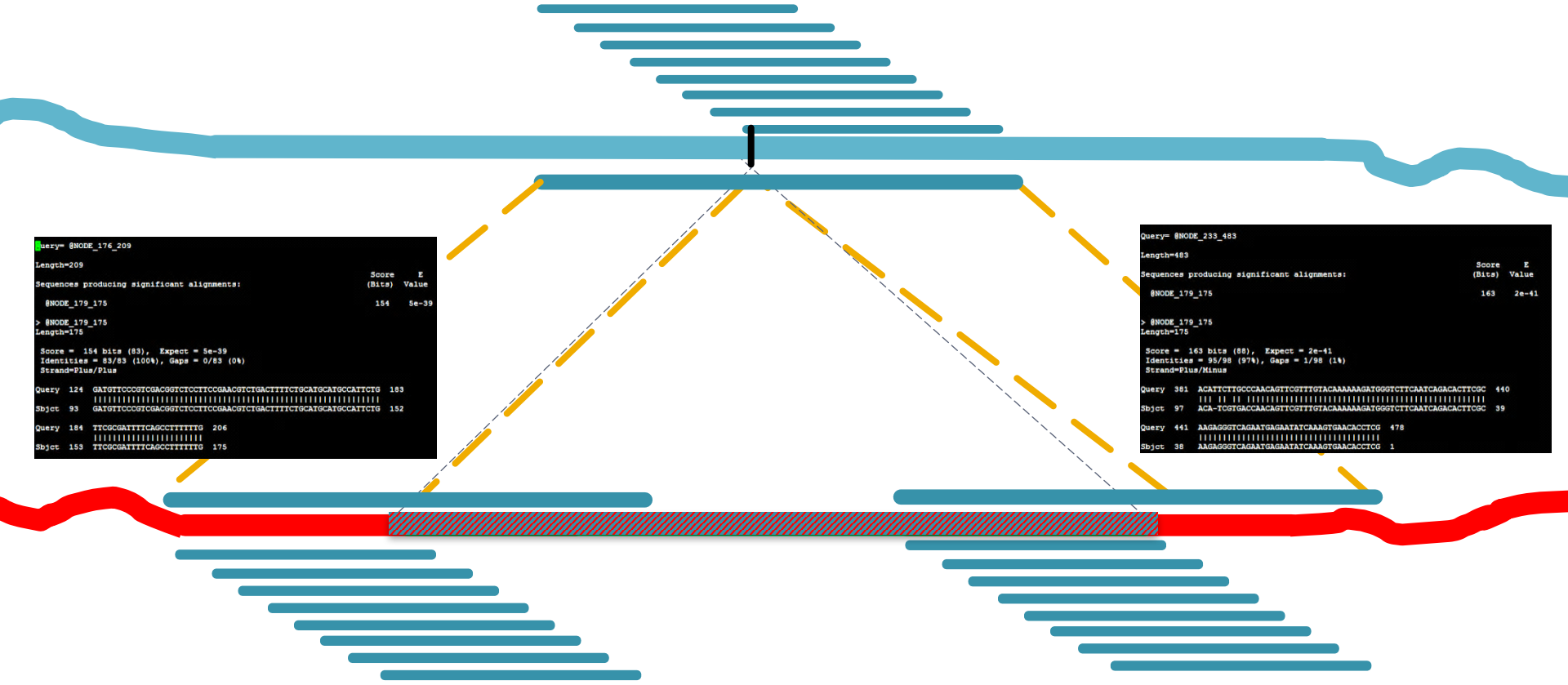
- RUFUS SNP call

```
FOUND HXTU... DELETION SITE gb|TGGT1_chrVIII:6,800,386-6,802,735
Query= @NODE_271_1607

Length=1607
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

  @NODE_59_169                                                         165    2e-41


> @NODE_59_169
Length=169

 Score =  165 bits (89),  Expect = 2e-41
 Identities = 89/89 (100%), Gaps = 0/89 (0%)
 Strand=Plus/Minus

Query  5    ATATTACCTCAGAACGGCTGATGCTACTGGTAAATCCGTATTCTGGAAAGATTTCCGGTG  64
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  169  ATATTACCTCAGAACGGCTGATGCTACTGGTAAATCCGTATTCTGGAAAGATTTCCGGTG  110

Query  65   GTGAGTTCAAGTGTAGTAACTCGTCGACG  93
            |||||||||||||||||||||||||||||
Sbjct  109  GTGAGTTCAAGTGTAGTAACTCGTCGACG  81


 Score =  156 bits (84),  Expect = 1e-38
 Identities = 84/84 (100%), Gaps = 0/84 (0%)
 Strand=Plus/Minus

Query  1524  GTCGACGATGTCCTCAACAATCAGAACGTGCTTGTCGCGAAAGATTGACAAGTCGTCGCT  1583
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  87    GTCGACGATGTCCTCAACAATCAGAACGTGCTTGTCGCGAAAGATTGACAAGTCGTCGCT  28

Query  1584  CAAGACGGTGAGCTGGCCTGTGCT  1607
              ||||||||||||||||||||||||
Sbjct  27    CAAGACGGTGAGCTGGCCTGTGCT  4
```



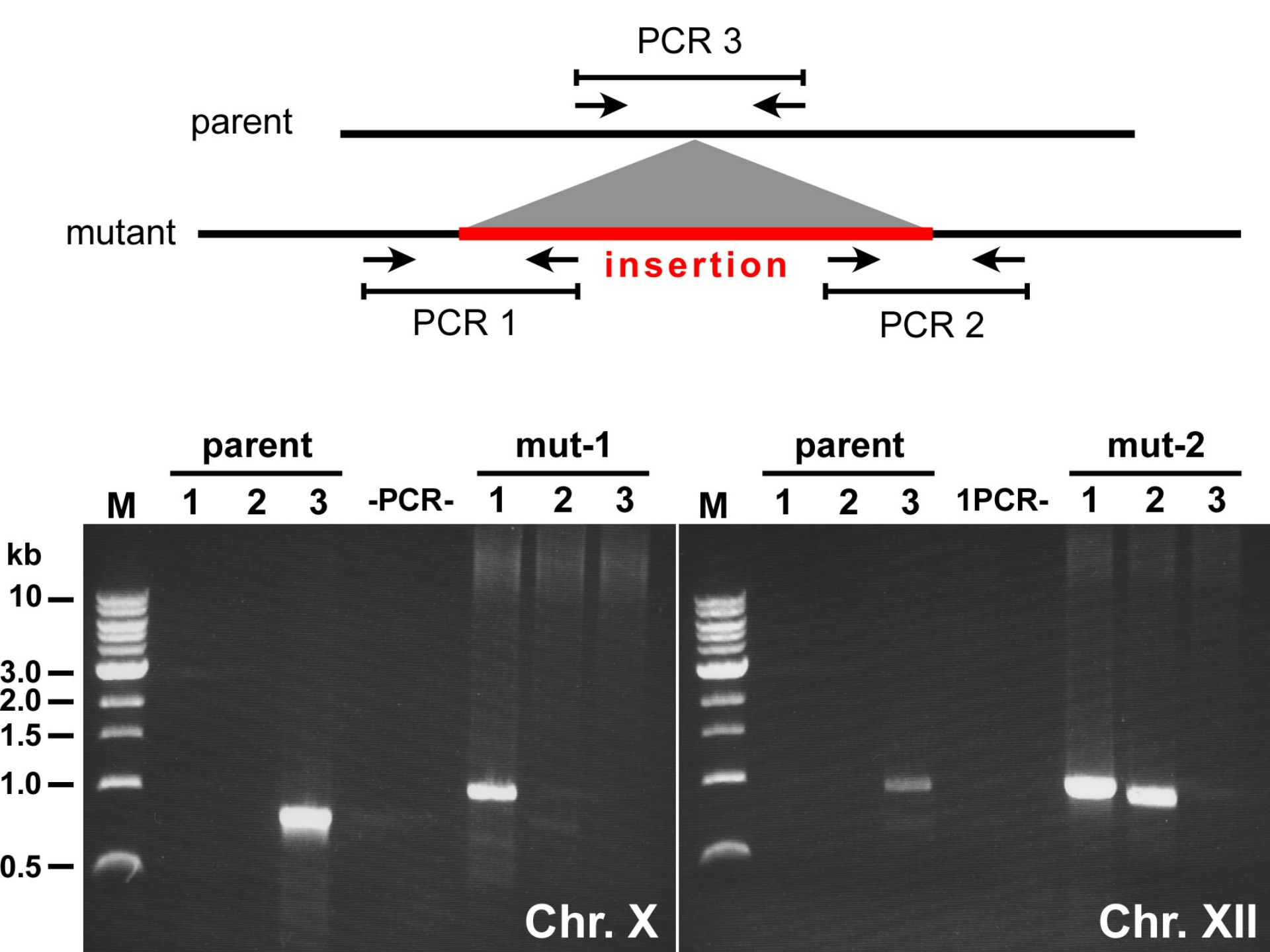
Plasmid Insertions





```
Query= @NODE_176_209

Length=209

Sequences producing significant alignments:                          Score     E
                                                                     (Bits)  Value

  @NODE_179_175                                                       154    5e-39

> @NODE_179_175
Length=175

 Score =  154 bits (83),  Expect = 5e-39
 Identities = 83/83 (100%), Gaps = 0/83 (0%)
 Strand=Plus/Plus

Query  124  GATGTTCCCGTCGACGGTCTCCTTCCGAACGTCTGACTTTTCTGCATGCATGCCATTCTG  183
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  93   GATGTTCCCGTCGACGGTCTCCTTCCGAACGTCTGACTTTTCTGCATGCATGCCATTCTG  152

Query  184  TTCGGGATTTTCAGCCTTTTTTG  206
             |||||||||||||||||||||||
Sbjct  153  TTCGCGATTTTCAGCCTTTTTTG  175
```

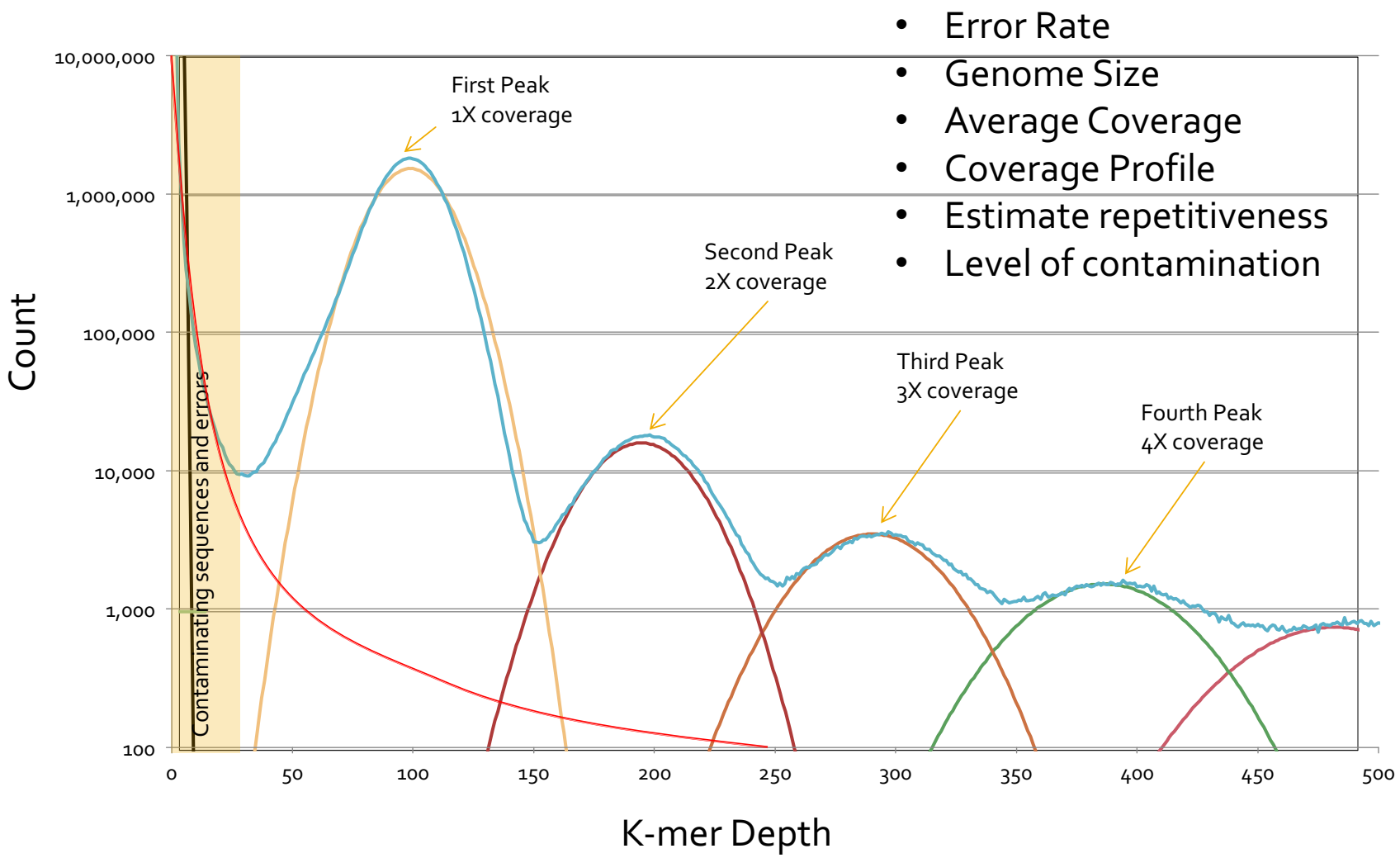
```
Query= @NODE_233_483

Length=483

Sequences producing significant alignments:                          Score     E
                                                                     (Bits)  Value

 @NODE_179_175                                                        163    2e-41


> @NODE_179_175
Length=175

 Score =  163 bits (88),  Expect = 2e-41
 Identities = 95/98 (97%), Gaps = 1/98 (1%)
 Strand=Plus/Minus

Query  381  ACATTCTTGCCCAACAGTTCGTTTGTACAAAAAAGATGGGTCTTCAATCAGACACTTCGC  440
            ||| || || ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  97   ACA-TCGTGACCAACAGTTCGTTTGTACAAAAAAGATGGGTCTTCAATCAGACACTTCGC  39

Query  441  AAGAGGGTCAGAATGAGAATATCAAAGTGAACACCTCG  478
            ||||||||||||||||||||||||||||||||||||||
Sbjct  38   AAGAGGGTCAGAATGAGAATATCAAAGTGAACACCTCG  1
```


PCR 3
parent
mutant
insertion
PCR 1
PCR 2
parent
mut-1
parent
mut-2
M
1
2
3
-PCR-
1
2
3
M
1
2
3
1PCR-
1
2
3
kb
10
3.0
2.0
1.5
1.0
0.5
Chr. X
Chr. XII

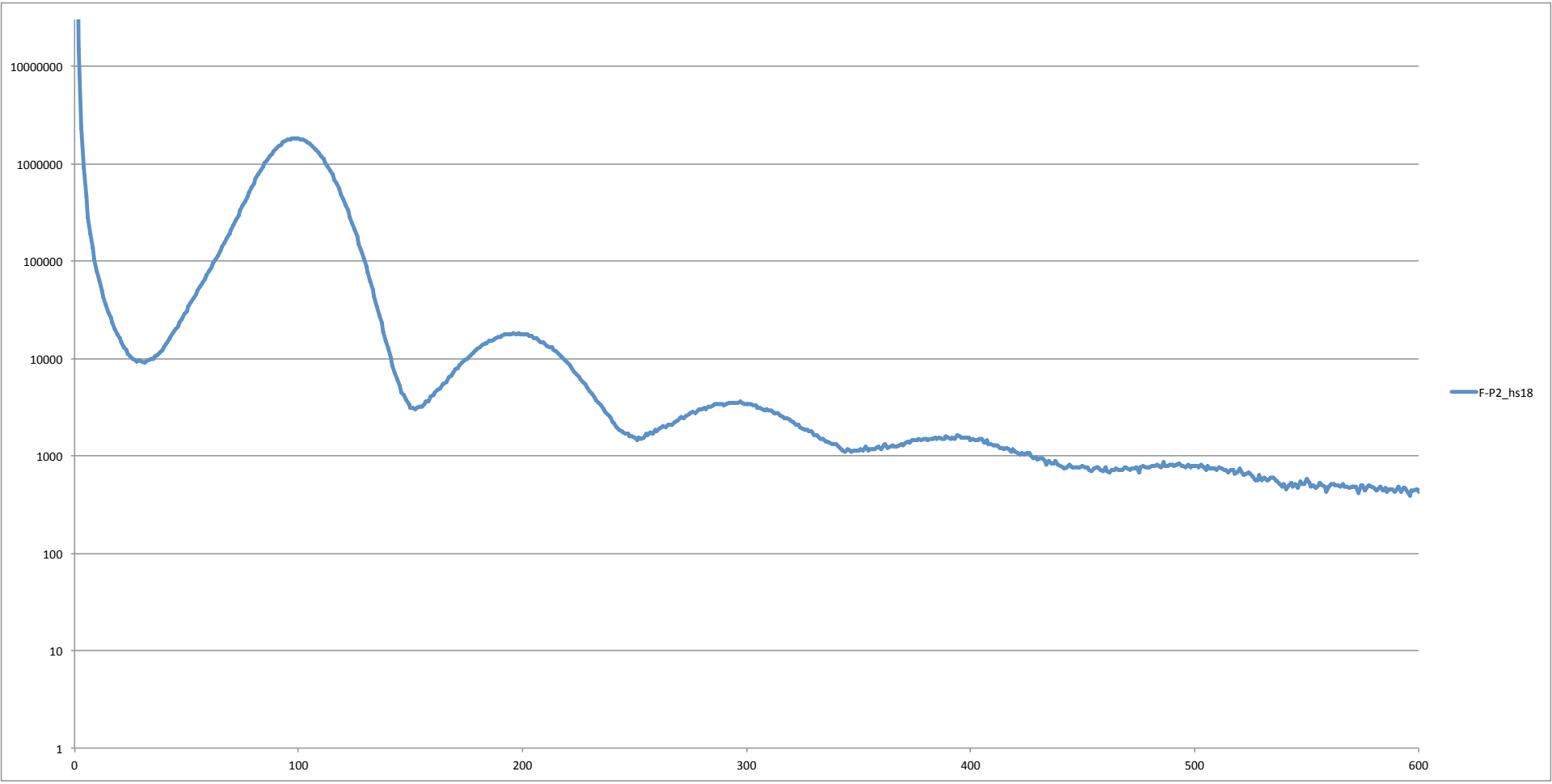
7 Not Mappable SNPs

- Missing From the GT1 Reference
  - NODE_53_176 : ME49 forkhead-associated domain-containing protein, mRNA
  - NODE_223_195: ME49 hypothetical protein, mRNA
- No BLAST hits in the NT database
  - NODE_127_183
  - NODE_138_182
- Repetitive Sequences
  - NODE_206_183: ME49 KH domain containing protein
  - NODE_163_189: ME49 hypothetical protein
  - NODE_85_169: Toxoplasma gondii repetitive DNA sequence

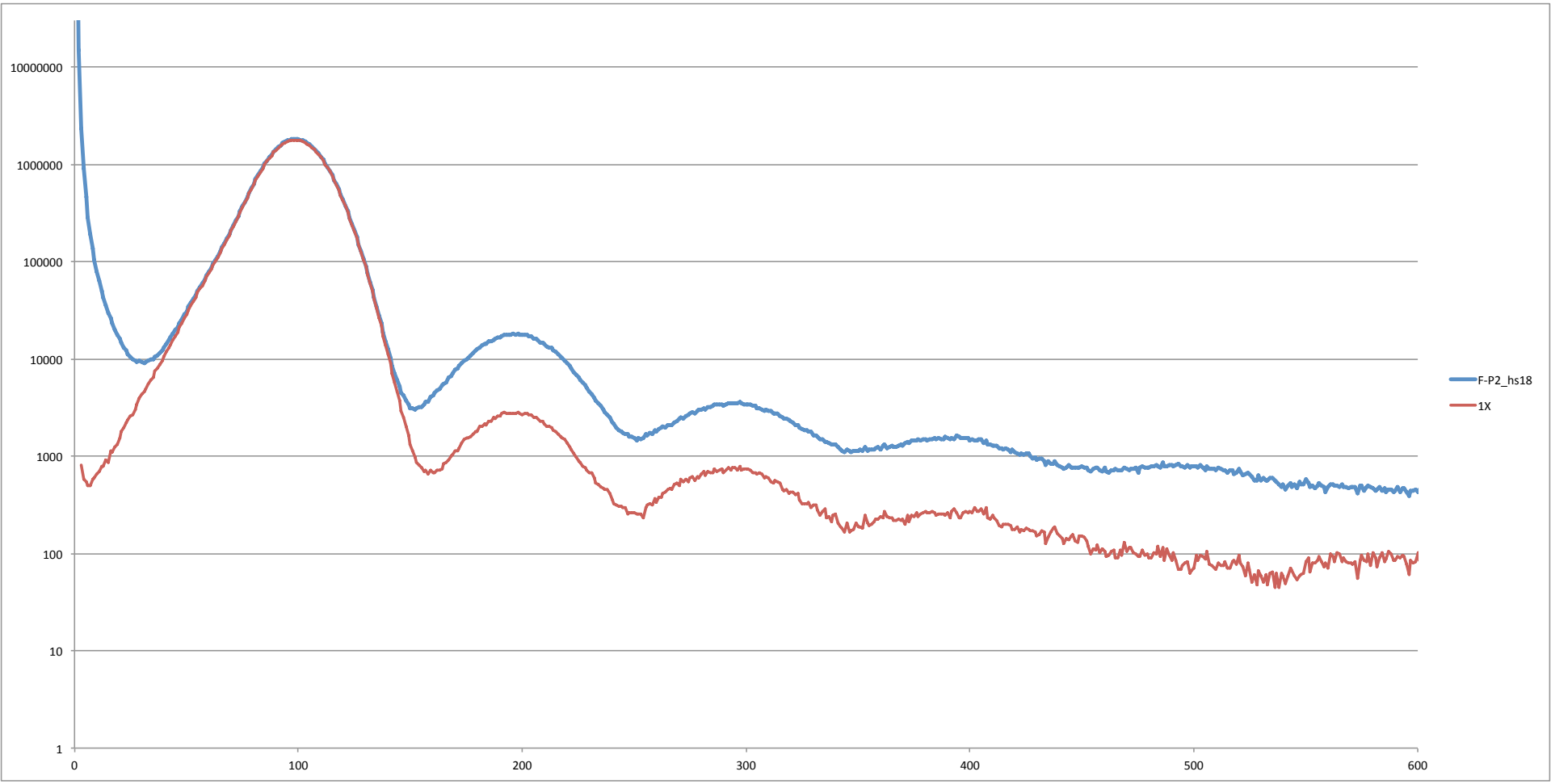
K-mer plot: Data Description



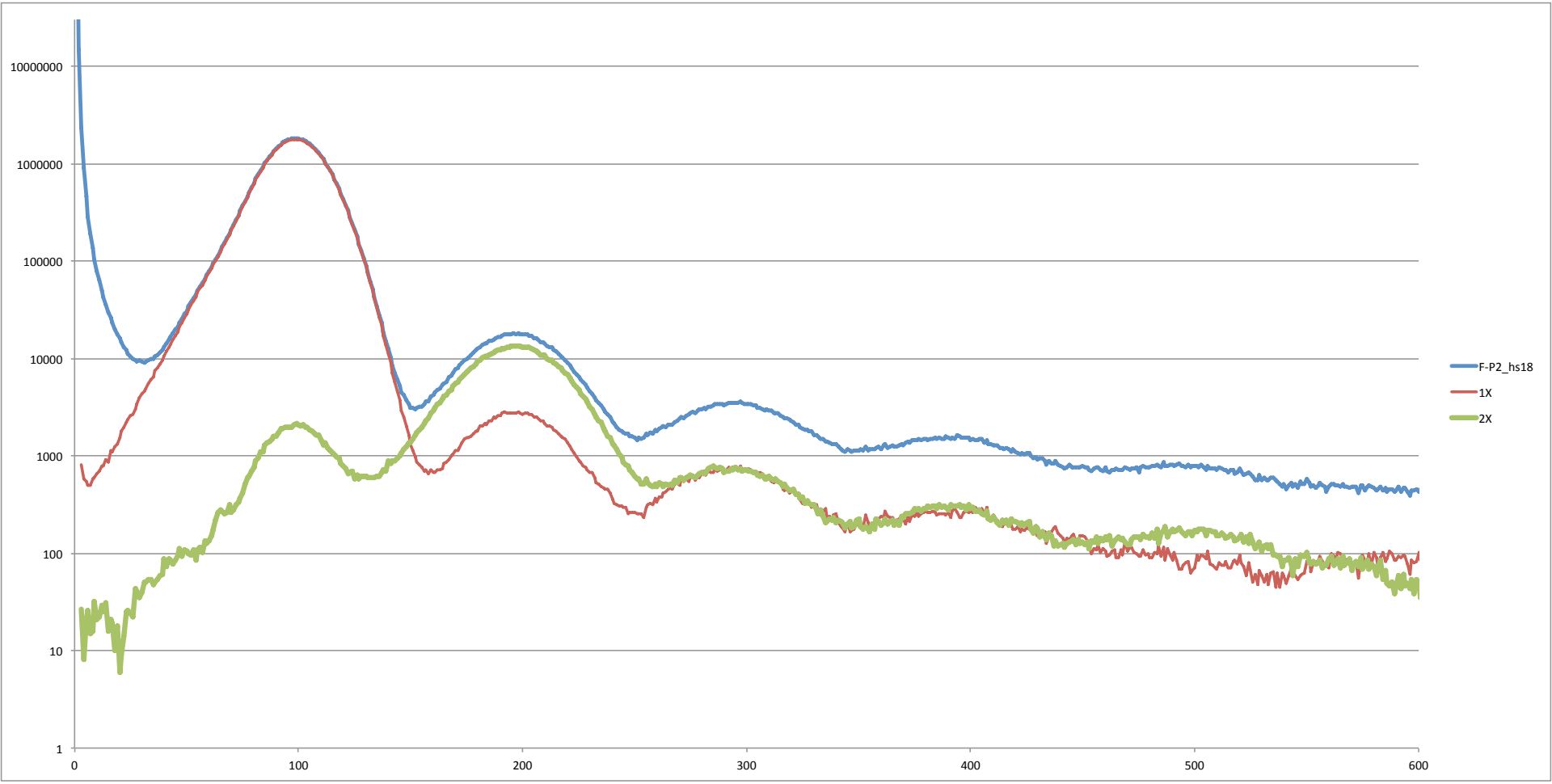
DOC2 – Hashes separated by K-mer count in the reference



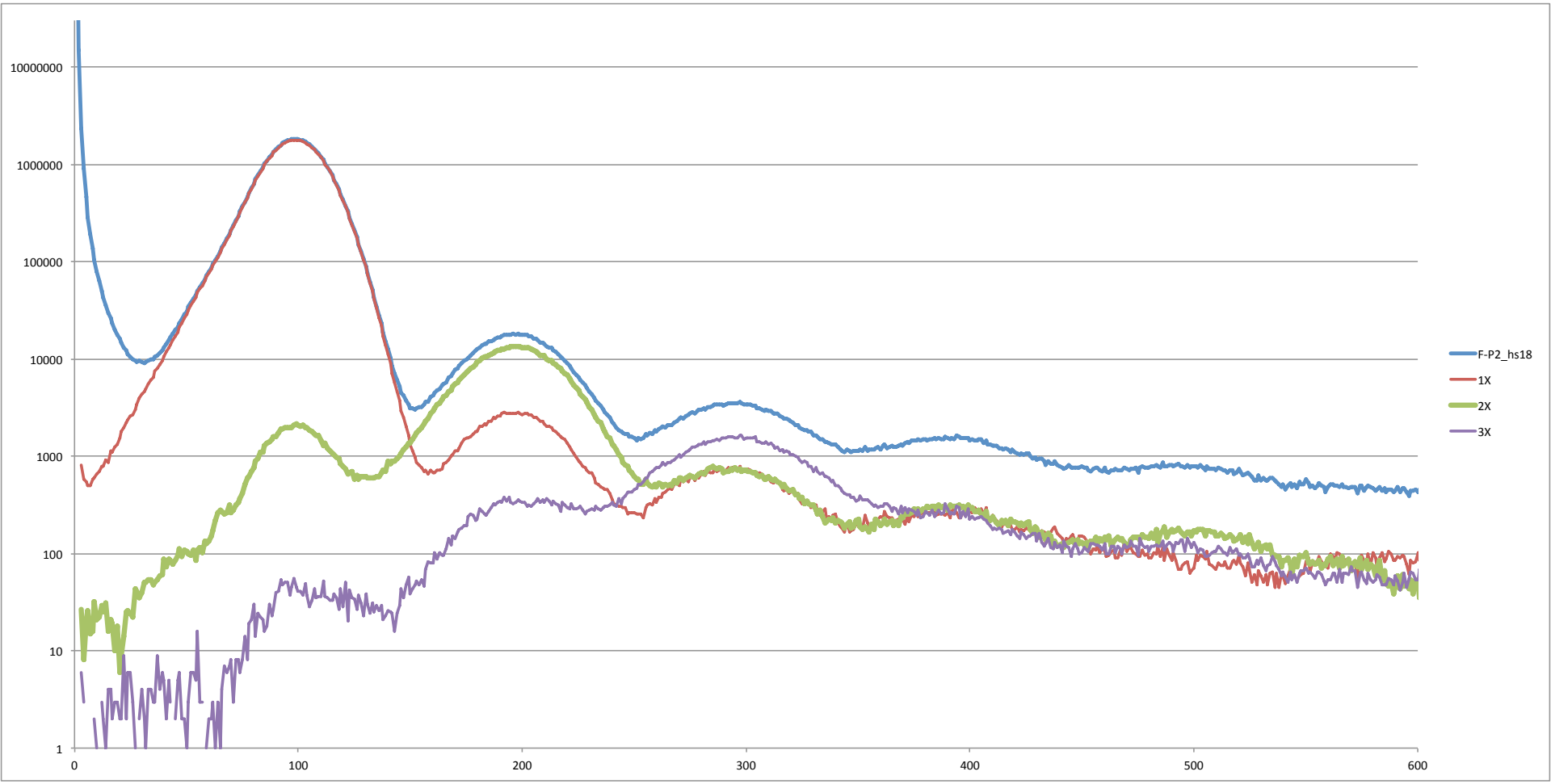
DOC2 – Hashes separated by copy number in the reference



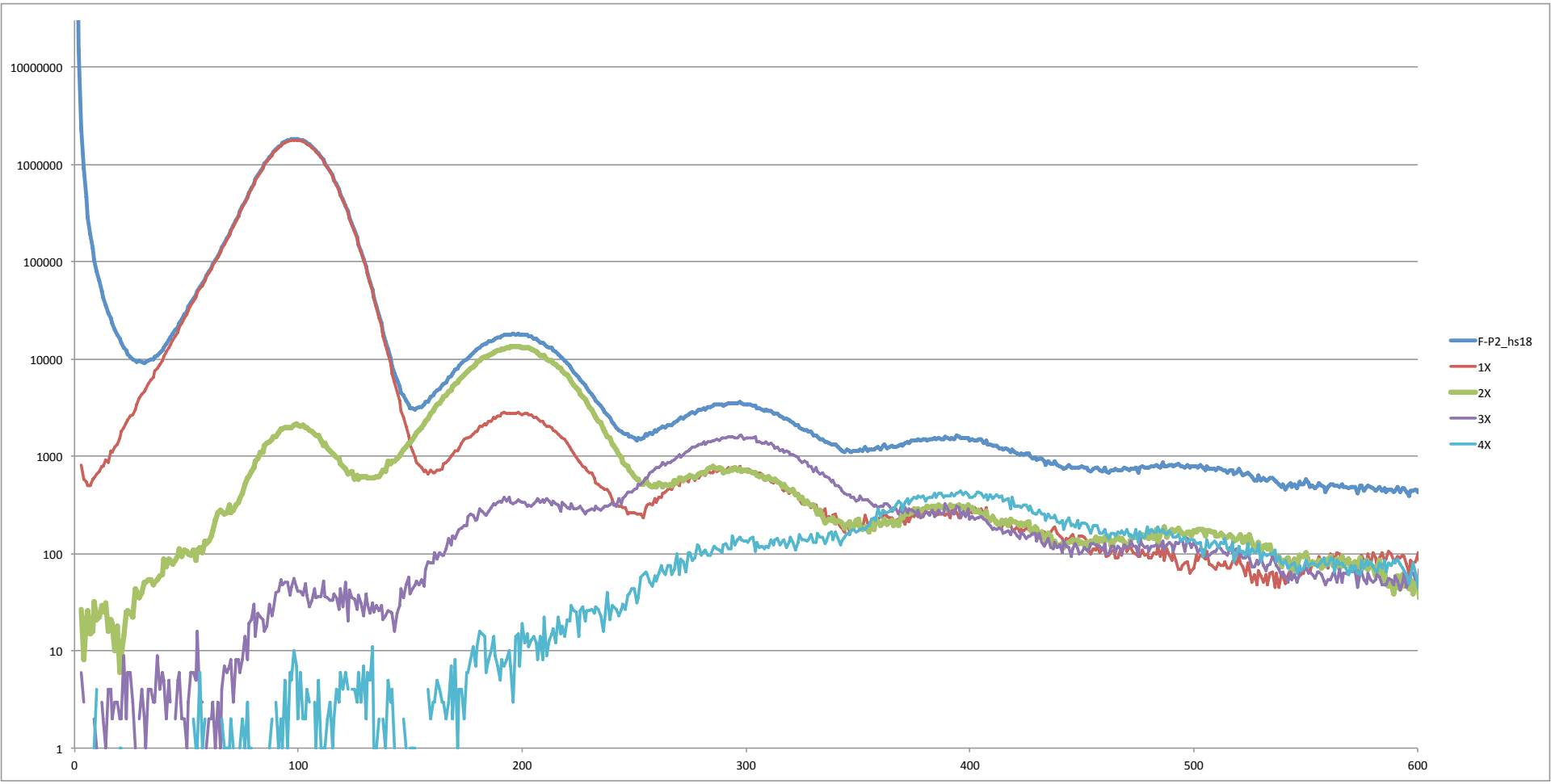
DOC2 – Hashes separated by copy number in the reference



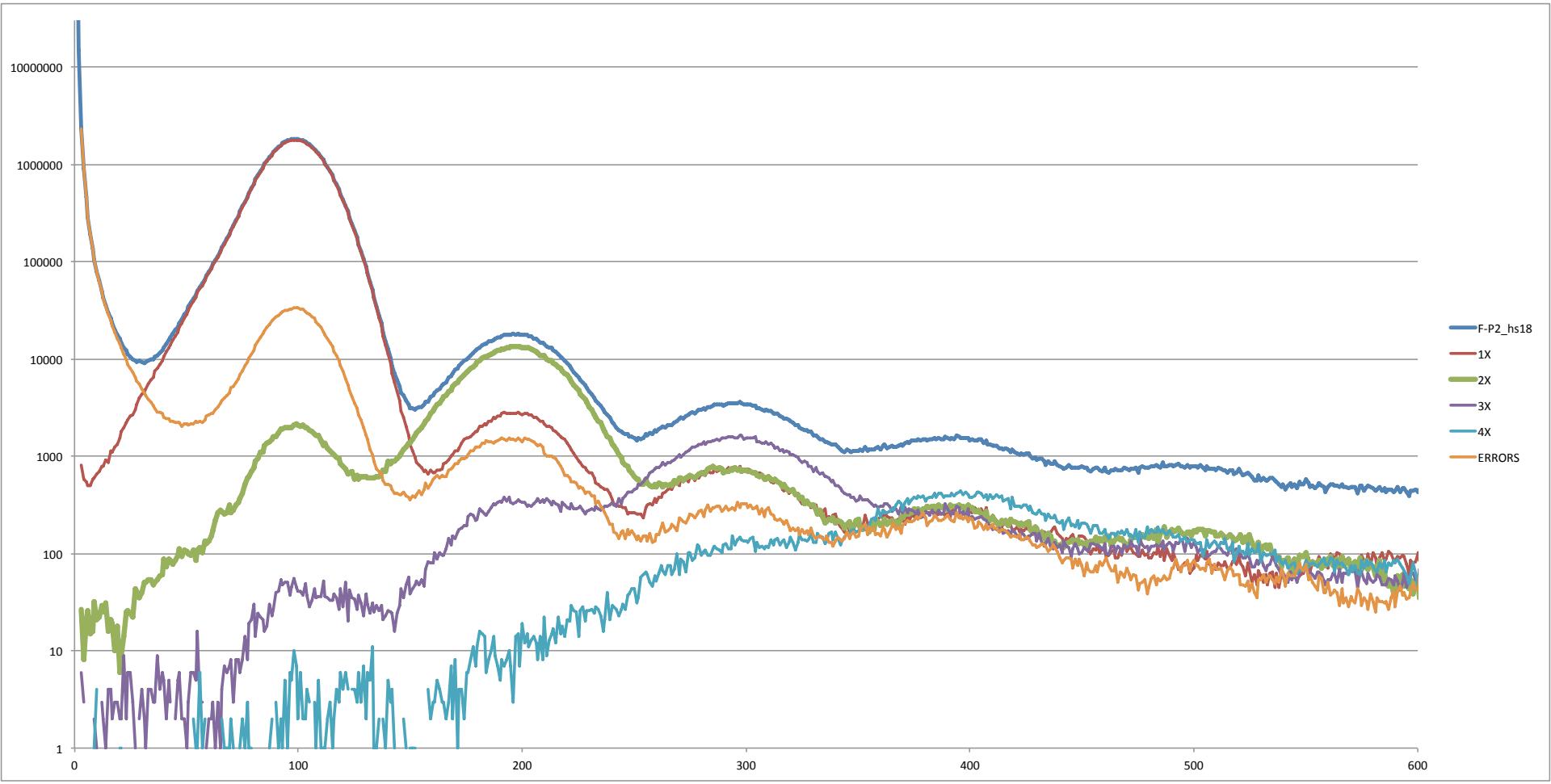
DOC2 – Hashes separated by copy number in the reference

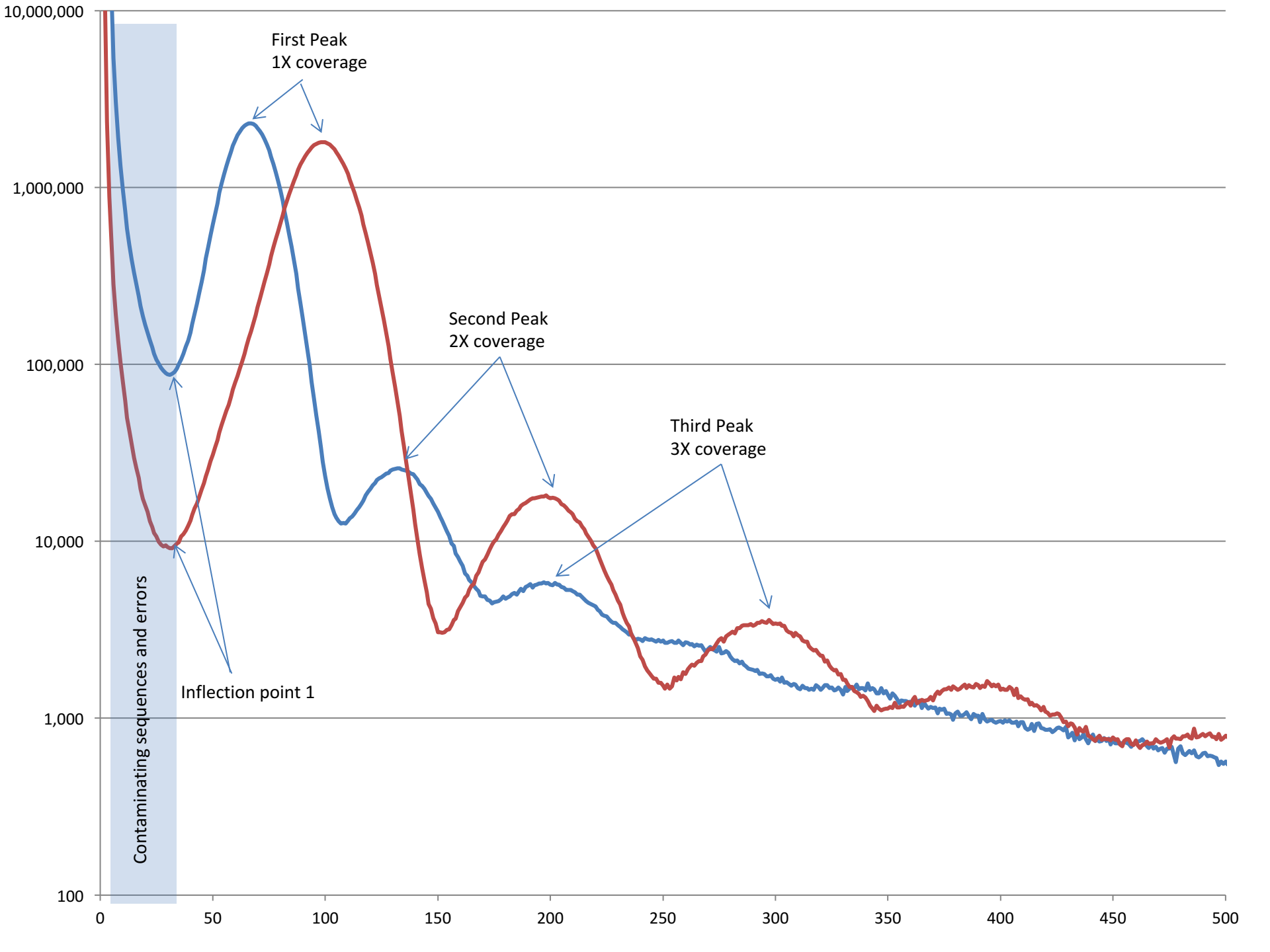


DOC2 – Hashes separated by copy number in the reference

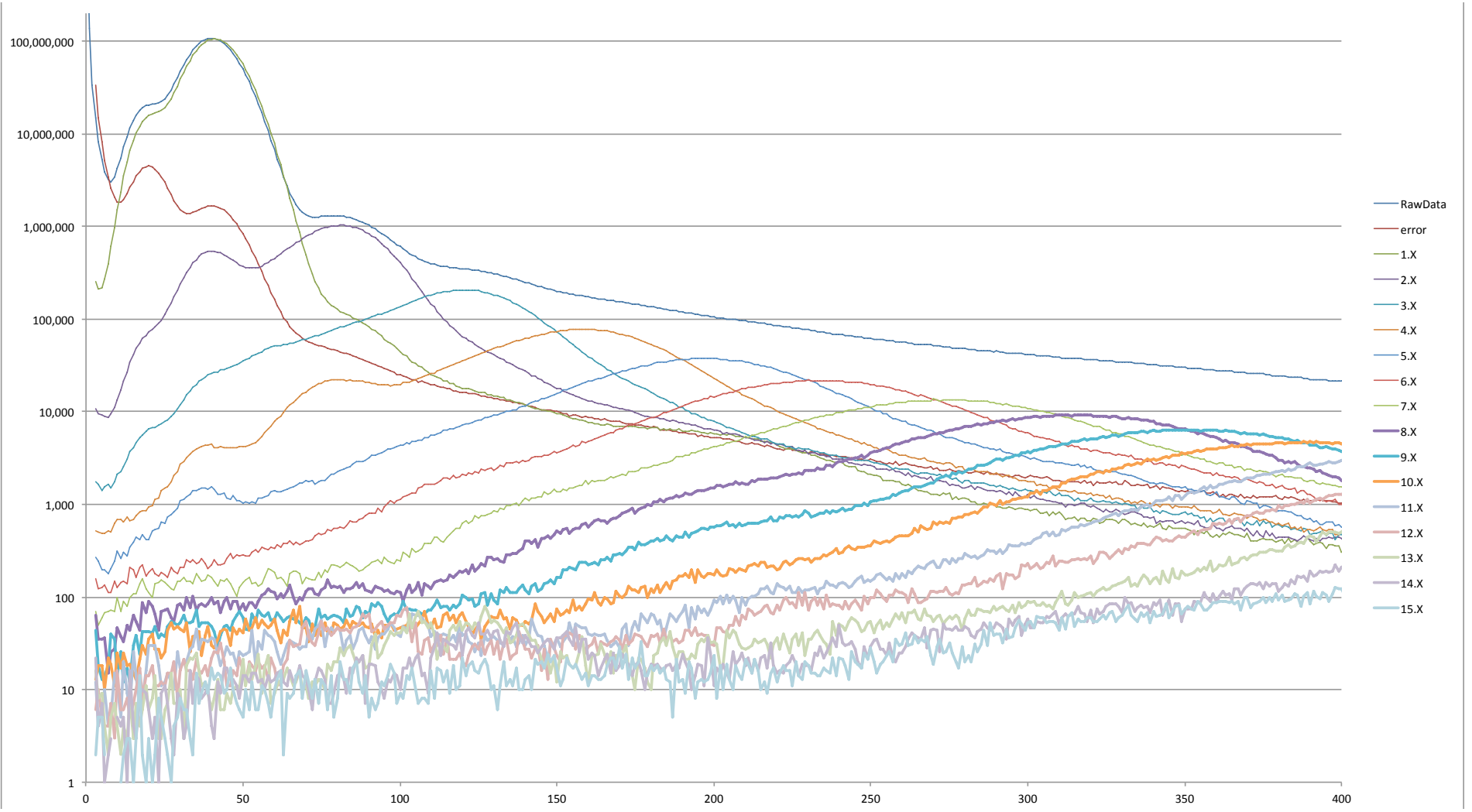


DOC2 – Hashes separated by copy number in the reference



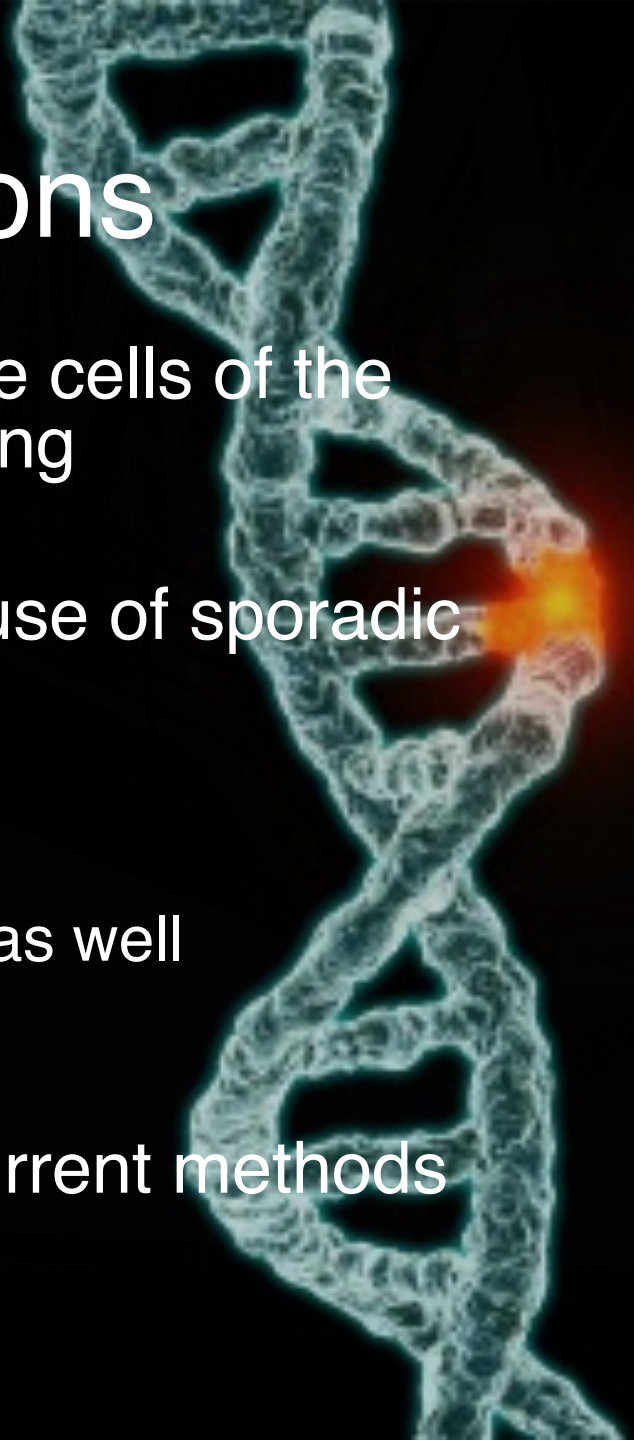

10,000,000
1,000,000
100,000
10,000
1,000
100
0
50
100
150
200
250
300
350
400
450
500
First Peak
1X coverage
Second Peak
2X coverage
Third Peak
3X coverage
Inflection point 1
Contaminating sequences and errors

Human – K-mers separated by copy number in the reference

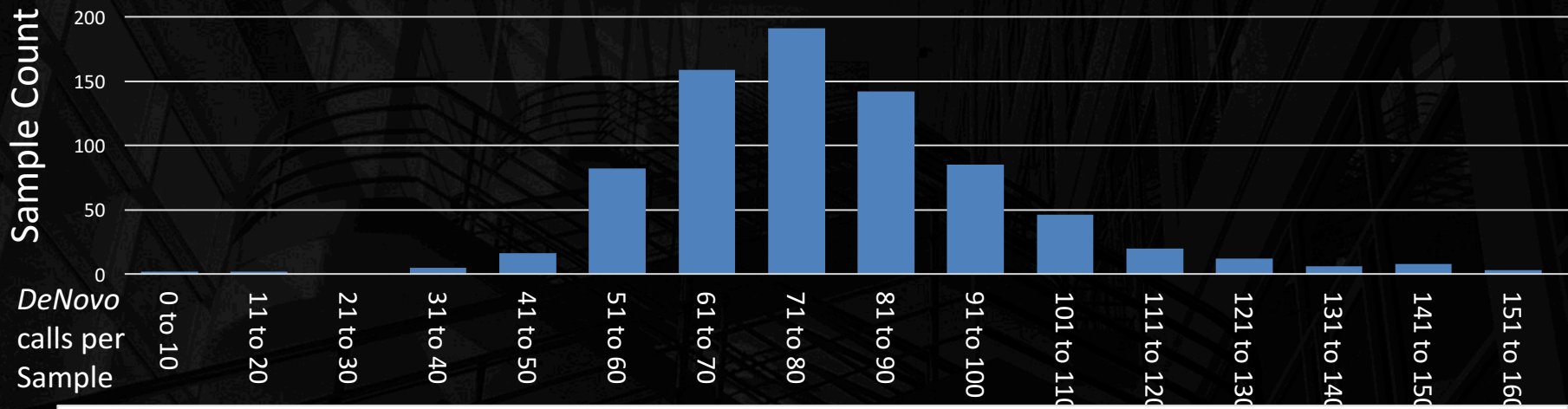


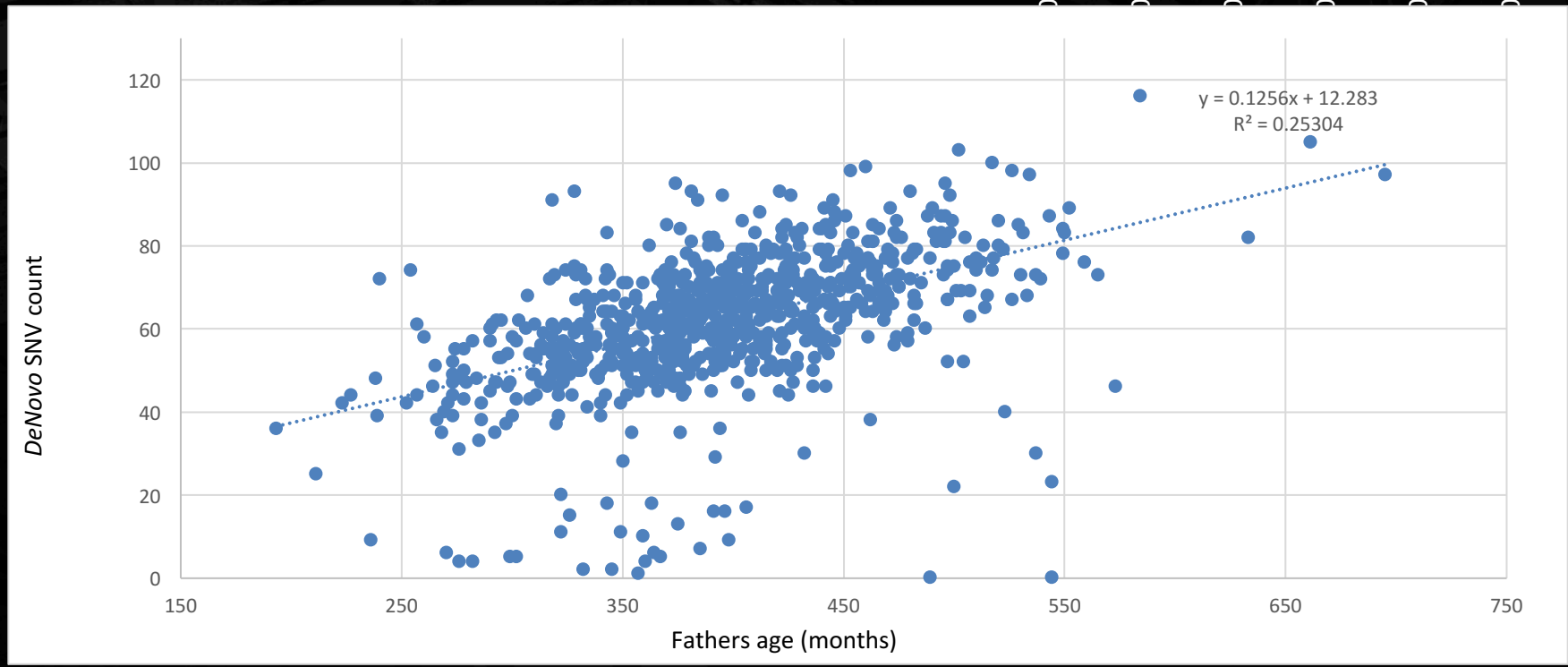
DeNovo Mutations

- Spontaneous mutations in germline cells of the parent that are passed onto offspring
- *DeNovo* mutations are a major cause of sporadic rare disease
- ~75 *DeNovo* SNV per generation
  - Insertion/deletions < 100kb are not as well characterized
- Extremely difficult to detect with current methods



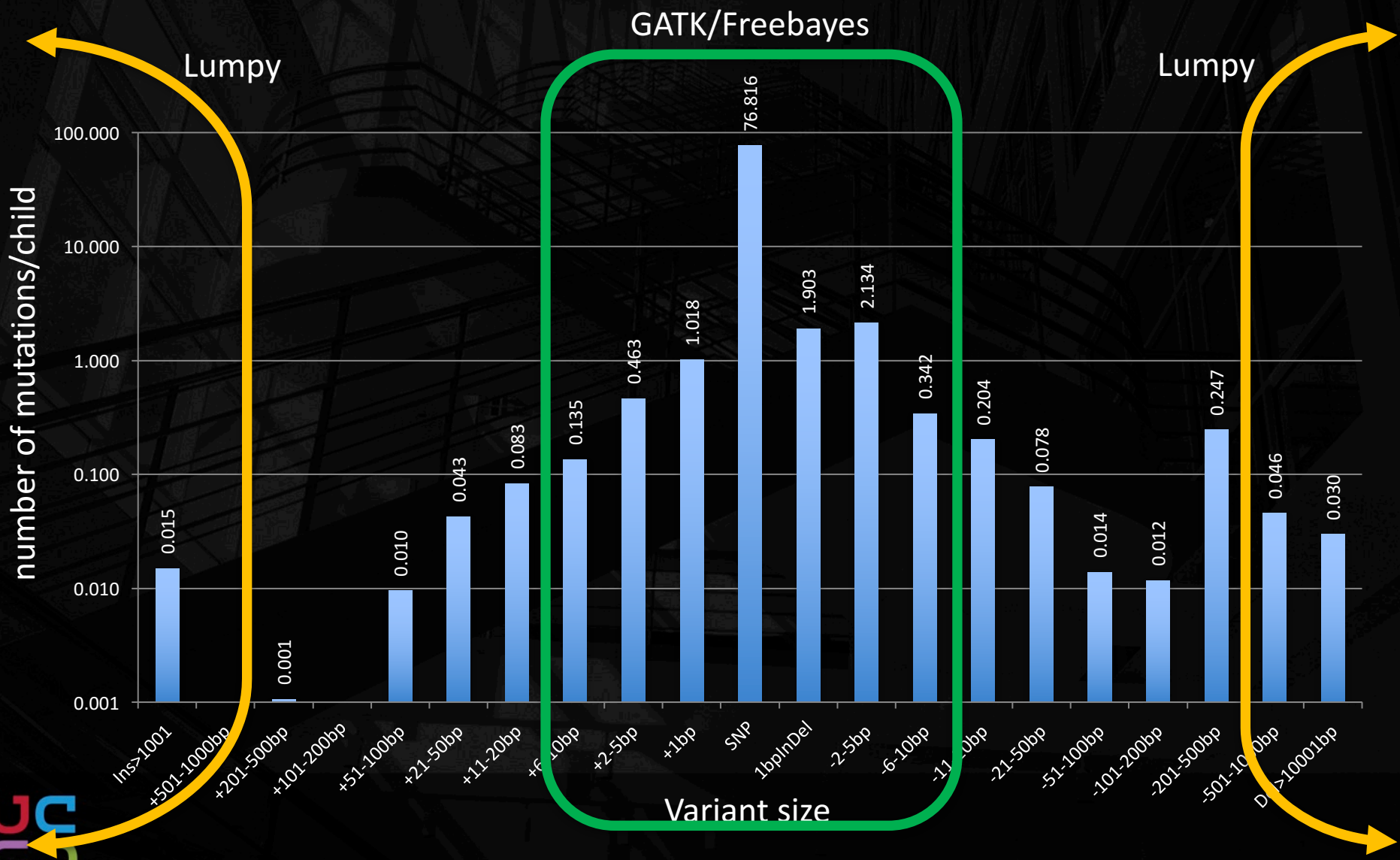
SFARI SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE





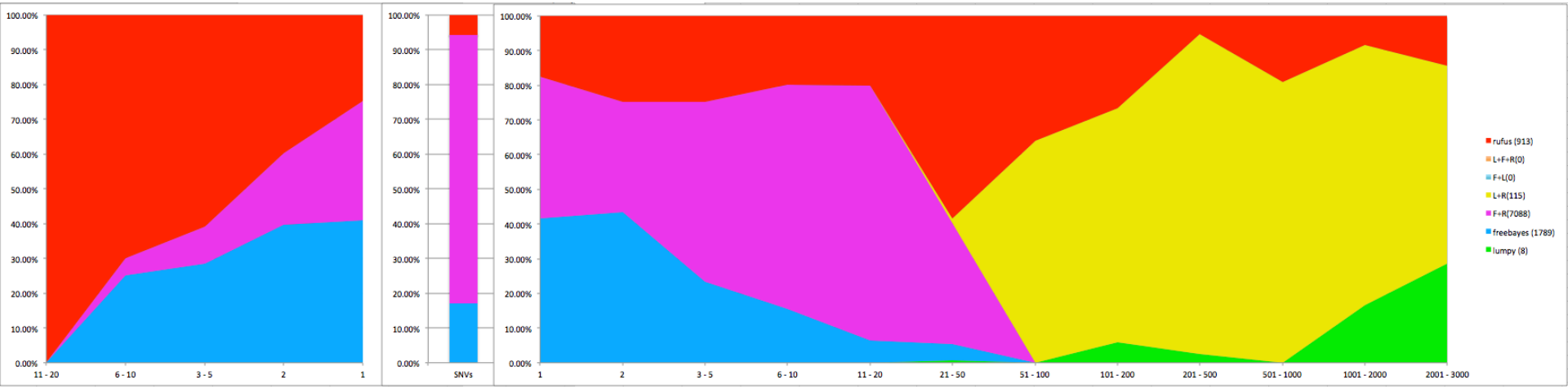
SFARI SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

DeNovo Variant Size Spectrum



RUFUS finds indels missed by mapping based methods

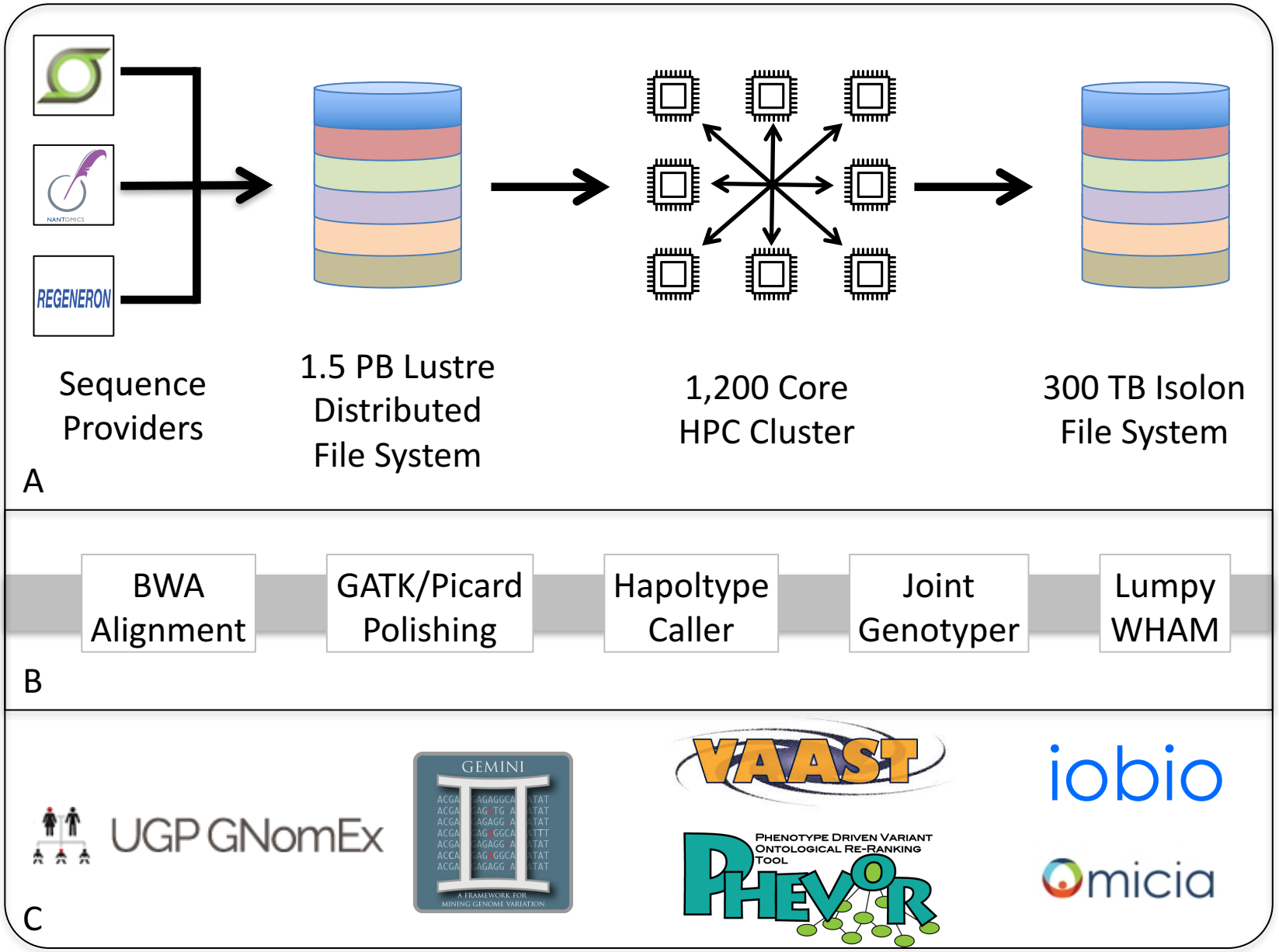
- WashU serial tumor sequencing project
- B1 vs Bo



Early Infantile Epileptic Encephalopathy

- Early onset seizure syndrome
  - Primarily caused by *DeNovo* mutations
- 15 trios from the University of Utah Hospital, sequenced to 60X
- All U Hospital sequencing projects are run though a pipeline based on BWA/GATK best practices

FastQForward: high-throughput variant calling pipeline



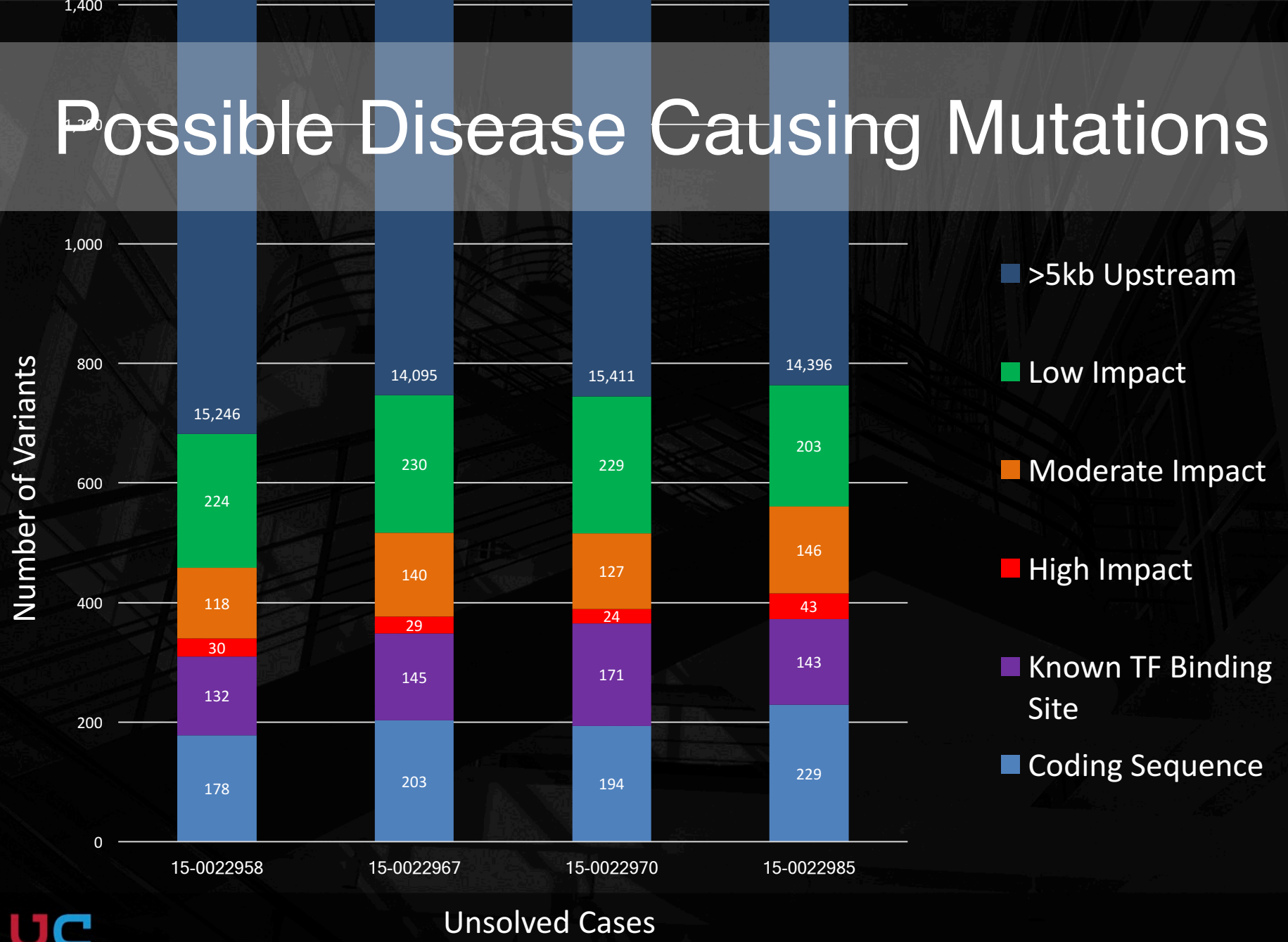
Cases solved with GATK + Annotation

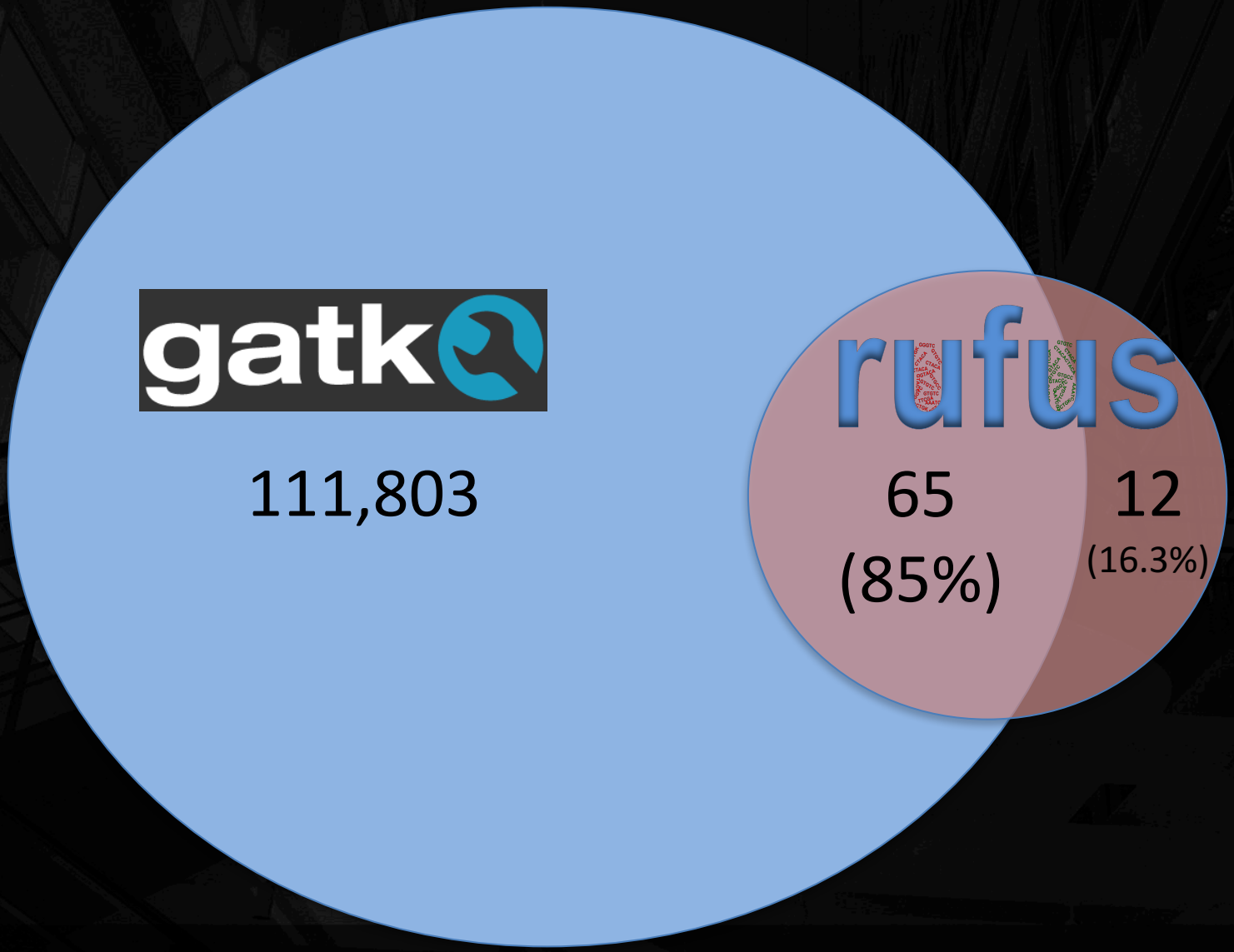
Mutations in known EIEE genes

Mutations in genes with associated phenotypes

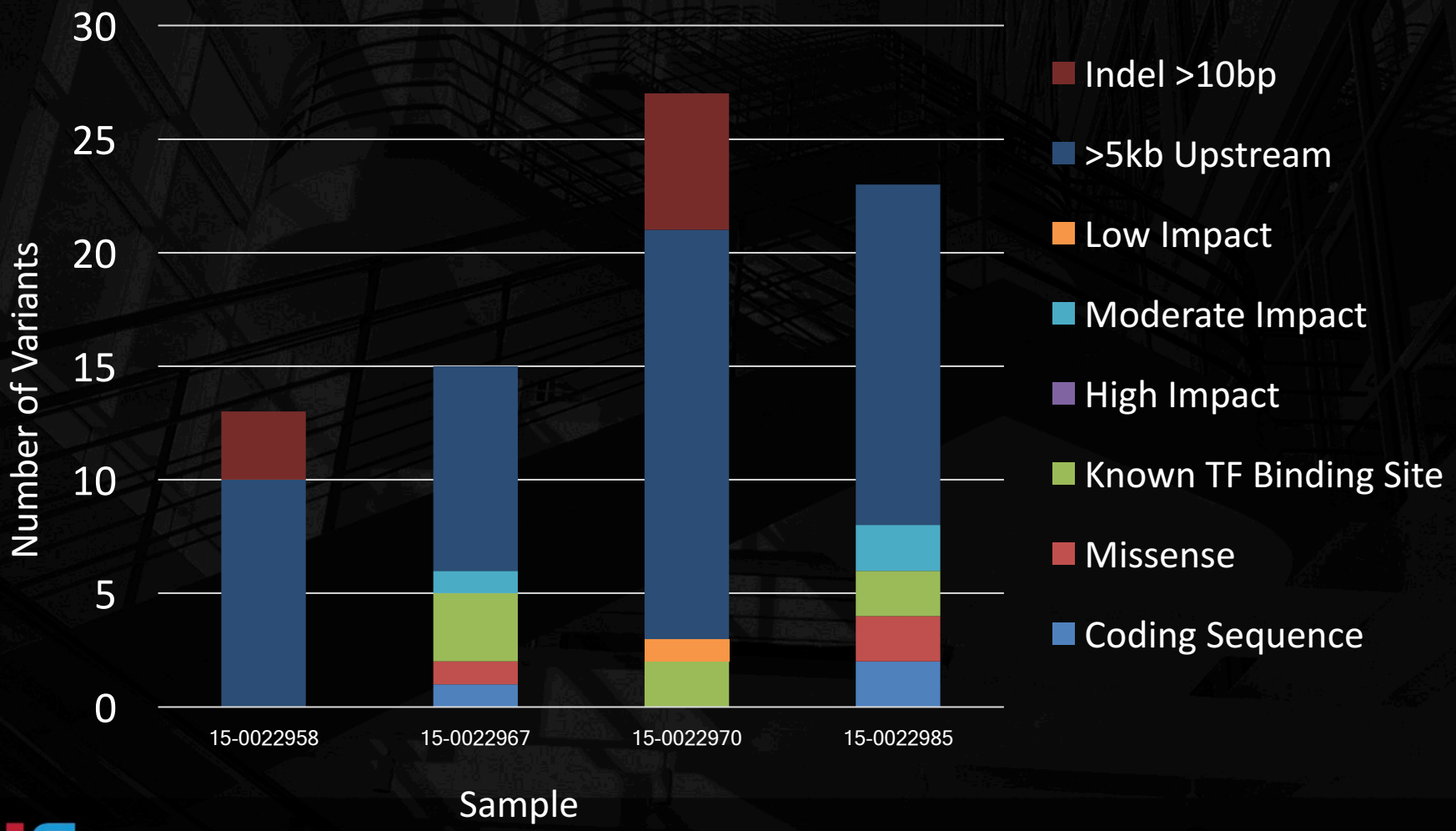
Undaignosed

Possible Disease Causing Mutations




gatk
111,803
rufus
65
(85%)
12
(16.3%)

RUFUS stats for EIEE

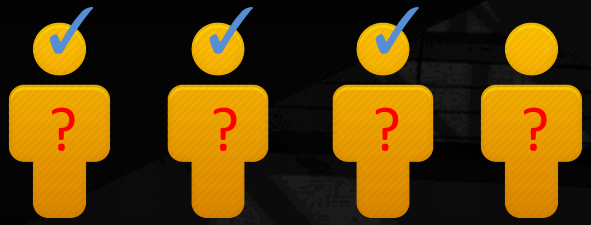


RUFUS results for EIEE

Mutations in known EIEE genes

Mutations in genes with associated phenotypes

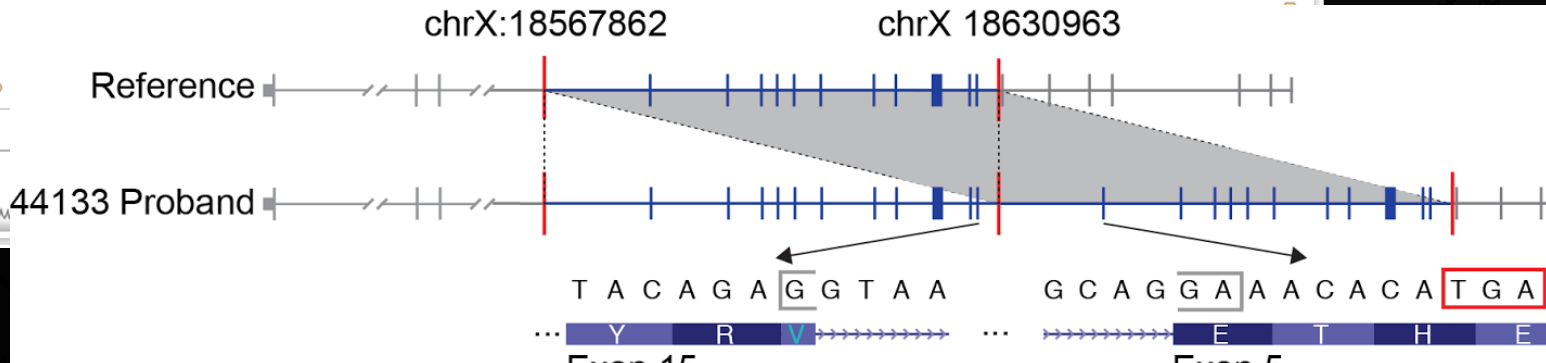
Undiagnosed



Unsolved Case 1

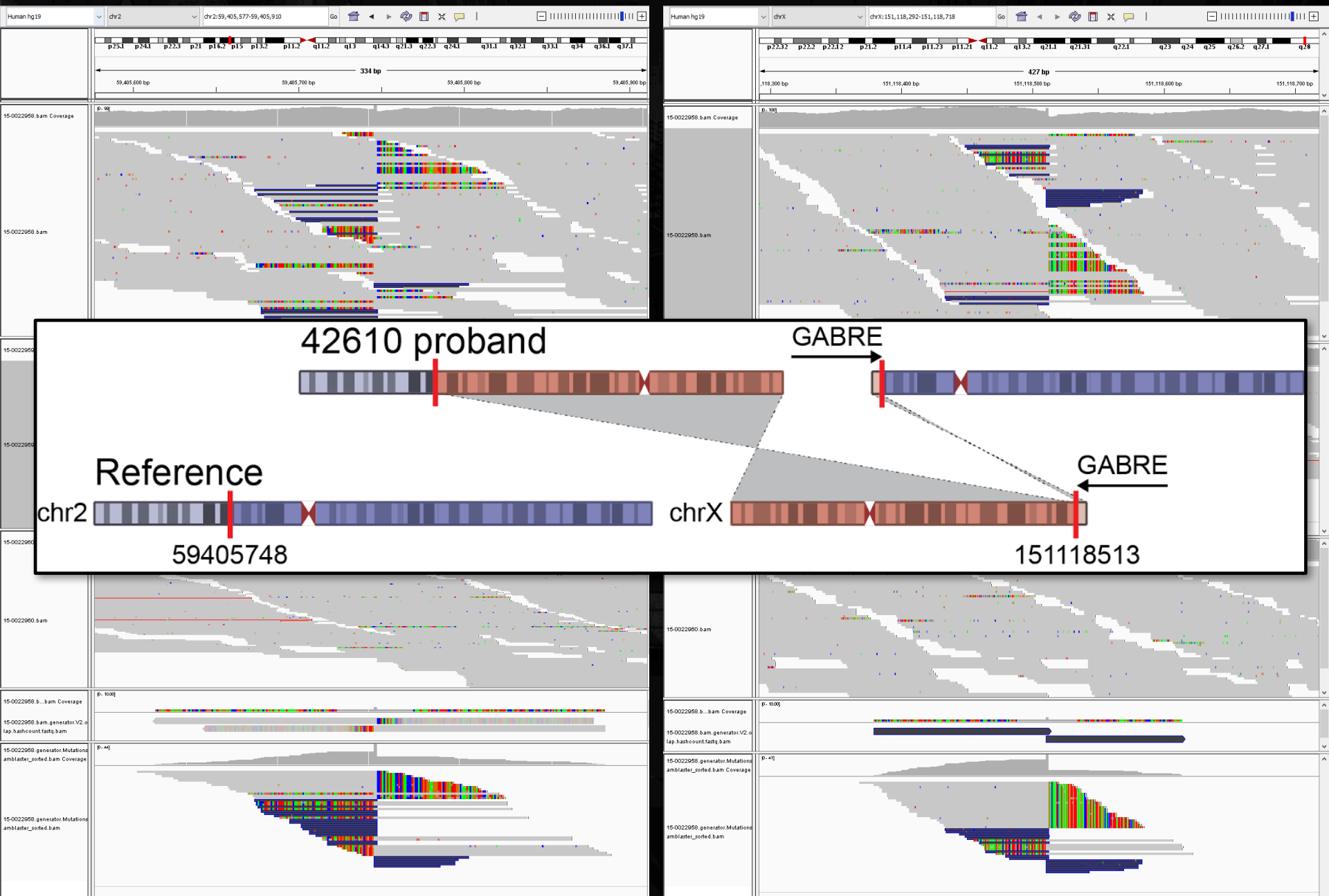
- Causal variant is on the X chromosome in a male proband called as homozygous, and was excluded in annotation for violating *DeNovo* model

Unsolved Case 2

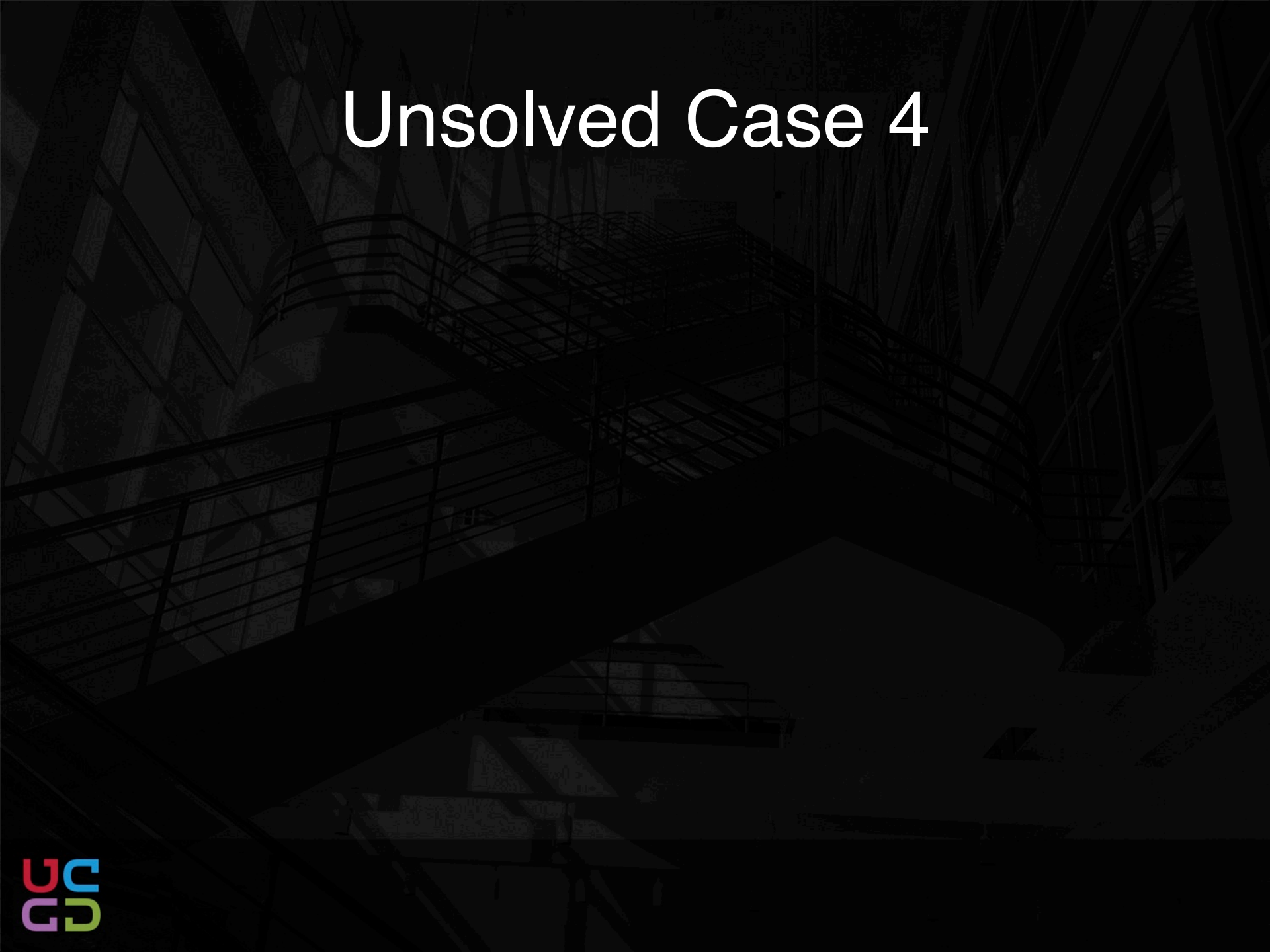


Unsolved Case 3

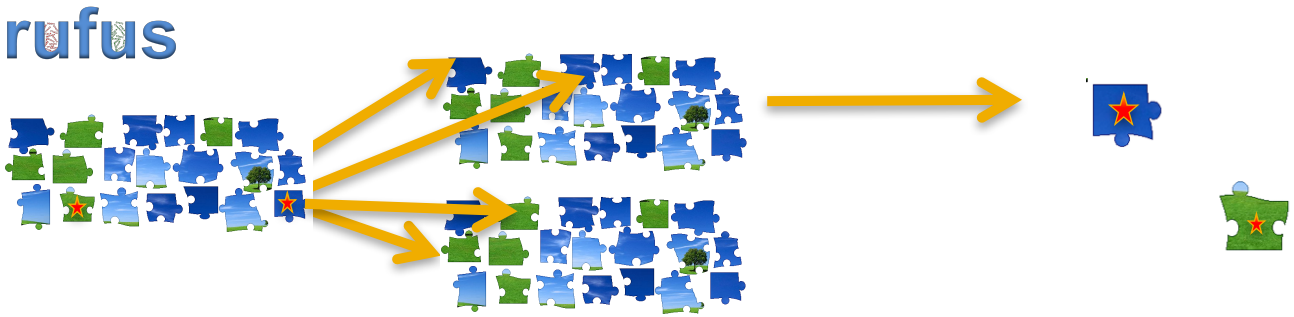
- 15-0022958
  - No variants in coding sequence of EIEE genes
  - RUFUS identifies two candidates
    - SNP 250bp upstream of a known EIEE gene in a pol2 biding site on chrX.
    - A balanced translocation between chrX and chr2 down stream of a known EIEE gene


42610 proband
GABRE
Reference
chr2
chrX
GABRE
59405748
151118513

Unsolved Case 4

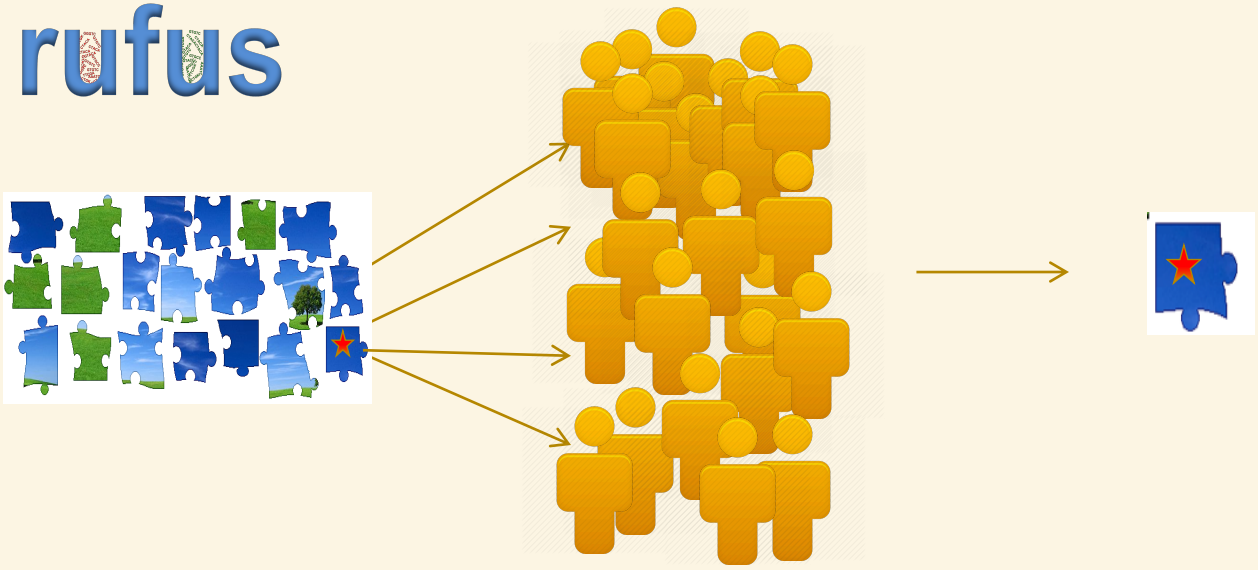


Whole Genome Analysis Methods



“Subtracting” variants that segregate in the general population at appreciable frequency

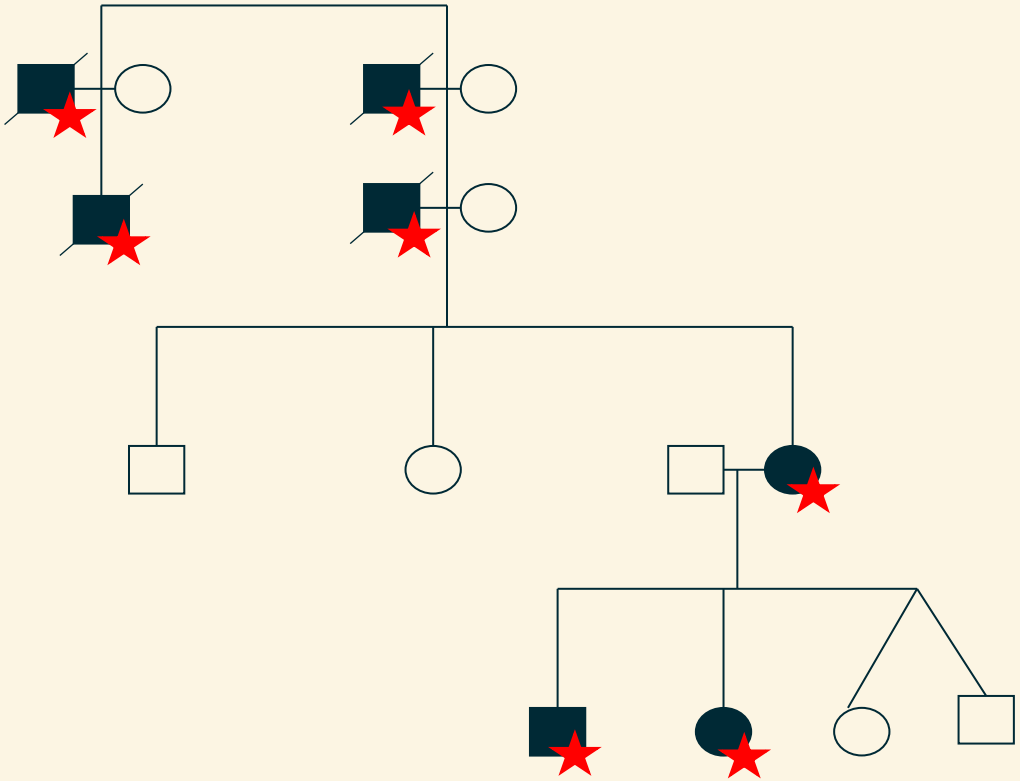
rufus

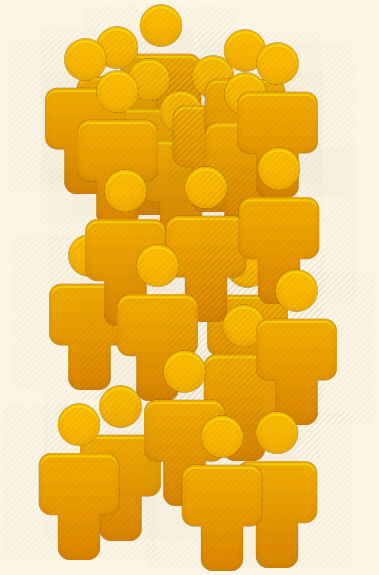


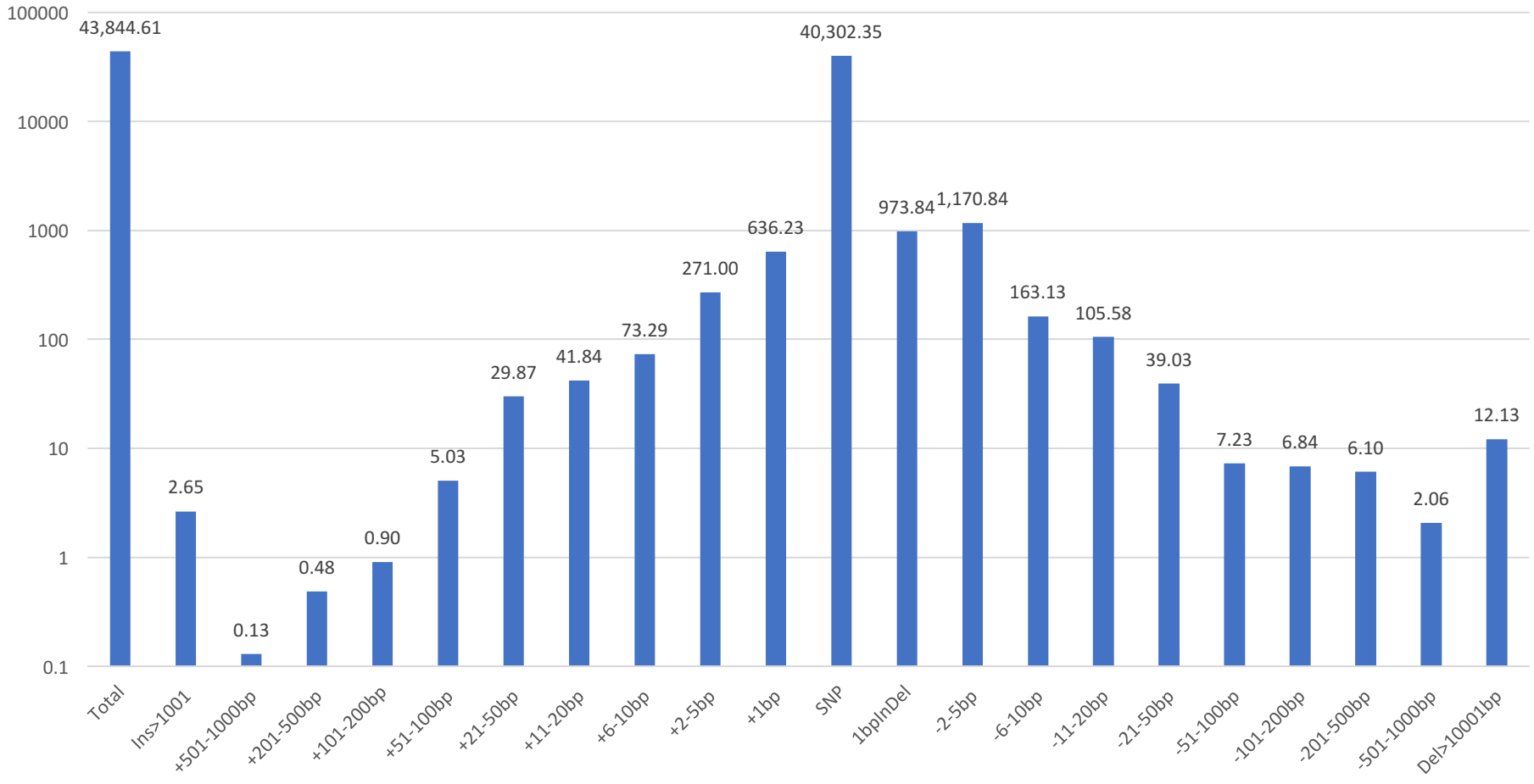
HHT: Hereditary Hemorrhagic Telangiectasia

- 37 Whole genome sequenced samples from 13 families
- Unlike out pervious work with RUFUS that focused on *DeNovo* diseases, the causative variant is believed to be segregating within the families
- To find rare inherited variants in each sample we used the raw 1000 genomes project reads as a reference k-mer set to call variants.
- Variants are then compared within each family to identify variation shared between all affected individual in that family, and if available, variations that are absent in unaffected relatives.

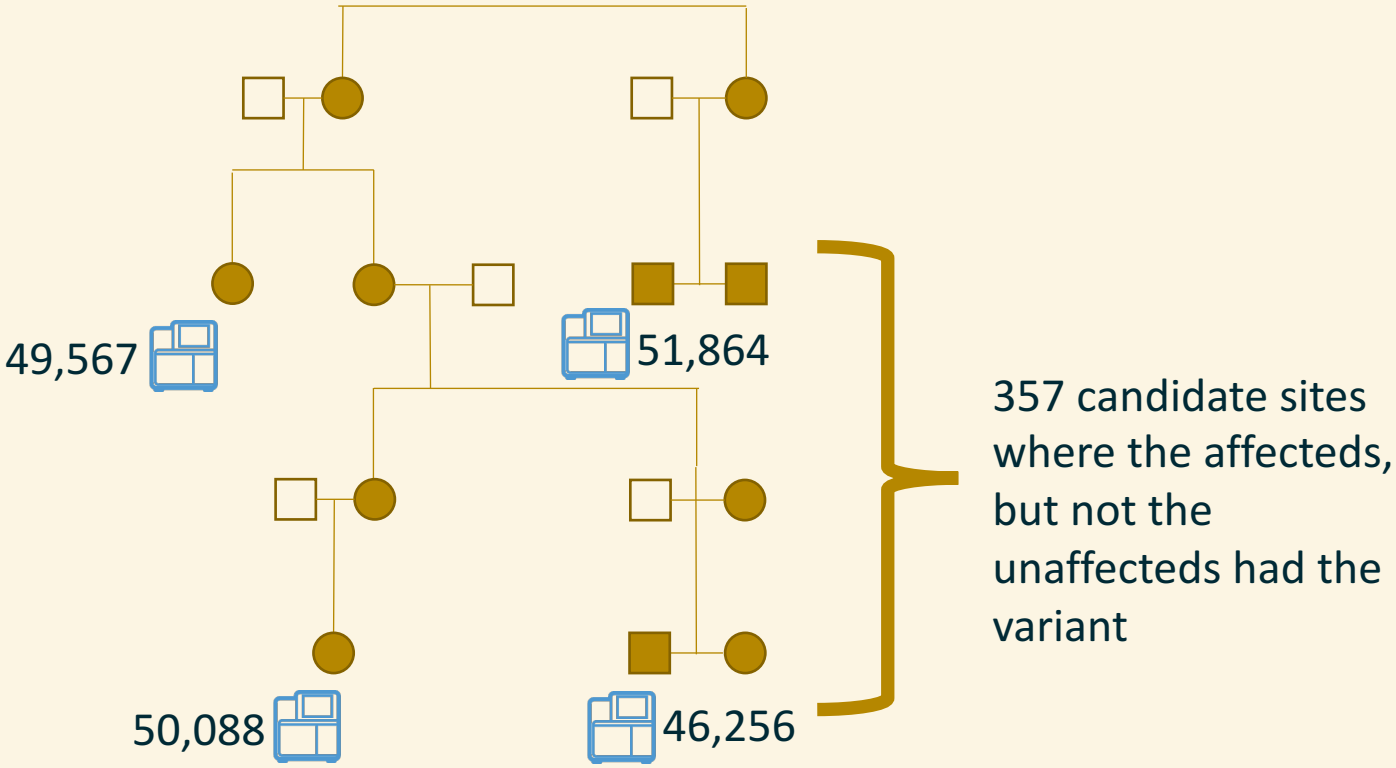
Family-private disease-causing variants



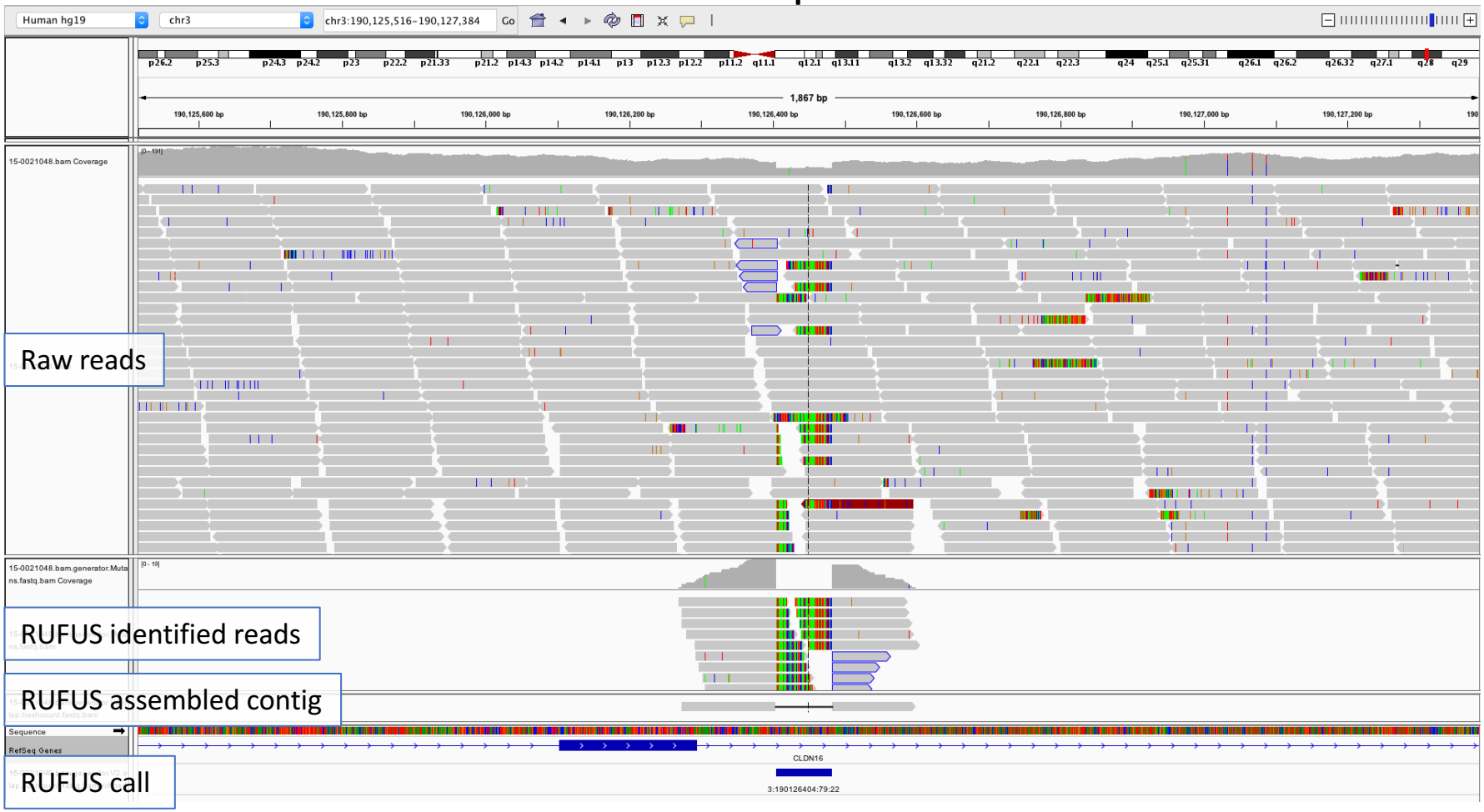



HHT, Rate of rare inherited variations by size per sample
100000
10000
1000
100
10
1
0.1
43,844.61
2.65
0.13
0.48
0.90
5.03
29.87
41.84
73.29
271.00
636.23
40,302.35
973.84
1,170.84
163.13
105.58
39.03
7.23
6.84
6.10
2.06
12.13
Total
Ins>1001
+501-1000bp
+201-500bp
+101-200bp
+51-100bp
+21-50bp
+11-20bp
+6-10bp
+2-5bp
+1bp
SNP
1bpInDel
-2-5bp
-6-10bp
-11-20bp
-21-50bp
-51-100bp
-101-200bp
-201-500bp
-501-1000bp
Del>10001bp

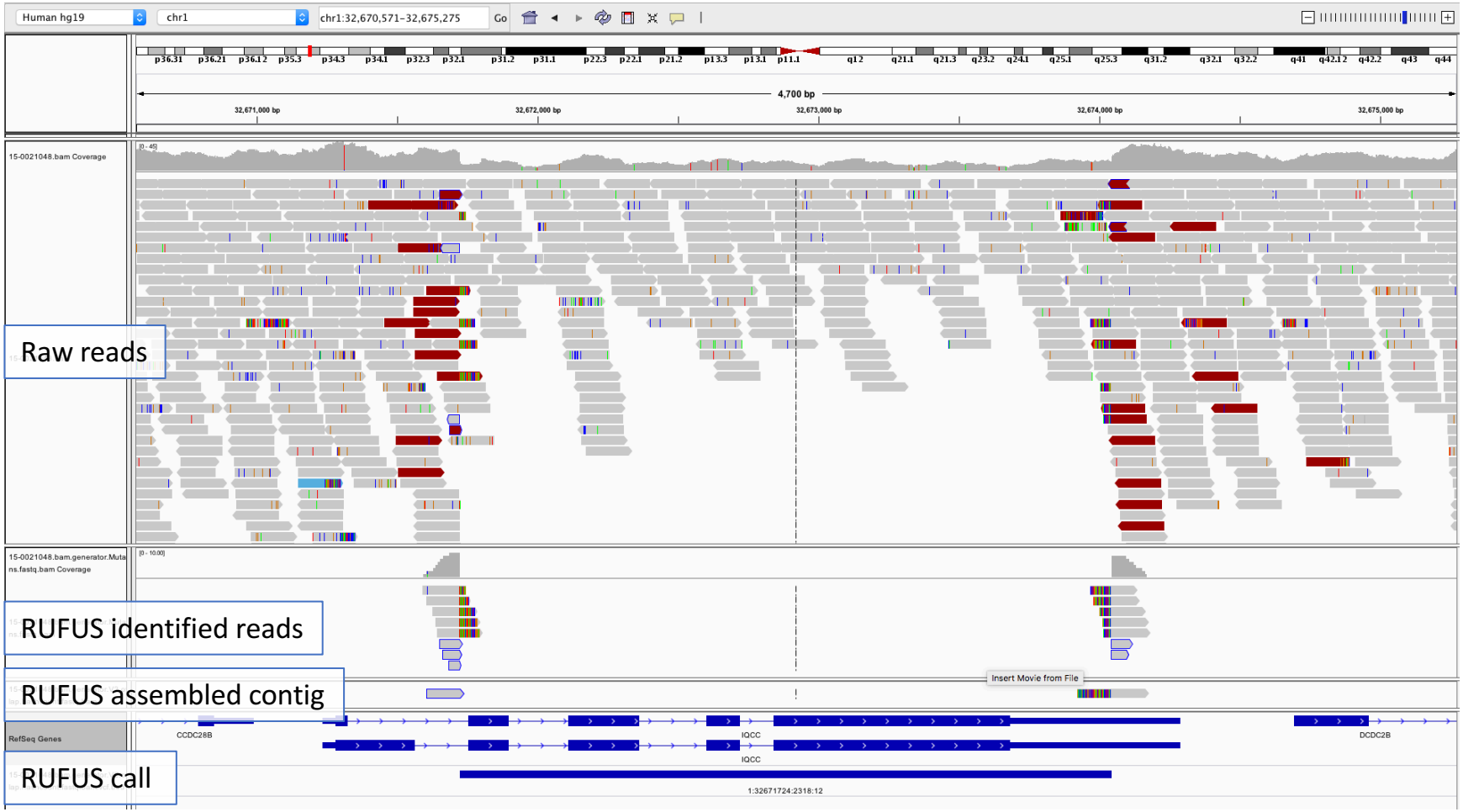
HHT Family C Pedigree



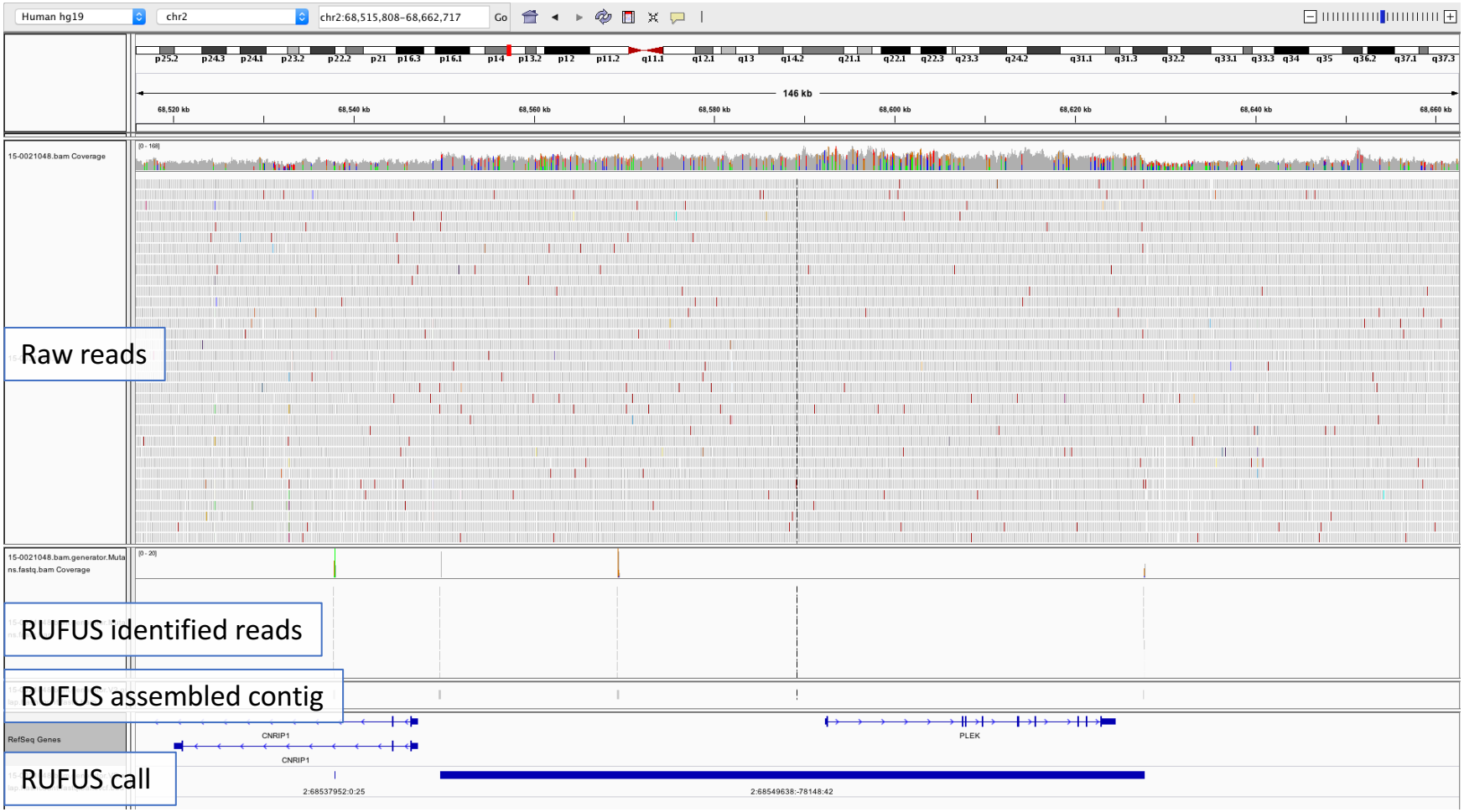
RUFUS rare inherited 79bp deletion call



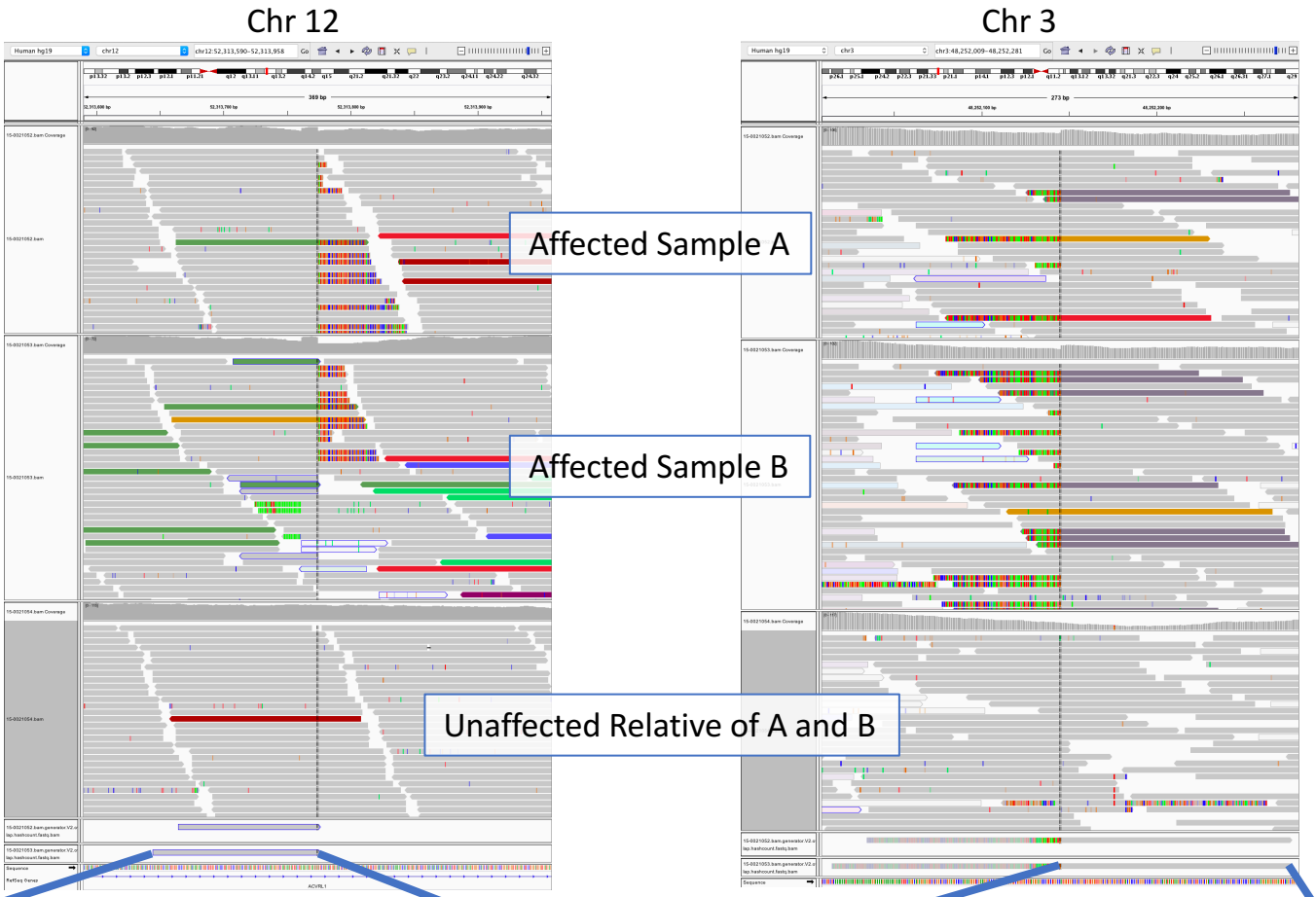
RUFUS rare inherited deletion call



RUFUS rare inherited tandem duplication call



RUFUS identifies a translocation between chromosomes 12 and 3 that disrupts a known HHT gene.

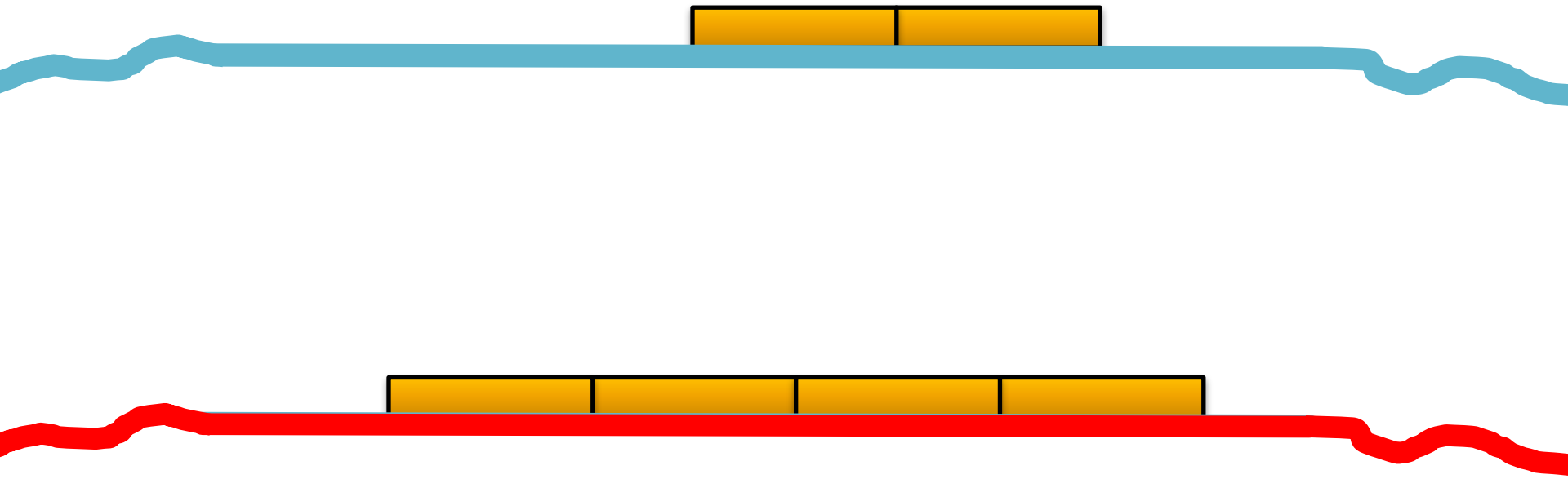


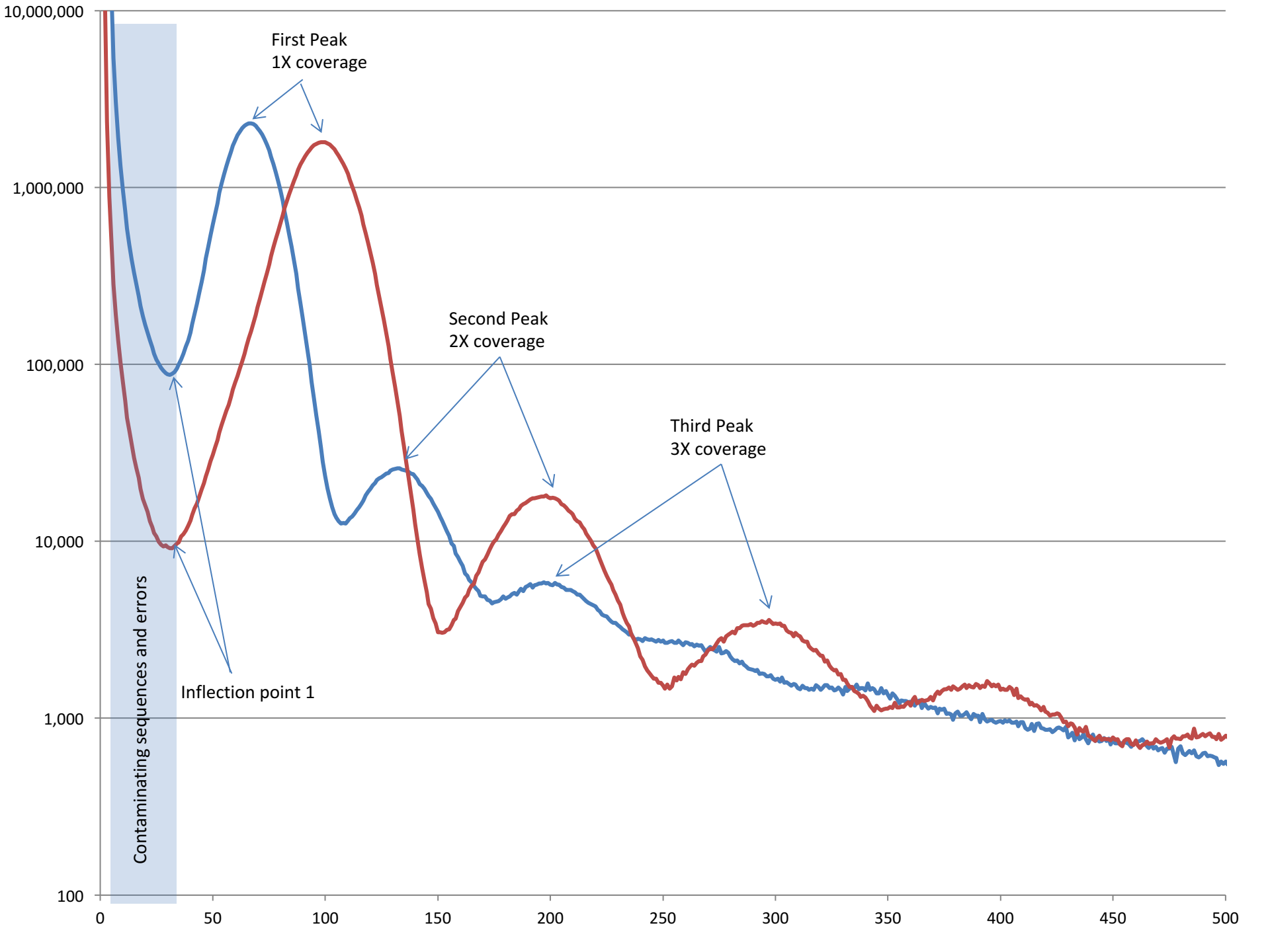
HHT Candidate variations per family.

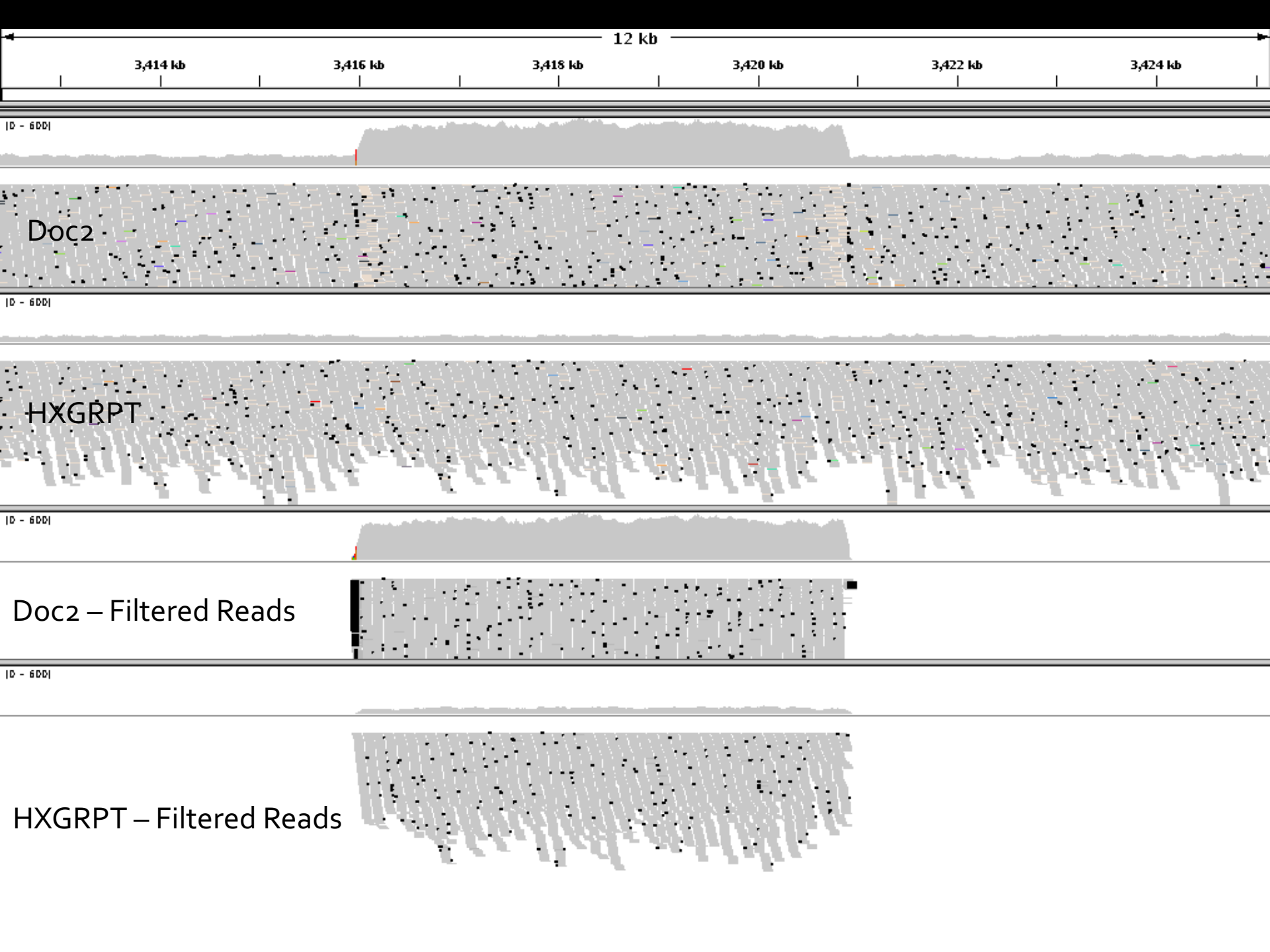
Due to the extremely accurate nature of RUFUS variant calling, we can compare inherited variations between affected members of the same families to reduce the number of possible variants

Group	Candidate variations shared among related affected individuals	Reduction in candidate variants	Genome-wide CDS Variants	Variation in Known HHT Genes
Family 5	28,362	37.41%	360	80
Family 6	8,709	80.78%	97	18
Family 7	11,394	74.85%	138	20
Family 9	6,054	86.64%	53	16
Family C	1,404	96.90%	13	6
Family H	15,489	65.82%	194	33
Family M	12,232	73.00%	161	48
Family NS	15,212	66.43%	165	38
Family N	13,349	70.54%	169	27
Family O	4,664	89.71%	61	10
Family RE	12,602	72.19%	136	18
Family R	7,104	84.32%	87	40

Copy Number Variant Detection




10,000,000
1,000,000
100,000
10,000
1,000
100
0
50
100
150
200
250
300
350
400
450
500
First Peak
1X coverage
Second Peak
2X coverage
Third Peak
3X coverage
Inflection point 1
Contaminating sequences and errors


12 kb
3,414 kb
3,416 kb
3,418 kb
3,420 kb
3,422 kb
3,424 kb
Doc2
HXGRPT
Doc2 – Filtered Reads
HXGRPT – Filtered Reads

Limitations of Kmer based methods

- Expansion and contraction of micro-satellites
- True homologous recombination with no target site mutations
- SNPs that create a sequence that already exists

Conclusions

- Completely reference free method
  - Can be done with any organism that can be sequenced
  - Eliminates bias towards the reference
- Ability find SNPs, novel sequences, copy number, and structural variants on the first pass
- Reduced false positive rate
  - No mapping errors
  - No reference errors