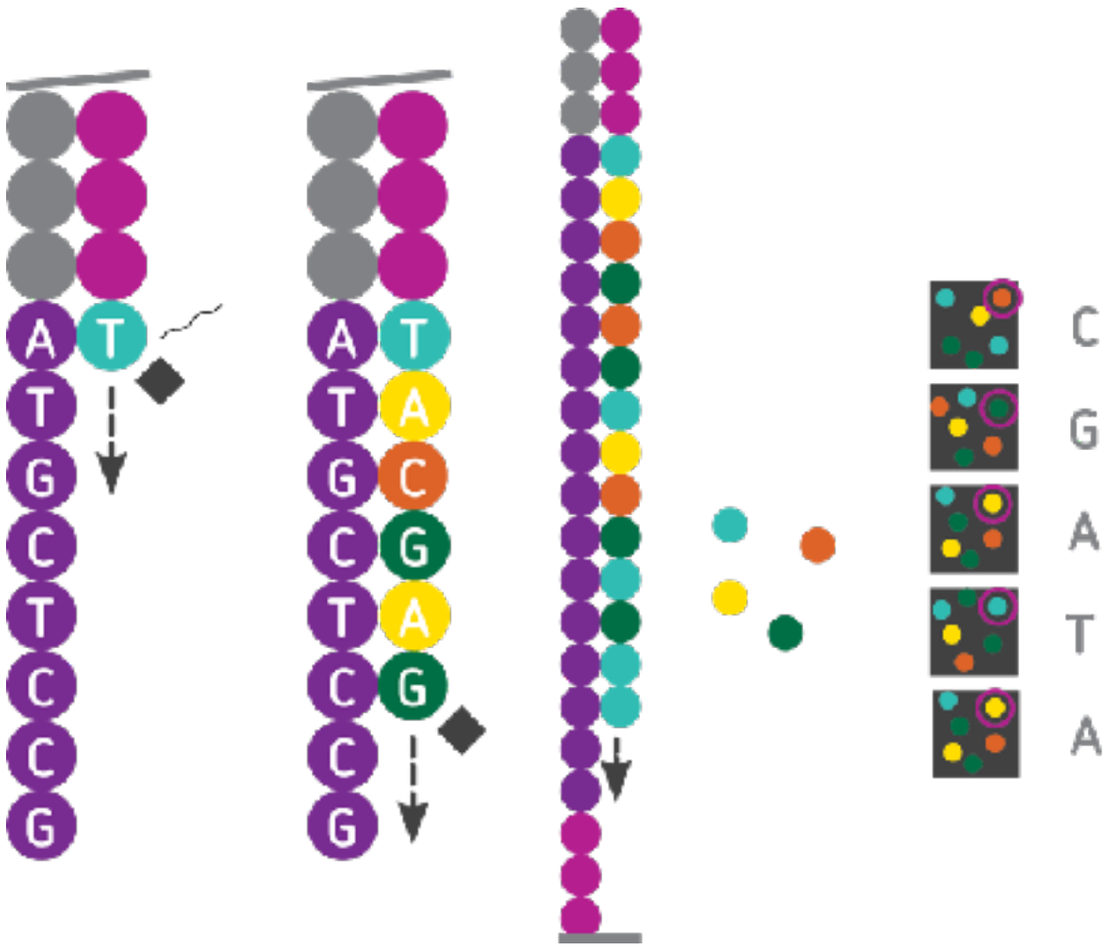
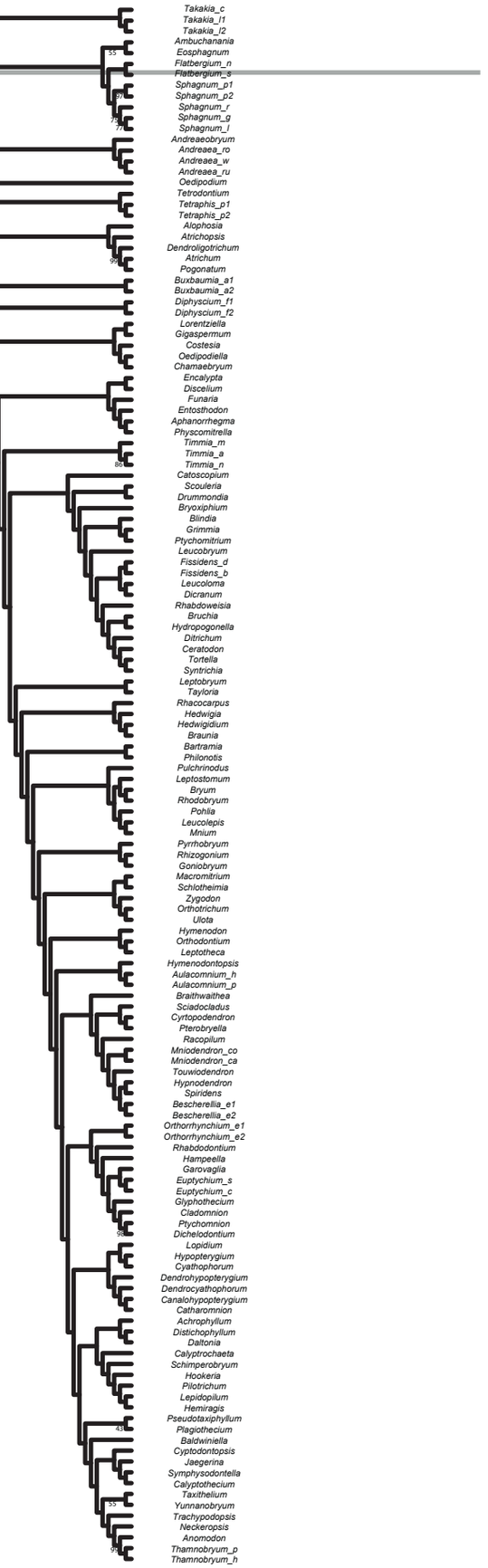
FROM READS TO SEQUENCES



?



ILLUMINA BASESPACE

BaseSpace® illumina®

Elliot Gardner (egardner@u.northwestern.edu) would like to share the run Elliot Gardner Hyb Seq Run 2 with you.

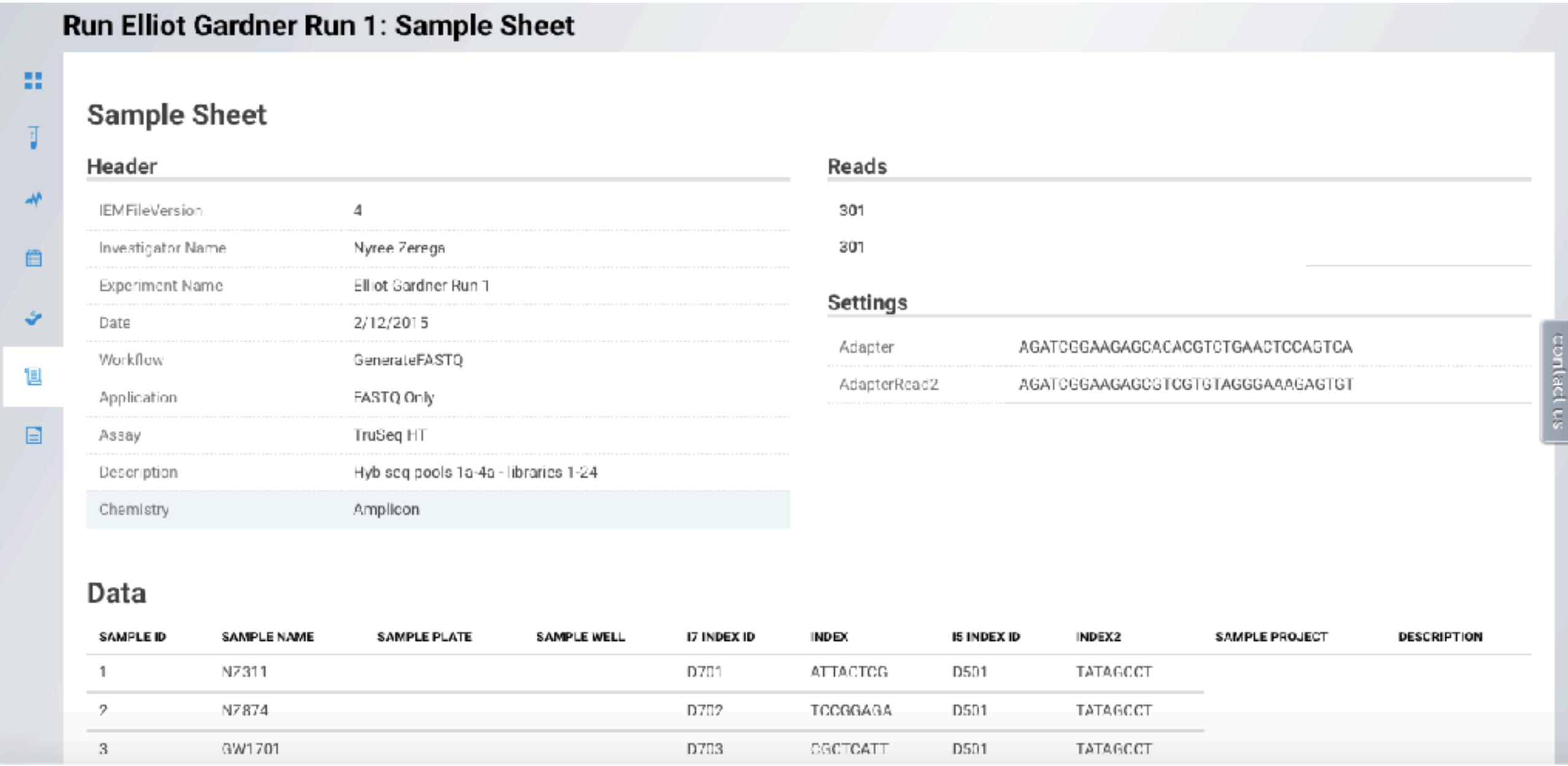
Here's our run!

Go to Elliot Gardner Hyb Seq Run 2 to accept the share.

Enjoy!

The BaseSpace Team

ILLUMINA BASESPACE



ILLUMINA BASESPACE

BaseSpace SEQUENCE HUB

DASHBOARD PREP RUNS PROJECTS APPS PUBLIC DATA

Elliot Gardner

illumina

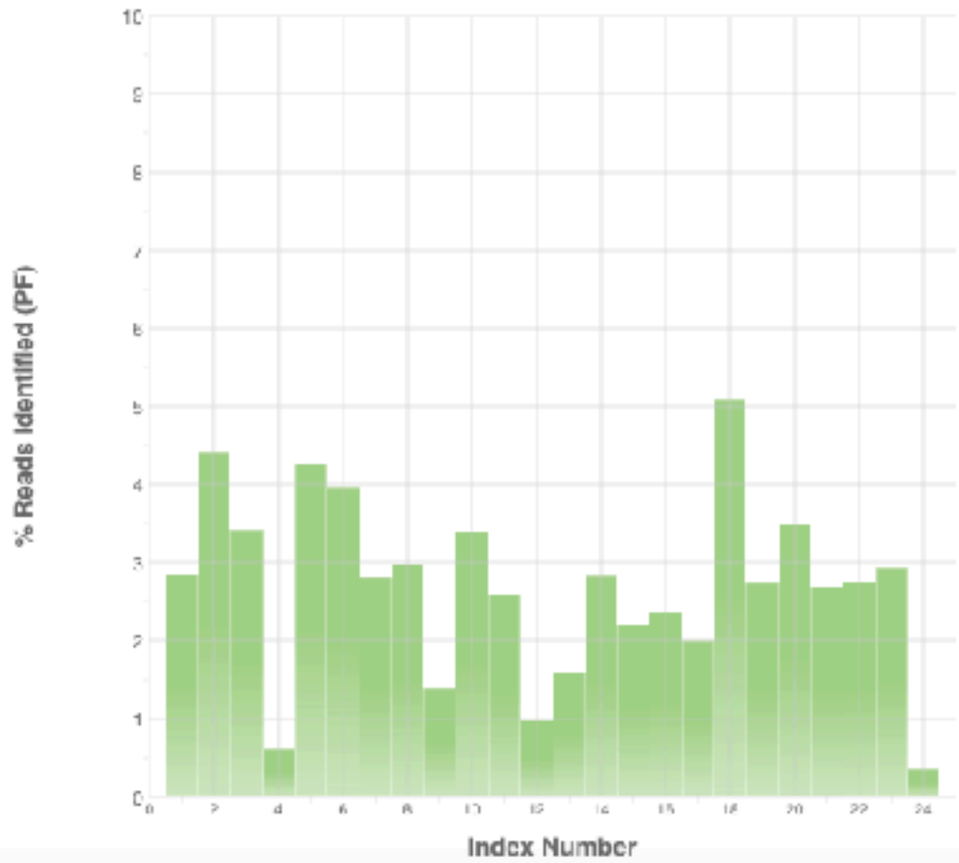
Run Elliot Gardner Run 1: Indexing QC

Lane 1

Reads Mapped to index ID

TOTAL READS	PF READS	% READS IDENTIFIED (PF)	CV	MIN	MAX
15491991	14893818	64.7902	0.4304	0.3630	5.0880

INDEX NUMBER	SAMPLE ID	PROJECT	INDEX 1 (I7)	INDEX 2 (I5)	% READS IDENTIFIED (PF)
1	7	NA	ATTACTCG	ATAGAGGC	2.8560
2	13	NA	ATTACTCG	CCTATCCT	4.4159
3	19	NA	ATTACTCG	GGCTCTGA	3.4150
4	1	NA	ATTACTCG	TATAGCCT	0.6166
5	11	NA	ATTCAGAA	ATAGAGGC	4.2578
6	17	NA	ATTCAGAA	CCTATCCT	3.9668
7	23	NA	ATTCAGAA	GGCTCTGA	2.8016
8	5	NA	ATTCAGAA	TATAGCCT	2.9860
9	9	NA	CGCTCATT	ATAGAGGC	1.3990



contact us

ILLUMINA BASESPACE

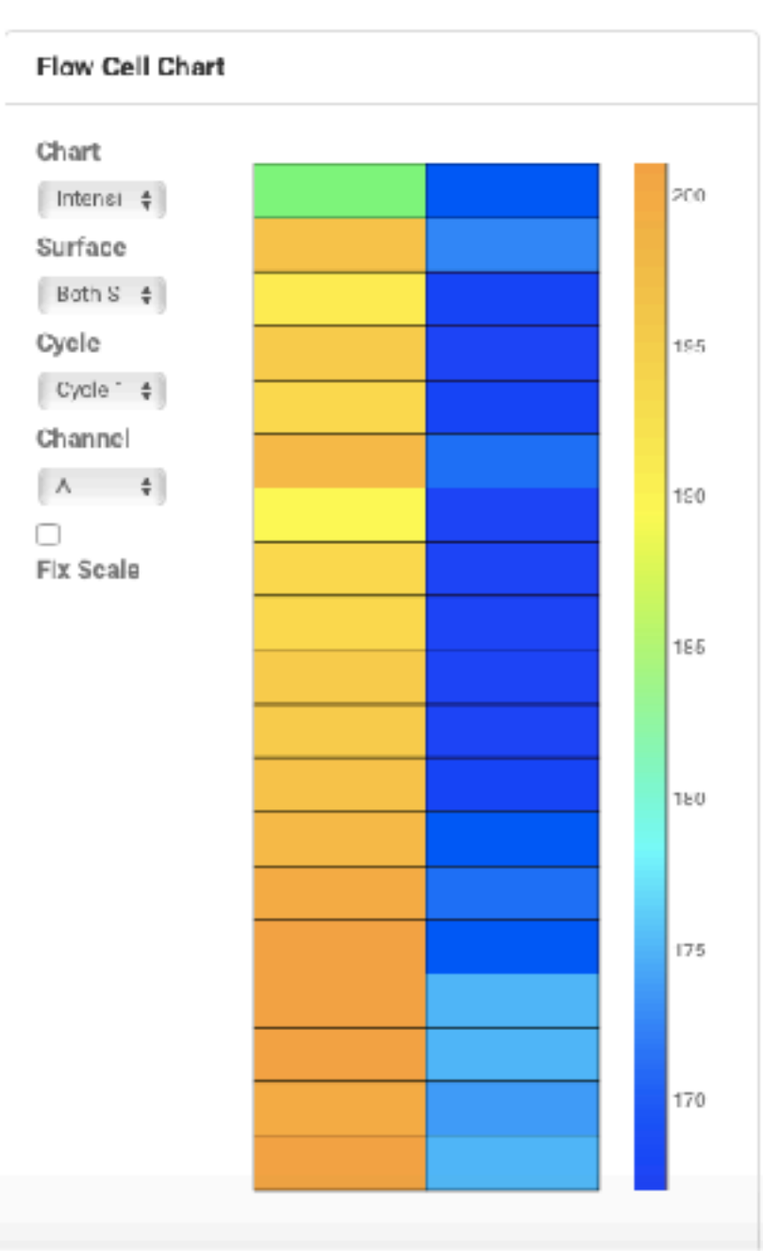
Run Elliot Gardner Run 1: Run & Lane Metrics

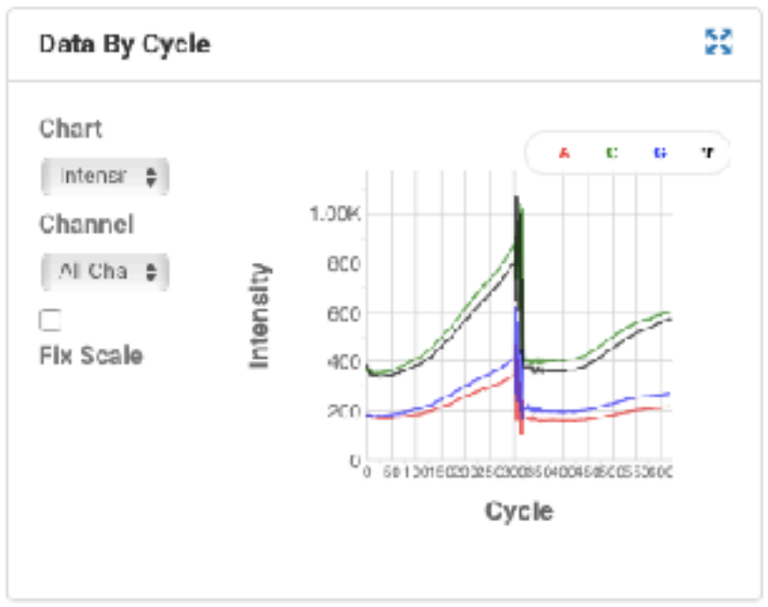
Samples

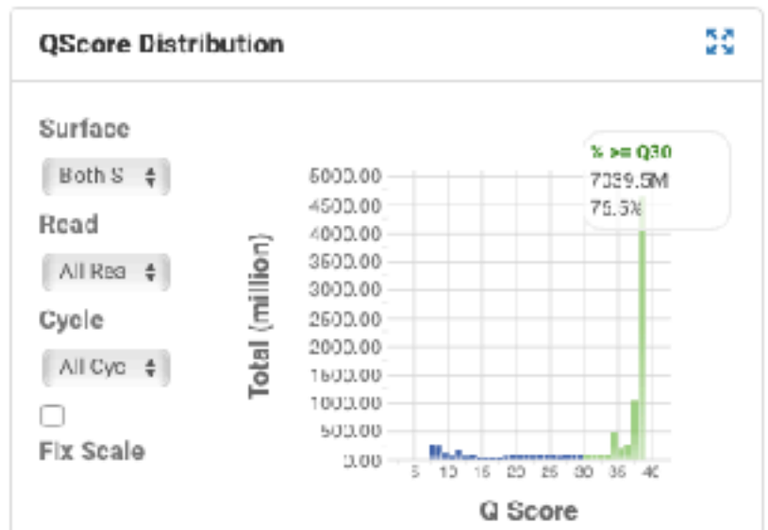
	CYCLES	YIELD	PROJECTED YIELD	ALIGNED (%)	ERROR RATE (%)	INTENSITY CYCLE 1	%≥Q30
Read 1	301	4.47 Gbp	4.47 Gbp	33.99	2.23	183	90.98
Read 2 (I)	8	104.26 Mbp	104.26 Mbp	0.00	0.00	469	69.32
Read 3 (I)	8	104.26 Mbp	104.26 Mbp	0.00	0.00	259	97.08
Read 4	301	4.47 Gbp	4.47 Gbp	33.77	5.58	185	62.08
Non-Index Reads Total	602	8.94 Gbp	8.94 Gbp	33.88	3.91	185	76.53
Totals	**618**	**9.14 Gbp**	**9.14 Gbp**	**33.88**	**3.91**	**275**	**76.68**

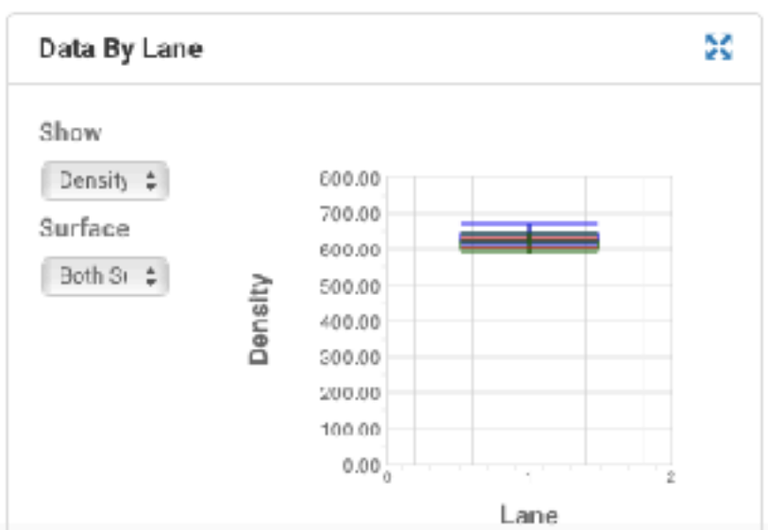
LANE	READ	TILES	DENSITY (K / MM²)	CLUSTER PF (%)	PHAS/PREPHAS (%)	READS	READS PF	%≥Q30	YIELD	CYCLES ERR RATED	ALIGNED (%)	ERROR RATE (%)	ERROR RATE 35 CYCLES (%)	ERROR RATE 75 CYCLES (%)	ERROR RATE 100 CYCLES (%)	INTENSITY CYCLE 1	COMMENTS	STATUS
1	1	38	634 ±14	96.14 ±0.33	0.180 / 0.087	15,491,991	14,893,818	90.98	4.47 Gbp	300	33.99 ±0.49	2.23 ±0.03	0.10 ±0.00	0.14 ±0.00	0.21 ±0.01	183 ±14		Initial
	2 (I)	38	634 ±14	96.14 ±0.33	0.000 / 0.000	15,491,991	14,893,818	69.32	104.26 Mbp	0	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	469 ±35		
	3 (I)	38	634 ±14	96.14 ±0.33	0.000 / 0.000	15,491,991	14,893,818	97.08	104.26 Mbp	0	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	259 ±15		
	4	38	634 ±14	96.14 ±0.33	0.148 / 0.093	15,491,991	14,893,818	62.08	4.47 Gbp	300	33.77 ±0.50	5.58 ±0.31	0.12 ±0.01	0.20 ±0.01	0.27 ±0.01	186 ±17		

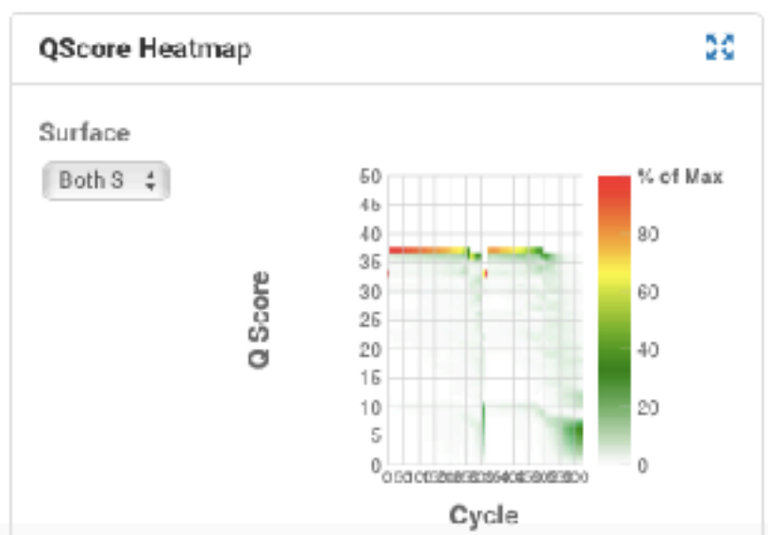
ILLUMINA BASESPACE



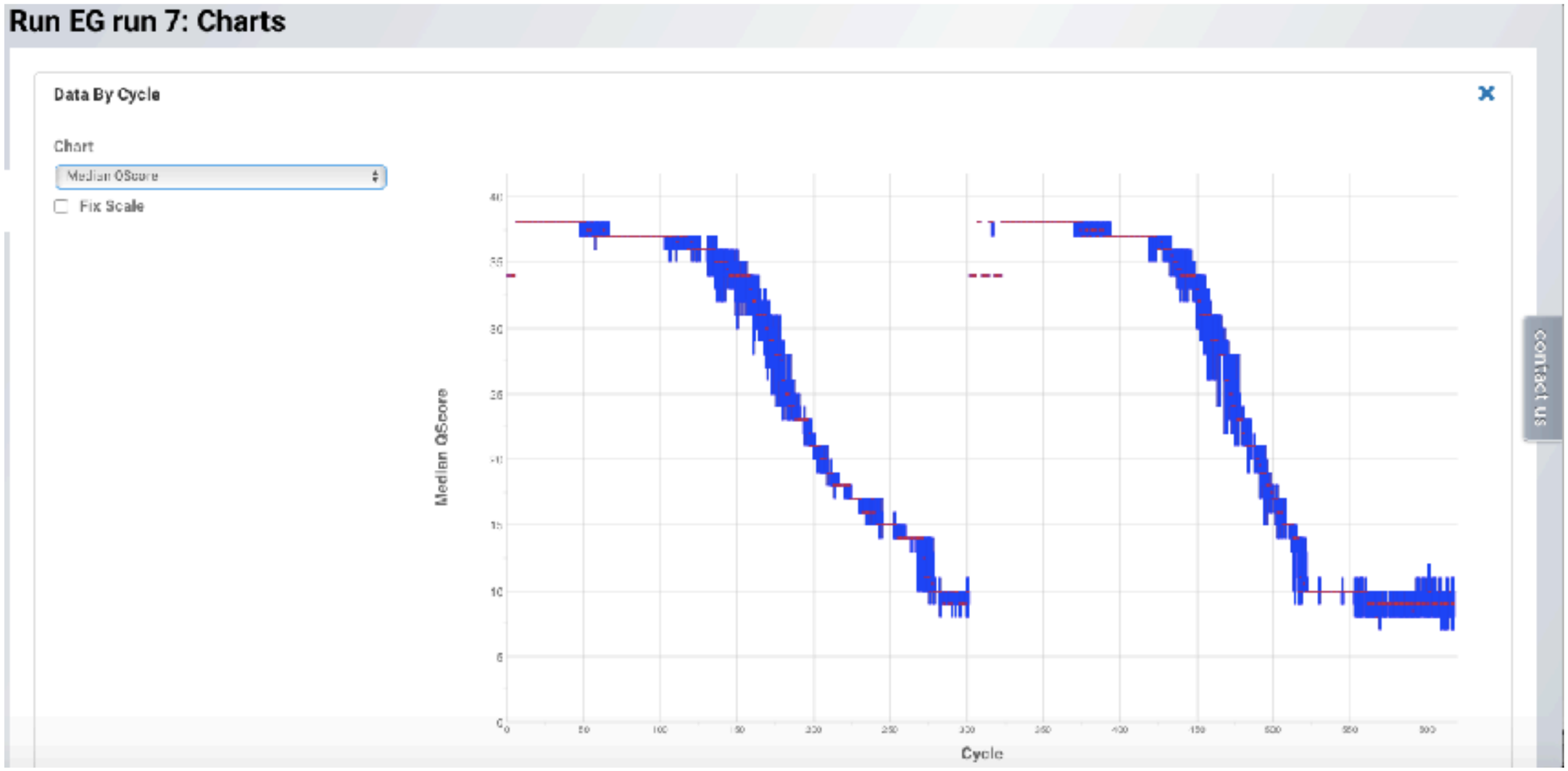




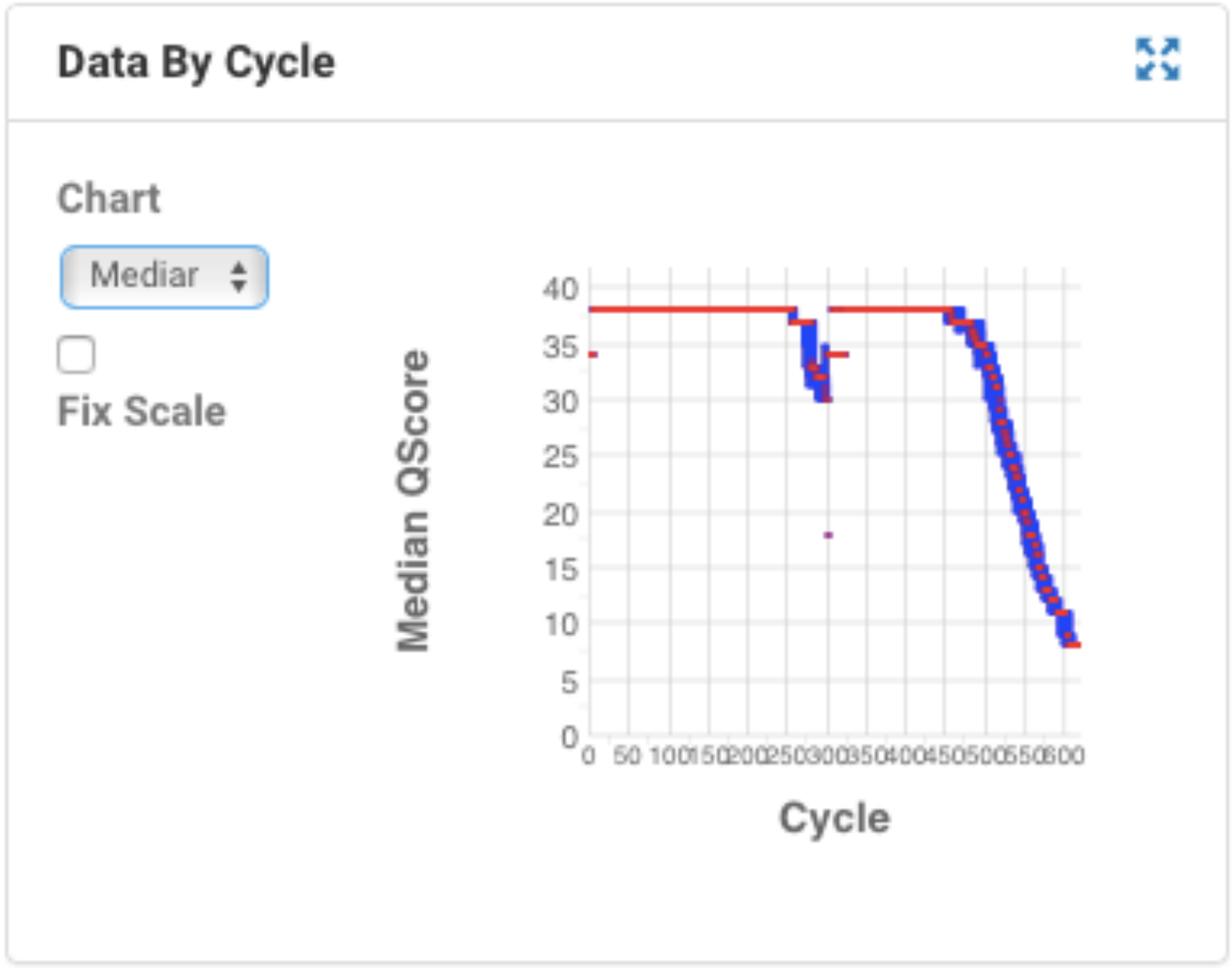




ILLUMINA BASESPACE



ILLUMINA BASESPACE



MiSeq 2x300: beware the reverse reads!

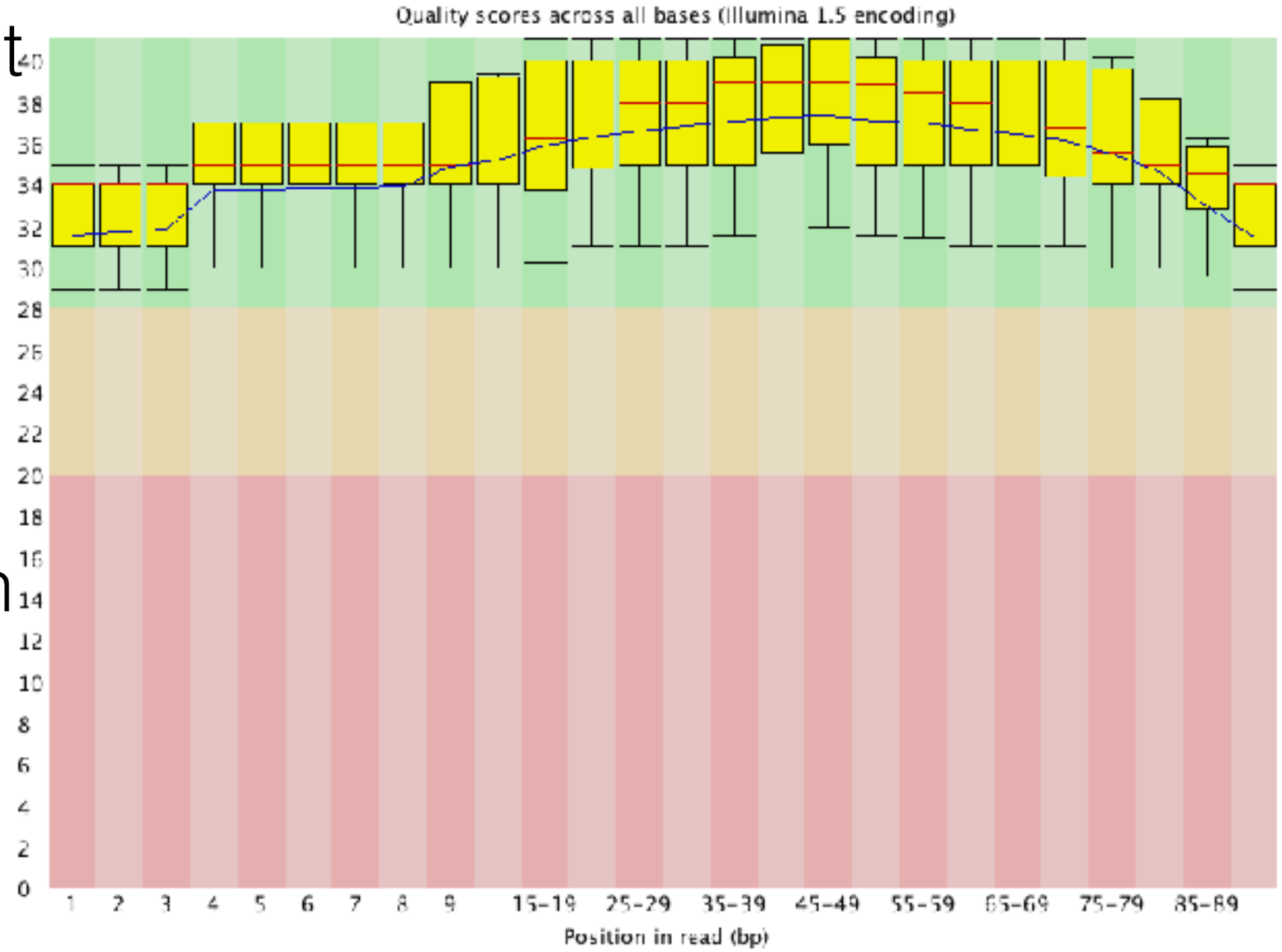
FASTQC

Summarize read content in FASTQ files

Identify Adapter Sequences

Discover Contamination

Suggest Trimming Strategy



https://www.bioinformatics.babraham.ac.uk/projects/fastqc

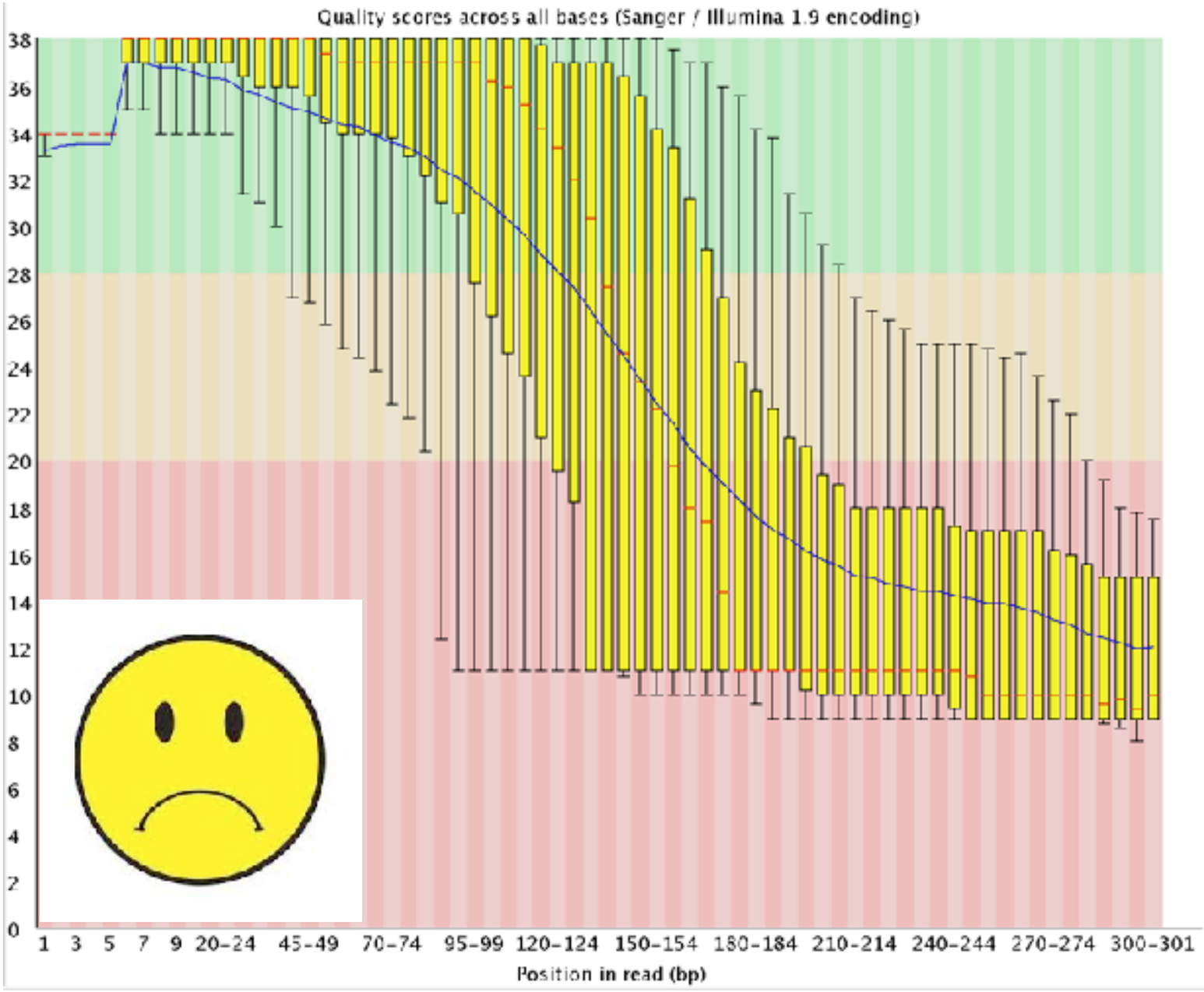
FASTQC

Summarize read content in FASTQ files

Identify Adapter Sequences

Discover Contamination

Suggest Trimming Strategy



https://www.bioinformatics.babraham.ac.uk/projects/fastqc

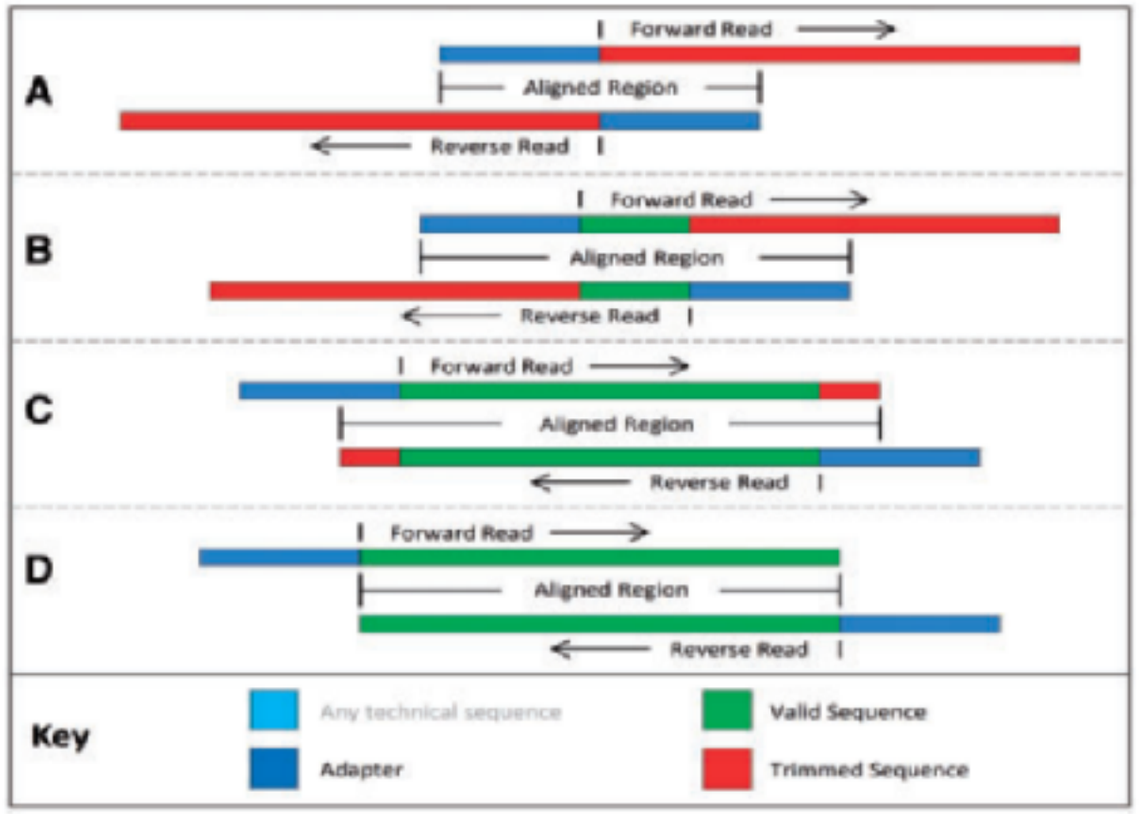
TRIMMOMATIC

Process FASTQ files

Remove adapter sequences

Trim bases by quality score

Maintains read pairs



Bolger et al., 2014, Bioinformatics

http://www.usadellab.org/cms/?page=trimmomatic

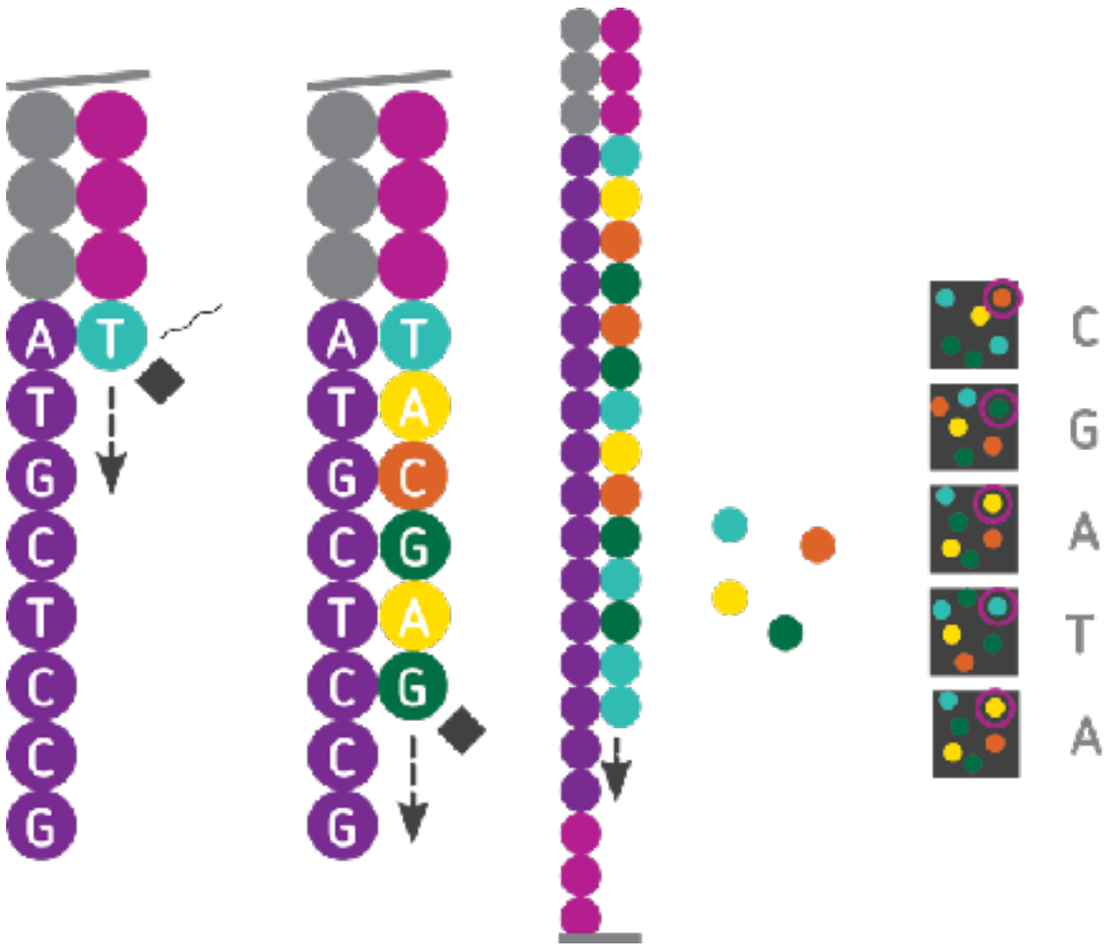
HANDS ON: PROCESSING SEQUENCES

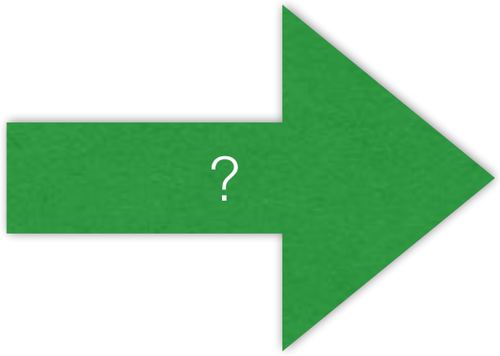
Analyze reads in FASTQC

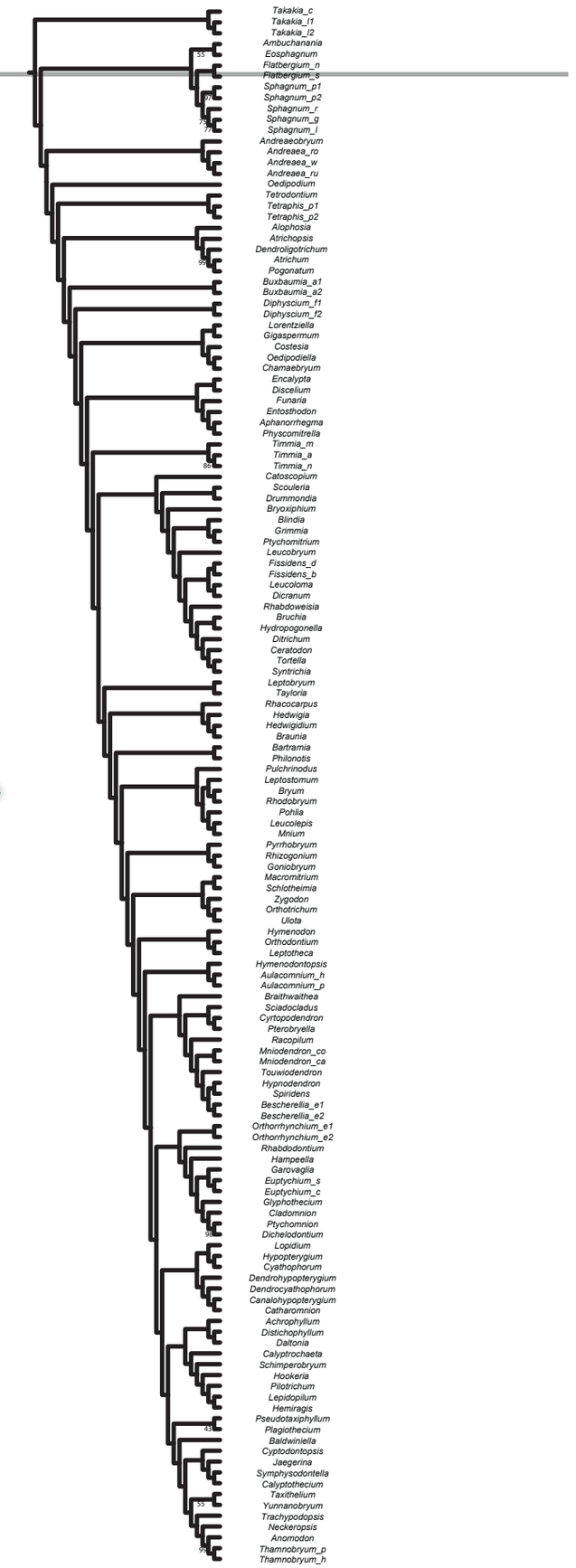
Trim Reads in Trimmomatic

Discuss trimming results

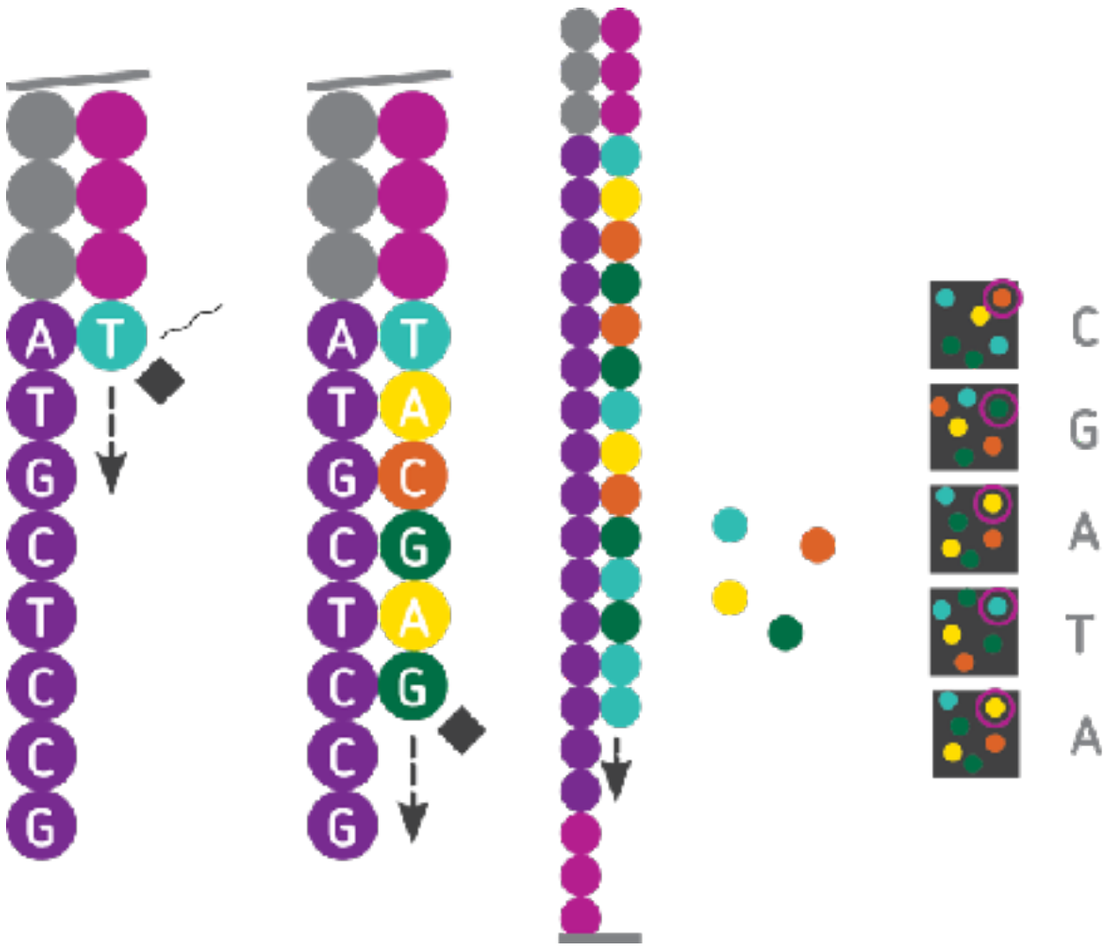
FROM READS TO SEQUENCES



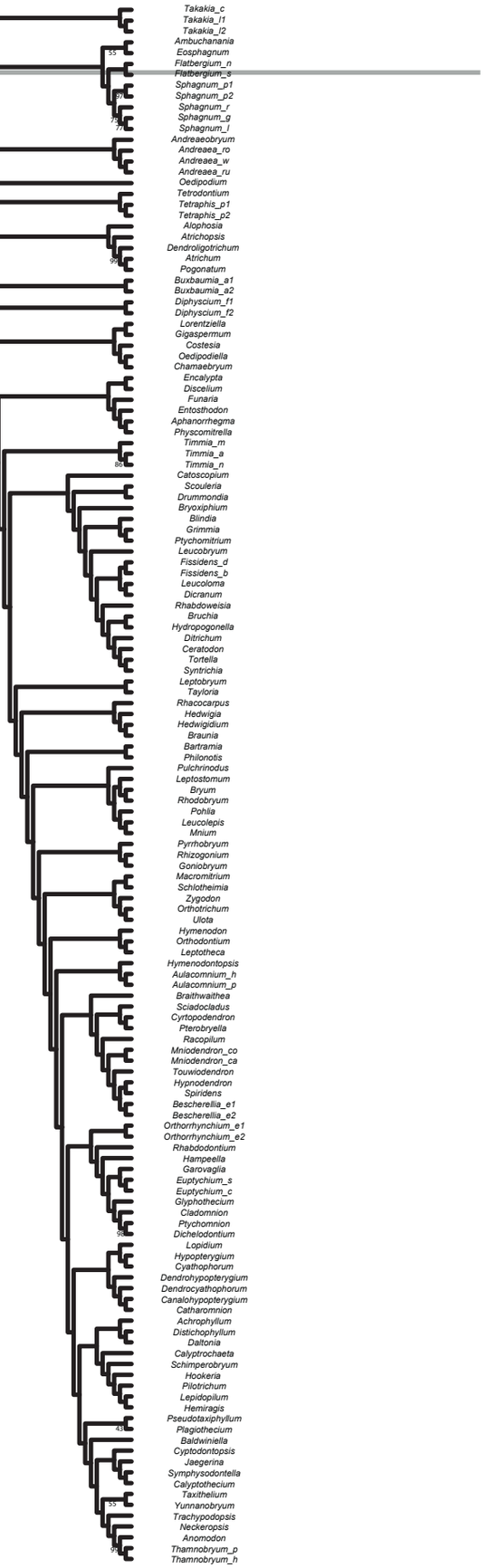


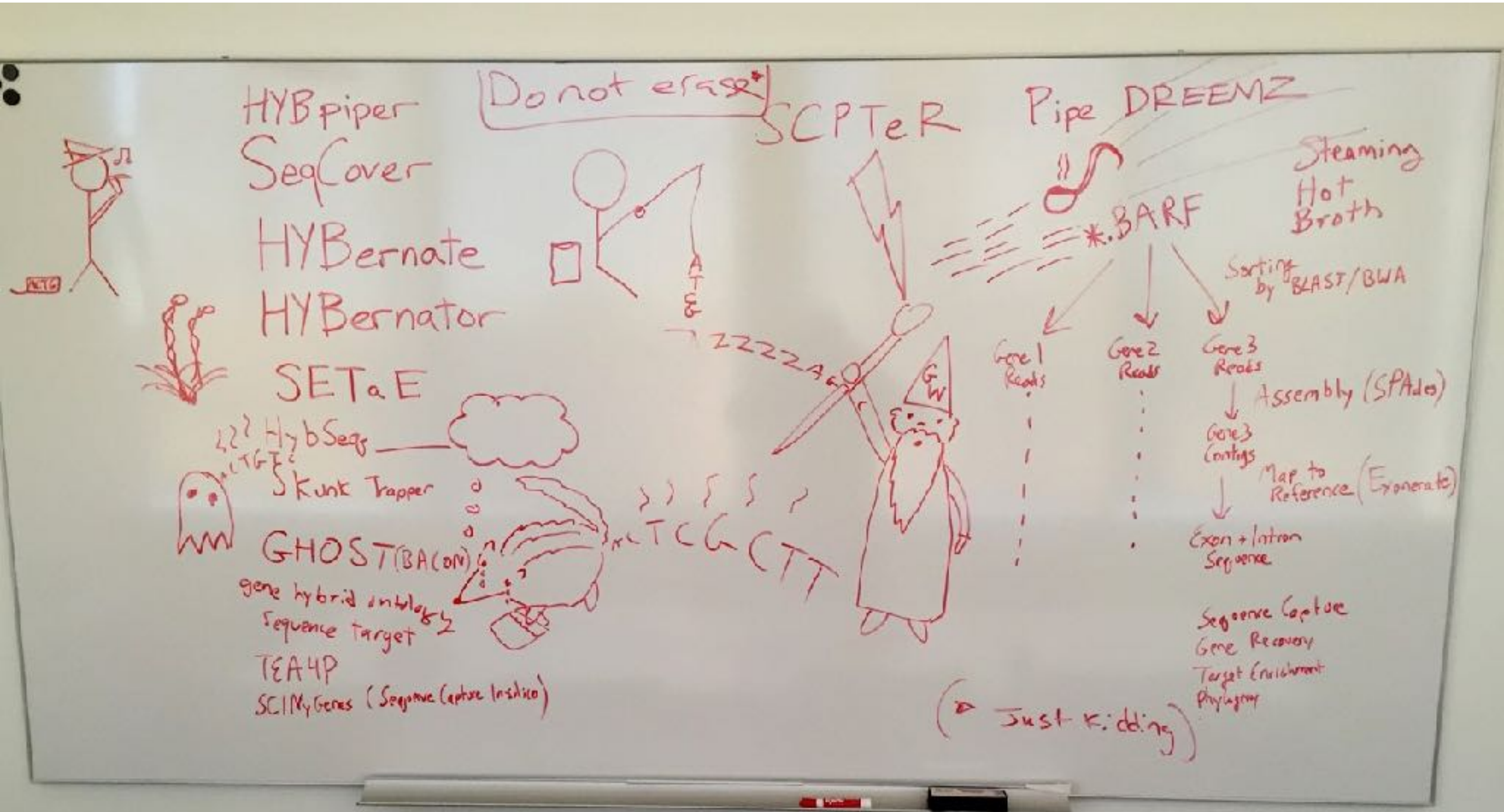


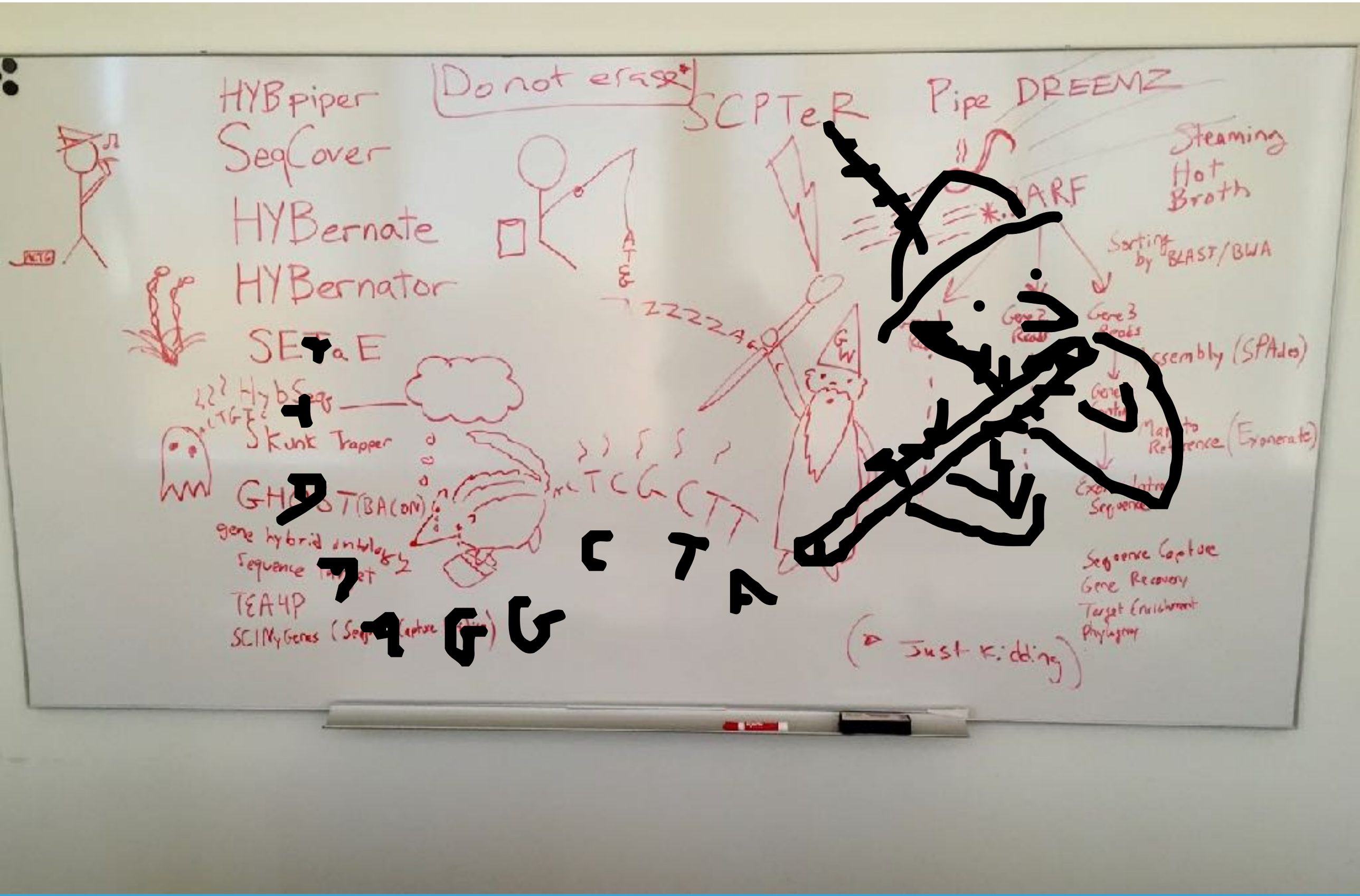
FROM READS TO SEQUENCES



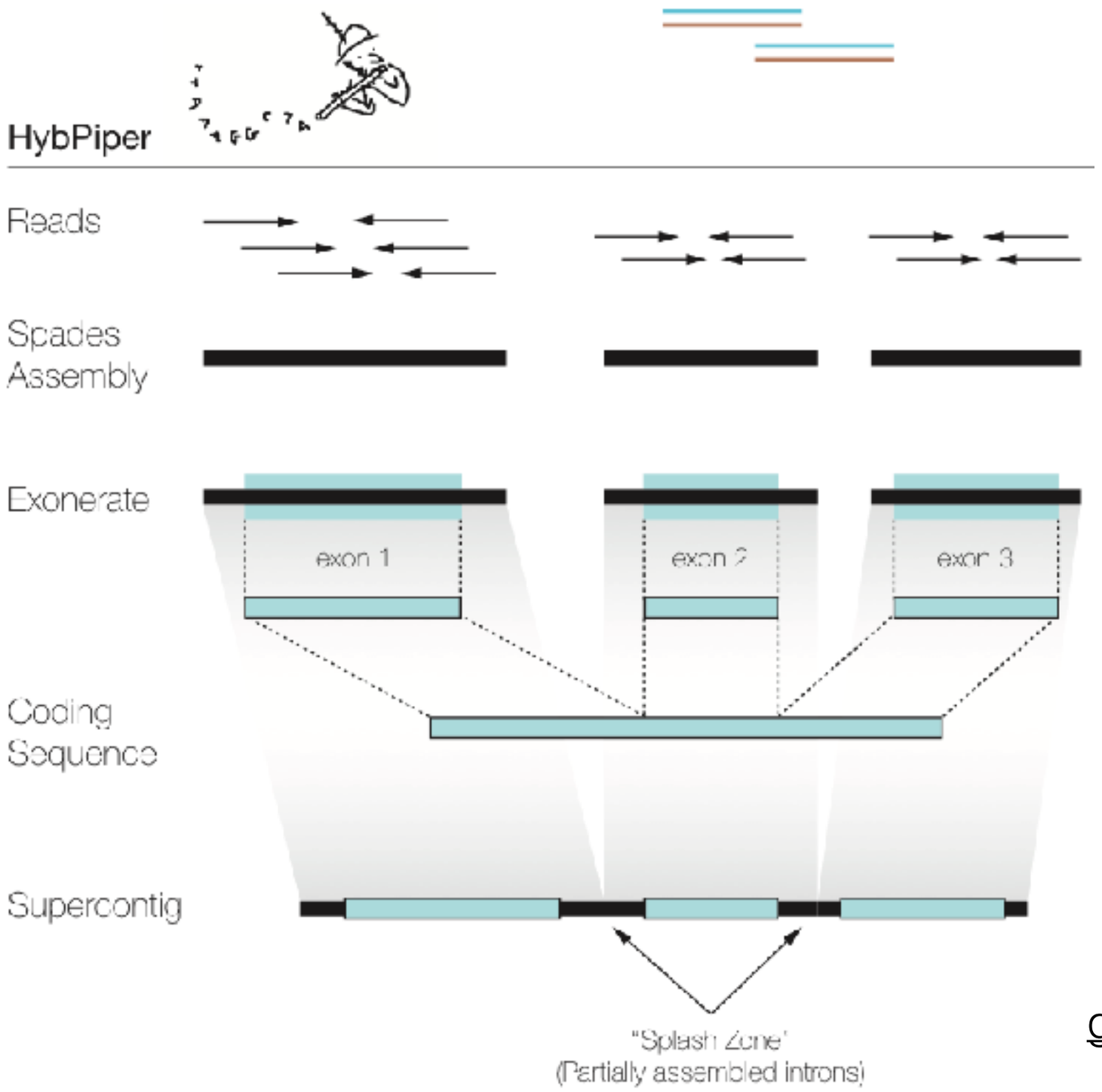
HybPiper!




HYBpiper
SeqCover
HYBernate
HYBernator
SETaE
Do not erase*
SCPTeR
Pipe DREEMZ
*.BARF
Steaming Hot Broth
Sorting by BLAST/BWA
Gene1 Reads
Gene2 Reads
Gene3 Reads
Assembly (SPAdes)
Gene3 Contigs
Map to Reference (Exonerate)
Exon + Intron Sequence
Sequence Capture
Gene Recovery
Target Enrichment
Phylogeny
ATG
ZZZZZ
GW
HybSeqs
Skunk Trapper
GHOST(BACON)
gene hybrid enrichment sequence target
TEAHP
SCINyGenes (Sequence Capture Insilico)
TCGCTT
(* Just Kidding)


HYBpiper
SeqCover
HYBernate
HYBernator
SETaE
Do not erase
SCPTeR
Pipe DREEMZ
Steaming Hot Broth
Sorting by BLAST/BWA
Junk Trapper
GHOST(BACON)
gene hybrid anthology
TEA4P
SCINyGenes
TCGCTT
Just Kidding
Sequence Capture
Gene Recovery
Target Enrichment
Phylogeny

THREE PHASES OF HYBPIPER



1. Sort Reads

2. Assemble

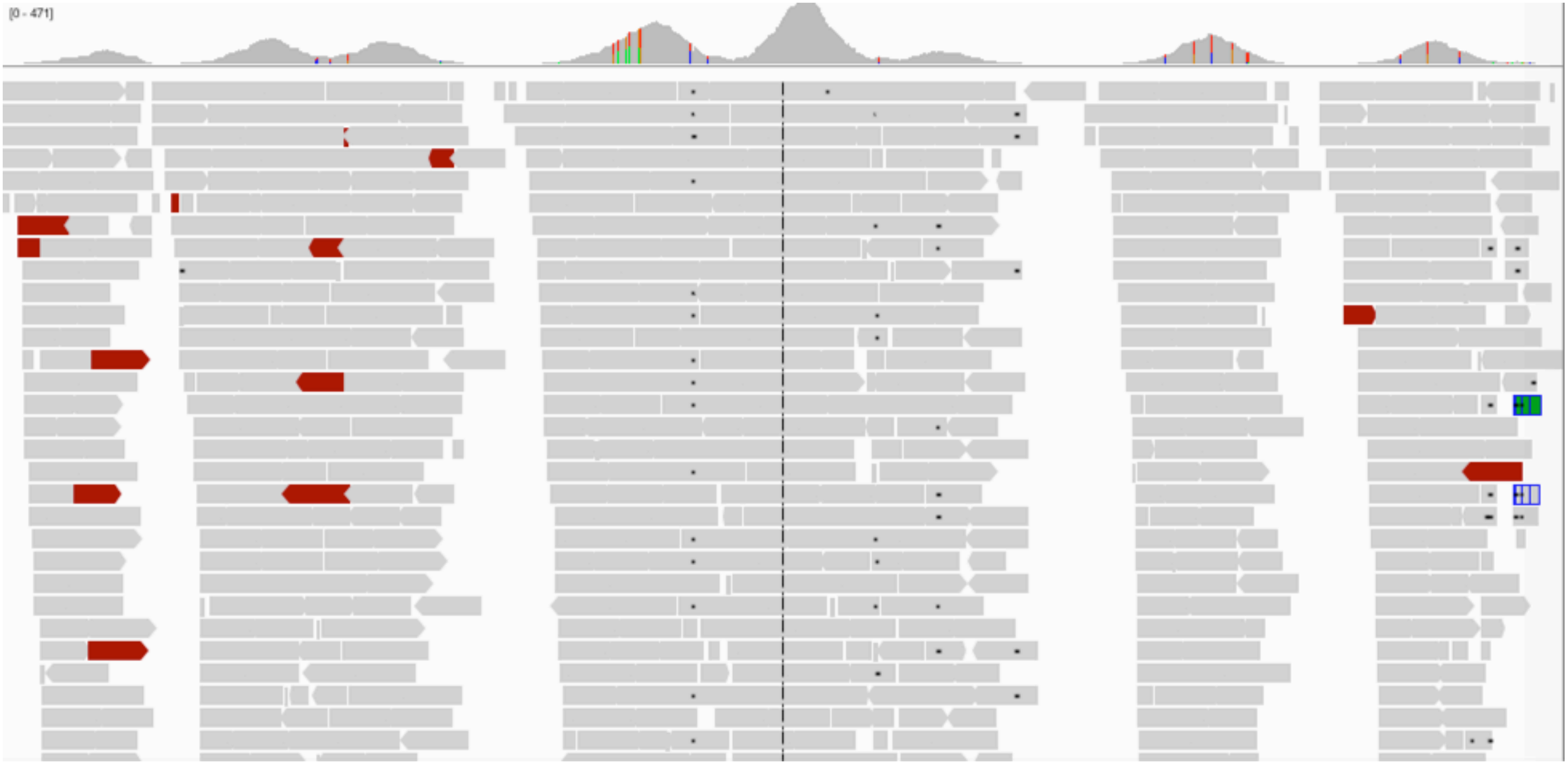
3. Extract Exons

github.com/mossmatters/HybPiper
Johnson et al. 2016 *APPS*

DEPTH OF COVERAGE

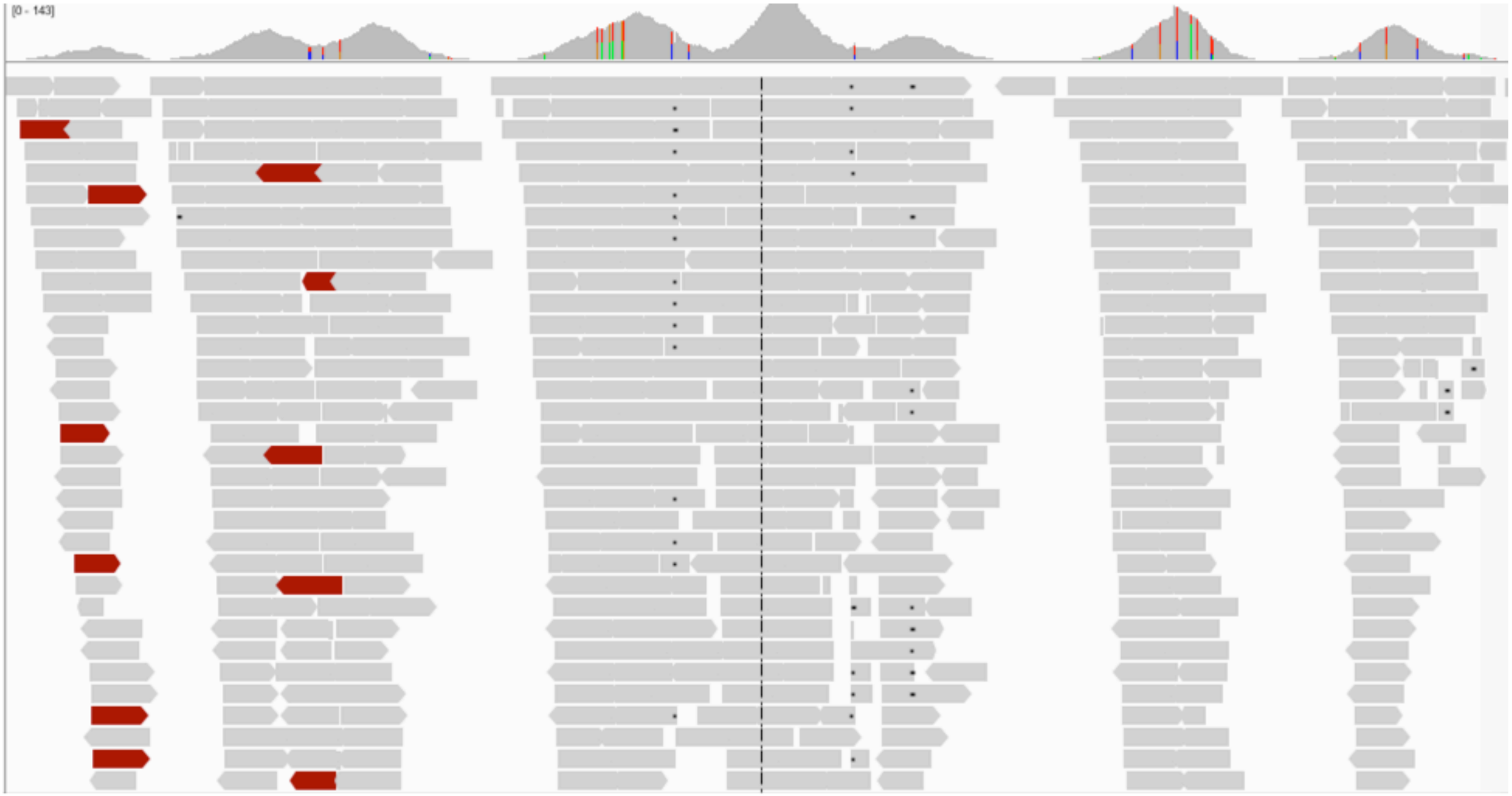
PCR Duplication is high

Depth is variable (highest in exons)



PCR DUPLICATES

Removing duplicates is not advised without mapping. This may accentuate sequencing errors!

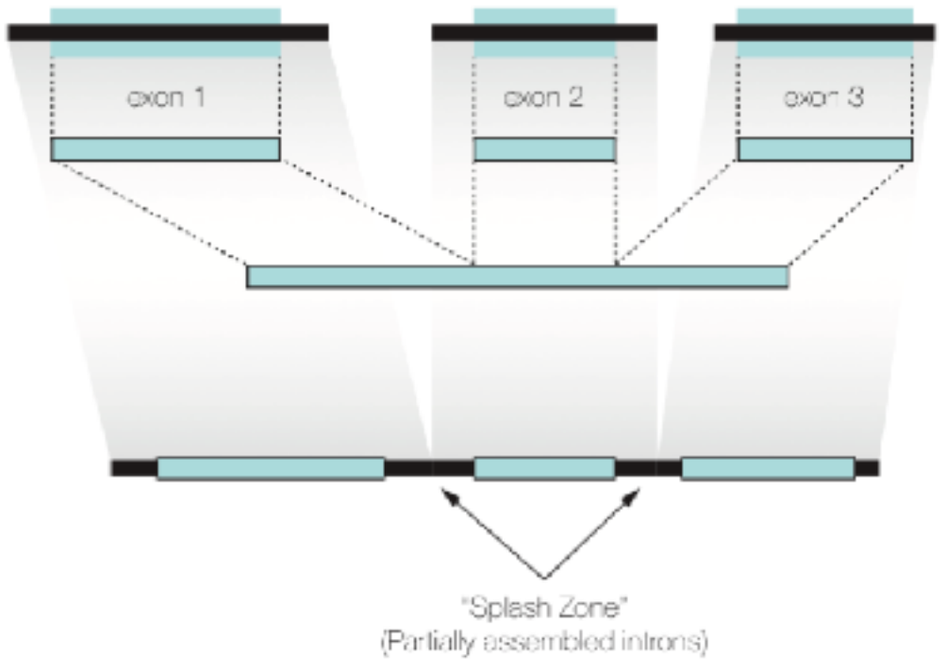


EXTRACTING EXONS WITH EXONERATE

Assembled contigs are aligned to target protein sequences

Introns are identified and cut from sequence

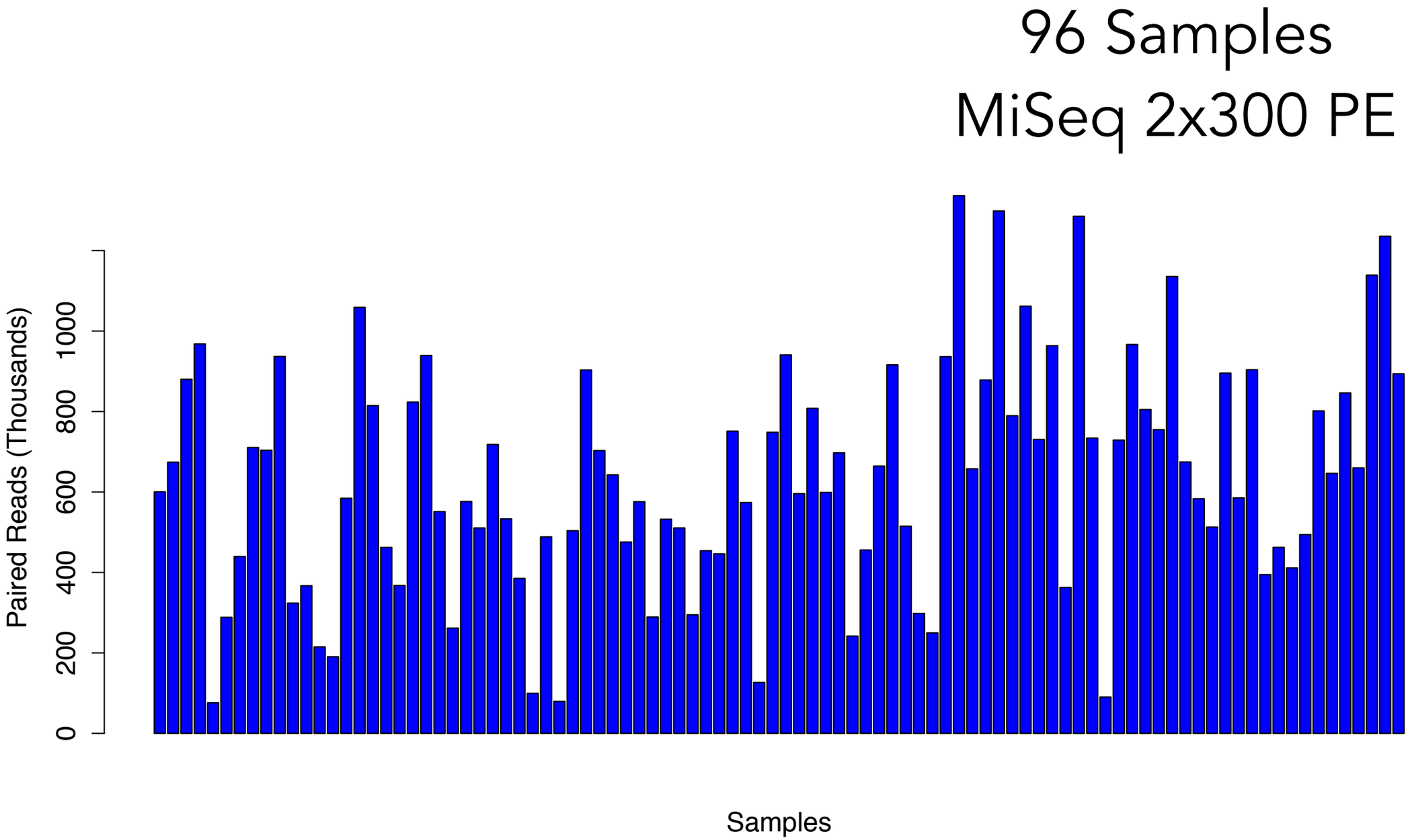
Non-contiguous hits are assembled into a "supercontig"




```
Command line: [exonerate -m protein2genome temp.prot.fa temp.contig.fa]
Hostname: [CBG002827.local]

C4 Alignment:
------------
         Query: Syntrichia-rpl16
        Target: Syntrichia-rpl16 <unknown description>
         Model: protein2genome:local
     Raw score: 636
   Query range: 2 -> 134
  Target range: 94 -> 1039

   3 : SerProLysArgThrLysPheArgLysGlnHisArgGlyArgMetLysGlyIleSerThrAr :   23
       !:!|||||||||||||||||||||||||||||||||||||||||||||||||||||:|||||||
       AsnProLysArgThrLysPheArgLysGlnHisArgGlyArgMetLysGlyIleAlaThrAr
  95 : AACCCTAAAAGAACAAAATTTCGTAAACAACATAGAGGAAGGATGAAAGGAATAGCTACTCG :  155

  24 : gGlyAsnSerIleCysPheGlyLysPheAlaLeuGlnAlaLeuGluProAlaTrpIleThrS :   44
       |||||||||||||..!||||||||||||||||||||||||||||||:!!||||||||||||
       gGlyAsnSerIleAlaPheGlyLysPheAlaLeuGlnAlaLeuGluProSerTrpIleThrS
 156 : AGGTAATTCTATTGCTTTTGGTAAATTTGCCCTTCAAGCACTTGAACCATCTTGGATTACAT :  218

  45 : erArgGlnIleGluAlaGlyArgArgAlaIleThrArgTyrAlaArgArgGlyGlyLysLeu :   64
       ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
       erArgGlnIleGluAlaGlyArgArgAlaIleThrArgTyrAlaArgArgGlyGlyLysLeu
 219 : CAAGACAAATAGAAGCCGGACGACGAGCAATTACACGTTATGCACGTCGTGGTGGAAAATTA :  278

  65 : TrpIleArgIlePheProAspLysProIleThrMetArgProAlaGluThrArgMetGlySe :   85
       ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
       TrpIleArgIlePheProAspLysProIleThrMetArgProAlaGluThrArgMetGlySe
 279 : TGGATACGTATATTTCCAGATAAGCCTATTACTATGCGACCTGCTGAAACACGTATGGGTTC :  341

  86 : rGlyLysGlySerProGluTyrTrpValSerValValLysProGlyArgIleLeuTyrGluI :  106
       ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
       rGlyLysGlySerProGluTyrTrpValSerValValLysProGlyArgIleLeuTyrGluI
 342 : AGGAAAAGGTTCACCAGAATATTGGGTATCTGTTGTTAAGCCAGGTAGAATATTATATGAAA :  404

 107 : leSerGlyValProGluThrValAlaArgAla{A}  >>>> Target Intron 1 >>>> :  117
       ||||||||||||||||!:!|||||||||||||{|}           549 bp
       leSerGlyValProGluSerValAlaArgAla{A}-+
 405 : TAAGCGGAGTACCTGAAAGTGTTGCTAGAGCA{G}ct..................... :  440

 118 :   {la}MetArgIleAlaAlaTyrLysMetProIleArgThrGlnPheIleThrSer :  134
       {||}||||||||||||||||||||||||||||||||||||||||||||||||||:!!
       ++{la}MetArgIleAlaAlaTyrLysMetProIleArgThrGlnPheIleThrAla
 441 : ag{CT}ATGAGAATTGCAGCTTATAAAATGCCTATTCGTACTCAATTTATTACTGCT : 1039
```

Exonerate: http://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate

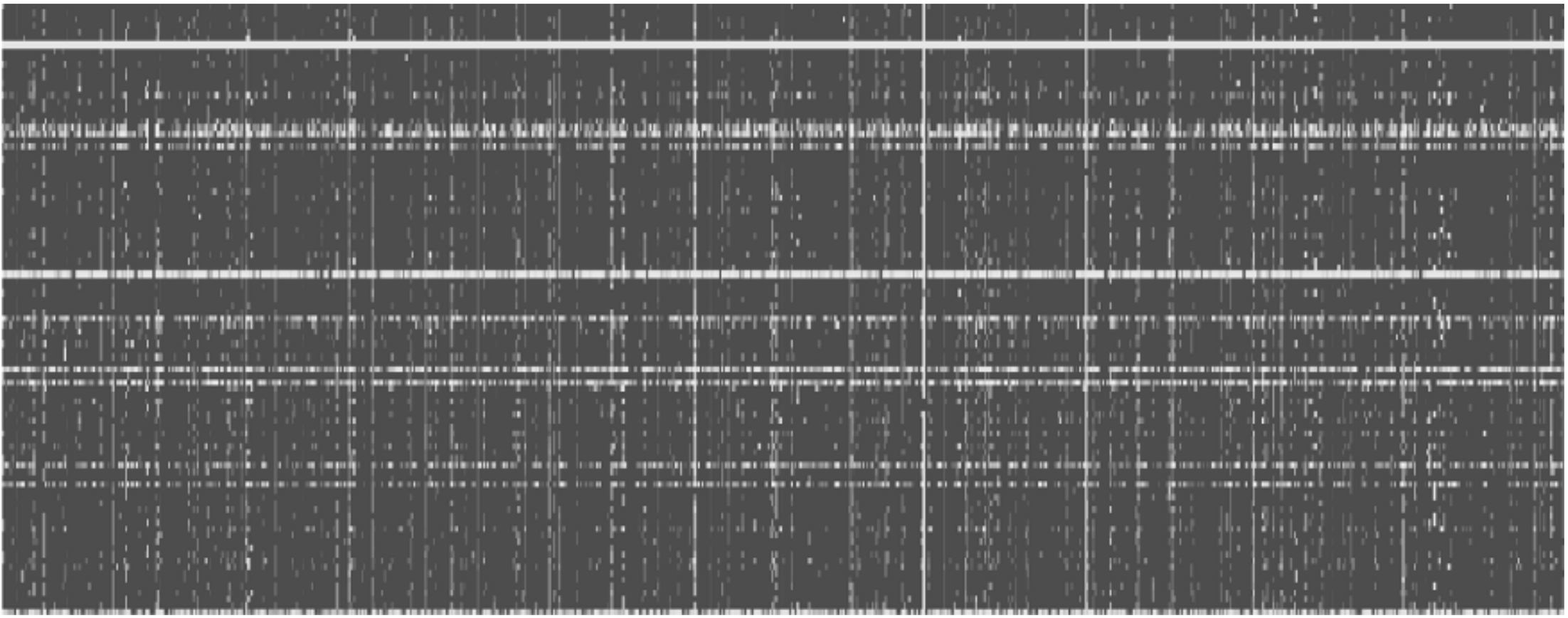
READS PER SAMPLE (MOSS TREE OF LIFE)



READS ON TARGET (MOSS TREE OF LIFE)

Mean: 92% On Target

GENE RECOVERY (MOSS TREE OF LIFE)

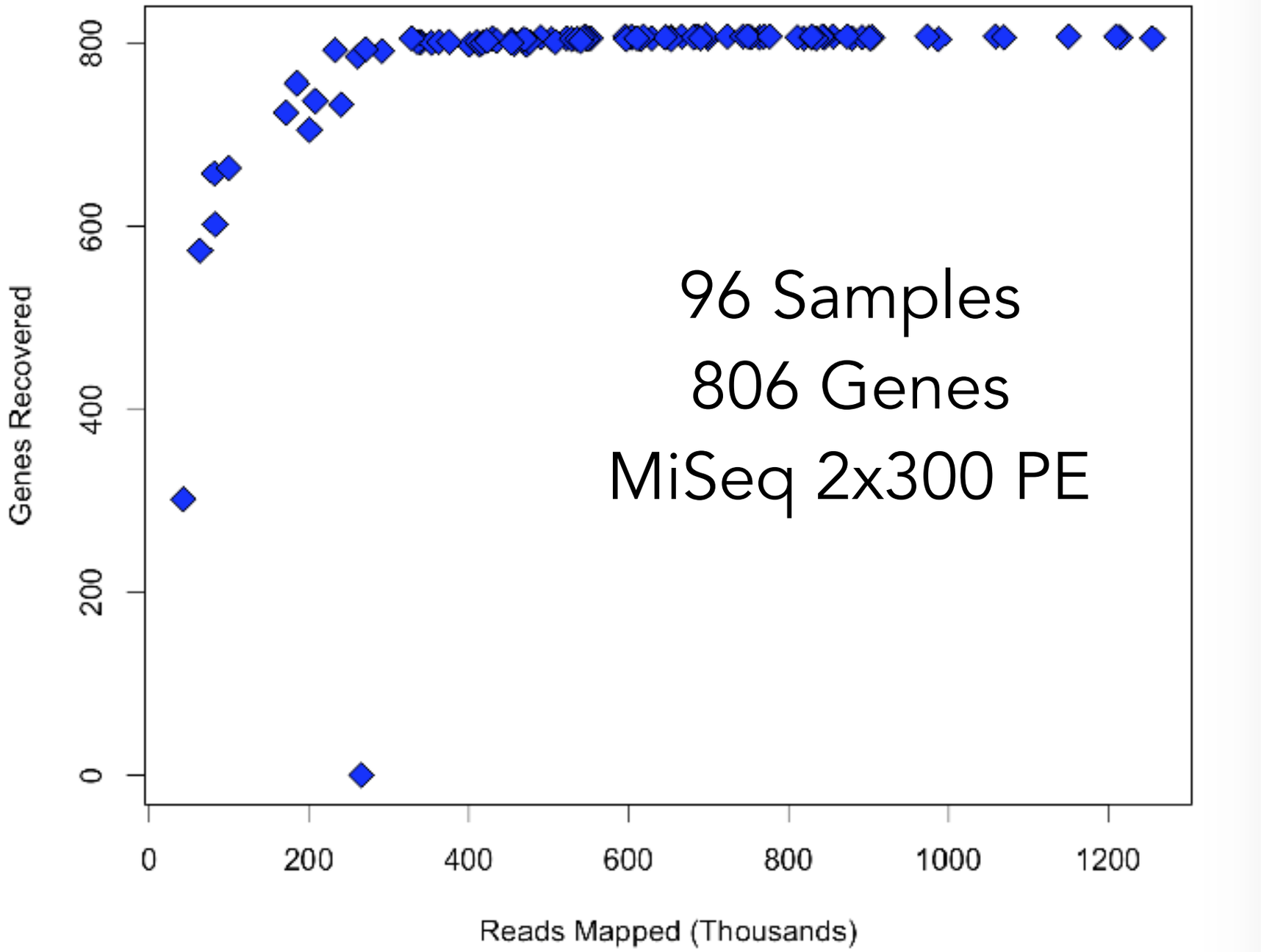


96 Samples

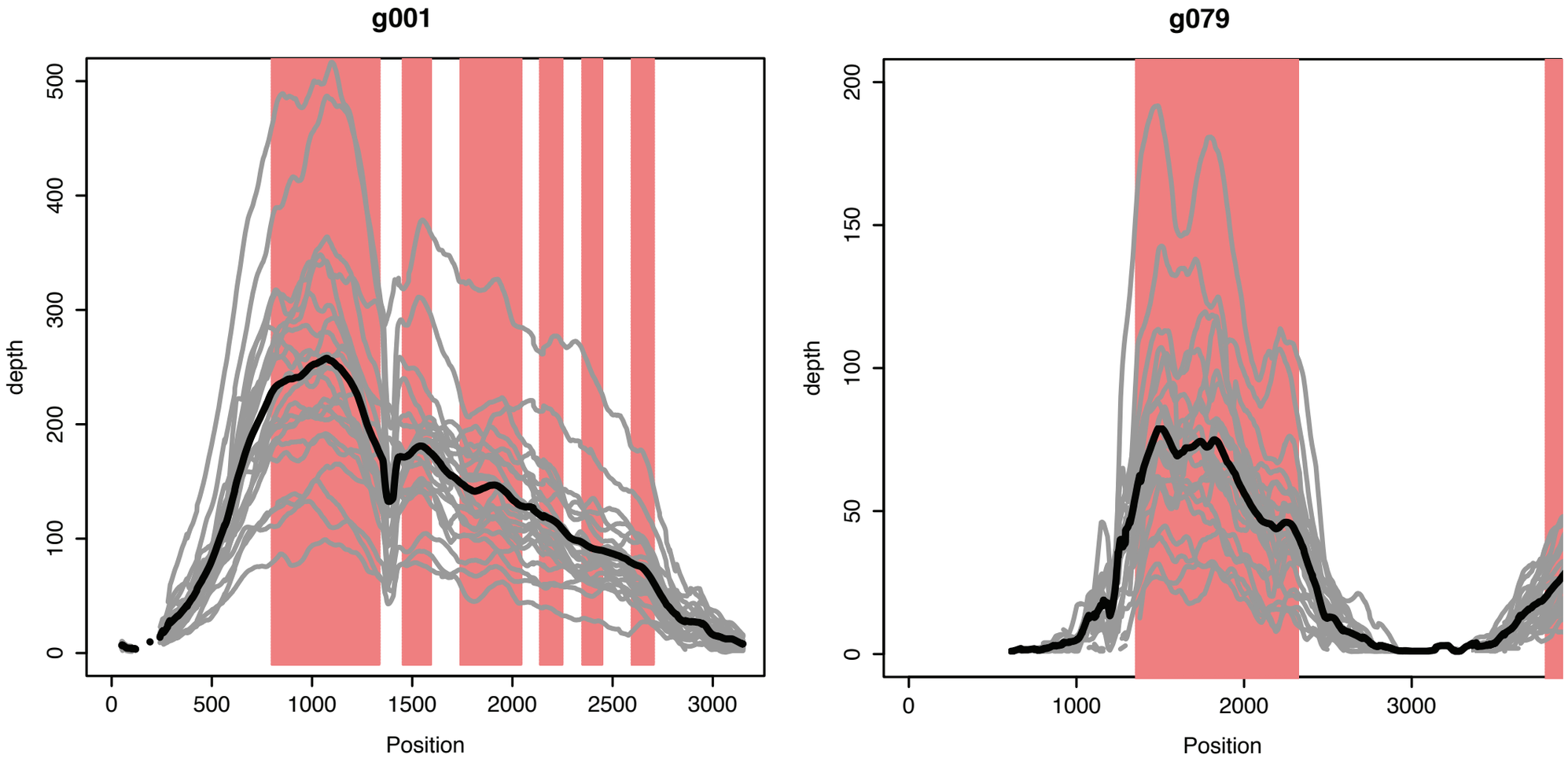
806 Genes

Percent Length vs. Target Gene

GENE RECOVERY (MOSS TREE OF LIFE)



RECOVERING INTRONS WITH HYBPIPER



Artocarpus camansi
(breadfruit)

Johnson et al., 2016, *Applications in Plant Sciences*

Exons

Introns

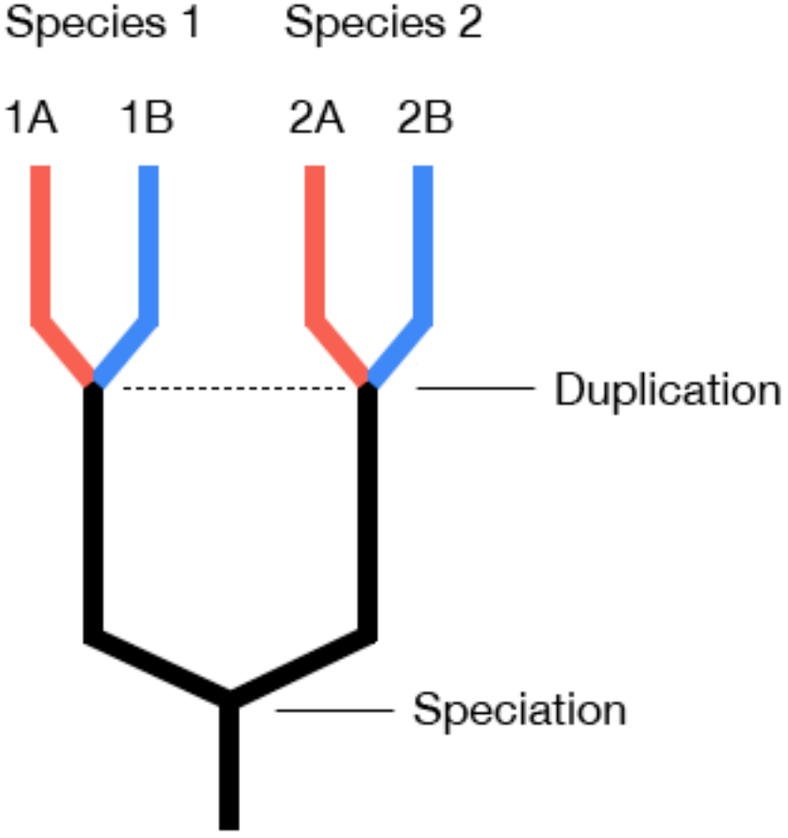
Sequencing Depth
(50 bp window)

Average Depth
(25 species)

DEALING WITH PARALOGS

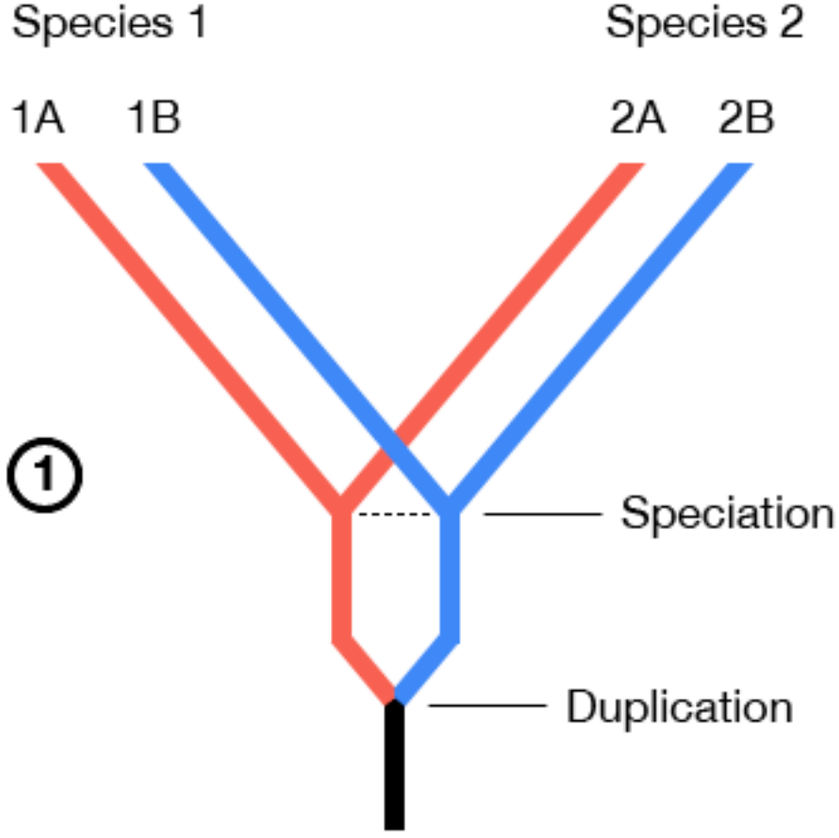
Paralogs: Gene copies arising via duplication

Orthologs: Gene copies arising via speciation



Orthologs are needed to reconstruct species phylogeny

Many paralogs at the same node: Whole Genome Duplication?

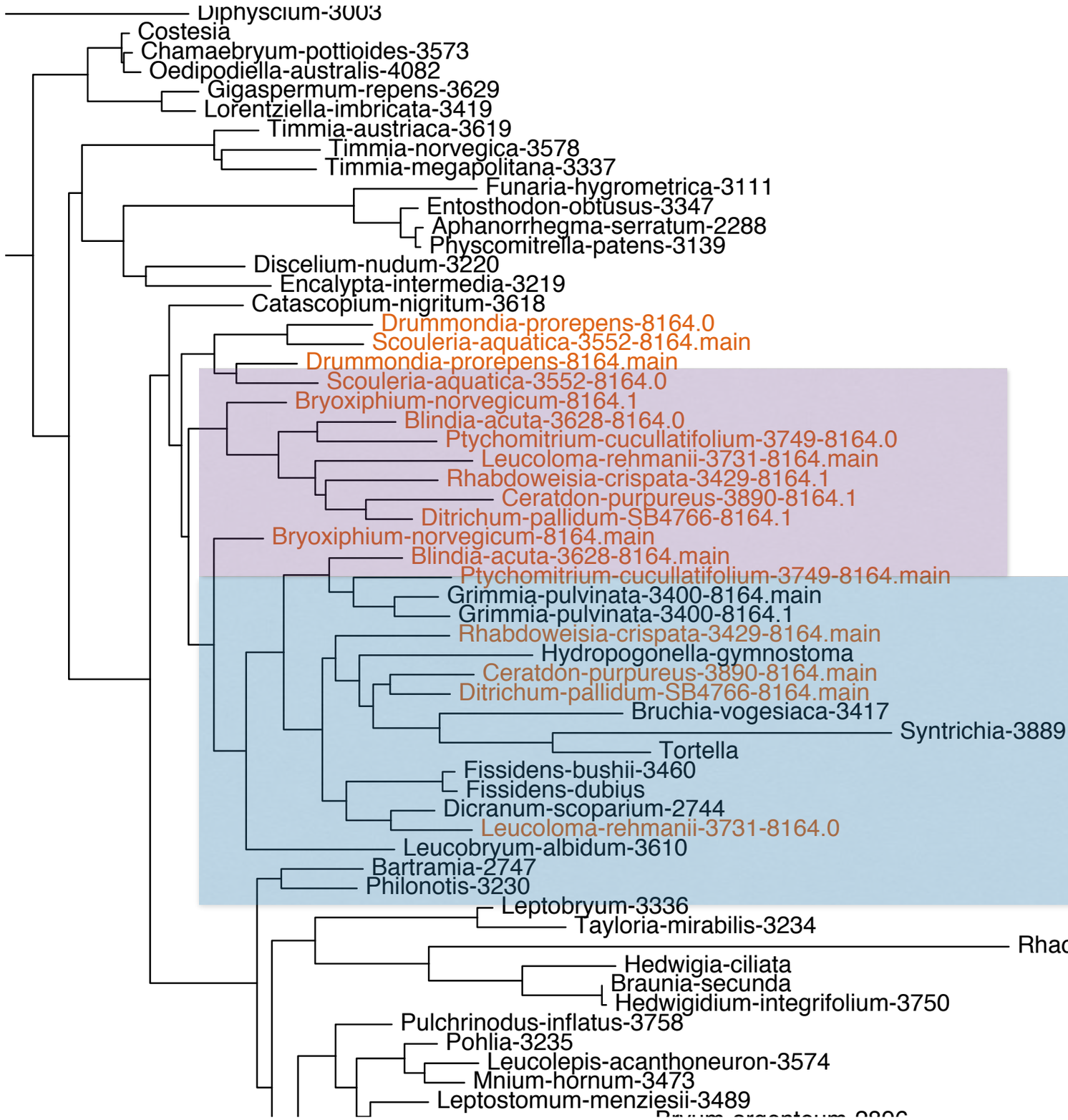


BAD PARALOGS

HybPiper can collect putative paralogs

Build initial gene trees

Identify previously unknown gene and genome duplications outside target taxa



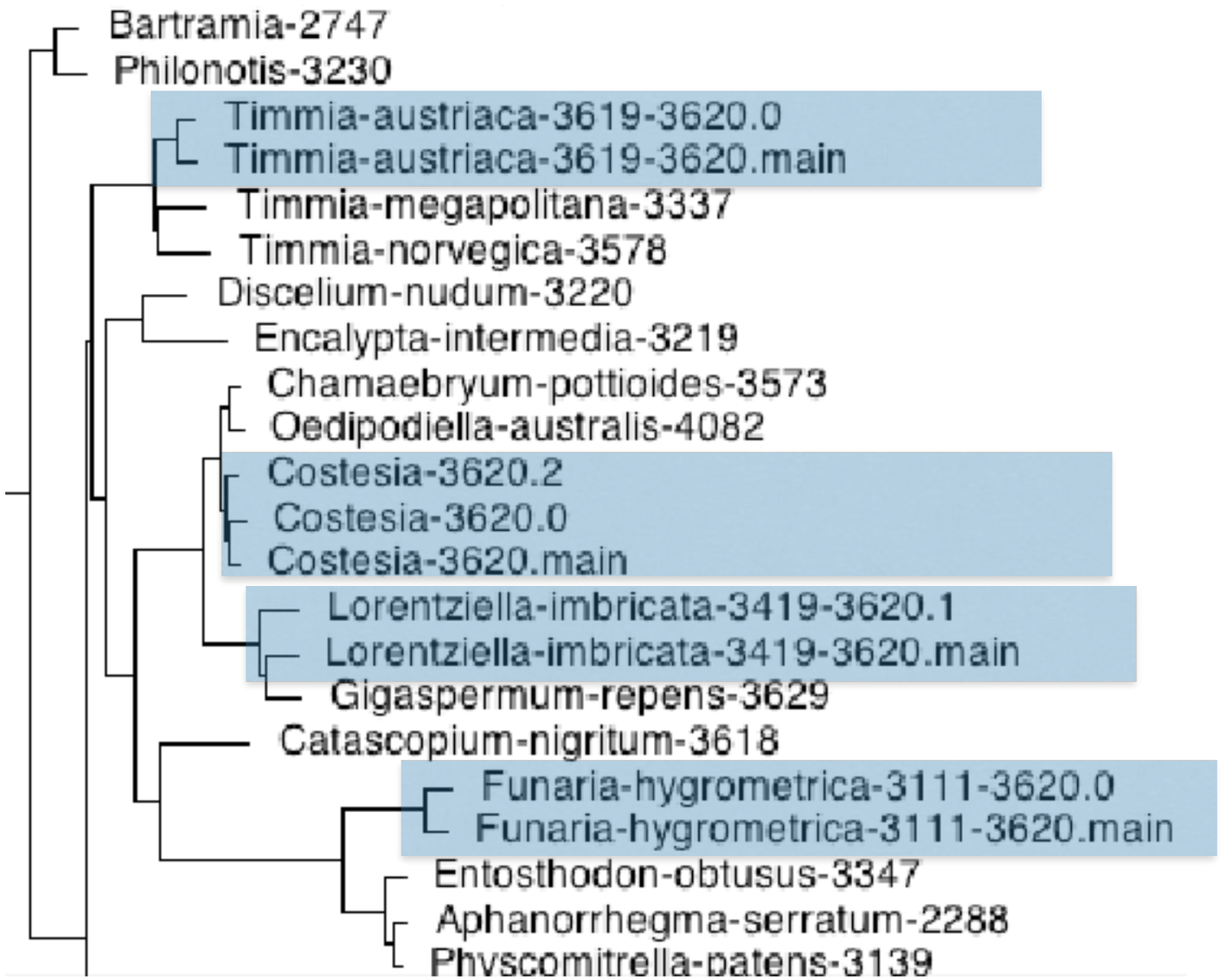
GOOD PARALOGS

HybPiper can collect putative paralogs

Build initial gene trees

Some duplications consistent with known WGD events

Recent duplications can be pruned



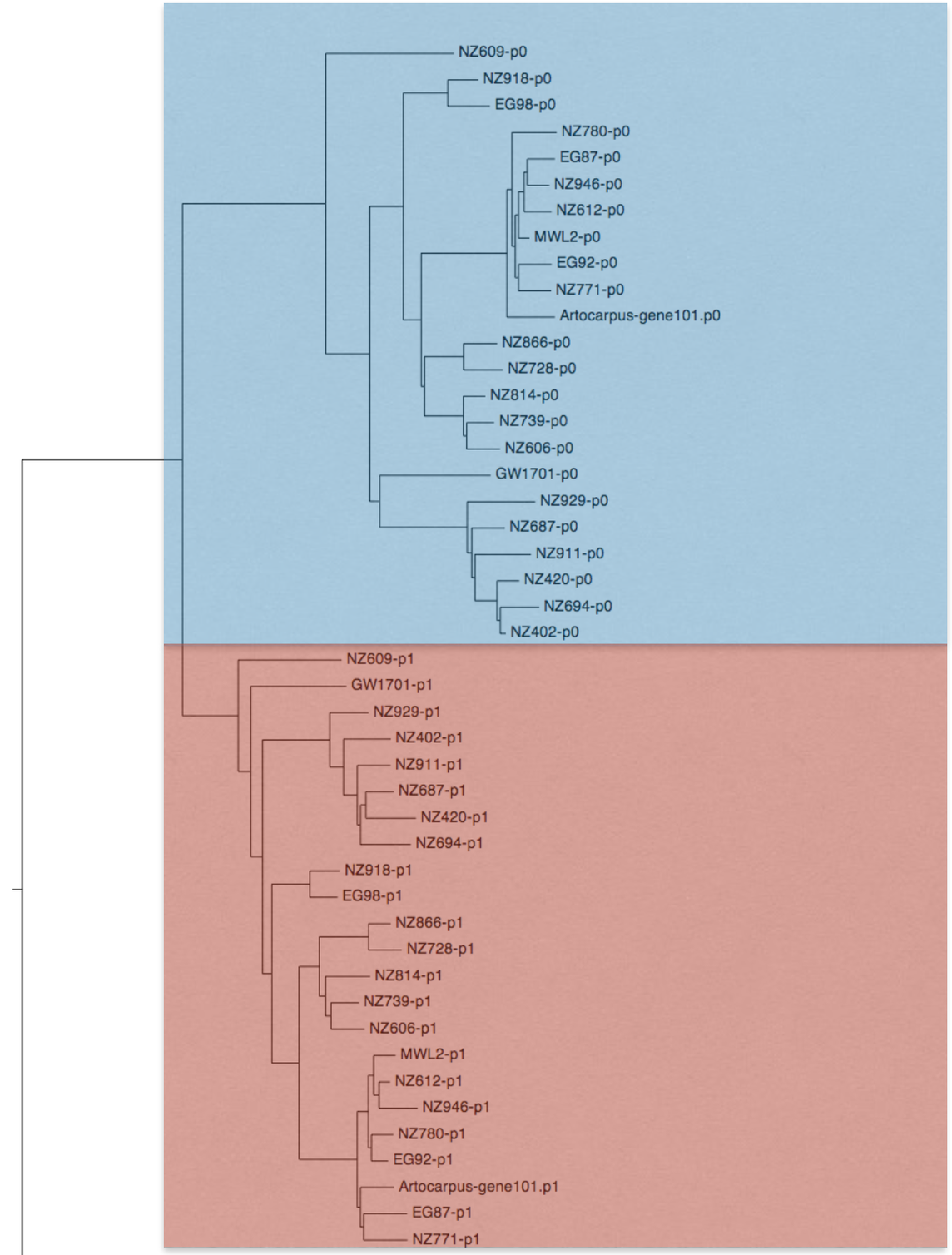
GREAT PARALOGS

HybPiper can collect putative paralogs

Build initial gene trees

Some duplications consistent with known WGD events

Increases the number of genes!



P0

P1

Outgroups

Johnson et al. 2016 APPS