BMS

Bioinformatics & Hybrid-Seq

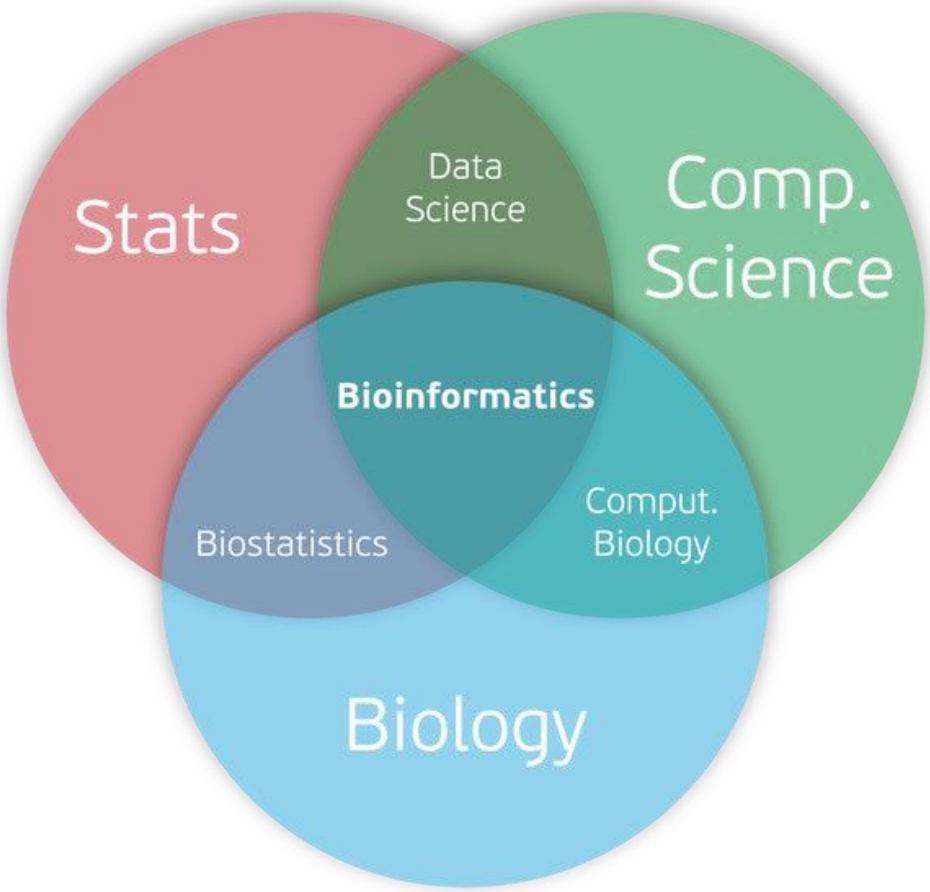
생물정보팀 전형린

Definition of Bioinformatics

- **Bioinformatics** is an interdisciplinary field that develops methods and software tools for understanding biological data.
- As an interdisciplinary field of science, bioinformatics combines biology, computer science, information engineering, mathematics and statistics to analyze and interpret biological data.
- Bioinformatics has been used for *in silico* analyses of biological queries using mathematical and statistical techniques.

https://www.wur.nl/en/Education-Programmes/BSc-Minors/List-of-BSc-minors/BSc-minor-Bioinformatics-WUBIF.htm

Filed of Bioinformatics



https://www.quora.com/Whats-the-difference-between-bioinformatics-and-biotechnology

Required skills

https://www.semanticscholar.org/paper/Bioinformatics-curriculum-development-and-skill-for-Wu-Palani/42d0e8d39dcd5ffbc04a16a2a3e04bb61c3ec4c9/figure/8

Required skills

- Programming
  - Python
  - Perl
  - R
  - JAVA

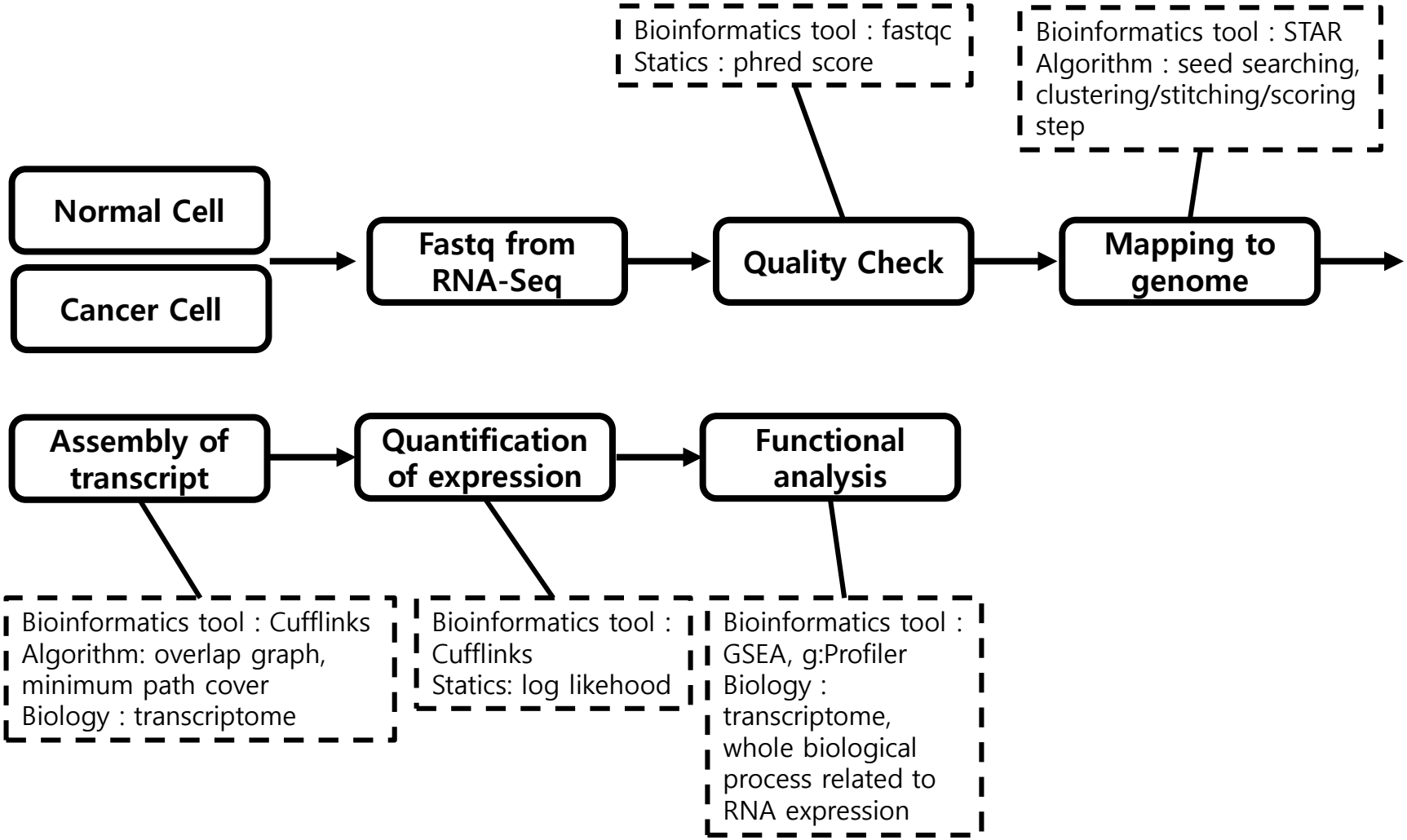
- Bioinformatics tools
  - Alignment
  - Variant Calls
  - Peak Calling
  - Quantification of expression

- Algorithm
  - Artificial neural network
  - Decision tree learning
  - Bayesian network
  - Support vector machine

- Statics
  - chi-squared distribution
  - F-distribution
  - Regression analysis
  - Multivariate statistics

- Biology
  - Genomics
  - Epigenomics
  - Transcriptomics
  - Proteomics
  - Metabolomics

Example of bioinformatics



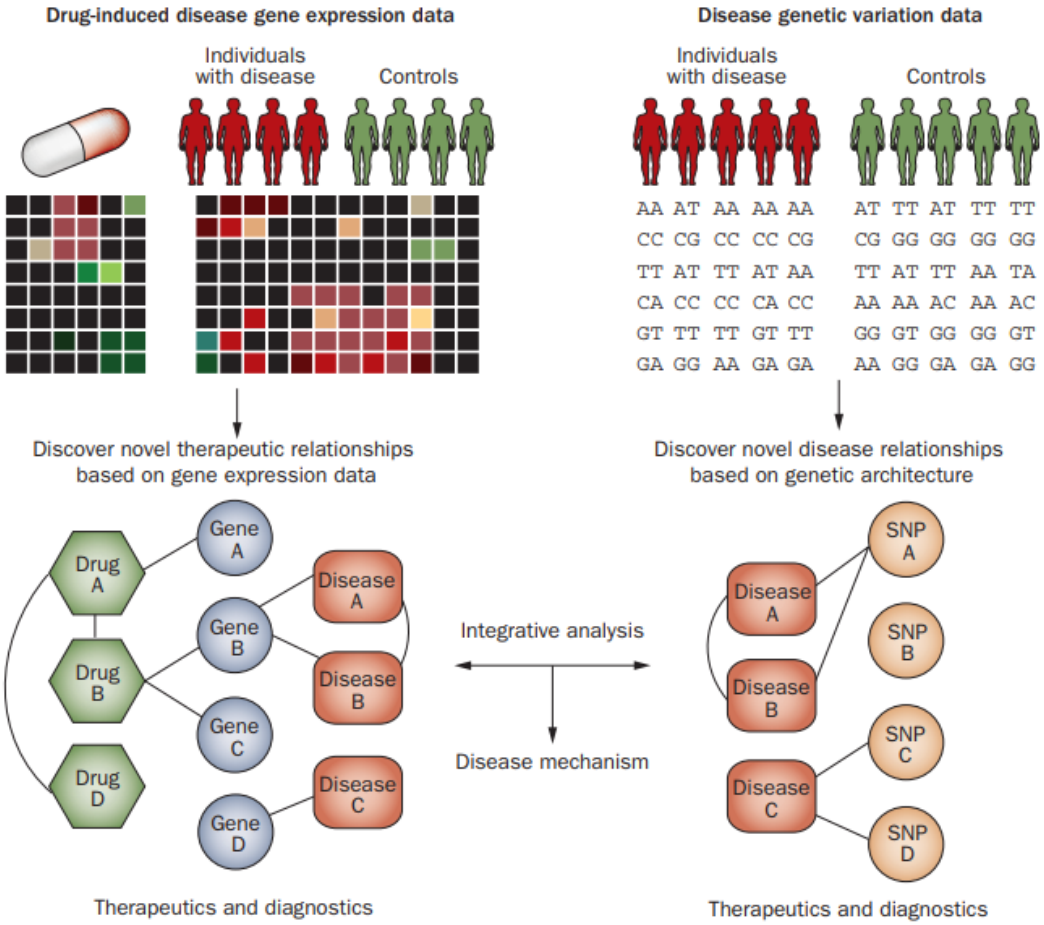
Example of bioinformatics

TGTACG

AGTGTAGTGTACGCCCAAT

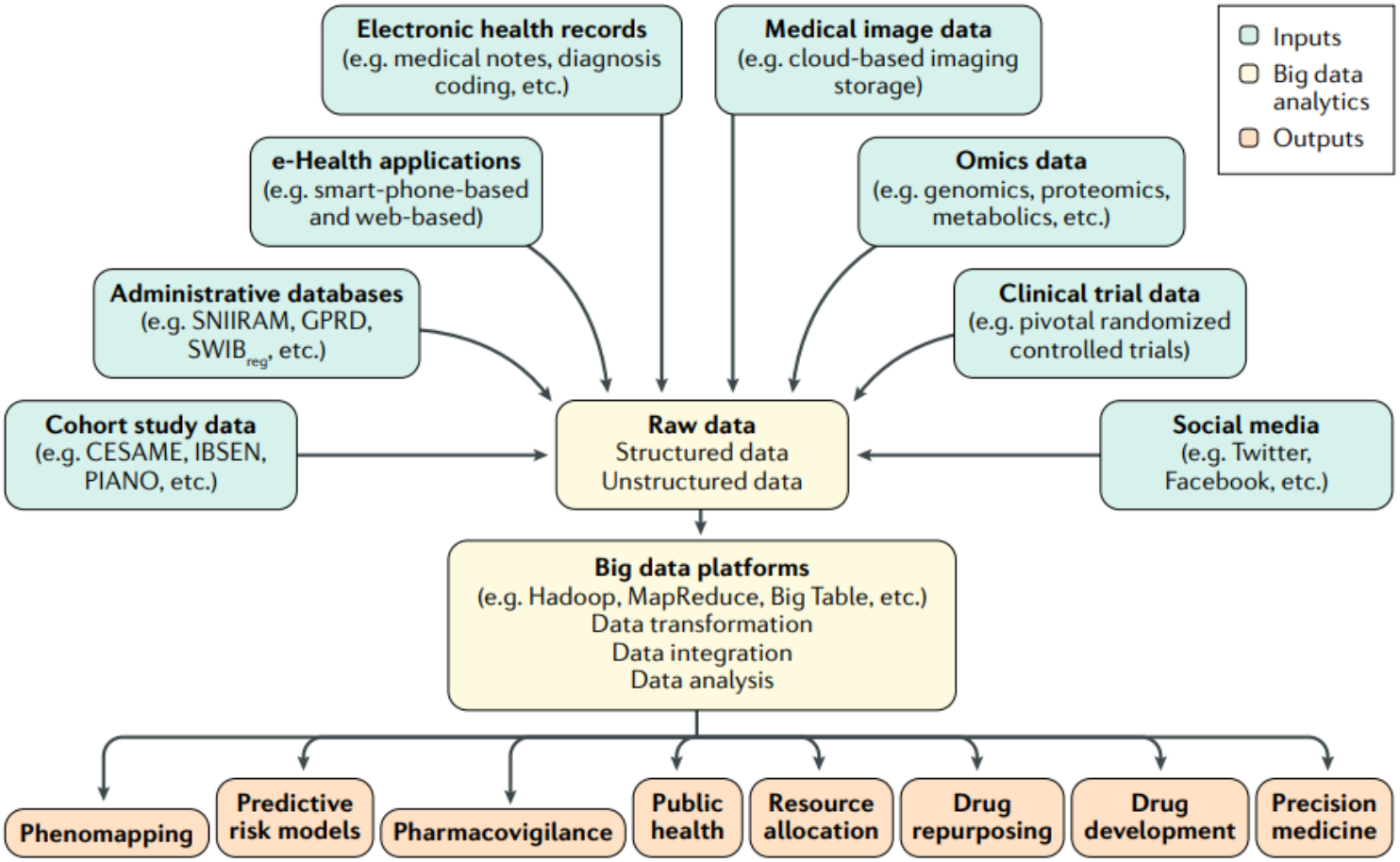
AGTGTAGTGTACGCCCAATTGTACGAGCATCATCGTGAAGGTCTGATGTTGTAC
GTGTACG

Example of bioinformatics



Current Biology, 2018

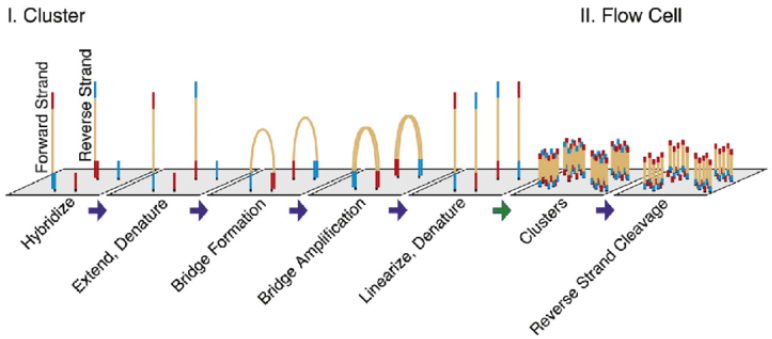
Future of bioinformatics



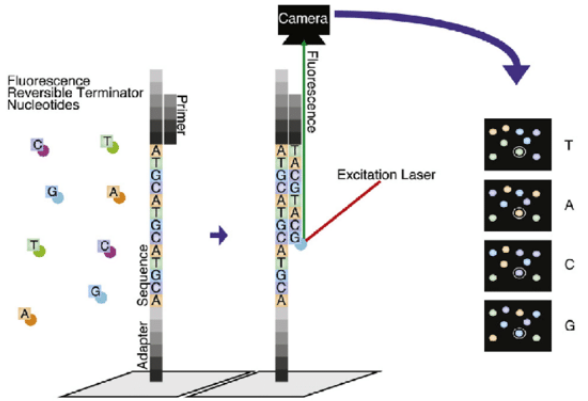
Nature Reviews Gastroenterology and Hepatology, 2019

Raw data from Illumina sequencer

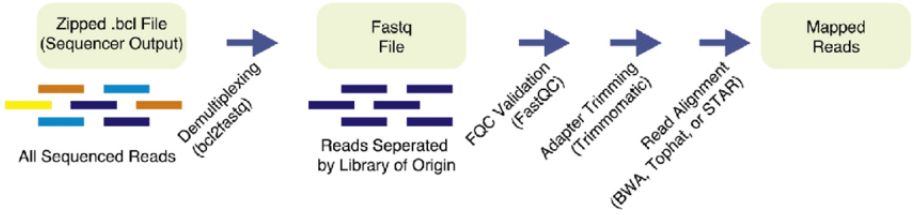
A. Clustering



B. High-throughput sequencing

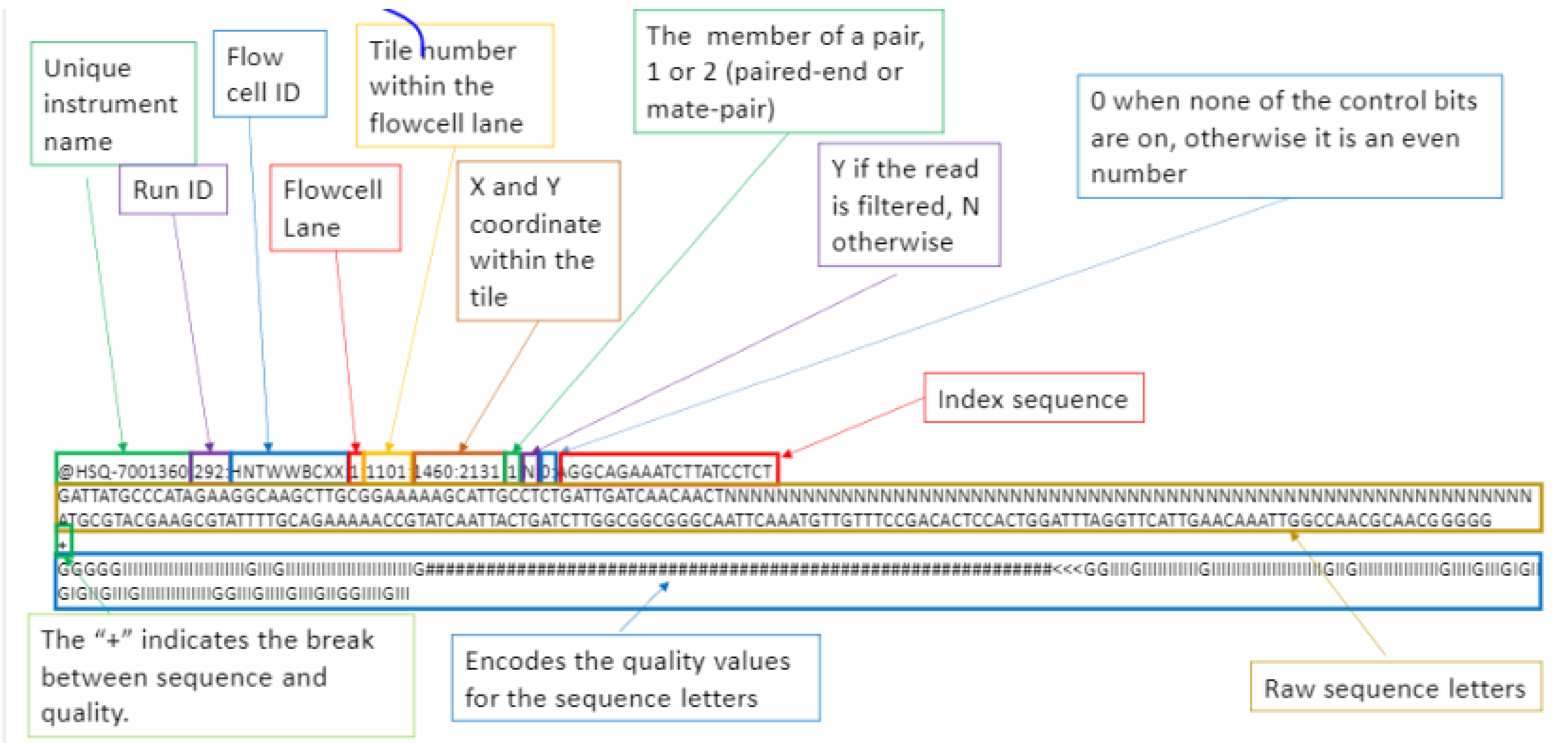


C. Demultiplexing samples and read mapping



https://kscbioinformatics.wordpress.com/2017/02/03/raw-illumina-sequence-data-files-for-dummies-part-1/

Fastq format



https://kscbioinformatics.wordpress.com/2017/02/03/raw-illumina-sequence-data-files-for-dummies-part-1/

phred score(Q score)

Q scores are defined as a property that is logarithmically related to the base calling error probabilities (P)[2].

$$Q = -10 \log_{10} P$$

Table 1: Quality Scores and Base Calling Accuracy

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf

ASCII table

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf
http://cn.voidcc.com/question/p-umzpfpil-qz.html

ASCII represent quality score

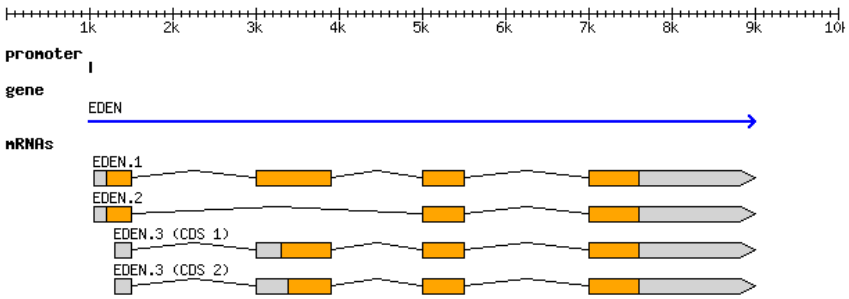
S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 41)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf
http://cn.voidcc.com/question/p-umzpfpil-qz.html

Fasta format

https://www.researchgate.net/figure/A-sample-of-the-Multi-FASTA-file_fig1_309134977

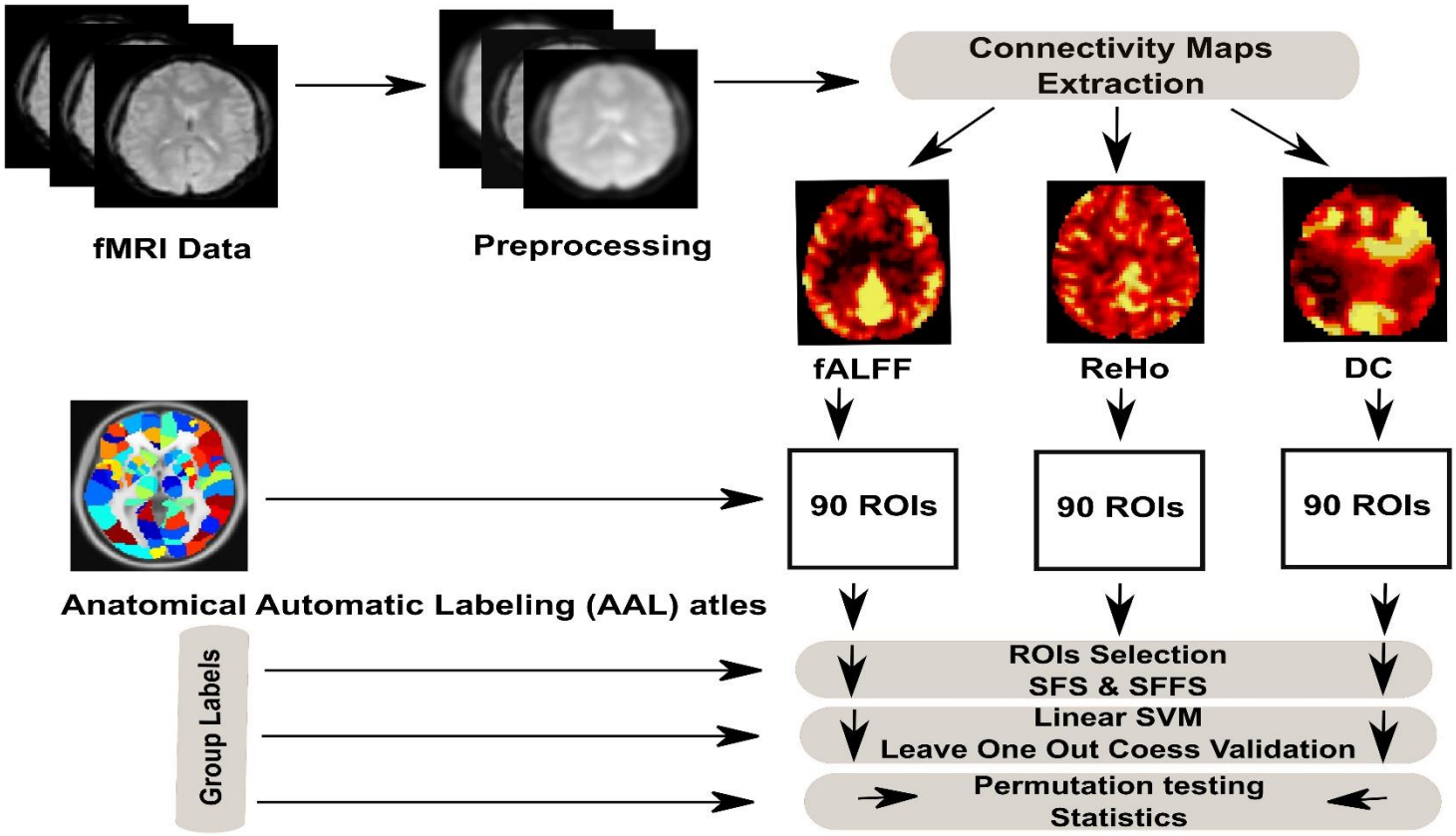
Gff format




```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS             5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS             7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

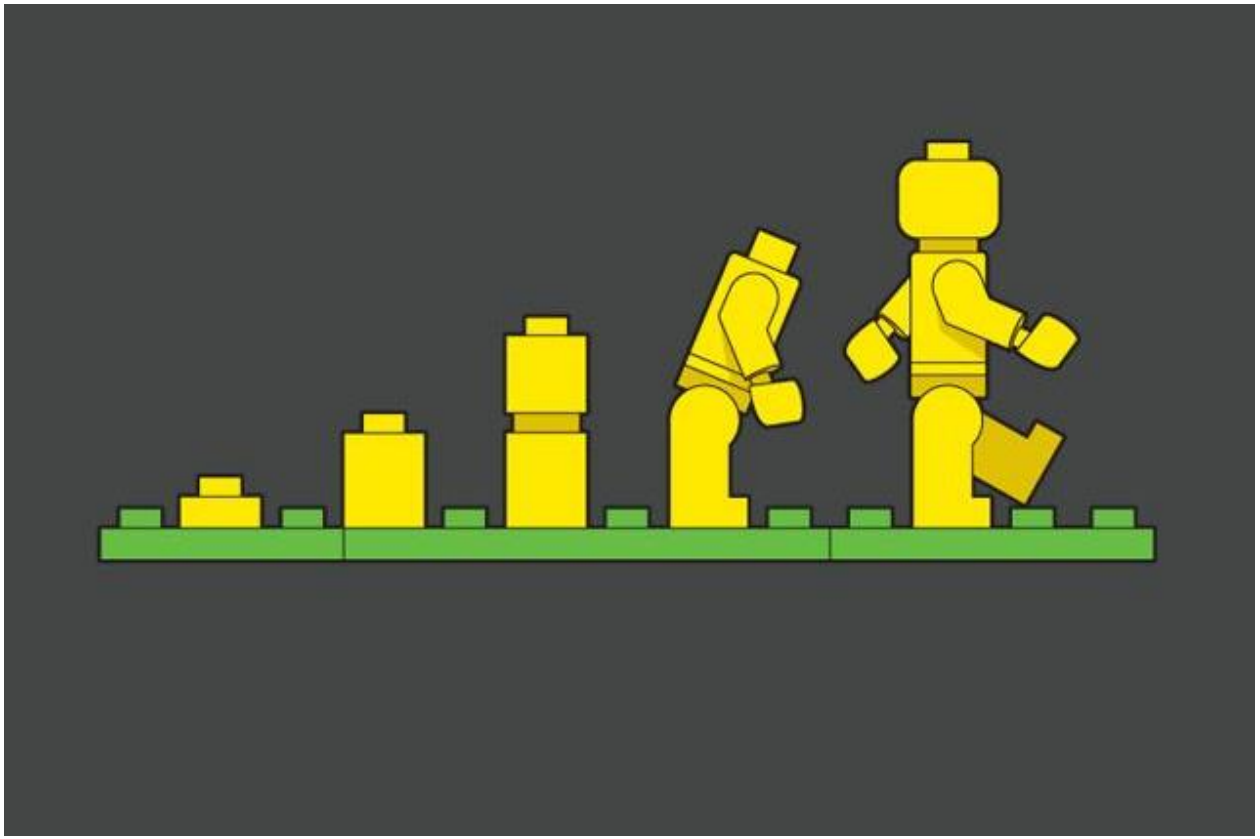
http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/950/index.php?manual=GFF3_format.html

Bioinformatics used in other filed



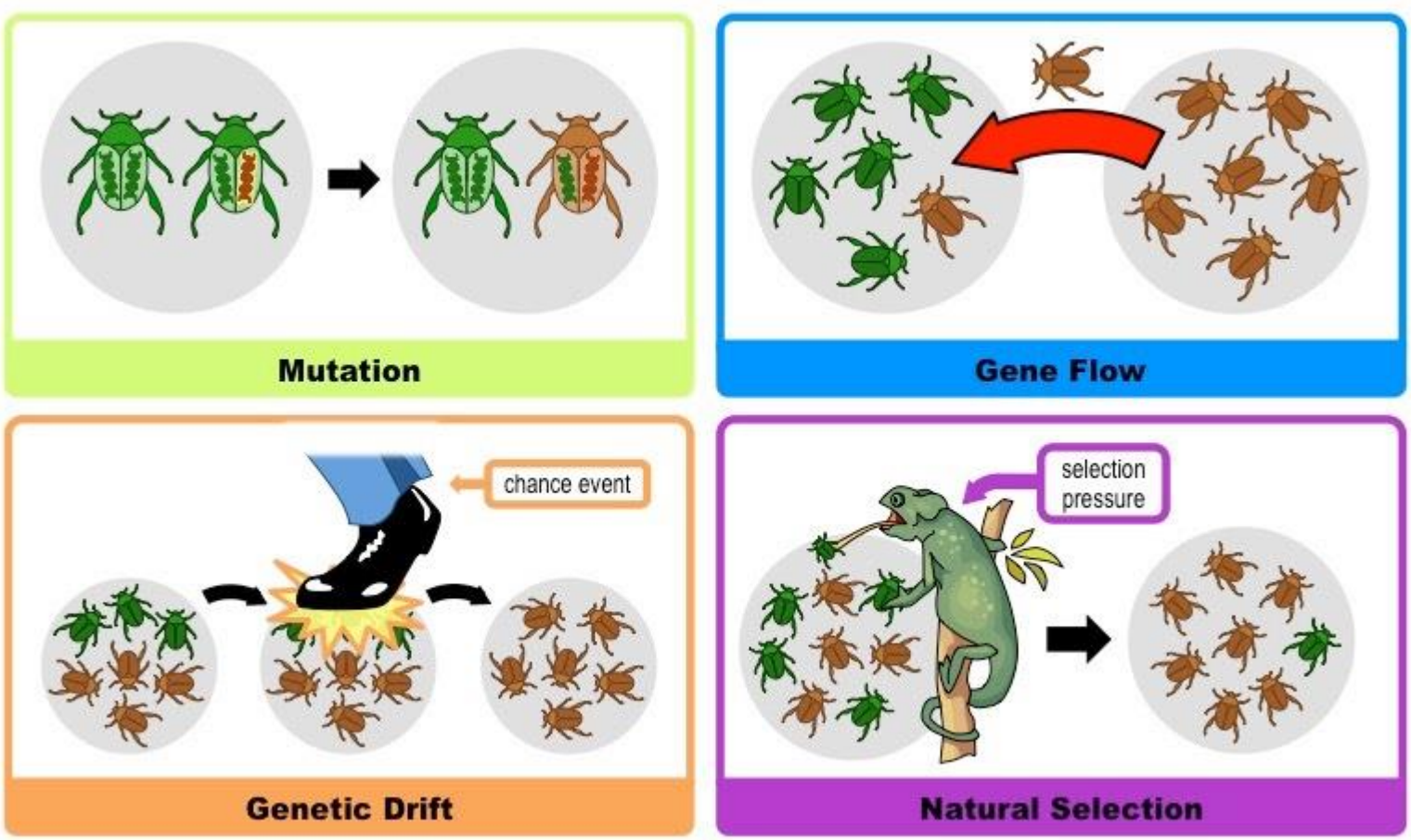
Frontiers in Human Neuroscience, 2019

Evolution



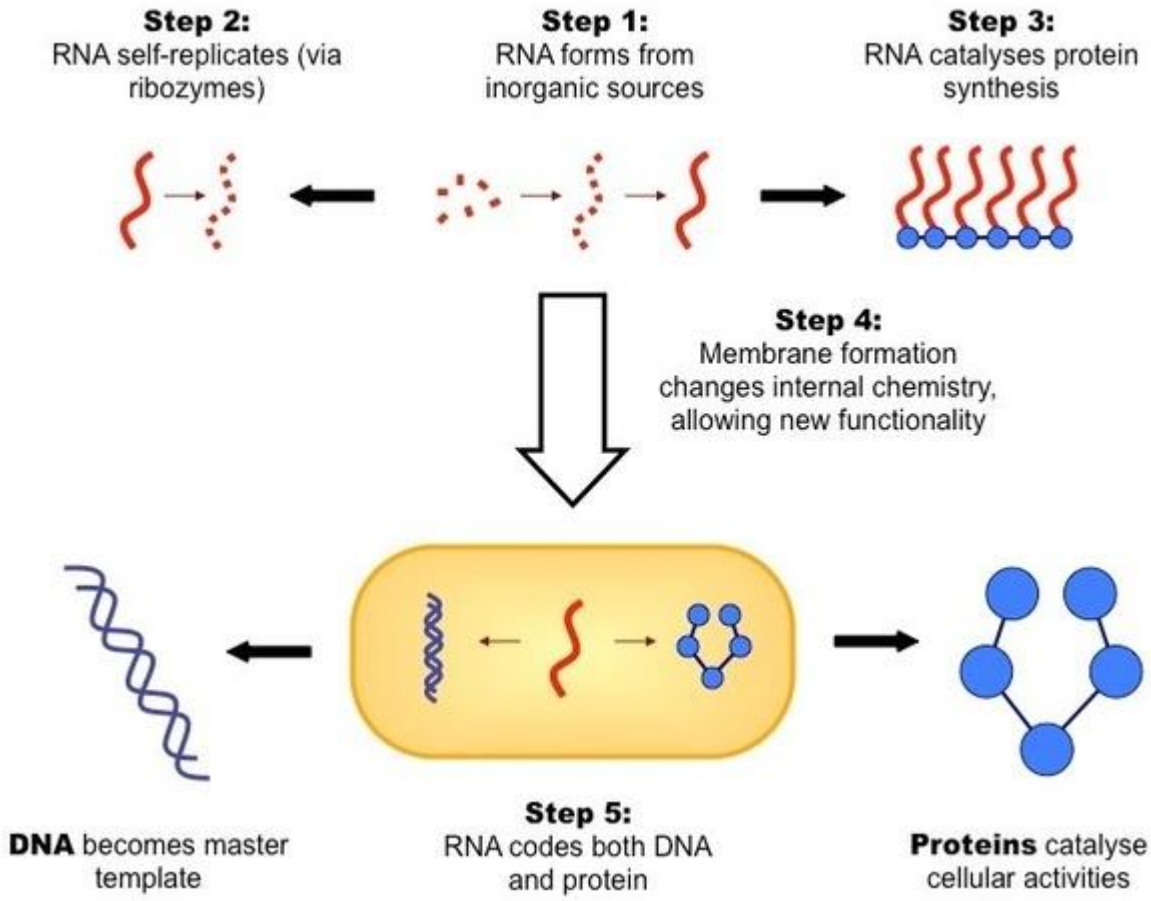
https://store.glennz.com/blogs/glennz-concepts/29641793-evolution

Evolution mechanism



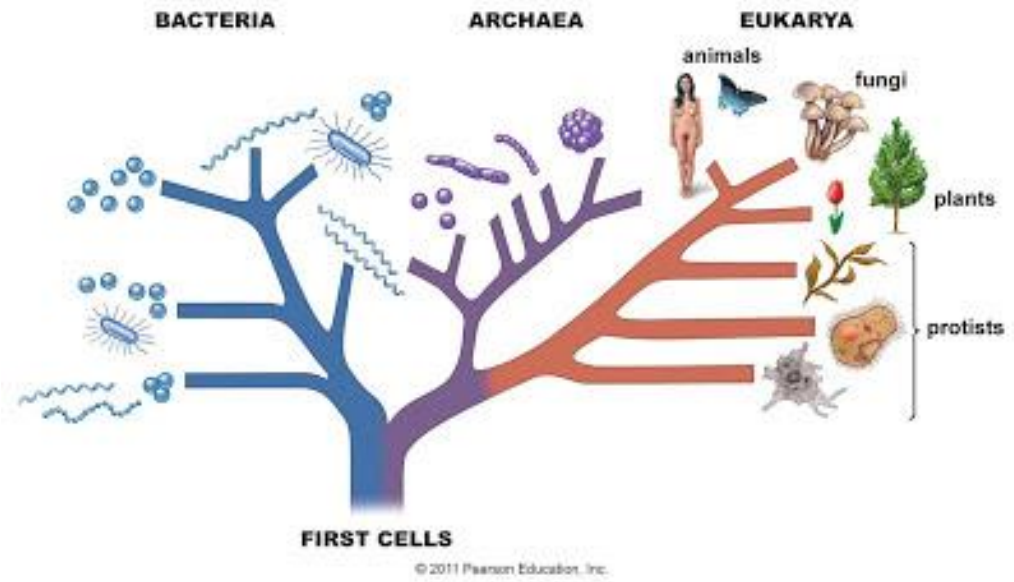
https://ib.bioninja.com.au/higher-level/topic-10-genetics-and-evolu/103-gene-pools-and-speciati/evolution.html

RNA world



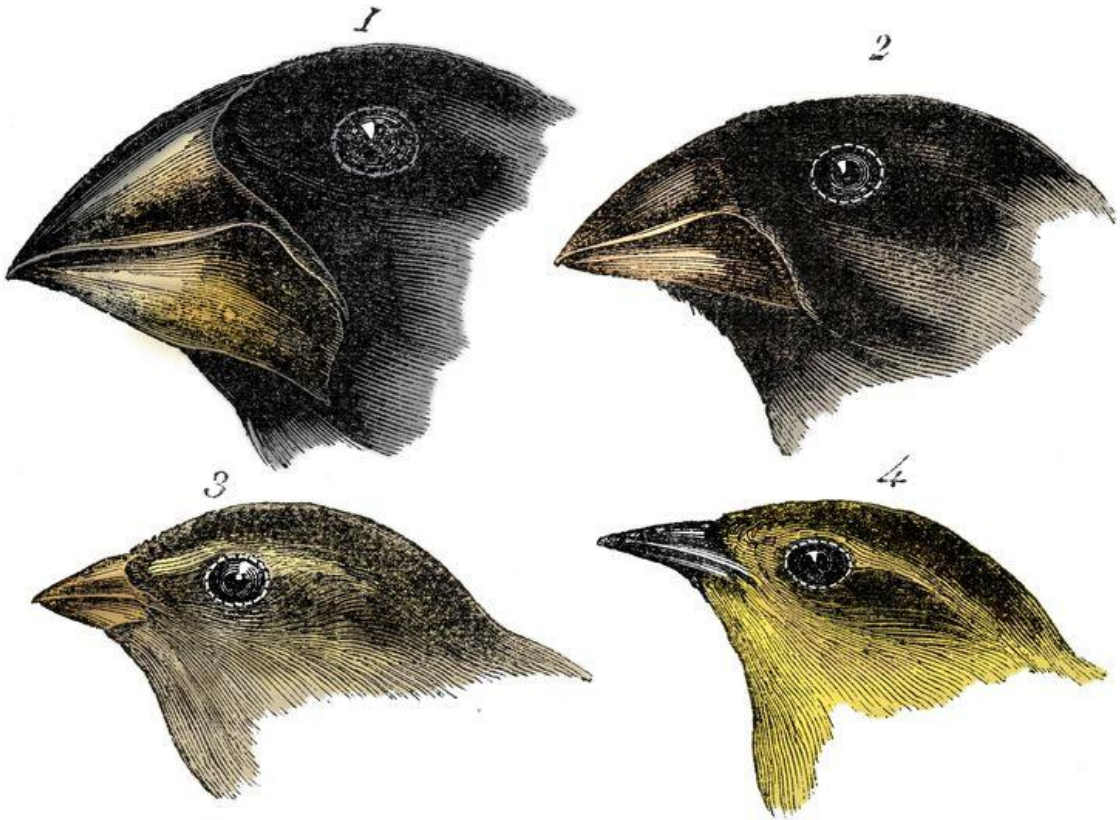
https://www.quora.com/What-are-the-leading-theories-as-to-how-ATP-arose-as-the-energy-carrier-in-cellular-respiration

Bioinformatics used in other filed



Frontiers in Human Neuroscience, 2019

Natural history - Darwin's finches

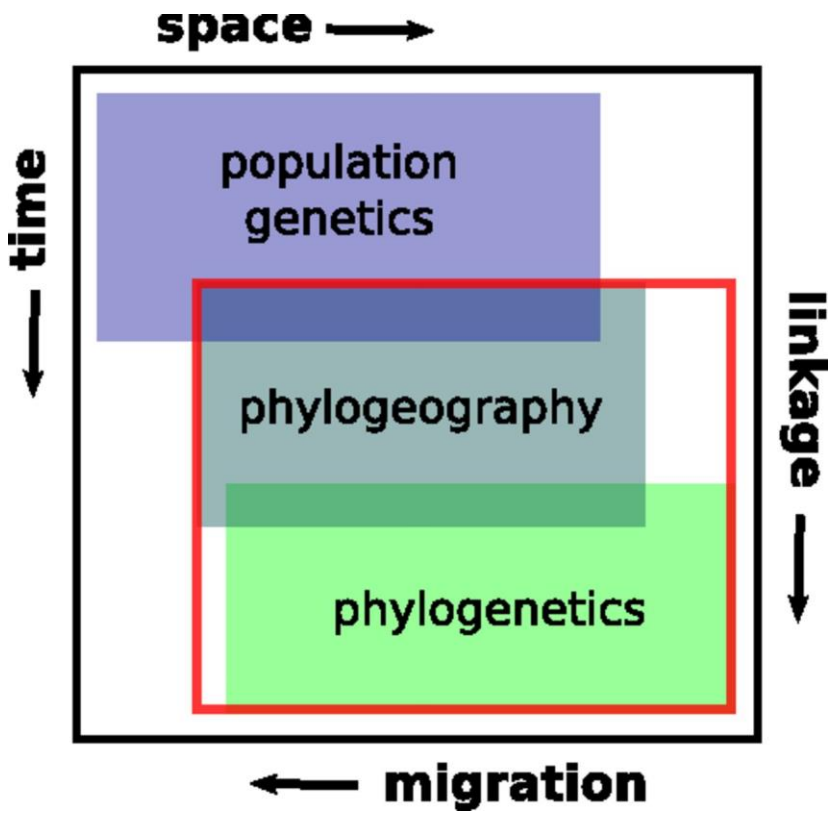


1. Geospiza magnirostris.
3. Geospiza parvula.

2. Geospiza fortis.
4. Certhidea olivasea.

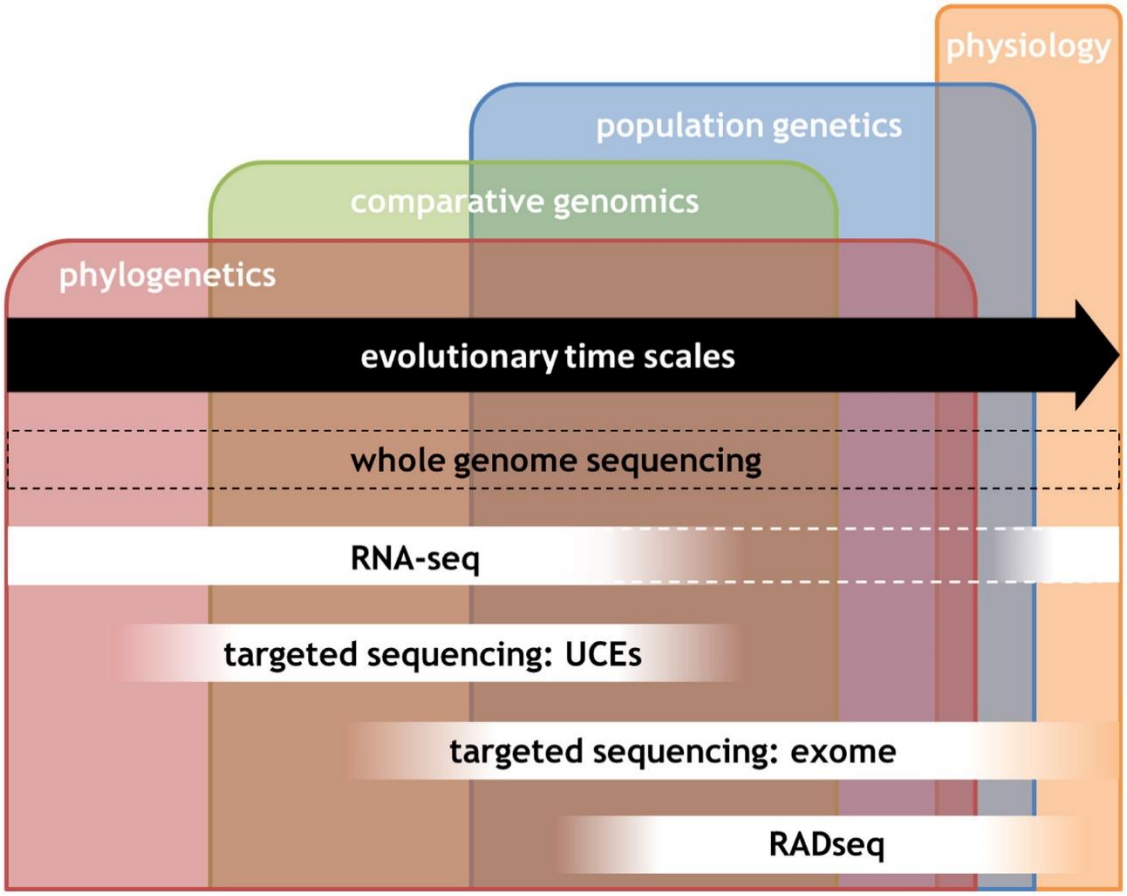
https://www.thoughtco.com/charles-darwins-finches-1224472

Population genetics, phylogeography, and phylogenetics



PNAS , 2016

Application of high-throughput sequencing methods



Marine Genomics, 2016

Application of high-throughput sequencing methods

Method	Other names or variants	Literature method	Literature examples
Amplicon sequencing	Multiplex PCR, parallel tagged sequencing	Binladen et al. (2007), Meyer et al. (2008), Tewhey et al. (2009b)	Chan et al. (2010), Griffin et al. (2011), Gunnarsdóttir et al. (2011); Morin et al. (2010), Parks et al. (2009)
Restriction-digest	Double-digest genome reduction, RAD sequencing (RAD-seq), complexity reduction of multilocus sequences (CRoPS), Genotyping by Sequencing (GBS)	Baird et al. (2008), Davey et al. (2011)	Andolfatto et al. (2011), Amaral et al. (2009), Bers et al. (2010), Emerson et al. (2010), Gompert et al. (2010), Hohenlohe et al. (2011), Hyten et al. (2010a,b), Kerstens et al. (2009), Ramos et al. (2009); Sánchez et al. (2009); Van Orsouw et al. (2007); Van Tassell et al. (2008), Wiedmann et al. (2008); Williams et al. (2010)
Target enrichment	Sequence capture, targeted resequencing, primer extension capture (PEC)	Albert et al. (2007), Gnirke et al. (2009), Hodges et al. (2007), Okou et al. (2007), Tewhey et al. (2009a), Maricic et al. (2010)	Briggs et al. (2009; Faircloth et al. in press)
Transcriptome	RNA-seq	Morin et al. (2008); Marioni et al. (2008)	Barbazuk and Schnable (2011), Cánovas et al. (2010), Chepelev et al. (2009); Geraldes et al. (2011), Hittinger et al. (2010)

***Molecular phylogenetics and evolution*, 2013**

Phylogenetic positons

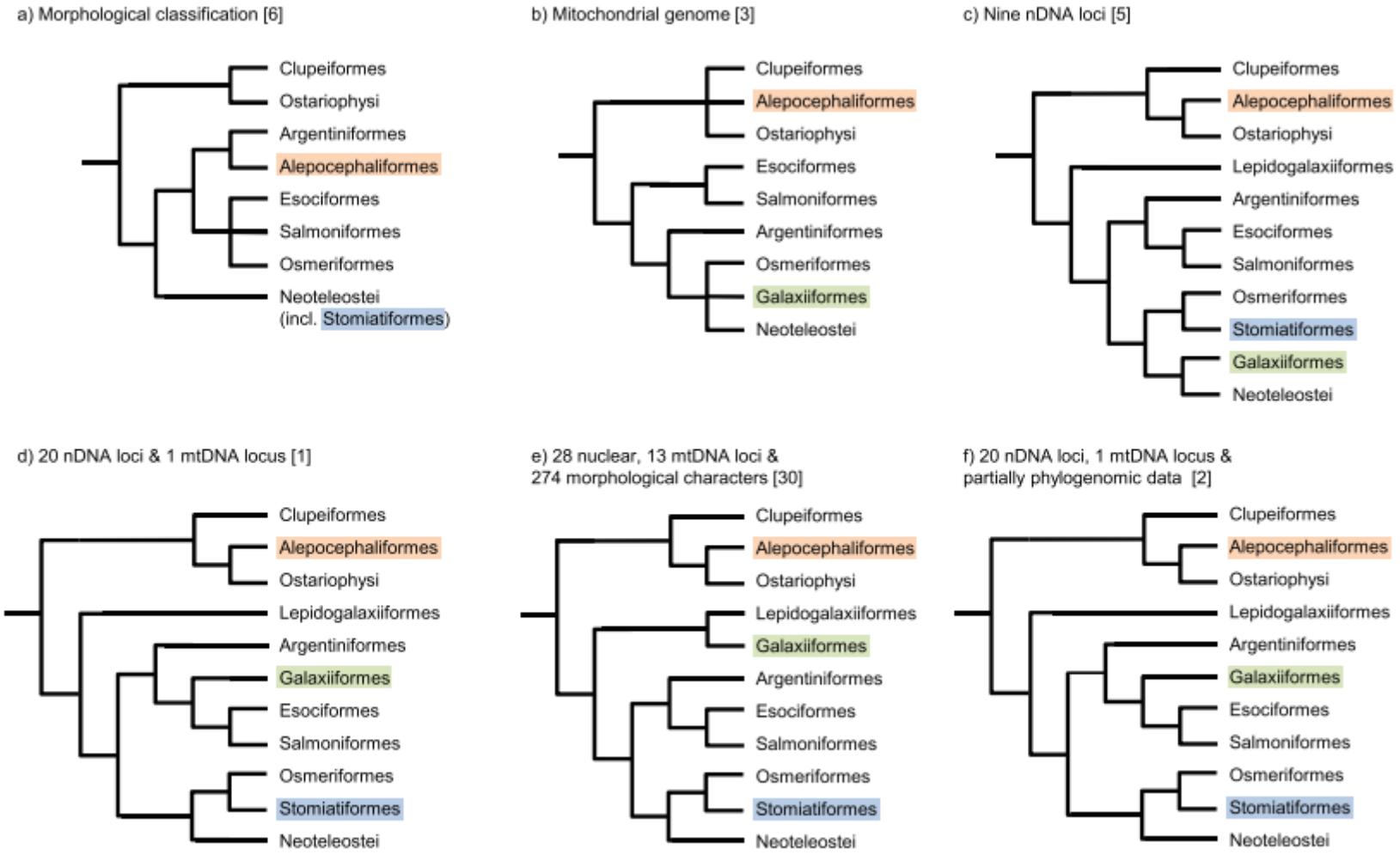
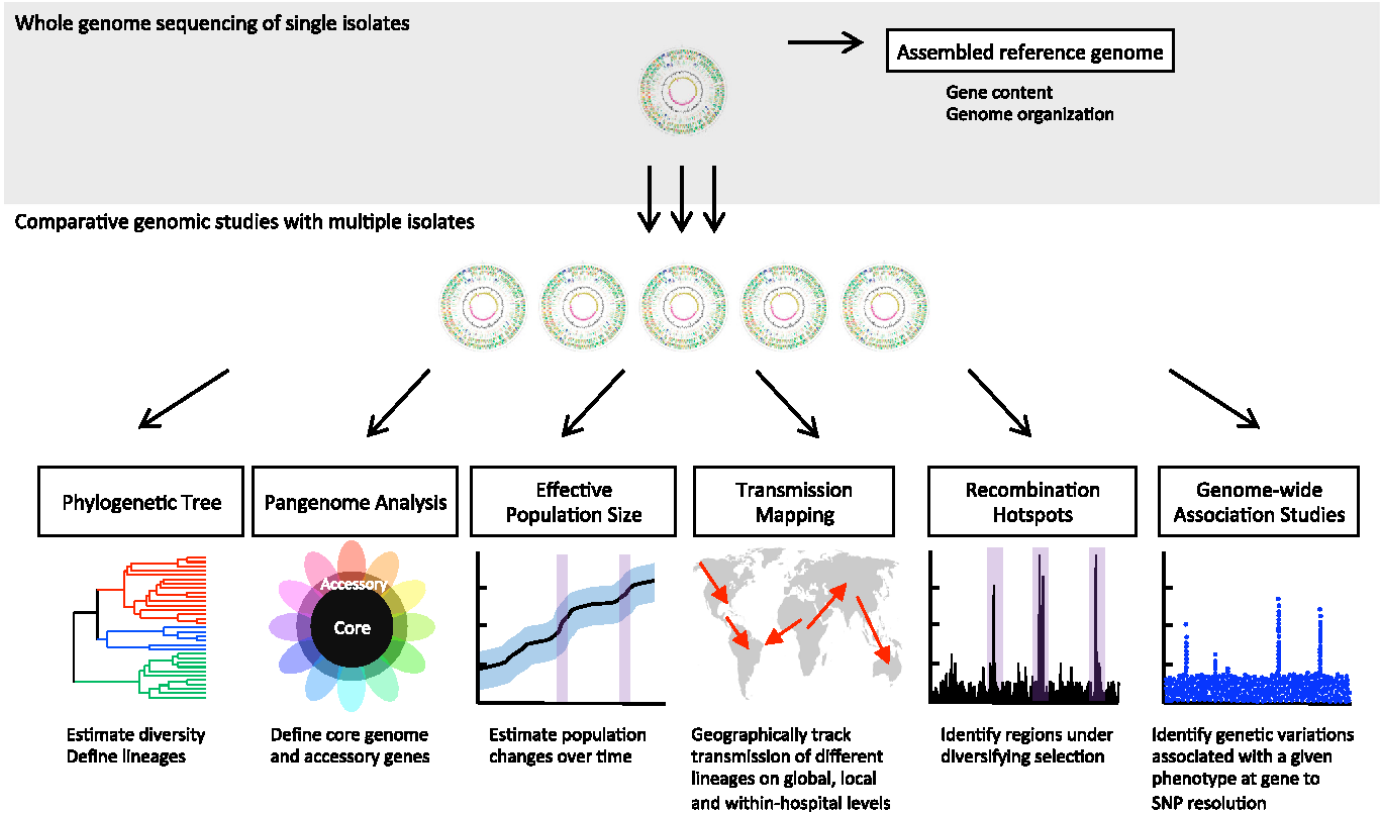


Fig. 1 Summary of previous phylogenetic estimates and classifications of Clupeocephalan fishes. Colours indicate taxa with variable phylogenetic positions

BMC Evolutionary Biology, 2018

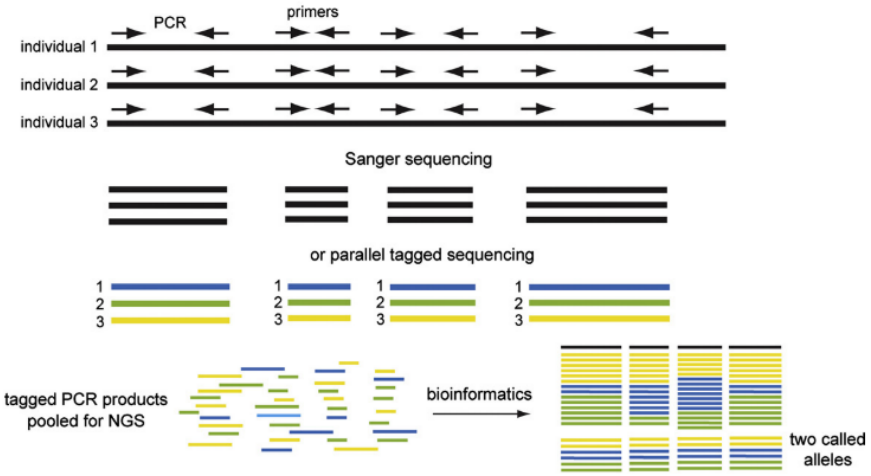
Whole genome sequencing for Phylogenetics



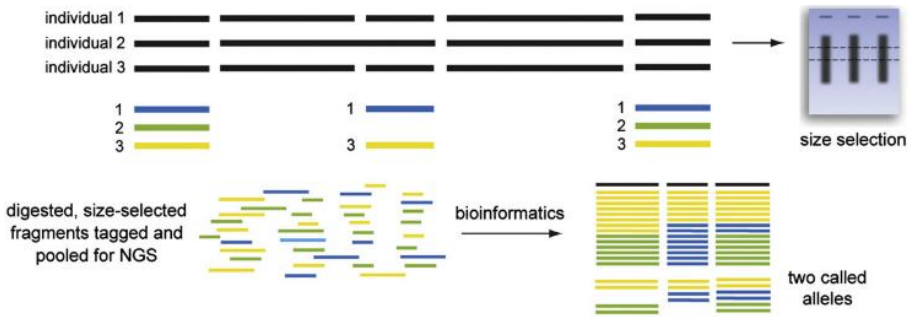
Cell host & microbe, 2016

Applications of NGS to phylogenetics

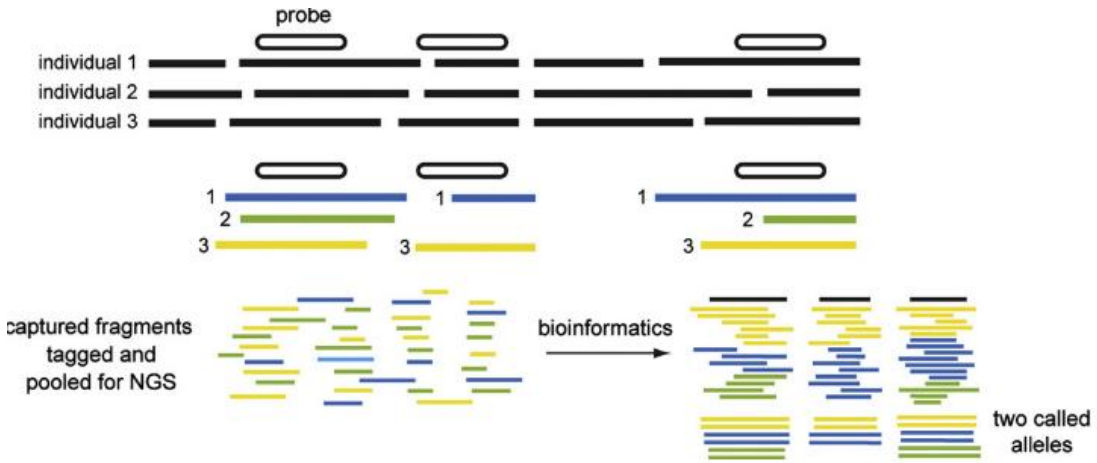
Amplicon based methods



Restriction-digest based methods

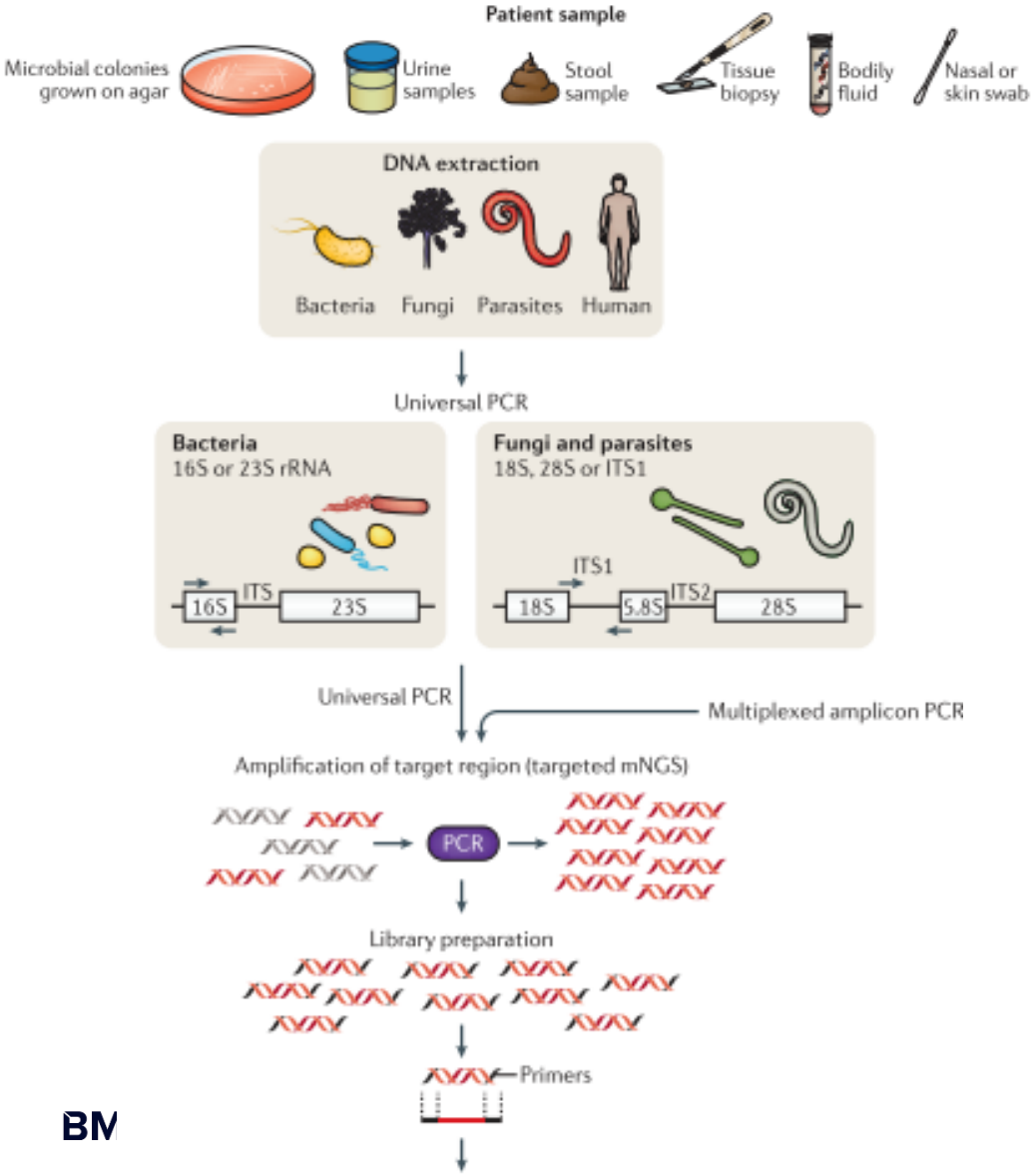


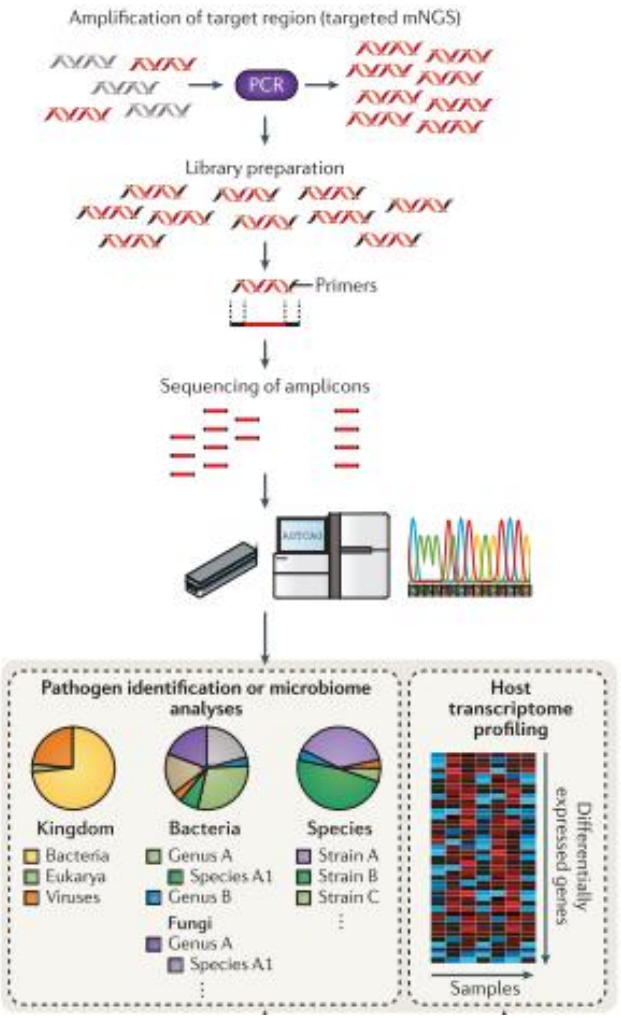
Probe based methods



Molecular phylogenetics and evolution, 2013

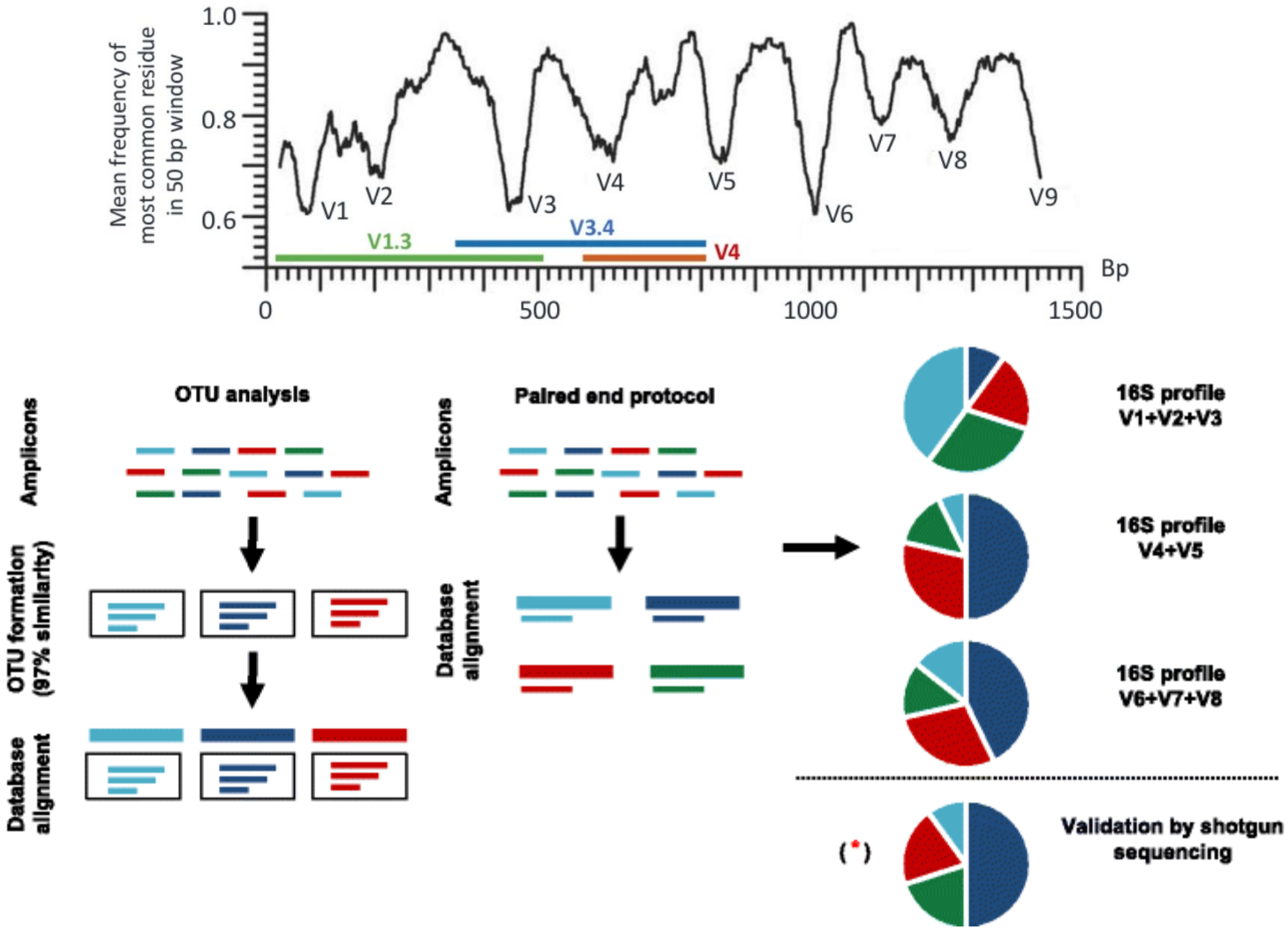
Amplicon sequencing for Phylogenetics





Nature Reviews Genetics, 2019

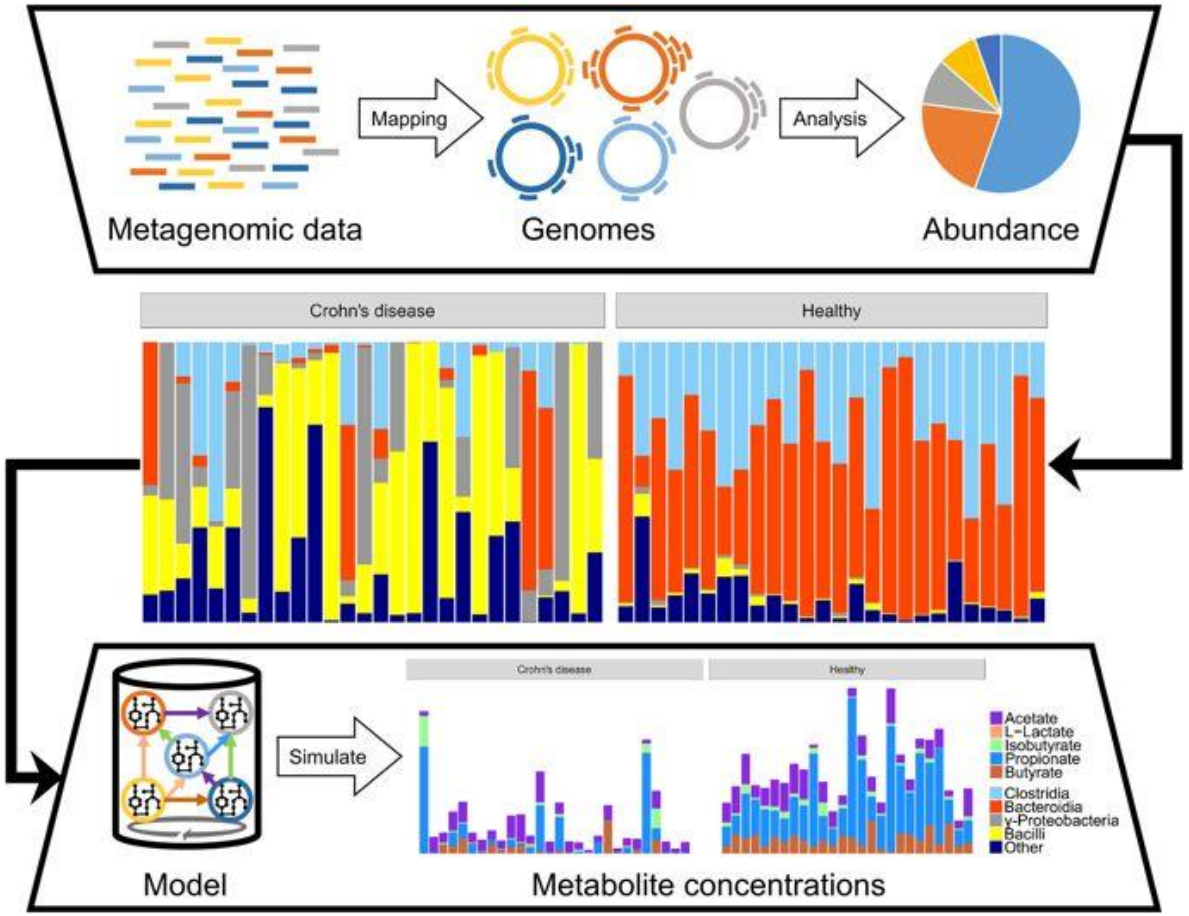
16s rRNA Amplicon-seq analysis



https://www.slideshare.net/MadsAlbertsen/130919-16-sworkshopintroduction

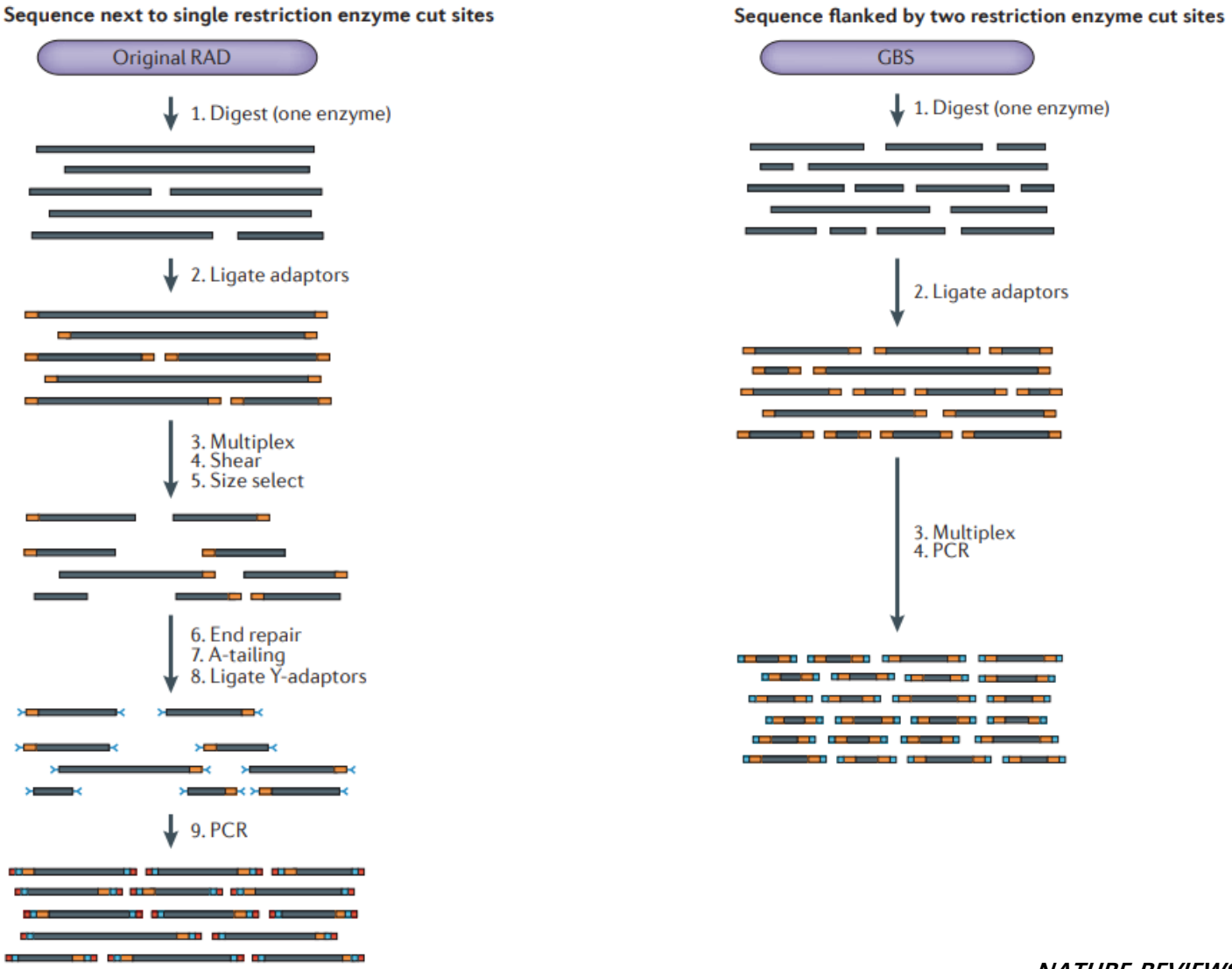
BMC Genomics, 2017

Diagnosis based on Amplicon-seq



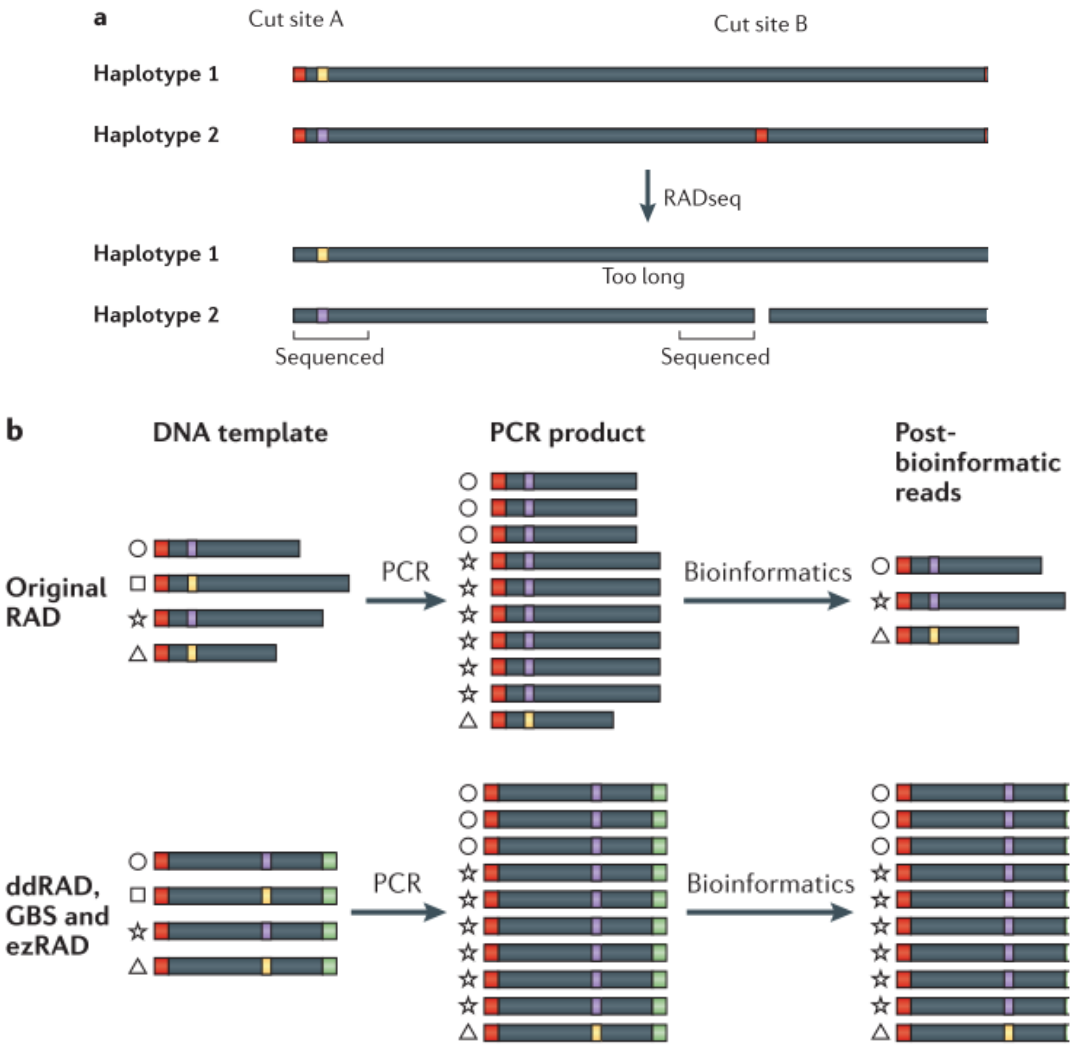
npj Systems Biology and Applicationsvolume, 2018

Restriction site-associated DNA sequencing



NATURE REVIEWS GENETICS, 2016

Restriction site-associated DNA sequencing



NATURE REVIEWS GENETICS, 2016

Hybrid-Seq

Syst. Biol. 61(5):727–744, 2012

DOI:10.1093/sysbio/sys049
Advance Access publication on May 17, 2012

Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics

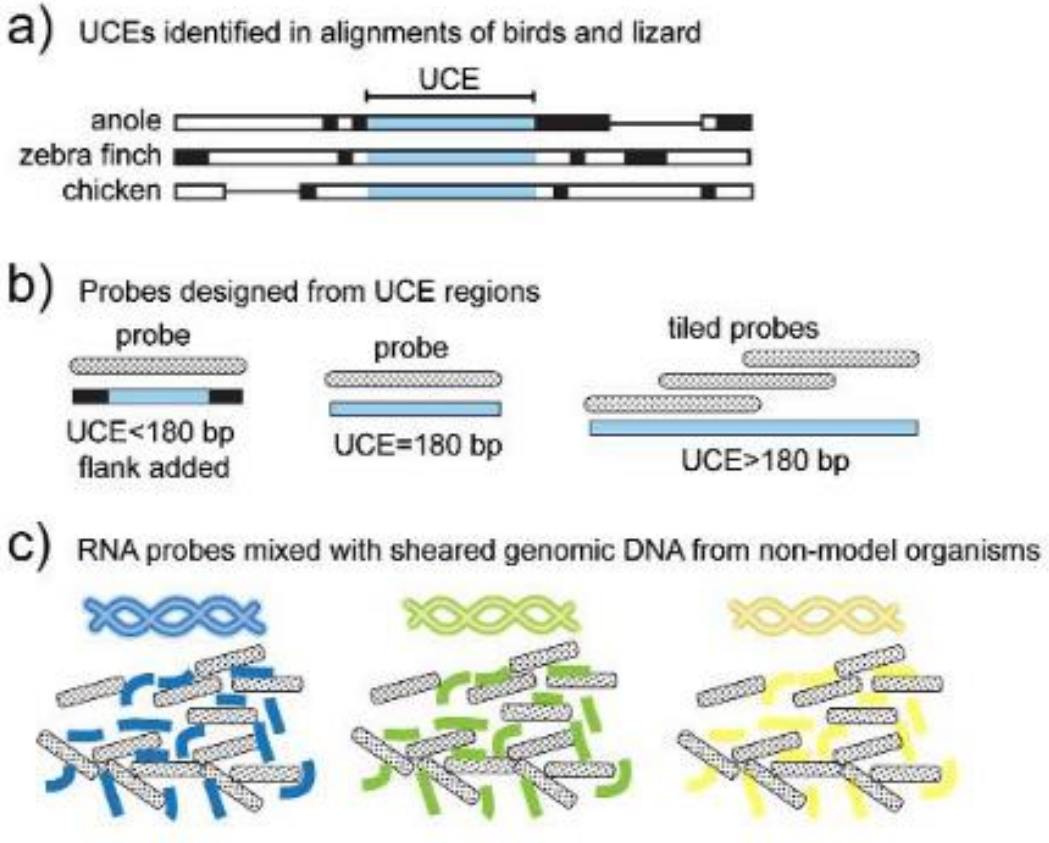
ALAN R. LEMMON[1,*], SANDRA A. EMME[2], AND EMILY MORIARTY LEMMON[2]

[1]*Department of Scientific Computing, Florida State University, Dirac Science Library, Tallahassee, FL 32306-4102, USA; and* [2]*Department of Biological Science, Florida State University, 319 Stadium Drive, PO Box 3064295, Tallahassee, FL, 32306-4295, USA;*
**Correspondence to be sent to: Department of Scientific Computing, Florida State University, Dirac Science Library, Tallahassee, FL 32306-4102; E-mail: alemmon@fsu.edu.*

Received 1 November 2011; reviews returned 19 January 2012; accepted 7 May 2012
Associate Editor: Bryan Carstens

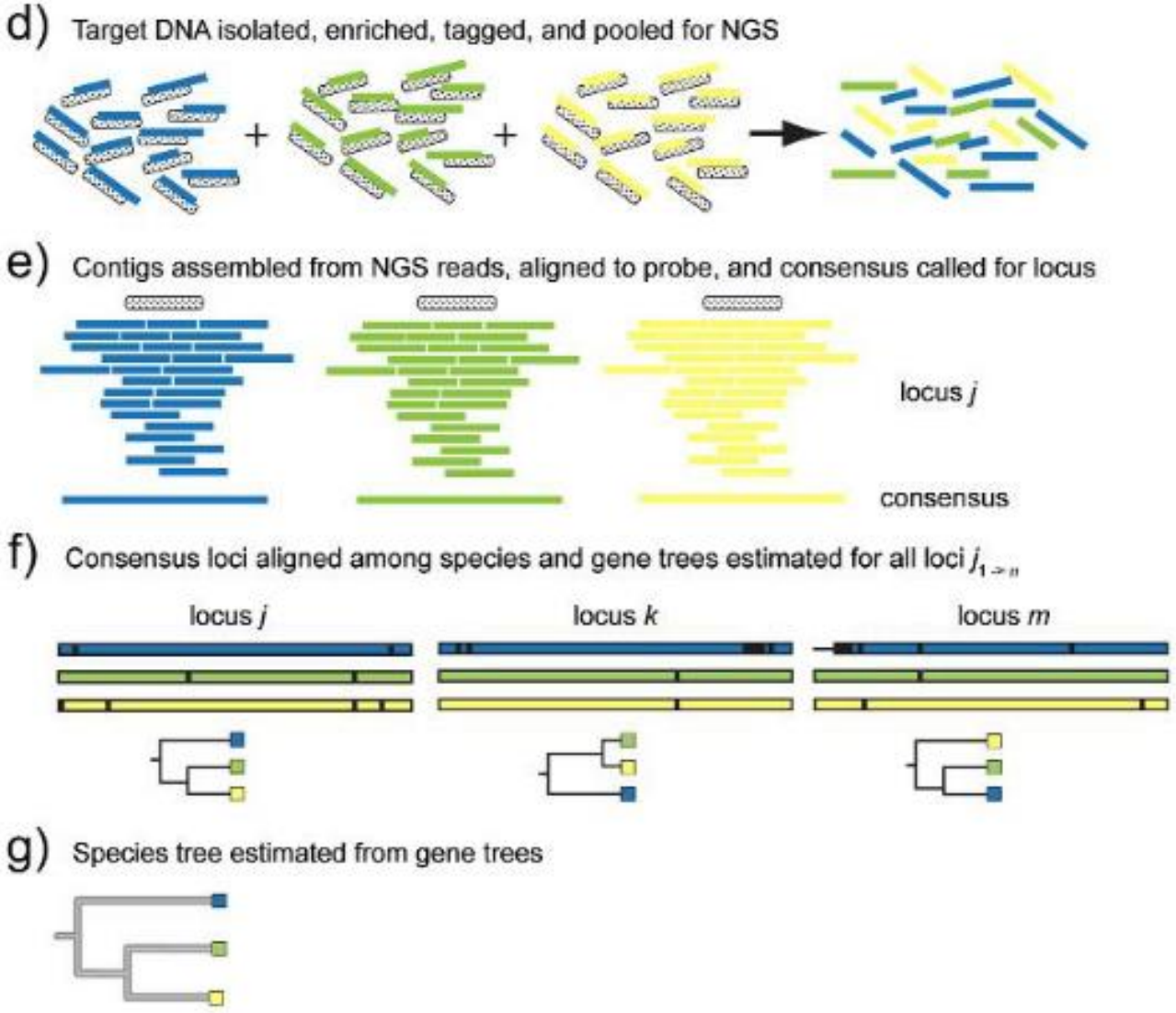
***NATURE REVIEWS GENETICS,* 2016**

Steps for Hybrid-Seq



Systematic Biology, 2012

Steps for Hybrid-Seq



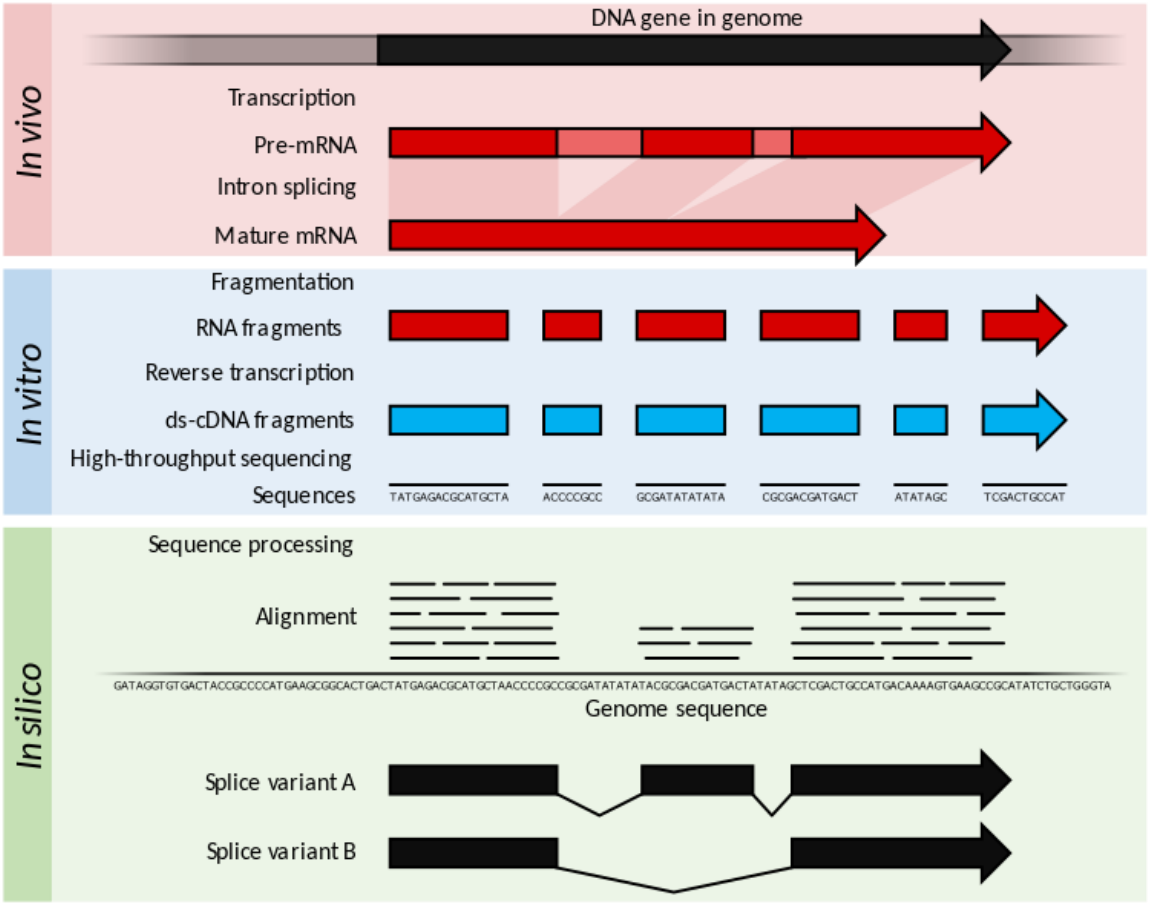
Systematic Biology, 2012

Locus Selection & Probe Design

- Locus Selection
  - ~500 "single copy" loci (typically long exons)
  - Conserved element (~20% divergence required)
  - Adjacent to less conserved regions
  - Loci are selected based on broad taxonomic group (e.g., vertebrates)
- Probe Design
  - Incorporate sufficient number of lineages
  - Tile probes across conserved region
  - Goal is to capture ~1500bp regions
  - Probe sets are designed for project-specific clade

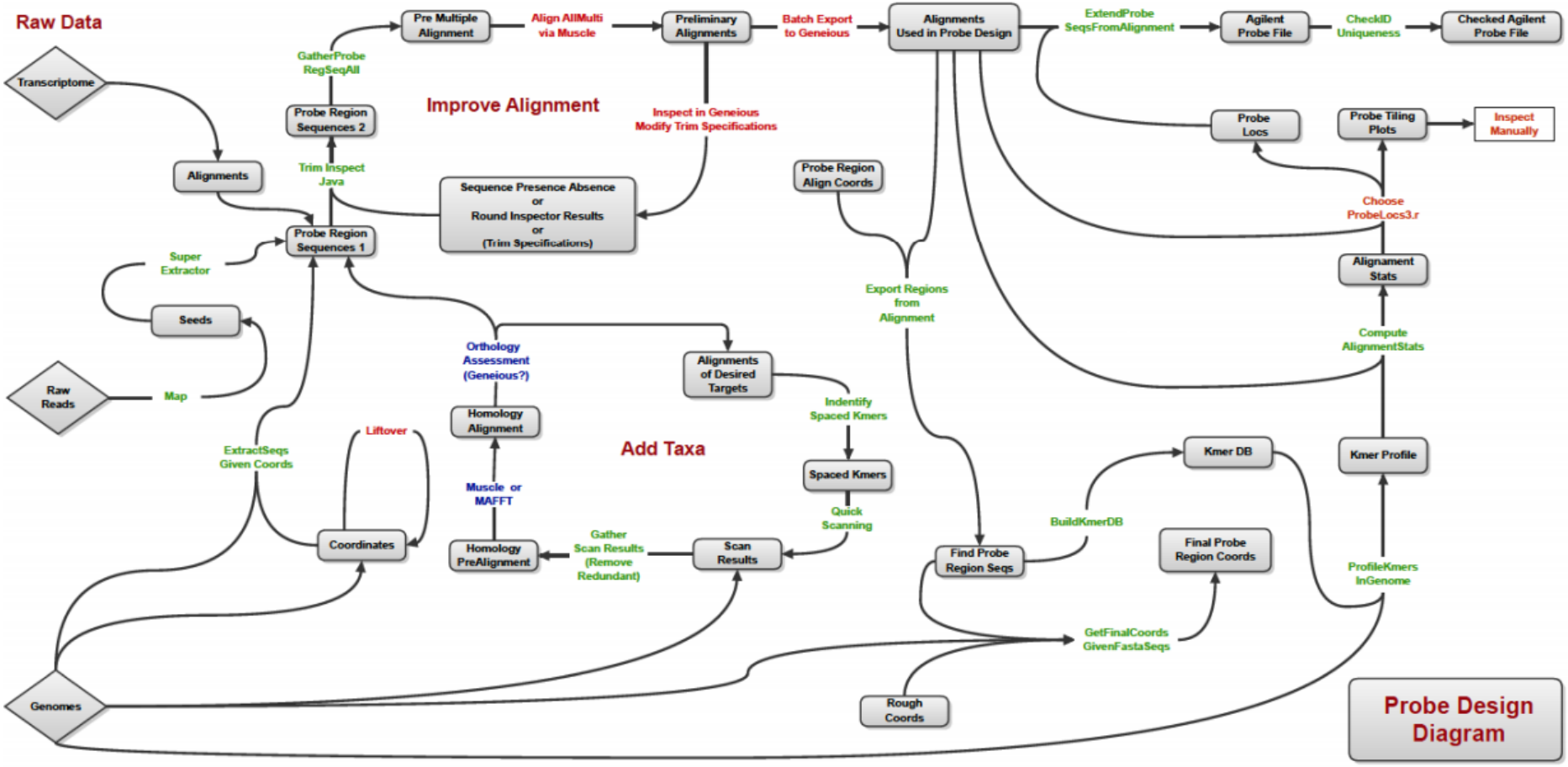
http://jemu.myspecies.info/sites/jemu.myspecies.info/files/EvoLunch_Hybrid%20enrichment.pdf

de novo transcriptome sequencing



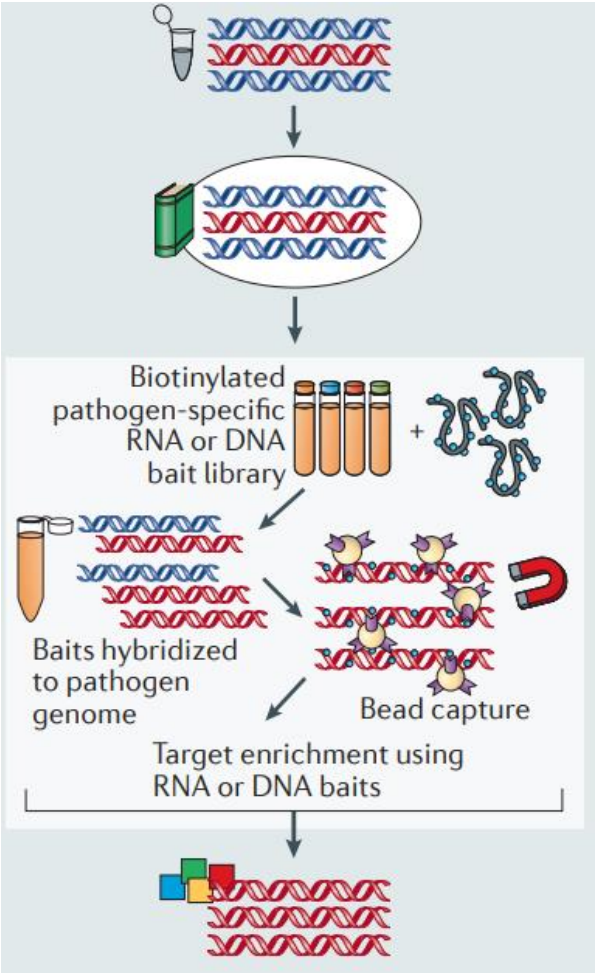
https://en.wikipedia.org/wiki/File:Summary_of_RNA-Seq.svg

Probe Design Workflow



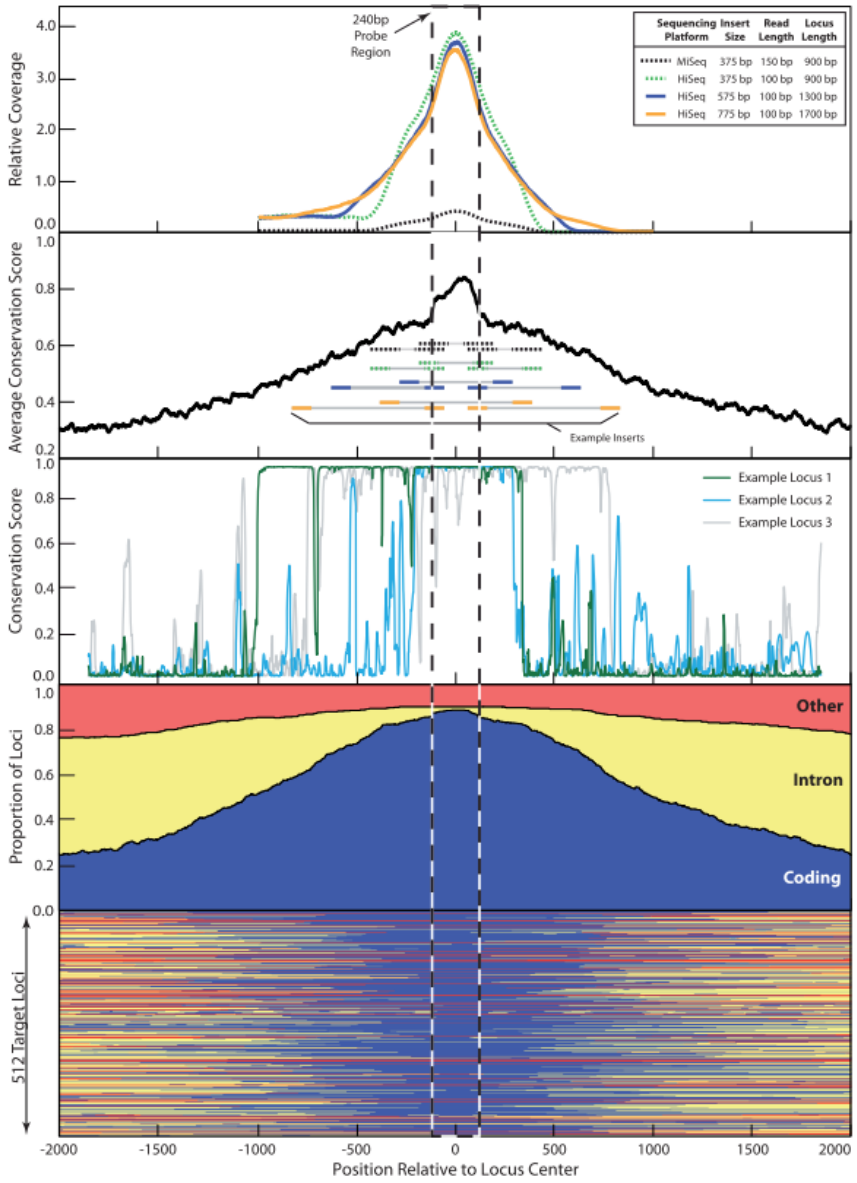
http://jemu.myspecies.info/sites/jemu.myspecies.info/files/EvoLunch_Hybrid%20enrichment.pdf

Experiment for Hybrid-Seq



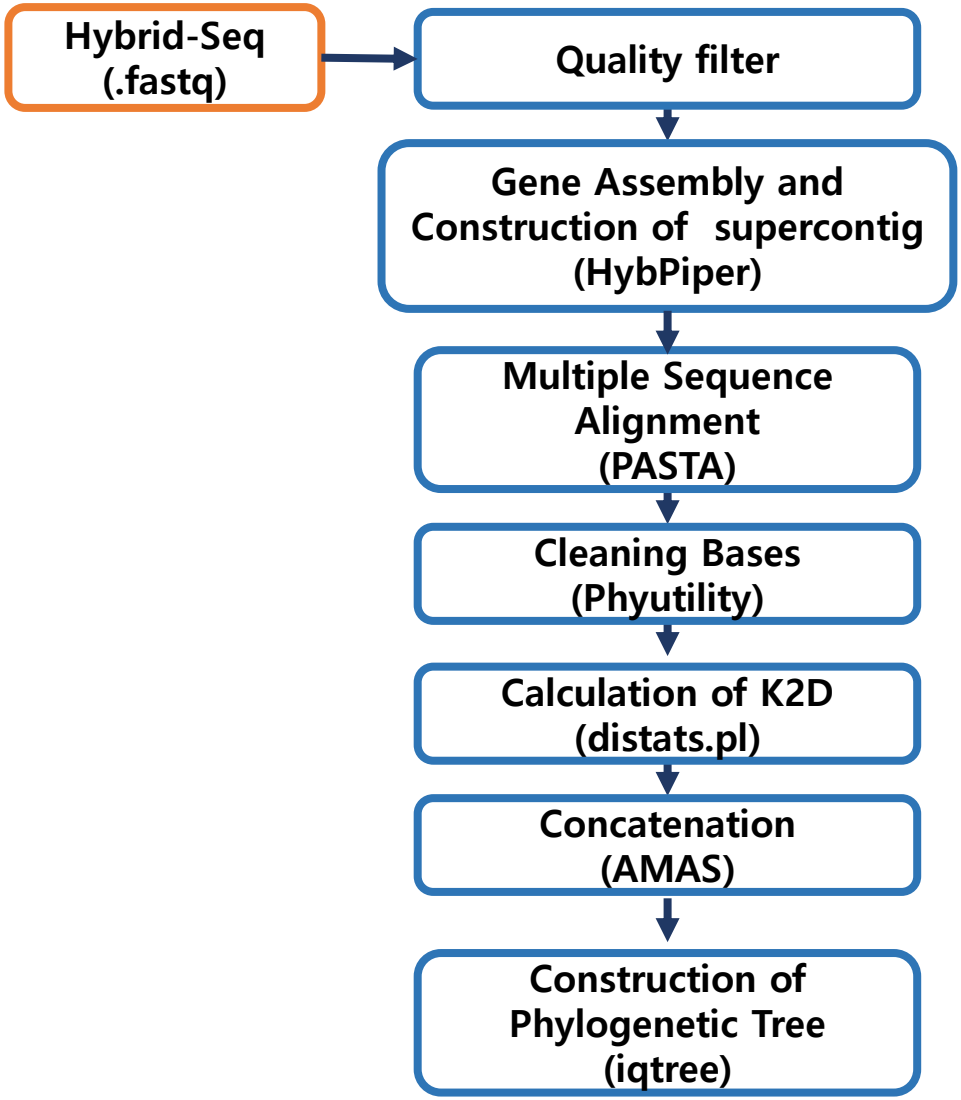
Nature Reviews Microbiology, 2017

Shallow-scale phylogenetic utility



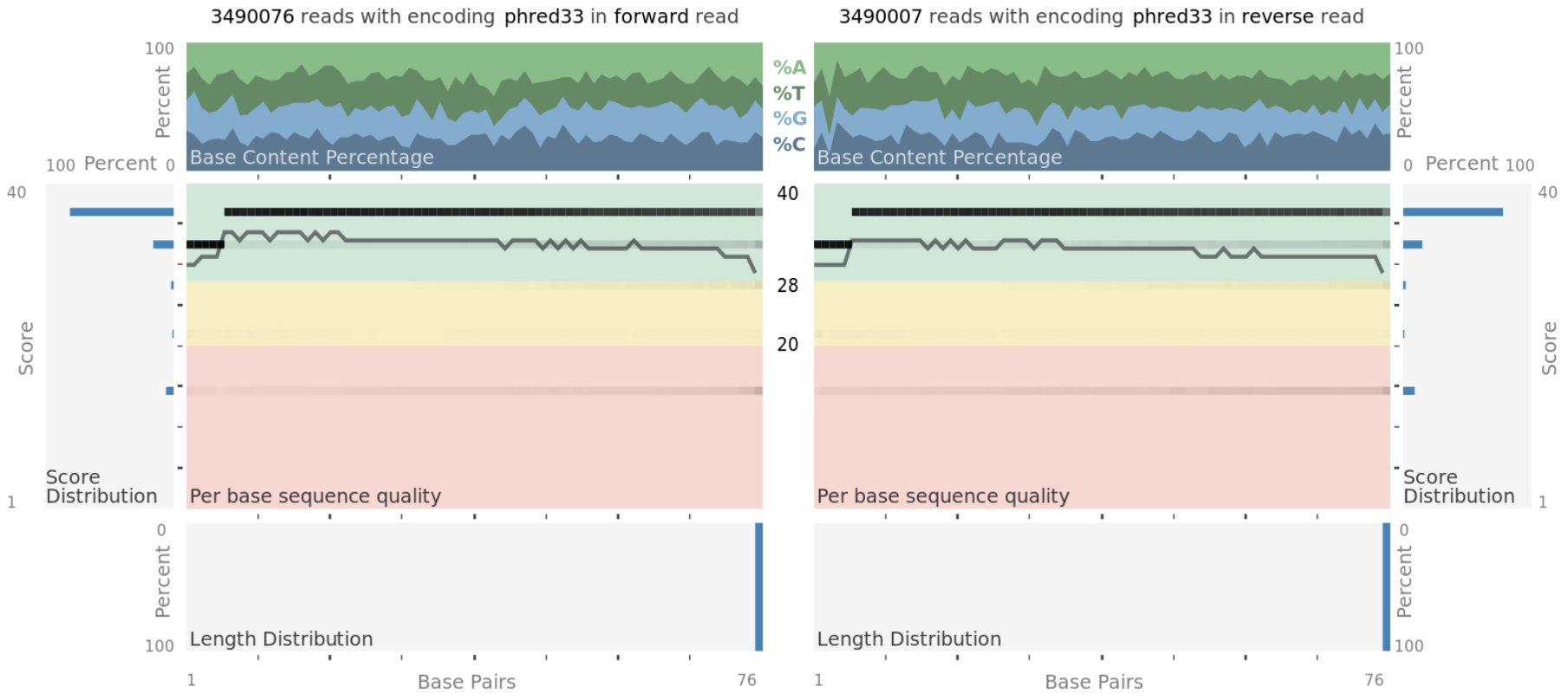
Systematic Biology, 2012

Pipeline to analysis Hybrid-Seq

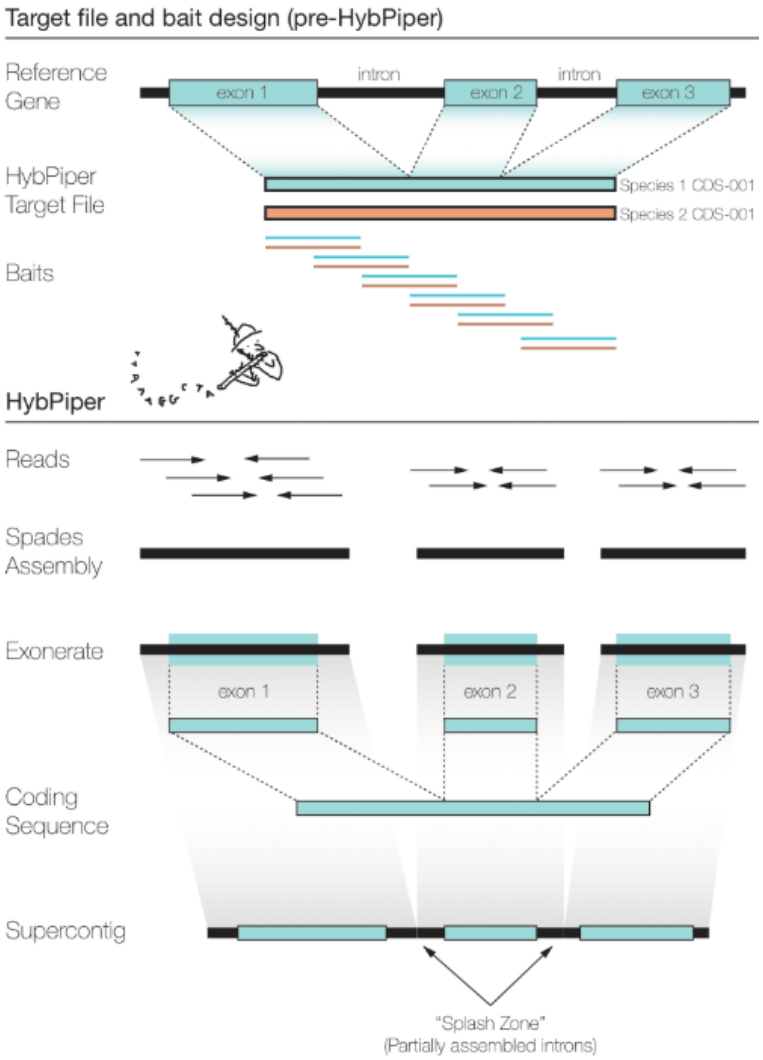


Sequencing quality

-

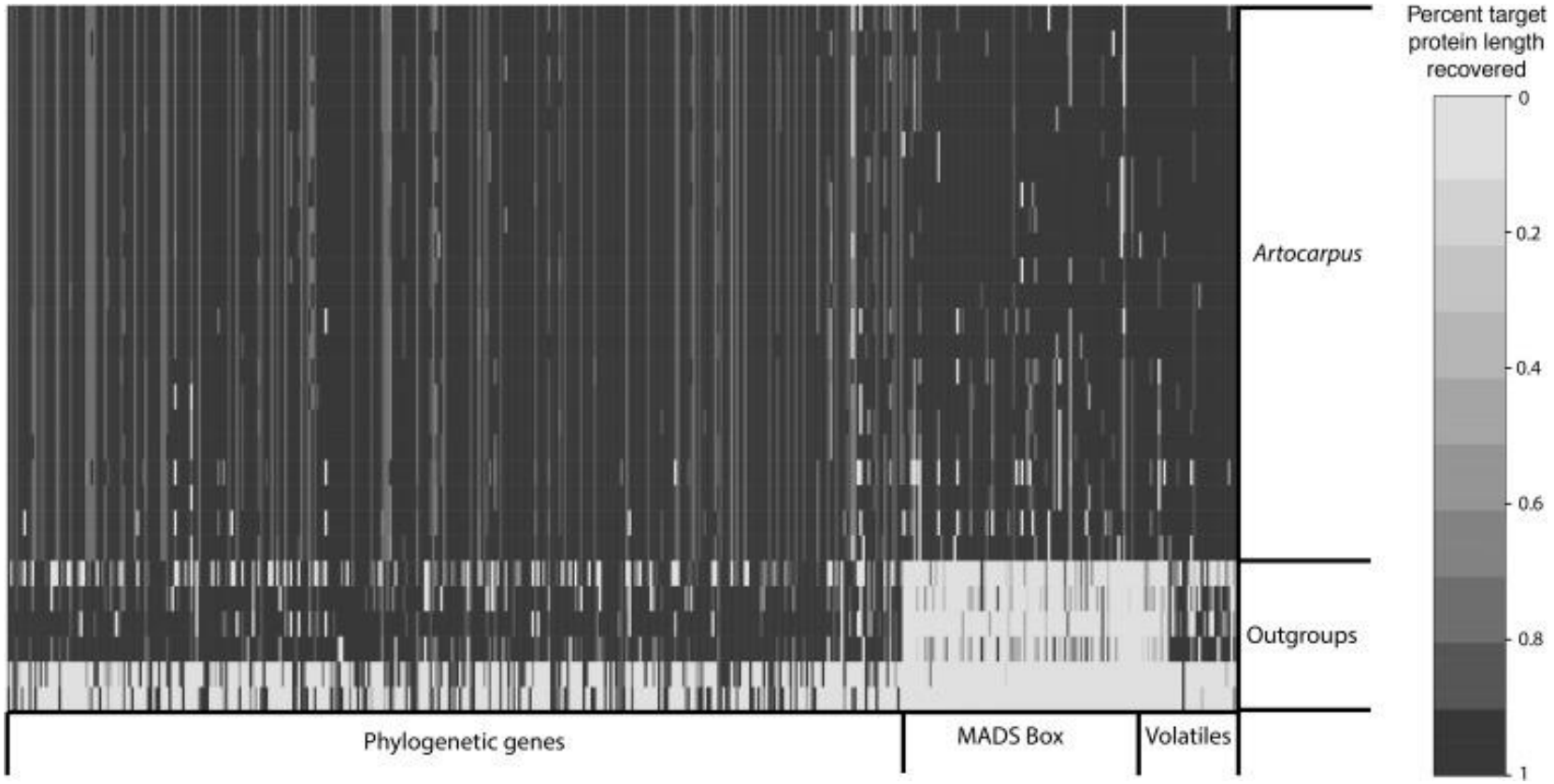


HybPiper



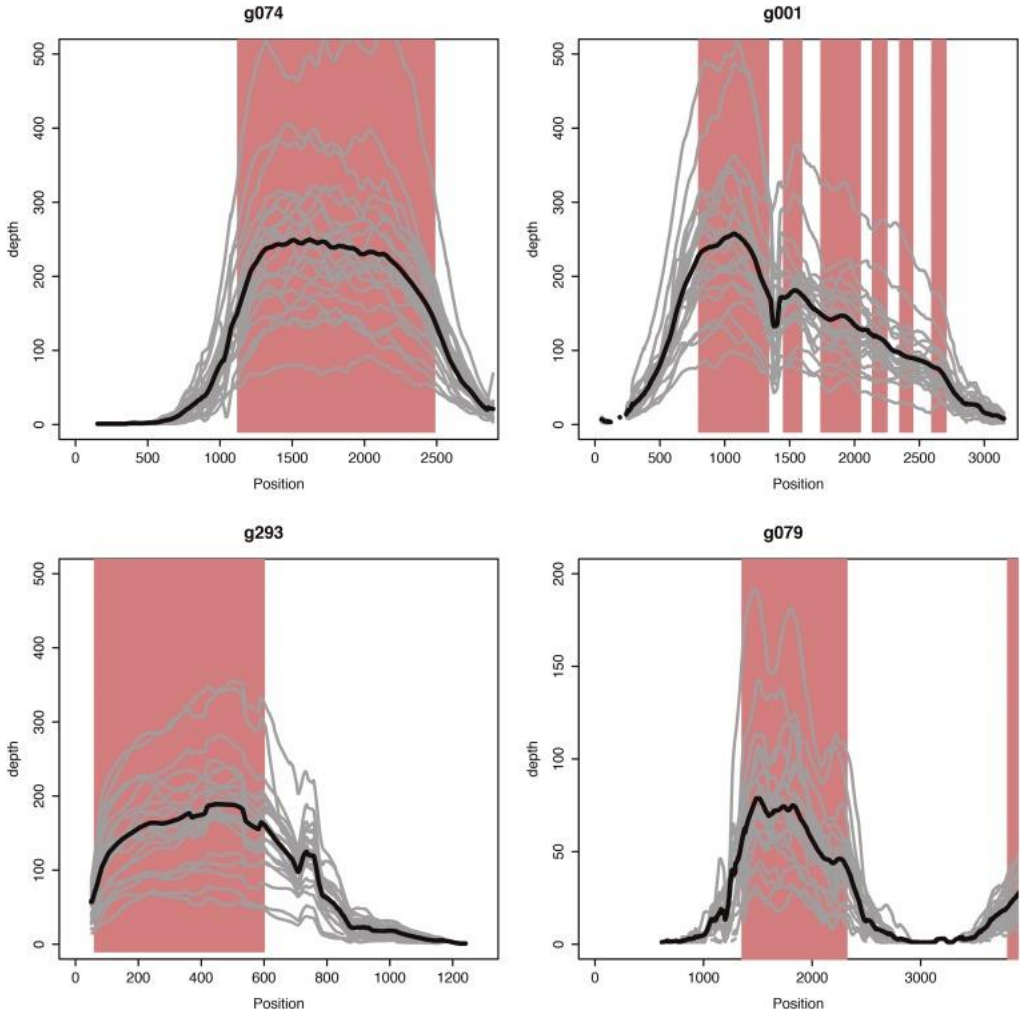
https://github.com/mossmatters/HybPiper/wiki

Recovery efficiency for 458 genes



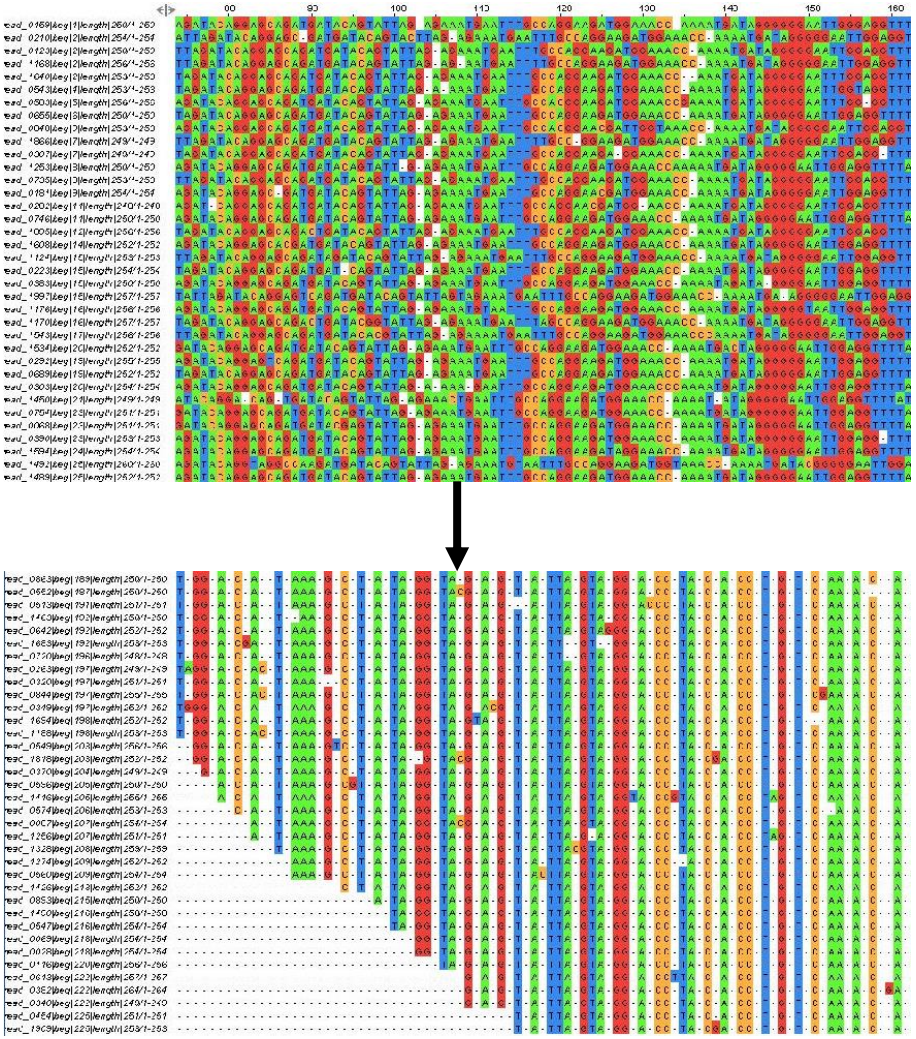
Appl Plant Sci., 2016

Depth-of-coverage plots



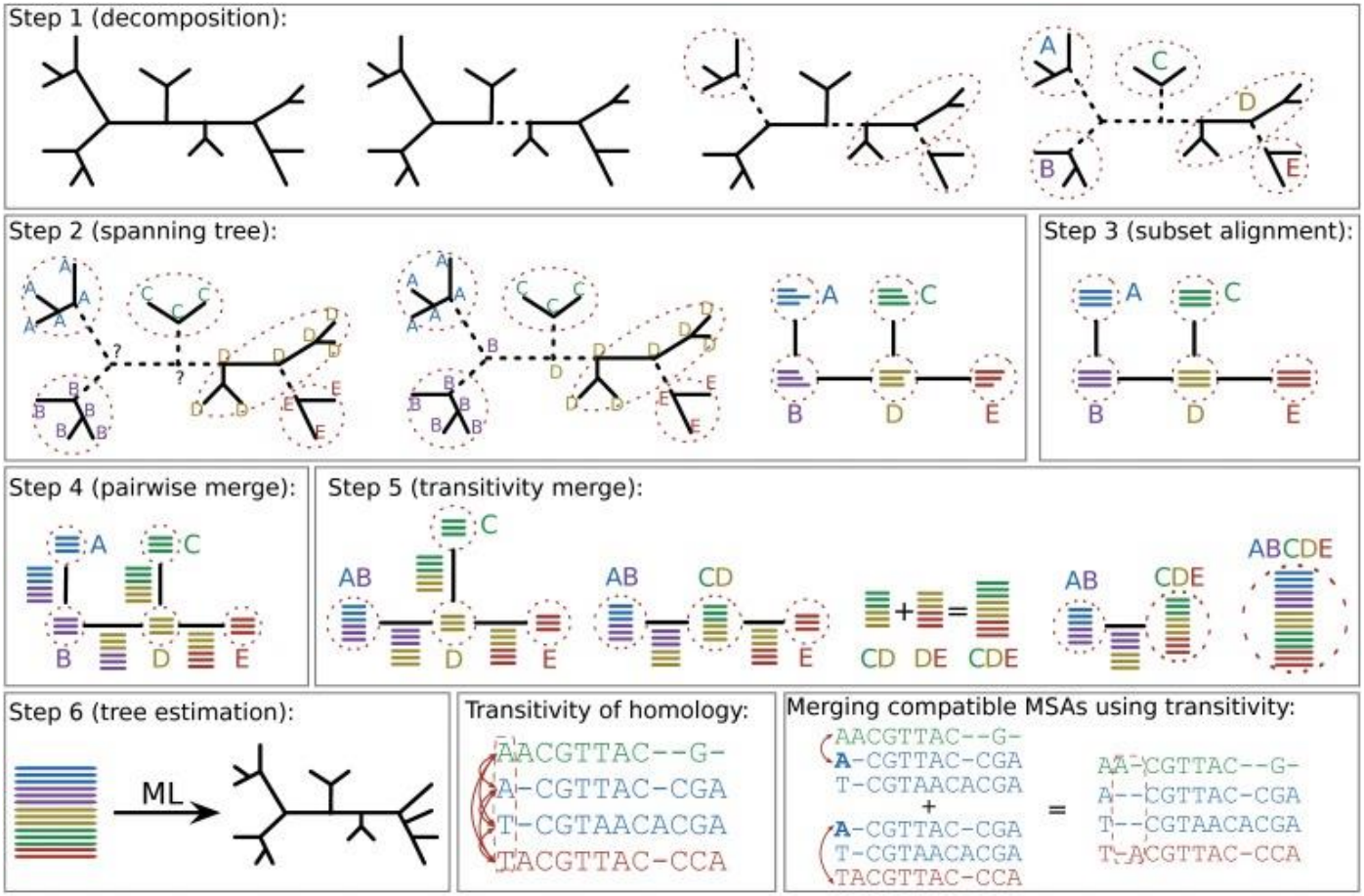
Appl Plant Sci., 2016

Multiple sequence alignment



Bioinformatics and Computational Biology, 2009

Alignment : PASTA



J Comput Biol., 2015

Alignment : mafft

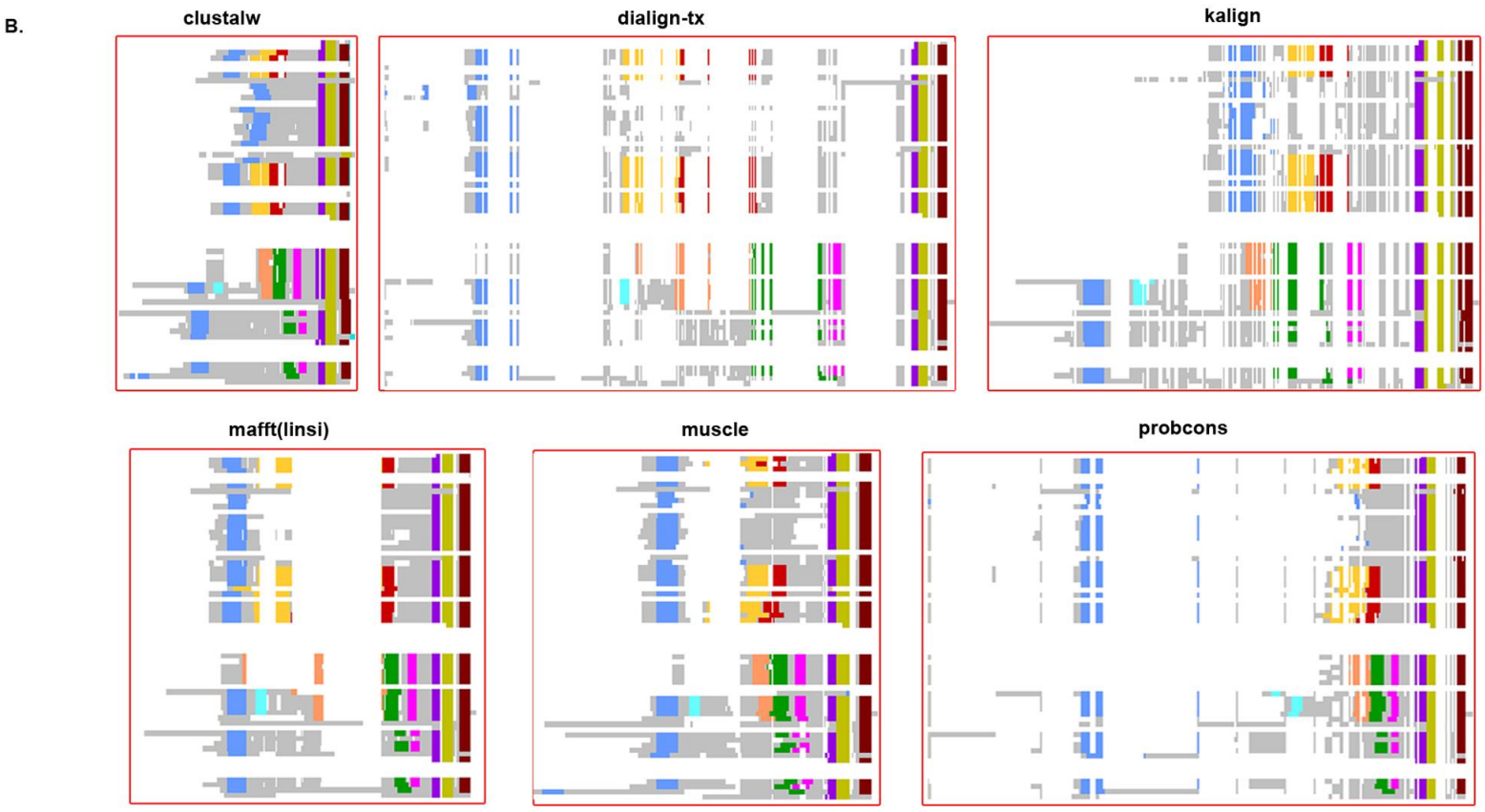
Option name	Command	
For a large-scale alignment (*N* > ~ 10 000). Progressive methods with the PartTree algorithm		
NW-NS-PartTree1	`mafft --parttree --retree 1`	Distance is by the 6mer method
NW-NS-PartTree2	`mafft --parttree --retree 2`	Distance is by the 6mer method. Guide tree is rebuilt
NW-NS-DPPartTree1	`mafft --dpparttree --retree 1`	Distance is estimated based on DP
NW-NS-DPPartTree2	`mafft --dpparttree --retree 2`	Distance is estimated based on DP. Guide tree is re-built
NW-NS-FastaPartTree1	`mafft --fastaparttree --retree 1`	Requires FASTA [40] to estimate distances
NW-NS-FastaPartTree2	`mafft --fastaparttree --retree 2`	Requires FASTA [40]. Guide tree is rebuilt
For a medium-scale alignment (~10 000 > *N* > ~200). Progressive methods		
FFT-NS-1	`mafft --retree 1`	Approximately two times faster than the default
FFT-NS-2	`mafft`	Default
For a small-scale alignment (*N* < ~200, *L* < ~10 000). Iterative refinement methods		
FFT-NS-i	`mafft-fftnsi`	Fastest of the four in this category. Uses WSP score only
G-INS-i	`mafft-ginsi`	Uses WSP score and consistency score from global alignments
L-INS-i	`mafft-linsi`	Uses WSP score and consistency score from local alignments
E-INS-i	`mafft-einsi`	Uses WSP score and consistency score from local alignments with a generalized affine gap cost
For a small-scale RNA alignment (*N* < ~50, *L* < ~1 000). Structural alignment methods		
Q-INS-i	`mafft-qinsi`	Requires no external structural alignment programs
X-INS-i-scarnapair	`mafft-xinsi --scarnapair`	Requires MXSCARNA (Tabei *et al.*, submitted for publication)
X-INS-i-larapair	`mafft-xinsi --larapair`	Requires LaRA [78]
X-INS-i-foldalignlocalpair	`mafft-xinsi --foldalignlocalpair`	Requires FOLDALIGN [79]. Uses the local alignment option
X-INS-i-foldalignglobalpair	`mafft-xinsi --foldalignglobalpair`	Requires FOLDALIGN [97]. Uses the global alignment option
If not sure which option to use		
Automatic	`mafft --auto`	Selects an appropriate option from FFT-NS-2, FFT-NS-i and L-INS-i, according to the size of input data

Briefings in Bioinformatics, 2008

Result from several programs

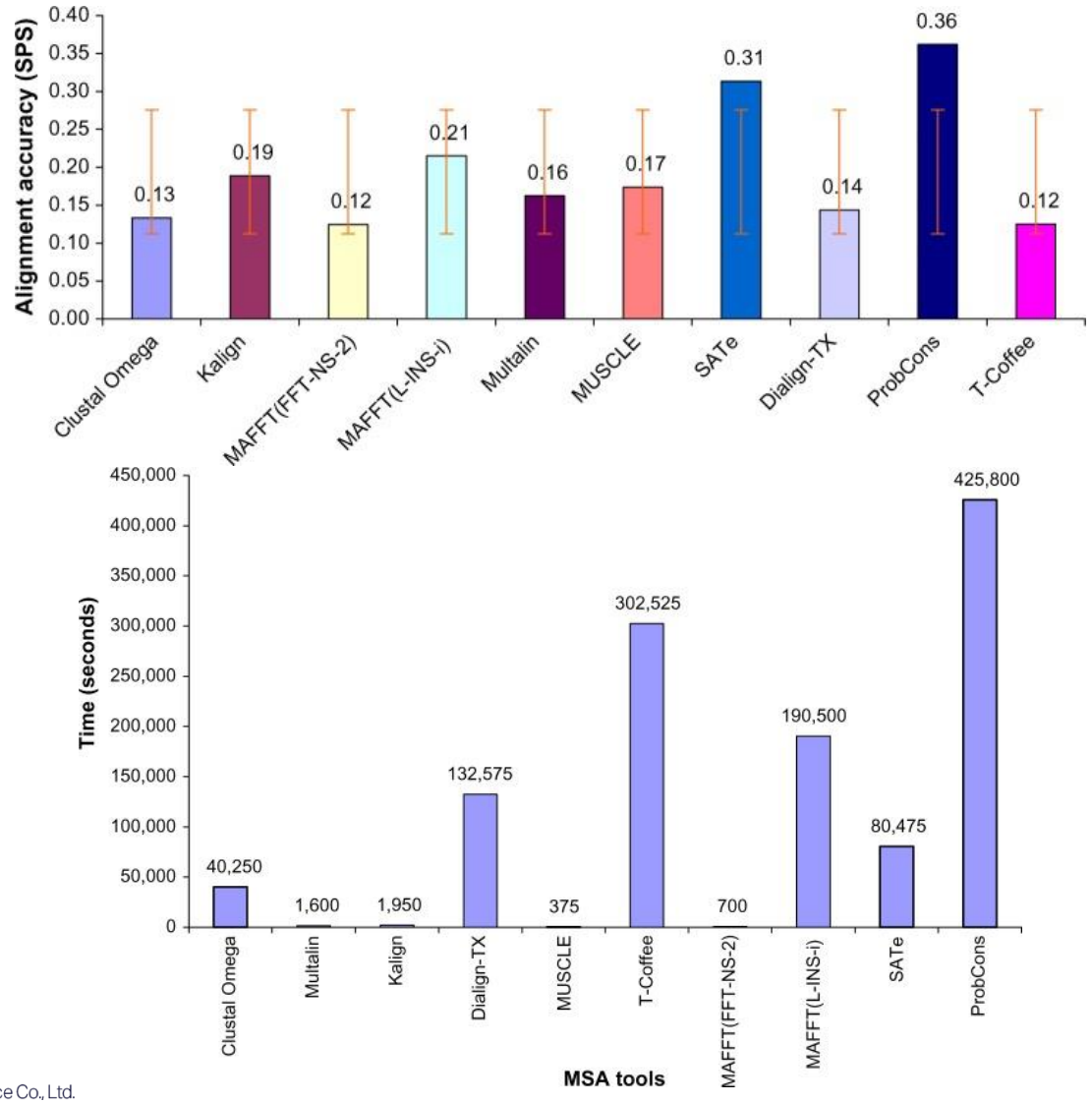
Plos one, 2011

Result from several programs



Plos one, 2011

Comparison among MSA tools



Evol Bioinform Online., 2014

AMAS : Concatenation

PeerJ

AMAS: a fast tool for alignment manipulation and computing of summary statistics

Marek L. Borowiec

Department of Entomology and Nematology, UC Davis, Davis, United States

A: concatenation

```
AMAS concat -i FILE_1.fas FILE_2.fas -f fasta -d dna --concat-out CONCATENATED_FILE.fas
```

FILE_1.fas

```
>taxon1
GGCGAATTCC
>taxon2
GGCGCATTCC
```

FILE_2.fas

```
>taxon1
AAATTTACCG
>taxon3
AAATAATCGG
```

→

CONCATENATED_FILE.fas

```
>taxon1
GGCGAATTCCAAATTTACCG
>taxon2
GGCGCATTCC??????????
>taxon3
??????????AAATAATCGG
```

partitions.txt

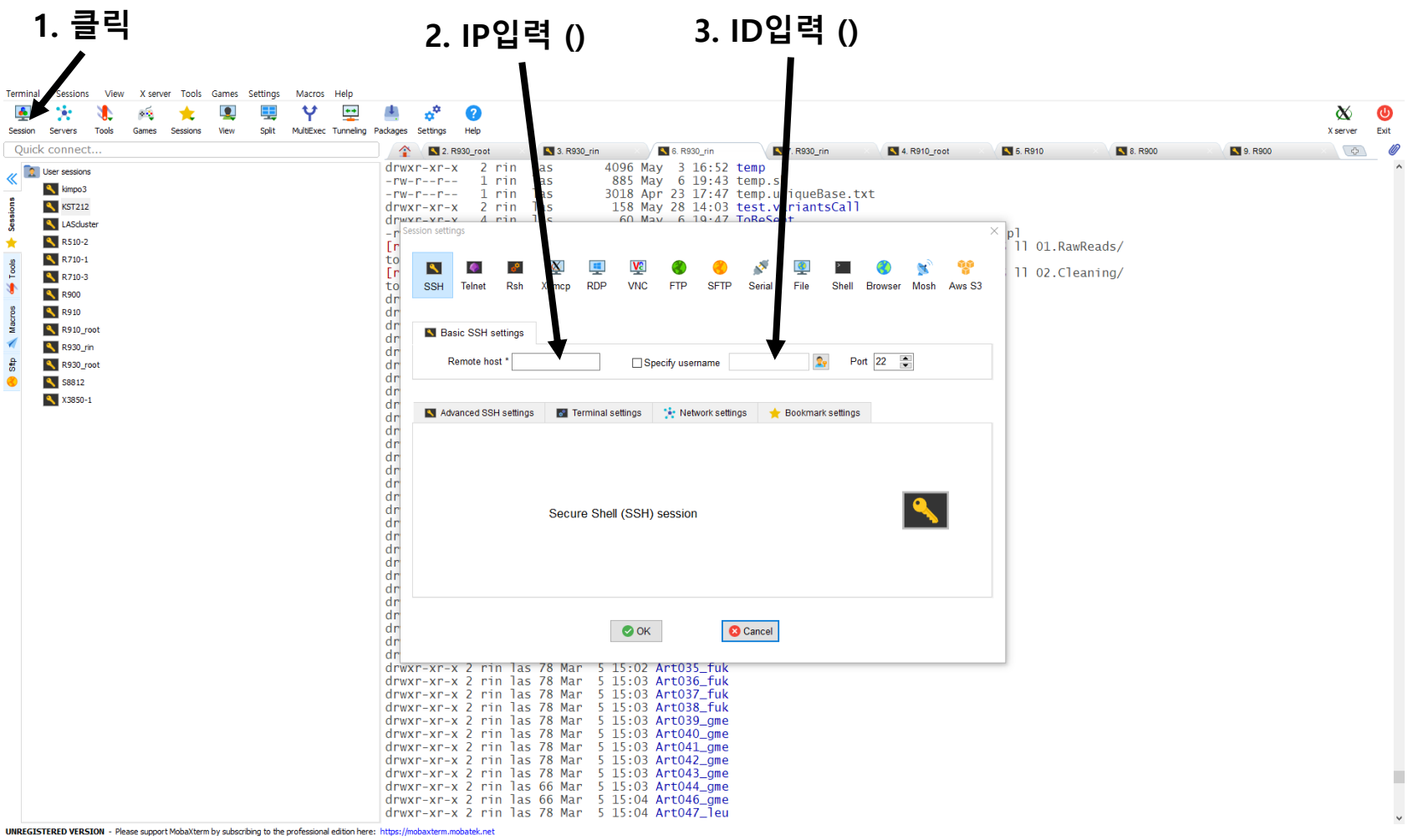
```
p1_FILE_1 = 1-10
p2_FILE_2 = 11-20
```

PeerJ, 2016

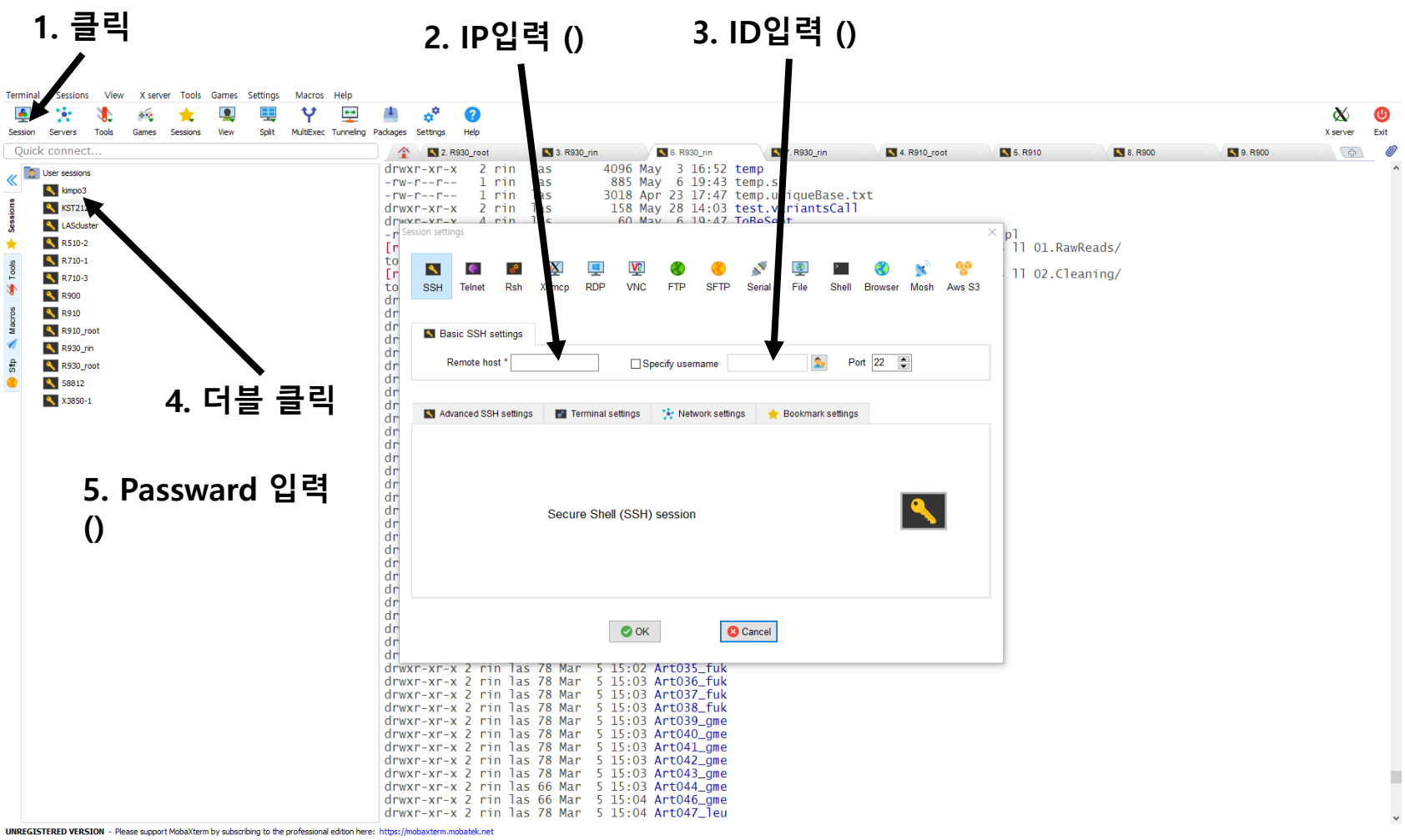
Positon of each gene on concatenated file

```
p1_DN12651_c0_g1_i1=1-1454
p2_DN68192_c0_g1_i2=1455-4746
p3_DN68970_c0_g1_i1=4747-9104
p4_DN68974_c0_g1_i1=9105-12393
p5_DN72723_c0_g1_i1=12394-16346
p6_DN75016_c0_g1_i5=16347-18196
p7_DN75231_c0_g1_i3=18197-21391
p8_DN75819_c0_g1_i2=21392-24880
p9_DN77438_c0_g1_i1=24881-27225
p10_DN77655_c0_g1_i1=27226-29364
```

서버 접속


1. 클릭
2. IP입력 ()
3. ID입력 ()
Session settings
SSH
Telnet
Rsh
RDP
VNC
FTP
SFTP
Serial
File
Shell
Browser
Mosh
Aws S3
Basic SSH settings
Remote host *
Specify username
Port 22
Advanced SSH settings
Terminal settings
Network settings
Bookmark settings
Secure Shell (SSH) session
OK
Cancel

서버 접속


1. 클릭
2. IP입력 ()
3. ID입력 ()
4. 더블 클릭
5. Password 입력 ()
Session settings
SSH
Telnet
Rsh
RDP
VNC
FTP
SFTP
Serial
File
Shell
Browser
Mosh
Aws S3
Basic SSH settings
Remote host *
Specify username
Port 22
Advanced SSH settings
Terminal settings
Network settings
Bookmark settings
Secure Shell (SSH) session
OK
Cancel

리눅스 기본 명령어

pwd (print working directory)

현재 작업중인 디렉토리 정보 출력

```
# pwd
/home/itholic
```

cd (change directory)

경로 이동

절대 경로와 상대 경로로 이동 가능하다.

```
# cd /home/itholic/mydir
# pwd
/home/itholic/mydir


# cd ..
# pwd
/home/itholic
```

ls (list)

디렉토리 목록 확인

```
# ls
testfile1  testfile2  testfile3


# ls -l
total 0
-rw-r--r-- 1 itholic 197121 0 11월  6 22:08 testfile1
-rw-r--r-- 1 itholic 197121 0 11월  6 22:08 testfile2
-rw-r--r-- 1 itholic 197121 0 11월  6 22:08 testfile3


# ls -a
./  ../  testfile1  testfile2  testfile3


# ls -al
total 4
drwxr-xr-x 1 itholic 197121 0 11월  6 22:08 ./
drwxr-xr-x 1 itholic 197121 0 11월  6 22:08 ../
-rw-r--r-- 1 itholic 197121 0 11월  6 22:08 testfile1
-rw-r--r-- 1 itholic 197121 0 11월  6 22:08 testfile2
-rw-r--r-- 1 itholic 197121 0 11월  6 22:08 testfile3
```

cp (copy)

파일 혹은 디렉토리를 복사

디렉토리를 복사할때는 -r 옵션을 주어야함

```
# ls
testdir/  testfile


# cp testfile1 testfile_cp
# ls
testdir/  testfile  testfile_cp


# cp -r testdir testdir_cp
# ls
testdir/  testdir_cp/  testfile  testfile_cp
```

리눅스 기본 명령어

mv (move)

파일 혹은 디렉토리 이동

실제로 원하는 위치로 이동할때도 사용하지만, 이름을 변경하는 용도로도 사용한다.

cp와는 달리 디렉토리를 이동할때도 별다른 옵션이 필요 없다.

```
# ls
testdir/  testfile


# mv testfile testfile_mv
# ls
testdir/  testfile_mv


# mv testfile_mv testdir/
# ls
testdir/


# ls testdir/
testfile
```

rm (remove)

파일이나 디렉토리를 삭제

디렉토리를 삭제할때는 r 옵션을 주어야 한다.

-f 옵션을 주면 사용자에게 삭제 여부를 묻지 않고 바로 삭제한다.

디렉토리를 삭제할 때에는 하위 디렉토리까지 모두 삭제되므로 유의하자.

```
# ls
testdir/  testfile1  testfile2


# rm -f testfile1
# ls
testdir/  testfile2


# rm -rf testdir/
# ls
testfile2
```

리눅스 기본 명령어

cat (concatenate)

cat 명령은 활용 방법이 꽤나 다양하다.

단순히 파일의 내용을 출력할 수도 있고,

파일 여러개를 합쳐서 하나의 파일로 만들 수도 있다.

그리고 기존 한 파일의 내용을 다른 파일에 덧붙일수도 있다.

새로운 파일을 만들때에도 사용된다.

file1, file2, file3 파일에는 각각 간단하게 숫자 1, 2, 3 이 적혀있다.

```
# ls
file1  file2  file3


# cat file1
1


# cat file2
2


# cat file3
3


# cat file1 file2 > file1_2
# ls
file1  file1_2  file2  file3


# cat file1_2
1
2
```

head

파일의 앞부분을 보고싶은 줄 수만큼 보여준다.

옵션을 지정하지 않으면 파일 상위 10줄을 보여준다.

```
# cat testfile
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15


# head -3 testfile
1
2
3
```

리눅스 기본 명령어

cat (concatenate)

cat 명령은 활용 방법이 꽤나 다양하다.

단순히 파일의 내용을 출력할 수도 있고,

파일 여러개를 합쳐서 하나의 파일로 만들 수도 있다.

그리고 기존 한 파일의 내용을 다른 파일에 덧붙일수도 있다.

새로운 파일을 만들때에도 사용된다.

file1, file2, file3 파일에는 각각 간단하게 숫자 1, 2, 3 이 적혀있다.

```
# ls
file1  file2  file3


# cat file1
1


# cat file2
2


# cat file3
3


# cat file1 file2 > file1_2
# ls
file1  file1_2  file2  file3


# cat file1_2
1
2
```

리눅스 기본 명령어

head

파일의 앞부분을 보고싶은 줄 수만큼 보여준다.

옵션을 지정하지 않으면 파일 상위 10줄을 보여준다.

```
# cat testfile
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15


# head -3 testfile
1
2
3
```

tail

파일의 뒷부분을 보고싶은 줄 수만큼 보여준다.

옵션을 지정하지 않으면 파일 하위 10줄을 보여준다.

참고로 -F 옵션을 주고 실행하면,

파일 내용을 화면에 계속 띄워주고 파일이 변하게되면 새로운 업데이트된 내용을 갱신해준다.

주로 실시간으로 내용이 추가되는 로그파일을 모니터링할때 유용하게 사용한다.

```
# cat testfile
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15


# tail -3 testfile
13
14
15
```

리눅스 기본 명령어

find

특정 파일이나 디렉토리를 검색한다

사용법이 앞의 명령어들에비해 살짝 복잡하므로, 기본 사용법을 언급하자면 다음과 같다.

> find [검색경로] -name [파일명]

파일명은 직접 풀 네임을 입력해도 되지만,

다음 예제처럼 특정 조건을 적용해 검색할수도 있다.

나같은 경우 주로 특정 확장자명을 찾기 위해 사용한다.

```
# ls
dir1/  dir3/  file1  file3  picture1.jpg  picture3.jpg
dir2/  dir4/  file2  file4  picture2.jpg  picture4.jpg


# find ./ -name 'file1'
./file1


# find ./ -name "*.jpg"
./picture1.jpg
./picture2.jpg
./picture3.jpg
./picture4.jpg
```

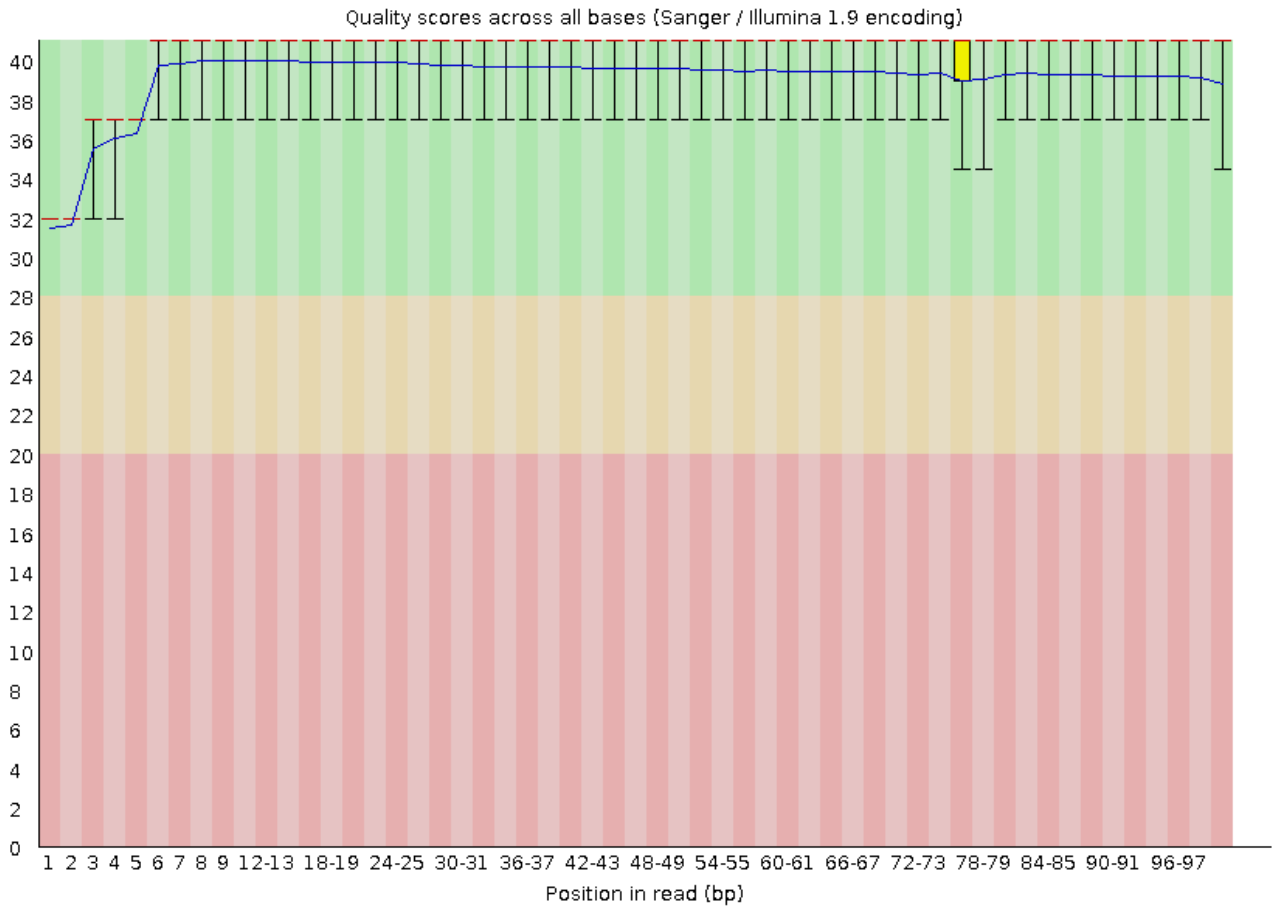
Quality check

fastqc [Input fastq]

```
fastqc RawData/Art020.R1.fq.gz
fastqc RawData/Art020.R2.fq.gz

fastqc RawData/Art025.R1.fq.gz
fastqc RawData/Art025.R2.fq.gz

fastqc RawData/Art069.R1.fq.gz
fastqc RawData/Art069.R2.fq.gz

firefox RawData/Art020.R1_fastqc.html
```

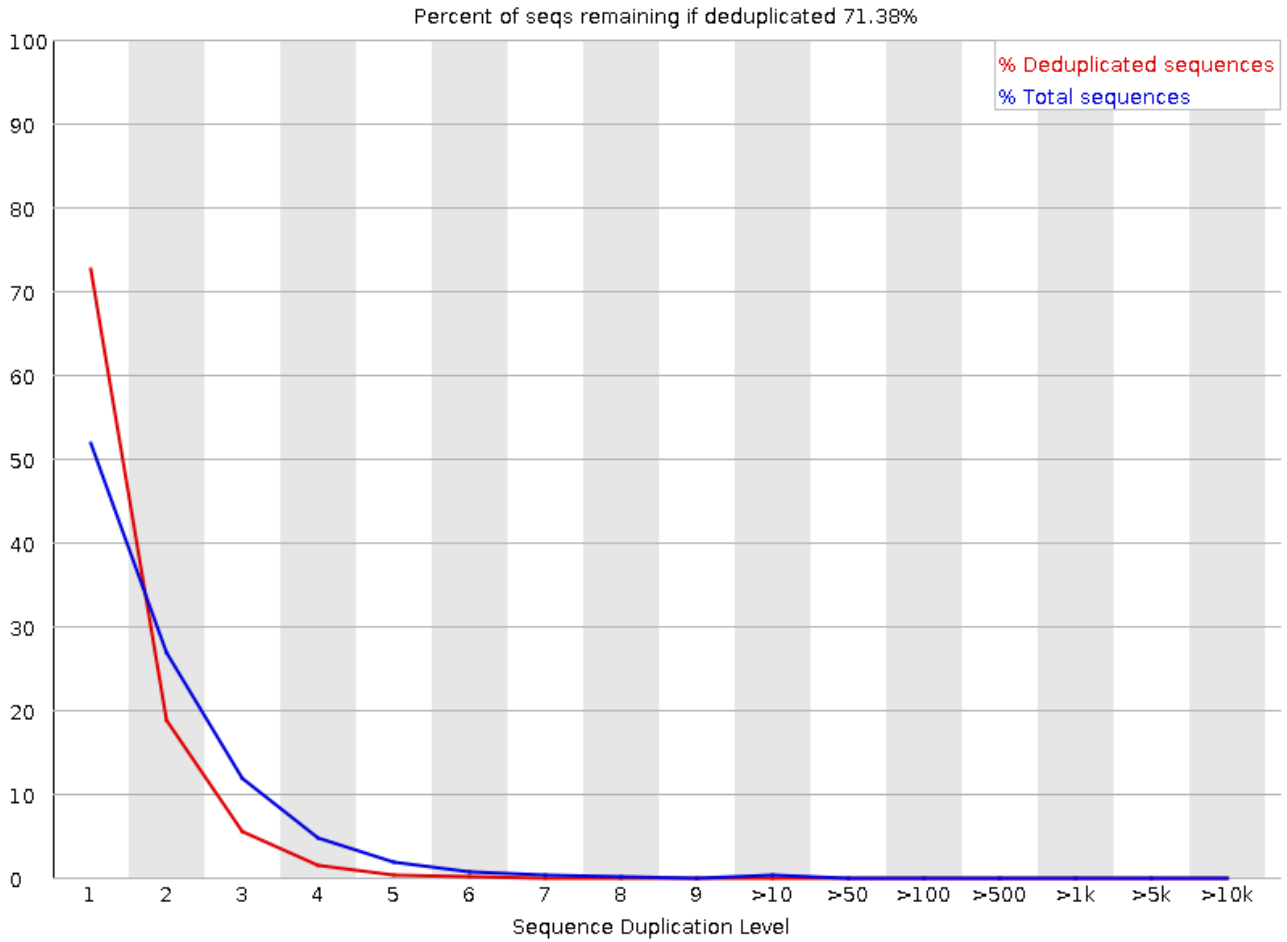
Quality check – Per base sequence quality

Per base sequence quality



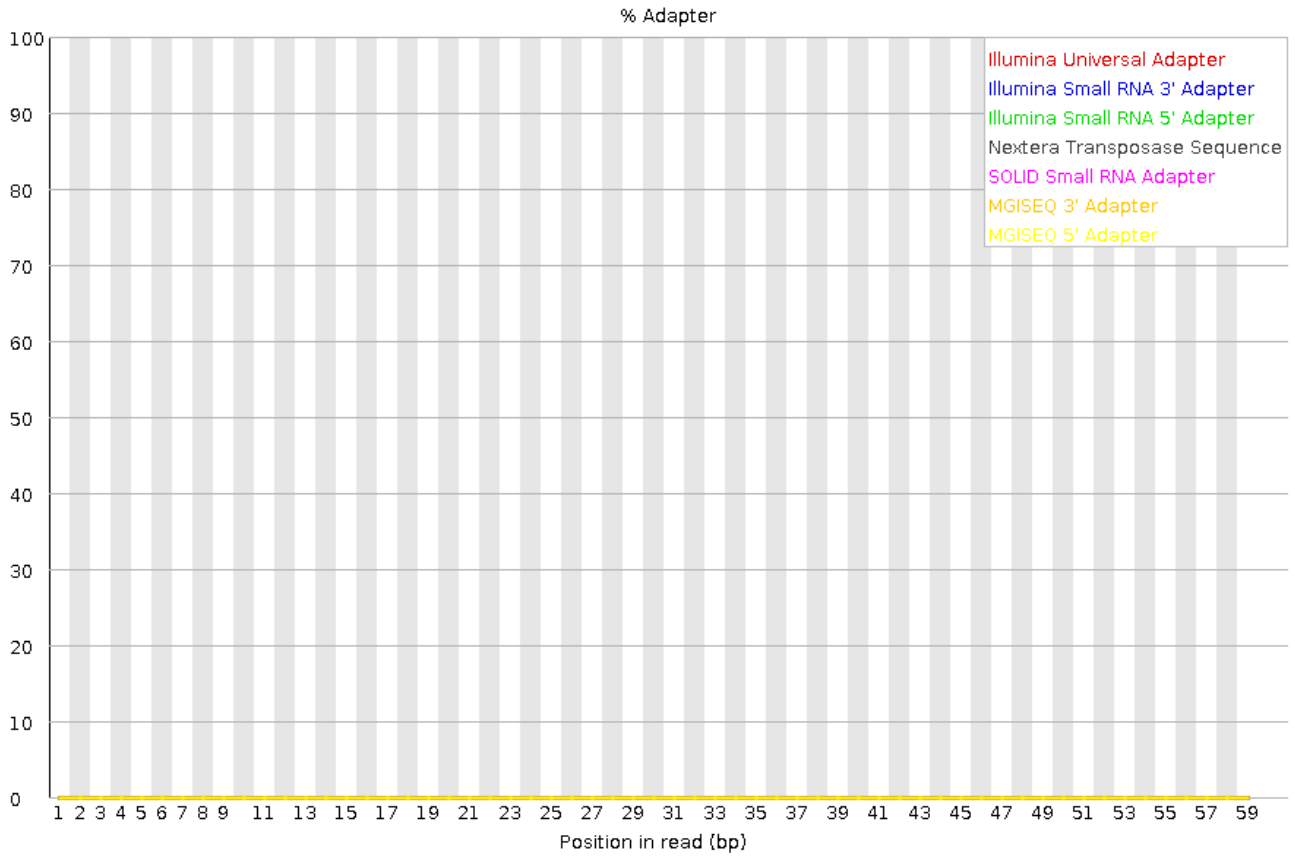
Quality check – Sequence duplication levels

Sequence Duplication Levels



Quality check – Adapter content

Adapter Content



Read cleaning

mkdir CleanRead

skewer -t 2 -m pe -q 30 -l 25 -k 3 -r 0.1 -d 0.1 -x AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -y AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -o CleanRead/Art020 RawData/Art020.R1.fq.gz RawData/Art020.R2.fq.gz

skewer -t 2 -m pe -q 30 -l 25 -k 3 -r 0.1 -d 0.1 -x AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -y AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -o CleanRead/Art025 RawData/Art025.R1.fq.gz RawData/Art025.R2.fq.gz

skewer -t 2 -m pe -q 30 -l 25 -k 3 -r 0.1 -d 0.1 -x AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -y AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -o CleanRead/Art069 RawData/Art069.R1.fq.gz RawData/Art069.R2.fq.gz

skewer -t 2 -m pe -q 30 -l 25 -k 3 -r 0.1 -d 0.1 -x AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -y AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -o CleanRead/Art102 RawData/Art102.R1.fq.gz RawData/Art102.R2.fq.gz

mq="30" # mq: minimum quality; quality to trim off the read terminal under $mq
st="3" # st: stringency; length of overlap with adapter sequence required to trim a sequence
er="0.1" # er: error rate; number of errors divided by the length of the matching region; default is '0.1'.
ml="25" # ml: minimum length; length to discard the under $ml after trimming
a1="AGATCGGAAGAGCACACGTCTGAACTCCAGTCA" # a1: Common Region of TruSeq Indexed Adapter of Read 1
a2="AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT" # a2: Common Region of TruSeq Indexed Adapter of Read 2

HybPiper

```
mkdir HybPiper
mkdir SuperContig

reads_first.py -b Reference/Single.Copy.GeneSequence.fa -r CleanRead/Art020-trimmed-pair?.fastq --prefix
HybPiper/Art020 --bwa --cpu 16
reads_first.py -b Reference/Single.Copy.GeneSequence.fa -r CleanRead/Art025-trimmed-pair?.fastq --prefix
HybPiper/Art025 --bwa --cpu 16
reads_first.py -b Reference/Single.Copy.GeneSequence.fa -r CleanRead/Art069-trimmed-pair?.fastq --prefix
HybPiper/Art069 --bwa --cpu 16
reads_first.py -b Reference/Single.Copy.GeneSequence.fa -r CleanRead/Art102-trimmed-pair?.fastq --prefix
HybPiper/Art102 --bwa --cpu 16


python /home/Programs/HybPiper/intronerate.py --prefix HybPiper/Art020
python /home/Programs/HybPiper/intronerate.py --prefix HybPiper/Art025
python /home/Programs/HybPiper/intronerate.py --prefix HybPiper/Art069
python /home/Programs/HybPiper/intronerate.py --prefix HybPiper/Art102
```

HybPiper

```
cd HybPiper/ ; python
/home/Programs/HybPiper/get_seq_lengths.py ../Reference/Single.Copy.GeneSequence.fa ../Reference/Name_lis
t.txt dna > Seq_Length.txt ; python /home/Programs/HybPiper/hybpiper_stats.py
Seq_Length.txt ../Reference/Name_list.txt > HybPiper_Stats.txt

while read i; do echo $i; python /home/Programs/HybPiper/paralog_investigator.py $i; done
< ../Reference/Name_list.txt &> paralog_investigator.log

python /home/Programs/HybPiper/retrieve_sequences.py ../Reference/Single.Copy.GeneSequence.fa .
supercontig > supercontigs.log


cd ../
perl Reference/singleline.pl HybPiper/c8590_g1_i1_supercontig.fasta >
SuperContig/c8590_g1_i1_supercontig.fasta
perl Reference/singleline.pl HybPiper/c46363_g1_i1_supercontig.fasta >
SuperContig/c46363_g1_i1_supercontig.fasta
perl Reference/singleline.pl HybPiper/c45242_g1_i2_supercontig.fasta >
SuperContig/c45242_g1_i2_supercontig.fasta
perl Reference/singleline.pl HybPiper/c36136_g2_i2_supercontig.fasta >
SuperContig/c36136_g2_i2_supercontig.fasta
```

Gene sequence summary

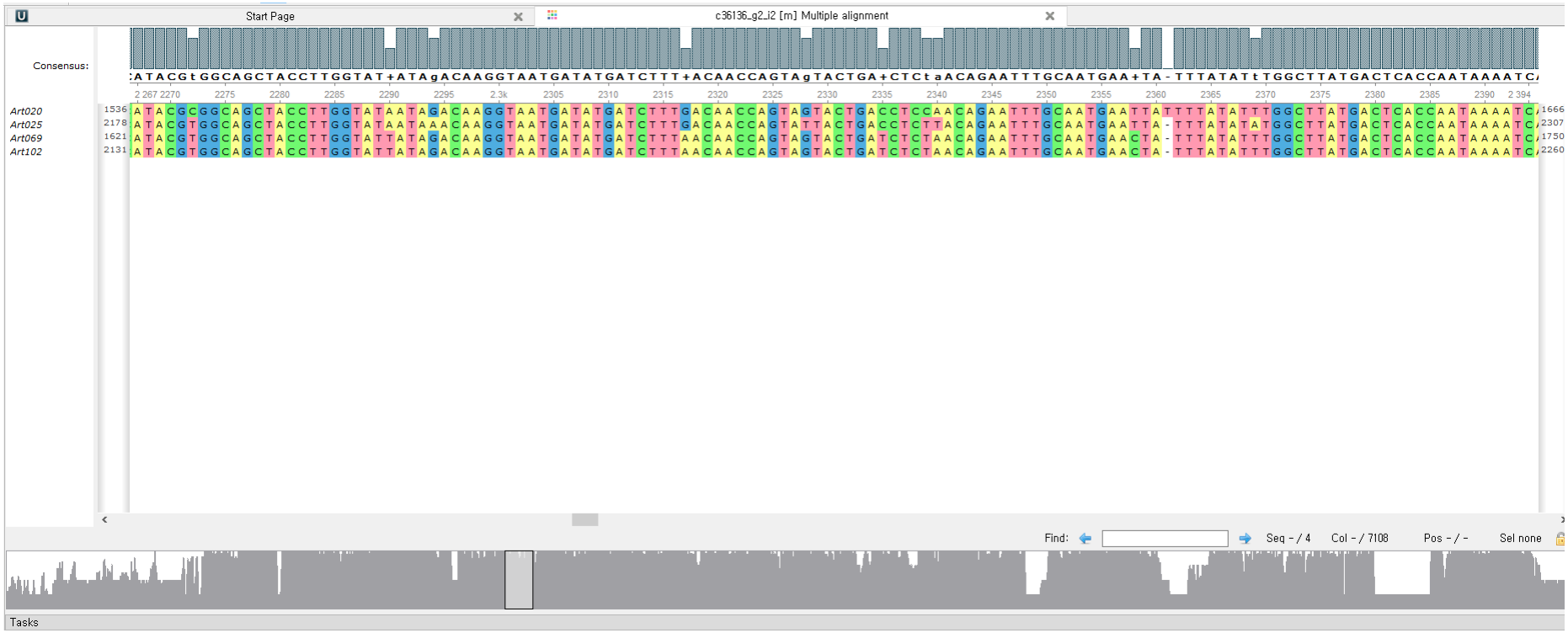
Alignment_name	No_of_taxa	Alignment_length	Total_matrix_cells	Undetermined_characters	Missing_percent	No_variable_sites	Proportion_variable_sites	Parsimony_informative_sites	Proportion_parsimony_informative
DN12651_c0_g1_i1.clean09.fas	4	1162	4648	0	0	76	0.065	8	0.007
DN68192_c0_g1_i2.clean09.fas	4	2805	11220	0	0	226	0.081	68	0.024
DN68970_c0_g1_i1.clean09.fas	4	3177	12708	0	0	161	0.051	40	0.013
DN68974_c0_g1_i1.clean09.fas	4	2345	9380	0	0	49	0.021	8	0.003
DN72723_c0_g1_i1.clean09.fas	4	2862	11448	0	0	110	0.038	16	0.006
DN75016_c0_g1_i5.clean09.fas	2	1850	3700	0	0	233	0.126	0	0
DN75231_c0_g1_i3.clean09.fas	4	3101	12404	0	0	85	0.027	19	0.006
DN75819_c0_g1_i2.clean09.fas	4	2892	11568	0	0	229	0.079	47	0.016
DN77438_c0_g1_i1.clean09.fas	1	2345	2345	0	0	0	0	0	0
DN77655_c0_g1_i1.clean09.fas	4	1898	7592	0	0	36	0.019	4	0.002
DN77838_c0_g1_i4.clean09.fas	4	2665	10660	0	0	140	0.053	27	0.01
DN78262_c0_g1_i2.clean09.fas	4	1510	6040	0	0	75	0.05	14	0.009
DN78320_c0_g1_i2.clean09.fas	3	2359	7077	0	0	79	0.033	0	0
DN78640_c0_g1_i4.clean09.fas	4	2611	10444	0	0	112	0.043	27	0.01
DN78677_c0_g1_i2.clean09.fas	2	2826	5652	0	0	52	0.018	0	0
DN78863_c0_g1_i1.clean09.fas	4	2568	10272	0	0	182	0.071	47	0.018
DN78946_c0_g3_i1.clean09.fas	4	1872	7488	0	0	35	0.019	4	0.002
DN79323_c0_g1_i2.clean09.fas	1	2217	2217	0	0	0	0	0	0
DN79610_c0_g1_i1.clean09.fas	2	2903	5806	0	0	81	0.028	0	0

Multiple sequence alignment

mkdir Alignment

```
mafft --thread 25 --adjustdirectionaccurately --leavegappyregion --maxiterate 0 --memsavetree --retree 1 SuperContig/c36136_g2_i2_supercontig.fasta > Alignment/c36136_g2_i2.fas

mafft --thread 25 --adjustdirectionaccurately --leavegappyregion --maxiterate 0 --memsavetree --retree 1 SuperContig/c45242_g1_i2_supercontig.fasta > Alignment/c45242_g1_i2.fas

mafft --thread 25 --adjustdirectionaccurately --leavegappyregion --maxiterate 0 --memsavetree --retree 1 SuperContig/c46363_g1_i1_supercontig.fasta > Alignment/c46363_g1_i1.fas

mafft --thread 25 --adjustdirectionaccurately --leavegappyregion --maxiterate 0 --memsavetree --retree 1 SuperContig/c8590_g1_i1_supercontig.fasta > Alignment/c8590_g1_i1.fas

sed -i 's/-c.*_g[0-9*]_i[0-9]*//'   Alignment/*
sed -i 's/_c.*_g[0-9*]_i[0-9]*//'   Alignment/*
```

MSA result


Start Page
c36136_g2_i2 [m] Multiple alignment
Consensus:
Art020
Art025
Art069
Art102
Find:
Seq - / 4
Col - / 7108
Pos - / -
Sel none
Tasks

Cleaning base with low coverage

mkdir CleanSeq

java -jar /home/Programs/phyutility-ver2.2.6/phyutility.jar -clean 0.9 -in Alignment/[GeneName] -out CleanSeq/[GeneName]

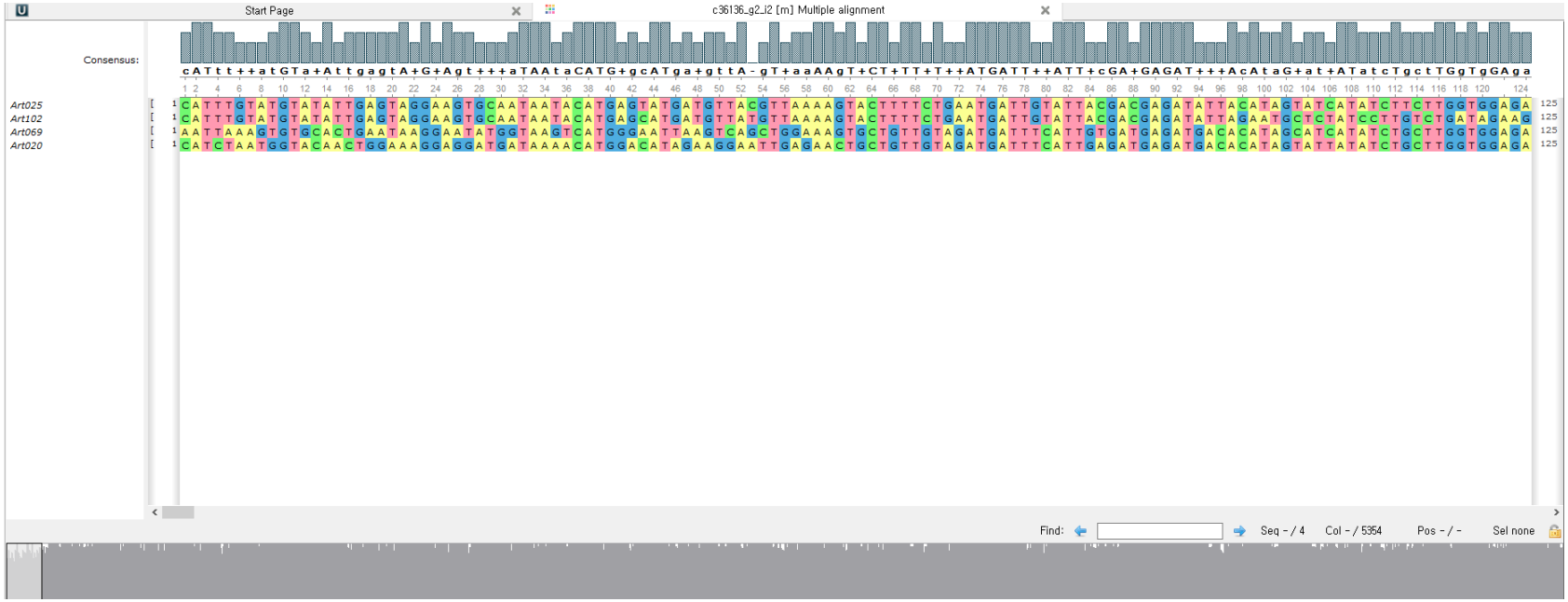
java -jar /home/Programs/phyutility-ver2.2.6/phyutility.jar -clean 0.9 -in Alignment/c36136_g2_i2.fas -out CleanSeq/c36136_g2_i2.fas

java -jar /home/Programs/phyutility-ver2.2.6/phyutility.jar -clean 0.9 -in Alignment/c45242_g1_i2.fas -out CleanSeq/c45242_g1_i2.fas

java -jar /home/Programs/phyutility-ver2.2.6/phyutility.jar -clean 0.9 -in Alignment/c46363_g1_i1.fas -out CleanSeq/c46363_g1_i1.fas

java -jar /home/Programs/phyutility-ver2.2.6/phyutility.jar -clean 0.9 -in Alignment/c8590_g1_i1.fas -out CleanSeq/c8590_g1_i1.fas

Cleaning result



Concatenation

mkdir ConcateSeq

/home/Programs/amas-0.98/amas/AMAS.py concat -i CleanSeq/*fas -f fasta -d dna --concat-out ConcateSeq/ConcatenatedSeq.fa -c 12

mv partitions.txt ConcateSeq/

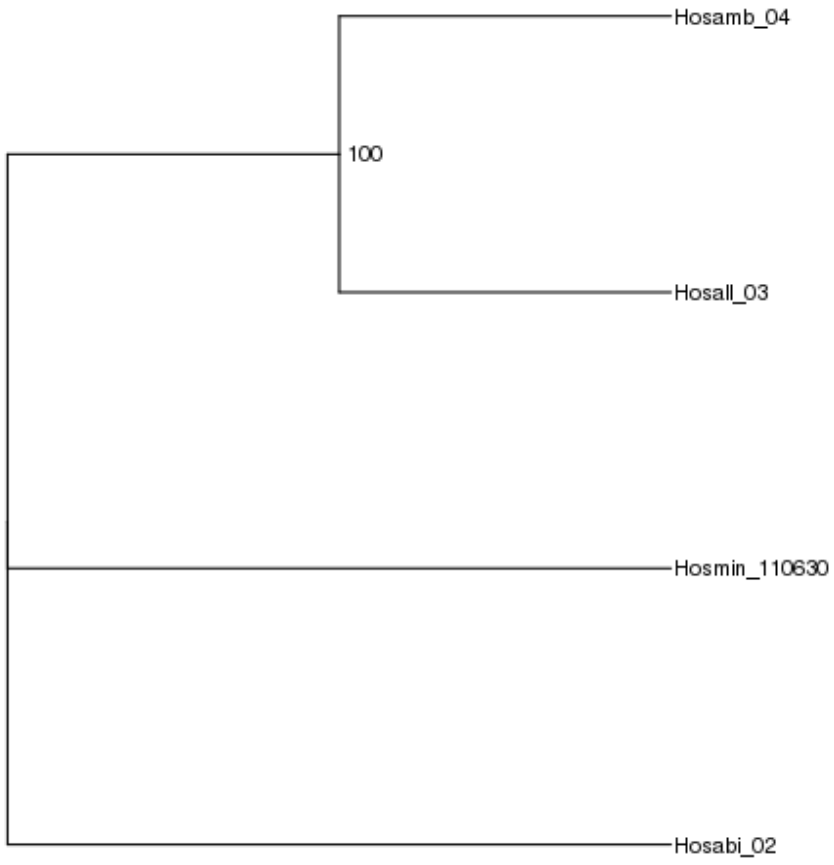
Tree construction

mkdir Tree

iqtree -s ConcateSeq/ConcatenatedSeq.fa -m GTR+G -nt 48 -pre Tree/Artemisia -alrt 1000 -bb 1000 -redo

cp Tree/Artemisia.treefile Tree/Artemisia.FinalTree.nwk
sed -i "s/[0-9][0-9][0-9]₩///gi" Tree/Artemisia.FinalTree.nwk
sed -i "s/[0-9][0-9]₩.[0-9]₩///gi" Tree/Artemisia.FinalTree.nwk
sed -i "s/[0-9][0-9]₩///gi" Tree/Artemisia.FinalTree.nwk
sed -i "s/[0-9]₩.[0-9]₩///gi" Tree/Artemisia.FinalTree.nwk
sed -i "s/[0-9]₩₩///gi" Tree/Artemisia.FinalTree.nwk

Phylogenetic tree



THANK YOU.

BMS