A primer for practical phylogenetic data gathering. Uconn EEB3899-007. Spring 2015 Session 4

# Sequence formats. Alignments



# A road map to phylogenetic data gathering



# Today: alignments



# Composing a sequence list

Never use a Word-type text editor for sequence manipulation, choose instead plain and versatile editors such as those used in programming with constant character width

Our choices: <u>Notepad++</u> (Win)

Wrangler (Mac)

<u>Vim</u> (Linux)



Sequence in fasta format (extension .fas or .fasta)

Name of the sequence after the ">"



There are several sequence and alignment formats (more on this later)

### Composing a sequence list

	😑 seq.f	as a	<b>↓</b>	😑 atp
ſ	1	>Funaria_hygrometrica_2_APRT-2660		182
l	2	TTTAGTTATAACTGTTGAAGCTCCAAAAGTACGGGAATCAGTTTTCATAAATA		183
l	3	TTGAATTGTAATTAAATTGAAAATTGAAAAAAGTTTACTGTTAGTTA		184
l	4	TAGAAAATAAAAGTACTTGGTTATTTTAAATTAAATTTATTATTATTATTGATA		185
l	5	ATATTTATAATTAAACAAAAATAACAAAAAATTGTAGTTCTTAAAAAAAGTCTT		186
l	6	TACTTTCTTTTTTTTTGTTTGGTAACCTTTTAGAATTAAATTTAATTAA		187
l	7	GATTTTTAGATTAATATTTAACAAAAATTATTAAAAAAATAATCTGAGTTGCAT		188
l	8	ACTGTTTTGTTATAAAAAAAATTACTTTGTCTAAAACTAGACAAAATTAAATA		189
l	9	AAAATCTTATTTGTCGAGTAGACCTCATCTTTTCAAGAGTTATTAATTGAGTT		190
	10	CGGAG		191
l				192

To generate the sequence list just copy and paste.

If you need to generate very long lists, you may want to learn how to do it using a shell command such as <u>cat</u>

### Save it as a fasta file

Н.	😑 афъл	
	182	ACTAATACGTCCTATATTTTGAGTTTTAGTTATAACTGTTGAAGCTCCAF
	183	TGAAATTGAAAAAATAATATAAATTTGAATTGTAATTAAATTGAAATTGA
	184	GATCAATTGGTAATTTTAGTTAATTAGAAAAATAAAAGTACTTGGTTATTI
	185	AACATAATTATATGTTATATAAATATATTTATAATTAAACAAAATAACAA
	186	ATAAAAAAAAAGATTTGACTTTATTATACTTTCTTTTTATTGTTTGGT#
	187	TTTAATAATTAGTTTAATATTTAAAAGATTTTTAGATTAATATTTAACA
	188	TGCTGTAGAAAATAATATACAATAATACTGTTTTGTTATAAAAAAAA
	189	CAATTAAGTTTTTTTTTATAAAAAAAAAAAAATCTTATTTGTCGAGTAGACC
	190	TAGGGAGGGATTTATGTCACCACAAACGGAGACTAAGG
đ	191	>Funaria_hygrometrica_1_APRT-1785
	192	TTTGACTAATACGTCCTATATTTTGAGTTTTAGTTATAACTGTTGAAGCI
	193	AAATTGAAATTGAAAAAAATAATATAAATTTGAATTGTAATTAAATTGAA
	194	TTTCGATCAATTGGTAATTTAGTTAATTAGAAAAATAAAAGTACTTGGTI
	195	ататаасатааттататдттатааатататттатааттааасаааат
	196	TTTAATAAAAAAAAGATTTGACTTTATTATACTTTCTTTTTTATTGTTI
	197	TTTATTTAATAATTAGTTTAATATTTAAAAGATTTTTAGATTAATATTT7
	198	CATCAAATGTAGAAAATAATATACAATAATACTGTTTTGTTATAAAAAA
	199	ATATCAATTAAGTTTTTTTTATAAAAAATAAAAAATCTTATTTGTCGAGT#
	200	GTTGTAGGGAGGGATTTATGTCACCACAAACGG
	201	>Funaria_hygrometrica_2_APRT-2660
	202	TTTAGTTATAACTGTTGAAGCTCCAAAAGTACGGGAATCAGTTTTCATA#
	203	TTGAATTGTAATTAAATTGAAAATTGAAAAAAGTTTACTGTTAGTTA
	204	TAGAAAATAAAAGTACTTGGTTATTTTAAATTAAATTTATTATTATTATTG
벽	205	ATATTTATAATTAAACAAAATAACAAAAAATTGTAGTTCTTAAAAAAAGI
	206	TACTTTCTTTTTTATTGTTTGGTAACCTTTTAGAATTAAATTTAATTAA
	207	GATTTTTAGATTAATATTTAACAAAAATTATTAAAAAAATAATCTGAGTTG
A	208	ACTGTTTTGTTATAAAAAAAATTACTTTGTCTAAAACTAGACAAAATTA#
Ë	209	AAAATCTTATTTGTCGAGTAGACCTCATCTTTTCAAGAGTTATTAATTGA
8	210	CGGAG
		182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210

8 211

# Alignment editor: <a href="https://www.editorscore.com">PhyDE</a>

- Intuitive, versatile and visual alignment editor
- Runs on java (system independent)
- Has several interesting plugins
- It's free!

### Sequence names

	Sequence	numes										
Ø			<	atpBr	rbcL_EX	(AMPLE_li	st.fas>					א נ
File E	dit Taxa Search	View Help										
Ð		+ L	abel 🔸	AT CG	Pro Pr tein te	n O Mixed	I	Mod	e. Align	10	~	CSE
^ #	Name	7			5	10	15	20	25	30	35	
1	Entosthodon_atten	uatus_Ireland_APR	T-2281	ТТ	GACT	AATAC	тсс	ТАТАТ	ТТТБА	) T T T T A (	G T T A G	AAC
2	Entosthodon_atten	uatus_USA_CA_AP	RT-2679	TG	АСТА	ATACGI	ссти	ΑΤΑΤΤ	ТТСАСТ	Г Т Т Т <mark>А</mark> G '	TAGA	ACT
3	Entosthodon_bergi	ianus_South_Africa	_APRT_Nic_E	BE CT	ATAT	ТТТБА	ттт.	ТАСТТ	AGAACT	G T T G A /	AGCTC	CAA
4	Entosthodon_bonp	landii_Mexico_API	RT-1085	ТТ	TGAC	TAATA	GTC	ТАТА	ТТТГСА	G T T T T J	AGTTA	GAA
5	Entosthodon_cf_fa	scicularis_APRT-31	57	ТТ	GACT	AATAC	тсс	ТАТАТ	ТТТСА	• T T T T A 🤇	G T T A G	AAC
6	Entosthodon_cf_fa	scicularis_APRT-31	58	ТТ	TGAG	ТТТА 🤇	ТТА	G A A C T	GTTGAA	GCTCC	A A A A <mark>G</mark>	TAC
7	Entosthodon_cf_fa	scicularis_APRT-31	59	AT	TTTG	A G T T T 1	AGT	TAGAA	СТСТТС	AAGCT	CAAA	AGT
8	Funaria_arctica_C	anada_1_APRT-23	25	GG	CTAA	TACGTO	CTA	Т А Т Т Т	Т G 🗛 G Т Т	Г Т Т <mark>А</mark> С Т Г	ΤΑΤΑΑ	C T G
9	Funaria_arctica_C	anada_2_APRT-23	26	ТТ	TGGC	TAATA	GTC	СТАТА	ТТТГСА	G T T T T J	AGTTA	TAA
10	Funaria_arctica_C	anada_3_APRT-23	23	CC	TTAG	тстссе	ттт	этсст	GACATA	AATCC	стссс	TAC
11	Funaria_arctica_U	SA_AK_1_APRT-23	324	ТТ	TGGC	ΤΑΑΤΑΟ	GTC	ТАТА	ТТТГС 🗚	G T T T T Л	AGTTA	TAA
12	Funaria_microston	na_Australia_APRT·	2084	ТТ	GACT	AATAC	тсс	ТАТАТ	ТТТСА	• T T T T A 🤇	G T T A G	AAC
13	Funaria_microston	na_Canada_1_APR	T-2883	TG	GCTA	ATACGI	ССТ	ΑΤΑΤΤ	ТТСАСТ	Г Т Т Т <mark>А</mark> С Г	ГТАТА	ACT
14	Funaria_hygromet	rica_USA_CA_1_AF	PRT-2529	TA	ATAC	GTCCTA	ТАТ	ТТТСА	GTTTTA	GTTAT	AACTG	TTG
15	Funaria_hygromet	rica_Egypt_APRT-2	844	CC	ТТАС	тстссе	ттт	этсст	GACATA	AATCC	стссс	TAC
16	Funaria_microston	na_China_1_APRT-	2680	ТТ	TGGC	TAATAC	GTC		ТТТТСА	GTTTT	AGTTA	TAA
17	Funaria_calvescer	ns_Bolivia_APRT-28	85	TG	GCTA	ATACGI	ССТ	ATATI	ттсАст	TTTAG	TATA	ACT
18	Funaria_flavicans_	USA_NC_1_APRT-	2014	T T	TGGC	TAATAC	GTC		ттттсА	GTTTT	AGTTA	TAA
19	Funaria_flavicans_	USA_NC_2_APRT-	1782	A C	TAAT	ACGTCO	TAT	ΑΤΤΤΤ	AGTTT	ΓΤΑΓΤΤ	ATAAC	төт
20	Funaria_hygromet	rica_1_APRT-1785		T T	TGAC	TAATAC	GTC	СТАТА	TTTGA	GTTTT	AGTTA	TAA
21	Funaria_hygromet	rica_2_APRT-2660		ТТ	TAGT	TATAA	TGT	GAAG	OTCCA A	A A G T A	GGGA	ATC
ve				> <								
	Sal: (voi	d)	NC: 1		tov	a: 21 / char	- 696					· ·
	Sel. (VUI	u)	NC. I		tax		5.000					

This is how a sequence list looks like

Alignment area

Mode button

# Aligner: MUSCLE

Alternatives:

Clustal

Mafft

Tcoffee

Although you can align manually easy sequence lists, it will save you time to use some aligner like MUSCLE

### Multiple Sequence Alignment

MUSCLE stands for **MU**Itiple Sequence Comparison by Log- Expectation. MUSCLE is claimed to achieve both better average accuracy and better speed than <u>ClustalW2</u> or <u>T-Coffee</u>, depending on the chosen options.



### Sequence list vs. Alignment: an important reminder



(Assuming there were no sequencing errors) these are plain, objective data

#	Name	80	)			85				90	)			9	5			10	0			10	5		1	10	)			11	5		
1	Entosthodon_cf_fascicularis_API	ТТ	С			-				Å	Т	A	A	A	ΤА	A	T	ТΪ	G	-		-	-		-	-	-	-			A	A	т
2	Entosthodon_bonplandii_Mexico	ΤТ	С			-				A	Т	A	A	A	ΤА	A	Т	ΤТ	G				-		-	-		-		-	A	А	т
3	Entosthodon_bergianus_South_#	ΤТ	С			-				A	Т	A	A	A	ΤА	A	Т	ΤТ	G				-		-	-		-		-	A	A	т
4	Entosthodon_attenuatus_Ireland	ΤТ	С			-				A	Т	A	A	A	ΤА	A	Т	ΤТ	G				-		-	-		-		-	A	A	т
5	Entosthodon_attenuatus_USA_C.	ΤТ	С							A	Т	A	A	A	ΤА	A	Т	ΤТ	G				-		-	-		-		-	A	A	т
6	Entosthodon_cf_fascicularis_API	ΤТ	С			-				A	Т	A	A	A	ΤА	A	A	ΤТ	G				-		-	-		-		-	A	A	т
7	Entosthodon_cf_fascicularis_API	ΤТ	С							A	Т	A	A	A	ΤА	A	А	ΤТ	G				-		-	-		-			A	A	т
8	Funaria_microstoma_China_1_A	ΤТ	С			-				A	Т	A	A.	A	ΤА	A	A	ΤТ	G	А	A A	Т	т	э.	A	A	A	A,	A A	١T	А	A	т
9	Funaria_flavicans_USA_NC_2_A	ΤТ	С			-				A	Т	А	A.	A	ΤА	A	A	ΤТ	G	А	A A	Т	т	3 -	A	A	A	A,	A A	١T	А	A	т
10	Funaria_microstoma_Australia_A	ΤТ	С			-				A	Т	А	A	A	ΤА	A	A	τт	G	А	A A	Т	т	÷ -	A	A	А	A	A A	١T	А	A	т
11	Funaria_flavicans_USA_NC_1_A	ΤТ	С			-				A	Т	А	A.	A	ΤА	A	A	ТΤ	G	А	A A	Т	т	3 -	A	A	A	A,	A A	١T	А	A	т
12	Funaria_hygrometrica_Egypt_AF	ΤТ	С							A	Т	A	A.	A	ΤА	A	A	ТΤ	G	А	A A	Т	т	Э A	۸A	A	A	A,	A A	١T	А	A	т
13	Funaria_calvescens_Bolivia_API	ΤТ	С	A	ΓA	A	A	т	ΑA	A	Т	А	A.	A	ΤА	A	A	ΤТ	G	А	A A	Т	т	э.	A	A	A	A,	A A	١T	А	A	т
14	Funaria_arctica_Canada_3_APF	ΤТ	С							A	Т	А	A.	A	ΤА	A	A	ΤТ	G	А	A A	Т	т	3 -	-	A	A	A,	A A	١T	А	A	т
15	Funaria_microstoma_Canada_1_	ΤТ	С							A	Т	A	A.	A	ΤА	A	A	ΤТ	G	А	A A	Т	т	э.	A	A	A	A,	A A	١T	А	A	т
16	Funaria_arctica_Canada_2_APF	ΤТ	С							A	Т	A	A.	A	ΤА	A	A	ТΤ	G	А	A A	Т	т	э.	-	A	A	A,	A A	١T	A	A	т
17	Funaria_arctica_USA_AK_1_AP	ΤТ	С							A	Т	A	A.	A	ΤА	A	A	ТΤ	G	А	A A	Т	т	э.	A	A	A	A,	A A	١T	A	A	т
18	Funaria_hygrometrica_1_APRT-	ТТ	С							A	Т	А	A.	A	ΤA	A	A	ТΤ	G	А	A A	Т	Т	э.	A	A	A	A,	A A	N T	A	A	т
19	Funaria_hygrometrica_USA_CA_	ТТ	С							A	Т	А	A.	A	ΤA	A	A	ТΤ	G	A	A A	Т	т	э.	A	A	A	A,	A A	١T	A	A	т
20	Funaria_hygrometrica_2_APRT-:	ТТ	С			-				A	Т	А	A.	A	ΤA	A	А	ТΤ	G	А	A A	Т	Т	э.	A	А	А	A	A A	١T	A	А	Т

However, an alignment is more tan plain data, it is also a statement on nucleotide homology: you are assuming (either if you are right or not) that each position (column) contains homologous nucleotides, and this will become VERY relevant in phylogenetic analysis. Be aware of it! This is why the manual adjustment or refining of the alignment is important

# Alignment elements: substitutions





Transversion

Transition

# Alignment elements: gaps

Gaps caused by INsertions or DELetions (indels)



Gaps are often ignored in phylogenetic analyses, but they can be evolutionary informative You can easily add an indel block to you alignment using the software <u>SeqState</u>

Read about different ways to code indels in: Simmons & Ochoterena 2000 Gaps as Characters in Sequence-Based Phylogenetic Analyses. *Syst. Bio.* 49: 369-381

If you want to acknowledge this information, you can code indels as binary characters in an adjacent block

### Alignment elements: inversions

# A T C T G A T C A A T G T A A A G A A A A G A A A A G A A A A G A A A A G A A A A A G A A A A G A A A A G A A A A G A A A A G A A A A G T A A A A A G A A A A G A A A A G A A A A G A A A A G A A A A G A A A A G A A A A G A A A A G A A A A

This is actually an inversion, not 8 substitutions



Exclude or code these features, but do not keep them in their raw state or they will overestimate the divergence of these two sequences

### Alignment elements: repeated elements

Some repeated features, such as microsatellites, may be of interest for phylogenetic analysis too, but confirm them, if necessary, with your raw data

A	С	A	A	Т	A	-	-	-	-	-	-	-	-	A	Т	A	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	С	A	А	Т	A	-	-	-	-	-	-	-	-	А	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	С	A	A	Т	A	-	-	-	-	-	-	-	-	А	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	А	Т
A	A	Т	A	A	Т	-	-	-	-	-	-	-	-	А	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	А	Т
A	A	Т	A	A	Т	-	-	-	-	-	-	-	-	A	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	A	Т	A	A	Т	-	-	-	-	-	-	-	-	A	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	A	Т	A	A	Т	-	-	-	-	-	-	-	-	А	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	А	Т
A	A	Т	A	A	Т	А	Т	A	Т	A	Т	A	Т	A	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	A	Т	A	A	Т	-	-	-	-	-	-	-	-	A	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	A	Т	A	A	Т	А	Т	A	Т	A	Т	A	Т	A	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	A	Т	A	A	Т	-	-	-	-	-	-	-	-	A	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	A	Т	A	A	Т	-	-	-	-	A	Т	A	Т	A	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	A	Т	A	A	Т	-	-	-	-	-	-	-	-	A	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	A	Т	A	A	Т	-	-	-	-	-	-	-	-	A	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	A	Т	A	A	Т	-	-	-	-	-	-	-	-	A	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	A	Т	A	A	Т	-	-	-	-	A	Т	A	Т	A	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	С	A	A	Т	A	-	-	-	-	-	-	-	-	A	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	С	A	A	Т	A	-	-	-	-	-	-	-	-	A	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	С	A	A	Т	A	-	-	-	-	-	-	-	-	A	Т	А	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т
A	С	A	A	Т	A	-	-	-	-	-	-	A	Т	A	Т	A	С	Т	G	Т	Т	Т	Т	G	Т	Т	A	Т

# Alignment elements: indels in coding regions

Be always suspicious about shifts in reading frames or very divergent sequences. You may need to check again your raw chromatograms or even re-sequence that sample

Т	Т	С	A	A	G	A	G	Т	Т	А	Т	Т	A	A	Т	Т	-	G	A	G	Т	Т	Т	Т	А	G	G	G	A	G	G	G
Т	Т	С	A	A	G	A	G	Т	Т	A	Т	Т	A	A	Т	Т	-	G	A	G	Т	Т	G	Т	A	G	G	G	A	G	G	G
Т	Т	С	A	A	G	A	G	Т	Т	A	Т	Т	A	A	Т	Т	-	G	A	G	Т	Т	G	Т	A	G	G	G	А	G	G	G
Т	Т	С	A	A	G	A	G	Т	Т	A	Т	Т	A	A	Т	Т	-	G	A	G	Т	Т	G	Т	G	G	G	G	A	G	G	G
Т	Т	С	A	A	G	A	G	Т	Т	А	Т	Т	A	A	Т	Т	-	G	A	G	Т	Т	G	Т	A	G	G	G	A	G	G	G
Т	Т	С	A	A	G	A	G	Т	Т	А	Т	Т	А	A	Т	Т	-	G	A	G	Т	Т	G	Т	A	G	G	G	А	G	G	G
Т	Т	Т	A	A	Т	A	G	Т	Т	А	Т	Т	А	A	Т	Т	Т	G	A	G	Т	Т	G	Т	Т	G	G	G	A	G	G	G
Т	Т	С	А	А	G	A	G	Т	Т	А	Т	Т	А	A	Т	Т	-	G	А	G	Т	Т	G	Т	А	G	G	G	A	G	G	G
Т	Т	С	A	А	G	A	G	Т	Т	А	Т	Т	А	A	Т	Т	-	G	А	G	Т	Т	G	Т	А	G	G	G	А	G	G	G
Т	Т	С	A	А	G	A	G	Т	Т	А	Т	Т	A	A	Т	Т	-	G	А	G	Т	Т	G	Т	A	G	G	G	A	G	G	G
Т	Т	С	A	А	G	A	G	Т	Т	А	Т	Т	A	A	Т	Т	-	G	A	G	Т	Т	G	Т	A	G	G	G	А	G	G	G
Т	Т	С	A	A	G	A	G	Т	Т	A	Т	Т	A	A	Т	Т	-	G	A	G	Т	Т	G	Т	A	G	G	G	A	G	G	G
Т	Т	С	A	A	G	A	G	Т	Т	A	Т	Т	A	А	Т	Т	-	G	A	G	Т	Т	G	Т	А	G	G	G	А	G	G	G
Т	Т	С	A	A	G	A	G	Т	Т	A	Т	Т	А	A	Т	Т	-	G	A	G	Т	Т	G	Т	А	G	G	G	А	G	G	G
Т	Т	С	А	А	G	А	G	Т	Т	А	Т	Т	А	А	Т	Т	-	G	А	G	Т	Т	G	Т	А	G	G	G	А	G	G	G

However, 3x nucleotide indels in coding regions are indeed a possibility (they would respect the reading frame)

А	A	Т	С	G	А	С	A	Т	С	Т	A	А	Т	A	A
А	A	Т	С	G	A	С	A	Т	С	A	A	A	Т	А	A
А	A	Т	С	G	A	С	A	Т	С	A	A	A	Т	А	A
А	A	Т	С	G	A	С	A	Т	С	A	A	A	Т	А	A
А	A	Т	С	G	A	С	A	Т	С	A	A	A	Т	А	A
А	A	Т	С	G	A	С	A	Т	С	Т	A	A	Т	А	A
А	A	Т	С	G	A	С	A	Т	С	A	A	A	Т	А	A
А	A	Т	С	G	A	С	A	Т	С	A	А	A	Т	А	A
А	A	Т	С	G	A	С	A	Т	С	A	А	A	Т	А	A
А	A	Т	С	A	G	G	A	G	С	-	-	-	Т	А	A
А	A	Т	С	A	G	G	A	G	С	-	-	-	Т	А	A
А	A	Т	С	A	G	G	A	G	С	-	-	-	Т	А	A
А	A	Т	С	A	G	С	A	G	С	-	-	-	Т	А	A
А	A	Т	С	A	G	С	A	G	С	-	-	-	Т	А	A
А	A	Т	С	A	G	С	A	G	С	-	-	-	Т	А	A
А	A	Т	С	A	G	С	A	G	С	-	-	-	Т	А	A
А	A	Т	С	A	G	С	A	G	С	-	-	-	Т	А	A
А	A	Т	С	A	G	С	A	G	С	-	-	-	Т	А	A
А	A	Т	С	A	G	С	A	G	С	-	-	-	Т	А	A
А	A	Т	С	A	G	С	A	G	С	-	-	-	Т	A	A
	A A A A A A A A A A A A A A A A A A A	<ul> <li>A</li> <li>A&lt;</li></ul>	A     A     T       A     A	A       A       T       C         A       <	A         T         C         G           A         T         C         G           A         T         C         G           A         T         C         G           A         T         C         G           A         T         C         G           A         A         T         C         G           A         A         T         C         G           A         A         T         C         G           A         A         T         C         G           A         A         T         C         G           A         A         T         C         G           A         A         T         C         A           A         A         T         C         A           A         A         T         C         A           A         A         T         C         A           A         A         T         C         A           A         A         T         C         A           A         A         T         C         A	A       T       C       G       A         A       T       C       G       A         A       T       C       G       A         A       T       C       G       A         A       T       C       G       A         A       T       C       G       A         A       T       C       G       A         A       T       C       G       A         A       T       C       G       A         A       T       C       A       G         A       T       C       A       G         A       T       C       A       G         A       T       C       A       G         A       T       C       A       G         A       T       C       A       G         A       T       C       A       G         A       T       C       A       G         A       T       C       A       G         A       T       C       A       G         A       T       C <td>AATCGACAATCGACAATCGACAATCGACAATCGACAATCGACAATCGACAATCGACAATCAGGAATCAGGAATCAGGAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAG<td>AATCGACAATCGACAAATCGACAAATCGACAAATCGACAAATCGACAAATCGACAAATCGACAAATCAGAAAATCAGAAAATCAGAAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCA&lt;</td><td>AATCGACATAATCGACATAATCGACATAATCGACATAATCGACATAATCGACATAATCGACATAATCGACATAATCAGGAGAATCAGGAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCA</td><td>AATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCAGGAGCAATCAGGAGCAATCAGGAGCAATCAGCAGCAATCAGCAGCAATCAGCAGCAATCAGCAGCAATCAGCAGCAATCAGCAGCAATCAGC<t< td=""><td>AATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCAGGATCAAATCAGGATCAAATCAGGAGGAAATCAGGAGGAAATCAGGAGGAAATCAGGAGGAAATCAGGAGGAAATCAGCAGGAG&lt;</td><td>AATCGACATCAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCAACATCAAAATCACACATCAAATCAGCACACAAAATCAGGAGCACAAAATCAGGAGCACAAAAATCAGGAGGAGCAAAA<td>AATCGATCTAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCACATCAATAAATCAGCATCAAAAAAATCAGGAGCAAAAAAATCAGGAGGCAAAAAAATCAGGAGGCAA</td></td></t<><td>AATCGACATAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAAAAAATCGACATCAAAAAAAATCGACATCAAAAAAAAATCGACATCAA<td< td=""><td>AATCGACATCAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAATAATAAATCGACATCAATAATAATAATAATAATAATAATAAATAAATAAAAAAAAAAAAAAAAA<td< td=""></td<></td></td<></td></td></td>	AATCGACAATCGACAATCGACAATCGACAATCGACAATCGACAATCGACAATCGACAATCAGGAATCAGGAATCAGGAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAGCAATCAG <td>AATCGACAATCGACAAATCGACAAATCGACAAATCGACAAATCGACAAATCGACAAATCGACAAATCAGAAAATCAGAAAATCAGAAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCA&lt;</td> <td>AATCGACATAATCGACATAATCGACATAATCGACATAATCGACATAATCGACATAATCGACATAATCGACATAATCAGGAGAATCAGGAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCA</td> <td>AATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCAGGAGCAATCAGGAGCAATCAGGAGCAATCAGCAGCAATCAGCAGCAATCAGCAGCAATCAGCAGCAATCAGCAGCAATCAGCAGCAATCAGC<t< td=""><td>AATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCAGGATCAAATCAGGATCAAATCAGGAGGAAATCAGGAGGAAATCAGGAGGAAATCAGGAGGAAATCAGGAGGAAATCAGCAGGAG&lt;</td><td>AATCGACATCAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCAACATCAAAATCACACATCAAATCAGCACACAAAATCAGGAGCACAAAATCAGGAGCACAAAAATCAGGAGGAGCAAAA<td>AATCGATCTAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCACATCAATAAATCAGCATCAAAAAAATCAGGAGCAAAAAAATCAGGAGGCAAAAAAATCAGGAGGCAA</td></td></t<><td>AATCGACATAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAAAAAATCGACATCAAAAAAAATCGACATCAAAAAAAAATCGACATCAA<td< td=""><td>AATCGACATCAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAATAATAAATCGACATCAATAATAATAATAATAATAATAATAAATAAATAAAAAAAAAAAAAAAAA<td< td=""></td<></td></td<></td></td>	AATCGACAATCGACAAATCGACAAATCGACAAATCGACAAATCGACAAATCGACAAATCGACAAATCAGAAAATCAGAAAATCAGAAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCAAATCAGCA<	AATCGACATAATCGACATAATCGACATAATCGACATAATCGACATAATCGACATAATCGACATAATCGACATAATCAGGAGAATCAGGAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCAGCAGAATCA	AATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCGACATCAATCAGGAGCAATCAGGAGCAATCAGGAGCAATCAGCAGCAATCAGCAGCAATCAGCAGCAATCAGCAGCAATCAGCAGCAATCAGCAGCAATCAGC <t< td=""><td>AATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCAGGATCAAATCAGGATCAAATCAGGAGGAAATCAGGAGGAAATCAGGAGGAAATCAGGAGGAAATCAGGAGGAAATCAGCAGGAG&lt;</td><td>AATCGACATCAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCAACATCAAAATCACACATCAAATCAGCACACAAAATCAGGAGCACAAAATCAGGAGCACAAAAATCAGGAGGAGCAAAA<td>AATCGATCTAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCACATCAATAAATCAGCATCAAAAAAATCAGGAGCAAAAAAATCAGGAGGCAAAAAAATCAGGAGGCAA</td></td></t<> <td>AATCGACATAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAAAAAATCGACATCAAAAAAAATCGACATCAAAAAAAAATCGACATCAA<td< td=""><td>AATCGACATCAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAATAATAAATCGACATCAATAATAATAATAATAATAATAATAAATAAATAAAAAAAAAAAAAAAAA<td< td=""></td<></td></td<></td>	AATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCGACATCAAATCAGGATCAAATCAGGATCAAATCAGGAGGAAATCAGGAGGAAATCAGGAGGAAATCAGGAGGAAATCAGGAGGAAATCAGCAGGAG<	AATCGACATCAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCAACATCAAAATCACACATCAAATCAGCACACAAAATCAGGAGCACAAAATCAGGAGCACAAAAATCAGGAGGAGCAAAA <td>AATCGATCTAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCACATCAATAAATCAGCATCAAAAAAATCAGGAGCAAAAAAATCAGGAGGCAAAAAAATCAGGAGGCAA</td>	AATCGATCTAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCGACATCAAAATCACATCAATAAATCAGCATCAAAAAAATCAGGAGCAAAAAAATCAGGAGGCAAAAAAATCAGGAGGCAA	AATCGACATAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAATAATCGACATCAAAAAAATCGACATCAAAAAAAATCGACATCAAAAAAAAATCGACATCAA <td< td=""><td>AATCGACATCAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAATAATAAATCGACATCAATAATAATAATAATAATAATAATAAATAAATAAAAAAAAAAAAAAAAA<td< td=""></td<></td></td<>	AATCGACATCAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAAATAAATCGACATCAATAATAAATCGACATCAATAATAATAATAATAATAATAATAAATAAATAAAAAAAAAAAAAAAAA <td< td=""></td<>

# Alignment elements: ambiguity

You cannot get rid of the ambiguity in many alignments. The best you can do is to exclude the areas of the alignment that are not trustworthy, either manually or using softwares such as <u>Gblocks</u>

It is always better to ignore dubious data than to introduce noise in your analysis. If you are not sure about the homology: exclude



### **Gblocks Server** Paste an alignment in NBRF/PIR or FASTA format: Or upload an alignment file: Less stringent Choose File No file chosen options allow Type of sequence: DNA 🔍 || Protein 💿 || Codons 🔾 more flexibility Options for a less stringent selection Allow smaller final blocks Allow gap positions within the final b You can increase Allow less strict flanking positions the stringency to Options for a more stringent selection: get a more Do not allow many contiguous nonconserved positions conservative Get Blocks Clear alignment

"Exclude" means that an area of the alignment will not be analyzed. <u>Never erase manually</u> any nucleotide, unless you confirm it with the raw chromatogram

### Alignment elements: ambiguity

Flank positions of the 10 selected block(s) Flanks: [10 81] [90 107] [109 193] [195 248] [250 341] [356 465] [470 481] [489 546] [555 719] [721 727] List of blocks to be kept

### Visualization



### **Parameter summary**

### Parameters used

Minimum Number Of Sequences For A Conserved Position: 11 Minimum Number Of Sequences For A Flanking Position: 11 Maximum Number Of Contiguous Nonconserved Positions: 8 Minimum Length Of A Block: 5 Allowed Gap Positions: With Half

You can download the corrected alignment without the exclusion blocks, however it is better to keep the sequences unedited. You will be able to exclude parts of the alignment before the analysis, just keep a record of the exclusion blocks.

## Sequence and alignment formats

### FASTA format

Extension: .fas or .fasta

Poor in information

Each sequence starts with ">"

Sequence name

📄 atp B	BrbcL_EXAMPLE_MUSCLE2.fas 🖾
1	>Intosthodon_cf_fascicularis_APRT-3157
2	TTGACTAATACGTCCTATATTTTGAGTTTTAGTTAGAACTGTTGAAGCTCCA
3	AAAGTACGGGAATCATTTTTCATAAATAATTTGAATAT
4	AGATTTGAATTGTAATTAAATCGAAATTGAAAAAAGTTTAATGTTAGTTA
5	GATCAATTAATAATTTTTAGTTAATGAGAAAATAAAAGTATTTGGTTATTTTAAA
6	CTTATTAT-TATATTGATATATAA
7	ATAAAATAACAAAAAATTCTAGTTGTTAAATTAATTATAAAA
8	AAGGTCTAGTTTTATTATACTTTTTTTTTTTTTTTGGTAACCTTTTATA-TTAGTTTT
9	CTTTATTATTAATTAGTTTAACATTCAAAATATTTTTAAAACAATGCTT
10	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
11	ACAATAATACTGTTTTGTTAT-AAAAAAATTACTTTGTCTAAAACTAGACAA
12	AATTAAATACCAATTAAGTTTTTTTAGTTTTATAAAAAAATCTTATTTGTCGAGCAGACC
13	TCATCTTTTCAAGAGTTATTAATTGAGTTTTAGGGAGGGA
14	NAN WANNANANANANANANANANANANANANANANANAN
15	
16	>Intosthodon bonplandii Mexico APRT-1085
17	TTTGACTAATACGTCCTATATTTTGAGTTTTAGTTAGAATTGTTGAAGCTCCA
18	AAAGTACGGGAATTATTTTTCATAAATAATTTGAATAT
19	AGATTTGAATTGTAATTAAATCGAAATTGAAAAAAGTTTAATGTTAGTTA
20	GATCAATTAATAATTTTTAGTTAATGAGAAAATAAAAGTATTTGGTTATTTTAAA
21	TTTATTAT-TATATTGATATAT-TATATTGATATATATAT
22	ATAAAATAACAAAAATTGTAGTTATTAAAAAAAGGTTTGAATTATAAAAAA
23	AAGGTCTAAGTTTATTATACTTT-TTTTTTTTTTTGGTAACCTTTTATA-TTAGTTTT
24	CTTTTTATTAATTAATTTAACATTCAAAAACATTTTTAAAACAATACTT
25	AAAAAAATAAAAAAATTATTAAAAAATAATCTGAGTTGCATCAAATGTAGAAAATAATAT
26	ACAATAATACTGTTTTGTTAT-AAAAAAATTATTTTGTCTAAAACTAGACAA
27	AATTCAATACCAATTAAGTTTTTTAGTTTTATAAAAAAATCTTATTTGTCGAGCAGACC
28	TCATCTTTTCAAGAGTTATTAATTGAGTTTTAGGGAGGGA
29	AGACTAAAGNNNNNNNNNNNNNNNNNNNNN
30	
31	>Intosthodon_bergianus_South_Africa_APRT_Nic_BER2
32	CTATATTTTGAGTTTTAGTTAGAACTGTTGAAGCTCCA

# Sequence and alignment formats

### **NEXUS** format

Extension: .nex

The whole file starts with #NEXUS

They are structured in blocks of information, each of them starting with "BEGIN" and ending with "END;" (Don't forget the ";"s)

Alignments are usual referred as "MATRIX". You can encode indels, exclude areas, include specific information for programs, etc

### Dimensions of the alignment are stated in advance #NEXUS This describes which symbols are BEGIN DATA: allowed DIMENSIONS NTAX=217 NCHAR=2789 FORMAT DATATYPE=DNA GAP=- MISSING=? MATRIX Pyramidula tetragona Morocco 1958 TTTGAGTAATACGT-CCTACAT----TTT Goniomitrium seroi Spain 2018 TTTGACTAATACGT-CCTACAT----TTT Goniomitrium acuminatum Australia 1 2017 TTTGACTAATACGT-CCTACAT-----Goniomitrium acuminatum Australia 2 2810 Goniomitrium acuminatum subsp Enerve Australia 2843 TTTGACTAATACGT-CCTACAT----TTT Funaria flavicans USA NC 1 2014 TTTGGCTAATACGT-CCTATATTTTGAGTTT Funaria flavicans USA NC 2 1782 TTTGACTAATACGT-CCTATATTTTGAGTTT Funaria hygrometrica 1 1785 TTTGACTAATACGT-CCTATATTTTGAGTTT Funaria hygrometrica 2 2660 Funaria hygrometrica Canada 583 Funaria hygrometrica Chile 1781 ???????????T-CCTATATTTTGA Funaria hygrometrica China 2 2622 Funaria hygrometrica Egypt 2844 TTTGACTAATACGT-CCTATATTTTGA Funaria hygrometrica Germany 2658 TTTGGCTAATACGT-CCTATATTTTGAGTTT Funaria hygrometrica Japan 2981 2222222222222222222222222222AGTTT Physcomitrium subsphaericum Mexico 1 2331 TTTGACTAATACGT-CCTATATTTTGAGTTT Physcomitrium subsphaericum Mexico 2 2893 ?????TAATACGT-CCTATATTTTGAGTTT

Fnd of matrix

# Sequence and alignment formats

### **PHYLIP** format

Extensions: .phy .phylipi

This format usually has a lot of restrictions (like a limit in the sequence name length), but some softwares only read PHYLIP

### **FASTAQ** format

Complex, data-rich format specific for High-Throughput Sequencing. It contains quality scores for each base



APRT2751	TTTGACTAATACGTCCTATATTTTGAGTTTTAGTTAGAACTGTTGAAGCT
APRT2362	?????????????????TATTTTGAGTTTTAGTTAGAACTGTTGAAGCT
APRT2342	TTTGACTAATACGTCCTATATTTTGAGTTTTAGTTAGAACTGTTGAAGCT
APRT1476	TTTGACTAATACGTCCTATATTTTGAGTTTTAGTTAGAACTGTTGAAGCT
APRT2991	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
APRT2992	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
APRT2993	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
APRT2994	?????TAATACGTCCTATATTTTGAGTTTTAGTTAGAACTGTTGAAGCT

@M00704:20:000000000-AAUNM:1:1101:15411:1391 1:N:0:42 TCTTTGAAATTCTCAGCCTCTCCGGCAACTCCATCCACTACAATACTCCTCGCAATTCCTTCT @M00704:20:000000000-AAUNM:1:1101:15531:1433 1:N:0:42 TCTATGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGTGATTTGTCTGGTTAATTCCGTTA @M00704:20:000000000-AAUNM:1:1101:17773:1531 1:N:0:42 @M00704:20:000000000-AAUNM:1:1101:14962:1595 1:N:0:42 GTATTAGCGGTGAACAGACCACTTCTACGAAGTAGTTAAGCGGGATTTAGAACTGCTCCATGGC @M00704:20:000000000-AAUNM:1:1101:14938:1649 1:N:0:42

### Format conversion

- PhyDE allows to export an alignment in fasta and nexus format
- Geneious provides a wider spectrum of formats and it is a good resource if you need a phylip file
- Mesquite is also a free, java-based program that you may use for a lot of purposes, including format conversion
- Alternatively, you can use the <u>format converter</u> of HIV database