Training course Molecular Epidemiology Leishmaniasis

# MANUAL *IN-SILICO* PROCEDURES



# 18 - 24 May 2009

Instituto Oswaldo Cruz Rio de Janeiro Brazil

# Table of contents

C. Computer based analyses

## 5. In silico sequence analyses

Protocol 5.1	Sequence retrieval from public domain data bases	p. 3
Protocol 5.2	Analysis of sequence chromatograms	p. 14
Protocol 5.3	Sequence alignments, primer design and in-silico RFLP	p. 18
Protocol 5.4	Phylogeny	p. 27

# 6. Multilocus sequence typing (MLST)

Protocol 6.1	Analysis of MLST data	p. 33
--------------	-----------------------	-------

#### 7. Multilocus microsatellite typing (MLMT)

Protocol 7.1	Population genetic analysis of Leishmania strains based on	p. 40
	MLMT data	p. 40

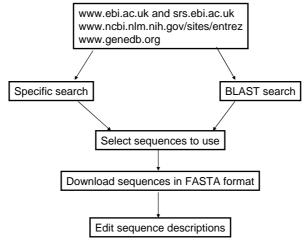
# C. COMPUTER-BASED ANALYSES

5. In-silico se	equence analyses
Protocol 5.1	Sequence retrieval from public domain data bases
Purpose	Retrieval of specific sequences for studying variability of <i>Leishmania</i> and for designing new assays and primers

#### A. Introduction:

In order to make optimal use of sequences in the public domain, retrieval of specific sequences is of great importance to study variability and to design new assays and primers. This can be done using similarity searches, or using sequence descriptions. As the workshop will deal with HSP70 RFLP, available sequences will be downloaded and used in further exercises. Also the ribosomal ITS1 region is used.

#### General workflow:



One way of collecting sequences from public domain data bases is by using searches with specific key words defined by the user.

**Advantage**: You will only retrieve sequences from which the submitter has indicated the name or function. In general this means that the submitter is quite sure about these, and you will in general collect only what you want. It is no guarantee for an error-free sequence.

**Disadvantage**: Your search results depend largely on the key words used. If not properly chosen, you will miss certain sequences. In addition, if the submitter made a typographical mistake during submission, a matching key word can be missed.

Method A: Retrieval of hsp70 nucleotide sequences from EBI (joint exercise)

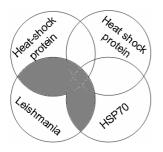
- 1. Open srs.ebi.ac.uk
- 2. Click the library page tab
- 3. Among nucleotide sequence databases, select EMBL
- 4. Click the query form tab
- 5. In the upper search box, select organism name from the pull-down menu
- 6. Type Leishmania in the search box
- 7. Click search

🖉 Standard Query Form - Windows Inl	ternet Explorer					
GO - 🕼 http://srs.ebi.ac.uk/srsb	oin/cgi-bin/wgetz					
Eile Edit View Favorites Iools E						
Links 🔚 ITG Intranet 👩 ITG web 📩 I						
Google	🛨 🚼 Zoeken 🔹 🕫 🍏	• 🕂 🖗 • 🖒	Bladwijzers *	Spelling control	eren 🔹 🎦 Autor	natisch aan
😪 🏟 😬 🗸 S Nucleotide - Sequer	nce 9 fro 🚯 Standard Query Form	×				
EMBL-EBI	tabases 🔄 Enter Text He	re	Go	Reset ⑦ Advanced Search	Give us feedback	
Databases Tools EBI G	Groups Training Industry	About Us	Help		Index <u>ನ</u> 🎒	
Quick Search Library Page	uery Form Tools	Results	Projects	Views	Databanks	HELP
Reset search EMBL						
	Fields you can search	Nounc	earch terms			
Search Options					III Searc	
Combine search terms	In a single field, you can sep	arate multiple va	ilues by: &,   or	ļ	Searc	n
with: & (AND)	Organism Name	<ul> <li>Leishmania</li> </ul>				
Use wildcards 🔽	1 AllText	•				
Get results of type:	1 AllText	•				
Entry	1 AllText	•				
Result Display Options	Create a view					
G. 17. 18. 1	Select the fields you want di	splayed in your v	iew and choose	the format		
View results using:	Channel an anna Ealdar					
EMBLSeqSimpleView  or	Choose 1 or more fields:	Dis	play As: 💿 Tal	ble O Lis	st	
or	ID Topology	1				
C Create a view	Molecule	Se	quence Format:	embl 💌	·]	
Show 30 💌	Data Class Division				-	
results per page	Sequence Length Accession Number	1				
	Accession withber	1				
Tips					III Searc	h
- i i i	ou/dog.moot.form/[0]./@uuuu0ogotio					

- 8. Click the query form tab
- 9. In the upper search box, select all text from the pull-down menu
- 10. Type HSP70 in the search box
- 11. Click search
- 12. Click the query form tab
- 13. Using the same technique, search for "heat-shock protein" (use the quotes)
- 14. Go to the query form
- Repeat the previous search, but now enter in the second box "heat shock protein" (no dash)
- 16. Repeat this search, but now select Combine search terms with **OR** from the left pulldown menu

🖉 Standard Query Form - Windows Internet Explorer	
🚱 🕞 🔻 🕼 http://srs.ebi.ac.uk/srsbin/kgi-bin/wgetz	🛨 🔸 🗙 Google
Elle Edit Yew Favorites Iools Help	
😪 🎄 🔡 🗸 🗧 Nucleotide - Sequence 9 fro 🚯 Standard Query Form 🗙 💋 GeneDB	🚹 • 🗟 • 🖶
EMBL-EBI	
Databases Tools EBI Groups Training Industry About Us Help Site Index 🕥 🏯	
Quick Search Library Page Query Form Tools Results Projects Views Databanks	
Reset search EMBL	
Search Options	* *
No entries found - please try aga	in. (// 🗋 🍾)
Combine search terms No entries found - please try again	in. 🥼 🚺
Combine search terms No entries found - please try again with: I (OR)	in.
with: I (OR)	in.
with: I (OR)  Use wildcards  Fields you can search Your search terms	
with: I (OR)  Viewildcards  Fields you can search Your search terms	
with: I (OR)  Viewildcards  Fields you can search Your search terms	
with: I (OR)  Use wildcards  Fields you can search Your search terms	
with: I (OR)  Viewildcards  Fields you can search Your search terms	
with: I (OR)  Viewildcards  Fields you can search Your search terms	
with:     [IOR]       Use wildcards     Image: Search for the search	
with: I (OR)  Viewildcards  Fields you can search Your search terms	
with:       [IOR]       Image: Constraint of the state of th	
with:     [IOR]       Use wildcards     Image: Second sec	
with:       [IOR]       Image: Constraint of the second se	
with:       [IOR]         Use wildcards       Image: Comparison of the state of the st	

18. Combine the previous searches to obtain the grey zone below: Q1 & (Q2 | Q4)



- 19. Select the entries you want to use.
- 20. Press the Save button
- 21. Select output of All to File using the FastaSeqs format
- 22. Press Save and choose a location and file name
- 23. Done, you have now retrieved *Leishmania* HSP70 sequences from the EMBL data base
- 24. Give each file entry a clear new name of maximum 40 characters. Include essential information such as the accession number and an abbreviation for the species or strain.

#### Additional remarks:

1. The above exercise is only an illustration of one possibility, it does not give a full overview of all search options. Retrieving all available sequences is a matter of proper (and sometimes inventive) selection of search terms. You can also look for particular accessions found in literature.

- 2. Selecting the proper set of sequences is often achieved by trial and error: some searches will result in too much junk, others will not get you all the sequences.
- 3. Searches can be combined in the Results page, but also already during the initial search in the Query form. The above exercise could be substituted by 1 search, when choosing Combine search terms with AND:

6	Organism Name	🔽 Leishmania
6	AllText	HSP70   "heat-shock protein"   "heat shock protein"
0	AllText	
6	AllText	

- 4. In order to select the proper sequences, it is important to critically read the sequence descriptions. A particular sequence might be listed as "putative", "similar", "precursor", "partial", or "mitochondrial". These might not always be relevant for your work.
- 5. Be aware that sometimes you will retrieve sequences that are the reverse complement of a coding region. In addition, not all sequences you retrieve will contain the entire fragment you are interested in, or some can contain a much larger fragment.
- 6. Make good use of search combinations, but be aware of their effect: "Q1 & Q2" (= "Q1 AND Q2") results in entries common to Q1 and Q2, in other words entries that came up in Q1 and in Q2. On the other hand, "Q1 OR Q2" (= "Q1 | Q2") takes entries that came up in Q1 or in Q2, in other words all entries from both searches. This may be counter-intuitive at times.
- Another possible combination is "NOT" (or "!"). E.g. in the above exercise "Q4!Q3" will result in entries containing the phrase "heat shock protein", excluding the ones with the dash.
- Always be on the lookout for missed entries, enlarge your searches if necessary. Sometimes searches do not result in what you expect because of the way search terms are checked against the entries. This you will find out only empirically.
- 9. You can also combine specific searches with resulting entries from BLAST.
- Use of a wildcard "\*": wildcards are used to find parts of a word in a description.
   Examples:
- 11. "shock" will find only "shock", but not "heat-shock", neither "shock-70"
- 12. "shock\*" will find "shock-70" but not "heat-shock"
- "\*shock\*" will find any word or phrase that contains shock in it, like "heat-shock-70 protein"
- 14. As a default, all searches are performed with a wildcard at the end of a word, as specified under search options:



- 15. In the Library page you will find an extensive list of libraries you can look in for finding sequences. Usually the general library used in the exercise is sufficient.
- 16. You can play around with different outputs besides FASTA, and search more specifically in particular entry fields. Trial-and-error is the best way of starting to deal with these options.
- 17. GenBank (www.ncbi.nlm.nih.gov/sites/entrez) offers the same possibilities for searching and search combinations, but sometimes comes up with quite unexpected results, e.g. a search for "donovani\*" ends up with a few thousand bacteria that do not have the word donovani anywhere in their description.
- 18. Similarly in GeneDB (www.genedb.org) BLAST and specific search options are foreseen.

#### Exercises in individual groups:

- Collect as many *Leishmania* HSP70 sequences as possible. Make sure you don't have duplicate entries by using the proper query combinations, e.g. instead of making separate FASTA files from different queries, first combine the queries and then make 1 FASTA file.
- **2**. If the sequences are bigger than the HSP70 coding regions, delete the extra nucleotides. Info on this can be found in the entries themselves.
- **3**. Put all sequences in one FASTA format file, which you identify as follows: "HSP70\_groupID.fsa".
- 4. Clearly identify each sequence with "GroupID\_accession\_short description":

```
> 1_ AF291716_bra_complCDS
```

Sequence

```
> 1_ AY423868_tar_complCDS
```

Sequence

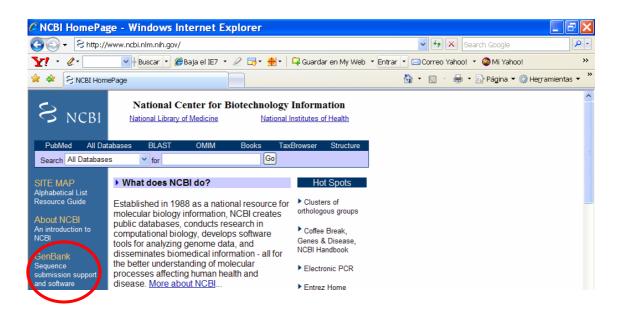
etc.

Beware: sequence names **must not exceed 40 characters** to be compatible with later programs!

 Perform the same exercise for rDNA ITS1 sequences, but replace the "HSP70" labels by "ITS1".

#### Method B: Retrieval of hsp70 nucleotide sequences from GenBank (joint exercise)

- 1. Open www.ncbi.nlm.nih.gov
- 2. Click on GenBank on the left-side menu



3. Click nucleotide on the search tab and search for Leishmania, then Go

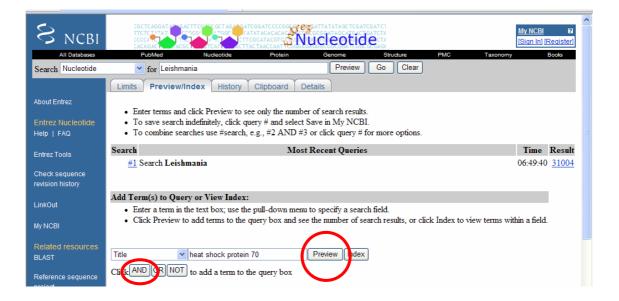
🥟 GenBank Ove	rview - Windows Inte	rnet Explorer				_ 7 🗙
(3) € http://	/www.ncbi.nlm.nih.gov/Genbank,	(index.html		🗸 🗲 🗙 Se	arch Google	<b>P</b> •
Y! · @·	💙 🔶 Buscar 🔻 🏉 Baja el 1	E7 • 🖉 🗔 • 🖶 •	🛱 Guardar en My Web	▼ Entrar ▼ ⊠Correo Yahoo!	🝷 🥹 Mi Yahoo!	>>
😭 🏟 🕞 GenBank	Overview			🟠 • 🗟 - 🖶 •	🔂 Página 🔻 🚫 Herran	nientas 🔻 👋
S NCBI	GenBank Overvie	w				^
PubMed	Entrez BLAST	OMIM Books	Taxonomy	Structure		
Search Nucleotide	🖌 for Leishmania	Go				
NCBI Home NCBI Site Map	▶ What is GenBank?					

4. Now you are redirected to the Entrez nucleotide resource, where you can find all *Leishmania* related nucleotide sequences. However, you need to refine your search to *Leishmania* heat shock protein 70. Then click on the Preview/Index tab.

SNCBI	My NCBi 2 [Sian In] [Reaister]
All Databases PubMed Nucleotide Protein Genome Structure	PMC Taxonomy Books
Search Nucleotide 🔽 for Leishmania Go Clear Save S	jearch
Limits Preview/Index History Clipboard Details	
Found 87995 nucleoide sequences. Nucleoide [31004] EST [26794] GSS [30195]	
Display Summary Show 20 V Sort By V Send to V	
All: 31004 Bacteria: 32 RefSeq: 24535 mRNA: 25058 🛠	
Items 1 - 20 of 31004 Page 1 of 1551 Next	▼ Top Organisms [ <u>Tree</u> ]
This search in Gene shows 25787 results, including:	Leishmania major <i>(9114)</i>
LmiF36.3860 (Leishmania major strain Friedlin): similar to leishmania major. I411.4-like protein LinJ36.2890 (Leishmania infantum JPCM5): similar to leishmania major. I411.4-like protein LmiF34.3440 (Leishmania major strain Friedlin): DNA topoisomerase IB, large subunit	Leishmania infantum (8714) Leishmania major strain Friedlin (8308) Leishmania braziliensis (8154) Leishmania infantum JPCMS (8028) All other taxa (12613)

5. The result of your first search (#1) is presented, you can add new terms to refine your search:

5a. Search field [title], Text box [heat shock protein 70]. Then click AND and Preview



With your new search (#2) now you have 13 results:

S NCBI	CGCTCAGGAT CACATTCOCCCCTAG AN ATCGCATCCCCCGCCCCCCAT TTCTCTTTTTTTTTTTTT	ATATAGCTCGATCGATCTA Ieotidecta	My NCBI PI [Sian In] [Register]
All Databases	PubMed Nucleotide Protein	Genome Structure PMC	Taxonomy Books
Search Nucleotide	✓ for Leishmania AND heat shock protein 70[Title]	Preview Go Clear Save Sea	rch
	Limits Preview/Index History Clipboard Details		
About Entrez Entrez Nucleotide Help I FAQ	<ul> <li>Enter terms and click Preview to see only the number of s</li> <li>To save search indefinitely, click query # and select Save</li> <li>To combine searches use #search, e.g., #2 AND #3 or c</li> </ul>	in My NCBI.	
Entrez Tools	Search Most Rec	ent Queries	Time Result
Endez Tools	#2 Search Leishmania AND heat shock protein 70[Tit	le]	09:02:59 13
Check sequence revision history	<u>#1</u> Search Leishmania		09:01:48 <u>31033</u>

Click on results **13**: And you will check that your *Leishmania* results have been filtered for heat shock protein 70.

C Leis	hmania AND heat shock protein 70[Title] - Nucleotide Results - V	Vindows	Internet Explorer	_ = X
$\bigcirc$				
<b>Y!</b> -	🖉 • 🔄 🗸 • Buscar • 🏈 Baja el IE7 • 🖉 🗔 • 🖶 • 📮 Guardar en My Web • E	intrar 💌 🖂	Correo Yahoo! 🝷 🥹 Mi Yahoo!	**
🚖 🏟	S Leishmania AND heat shock protein 70[Title]	🟠 -	🔊 🔹 🖶 🝷 🔂 Página 🔻 🍈 Herra	mientas 🕶 🎽
2:	Leishmania major heat shock protein 70, putative (LmjF28.2820) partial mRNA	Links		. (13)
3:	Leishmania braziliensis MHOM/BR/75/M2904 heat shock protein 70, putative (LbrM28_V2.3 partial mRNA			
□4:	Leishmania donovani strain Dd8 heat shock protein 70 gene, partial cds	Links		

5b. You can improve your search by changing the text in the query

- Search field [title], Text box [heat-shock protein 70]
- Search field [title], Text box [hsp70]

About Entrez Entrez Nucleotide Help   FAQ	<ul> <li>Click to see the items you have collected</li> <li>Enter terms and click Preview to see only the number of search results.</li> <li>To save search indefinitely, click query # and select Save in My NCBI.</li> <li>To combine searches use #search, e.g., #2 AND #3 or click query # for more options.</li> </ul>		
Entrez Tools	Search Most Recent Queries	Time	Result
210013	#4 Search Leishmania AND hsp70[Title]	09:06:1	5 <u>21</u>
Check sequence	#3 Search Leishmania AND heat-shock protein 70[Title]	09:05:5	4 <u>13</u>
revision history	#2 Search Leishmania AND heat shock protein 70[Title]	09:02:5	9 <u>13</u>
LinkOut			

5c. Or you can also combine searches:

Either by writing in the upper search text box:

leishmania[Organism]ANDheatshockprotein70[Title]ORleishmania[Organism]ANDheat-shockprotein70[Title]ORleishmania[Organism]ANDhsp70[Title]

Or by combining the AND/OR buttons and the search fields and text box

Entrez Nucleotide Help   FAQ	<ul> <li>Enter terms and click Preview to see only the number of search results.</li> <li>To save search indefinitely, click query # and select Save in My NCBI.</li> <li>To combine searches use #search, e.g., #2 AND #3 or click query # for more options.</li> </ul>		
Entrez Tools	Search Most Recent Queries	Time	Result
Enaczinos	#5 Search leishmania[Organism] AND heat shock protein 70[Title] OR leishmania[Organism] AND heat	t- 09:08:03	21
Check sequence	shock protein 70[Title] OR leishmania[Organism] AND hsp70[Title]		
revision history	#4 Search Leishmania AND hsp70[Title]	09:06:15	21
	#3 Search Leishmania AND heat-shock protein 70[Title]	09:05:54	13
LinkOut			_

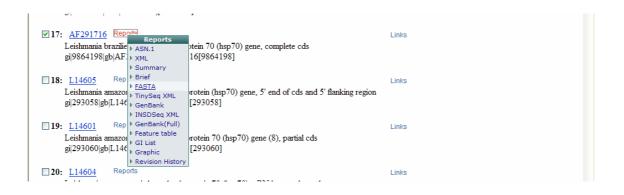
6. Check your entries by clicking the result and...:

<b></b> 1:	XM_001684512 Reports Leishmania major heat-shock protein hsp70, putative (LmjF28.2780) partial mRNA gi 157872029 ref XM_001684512.1 [157872029]	Links
2:	XM_001684511 Reports Leishmania major heat-shock protein hsp70, putative (LmjF28.2770) partial mRNA gi 157872027 ref XM_001684511.1 [157872027]	Links
3:	XM 001566275 Reports Leishmania braziliensis MHOM/BR/75/M2904 heat-shock protein hsp70, putative (LbrM28_V2.2990) partial mRNA gi 154340736 ref XM_001566275.1 [154340736]	Links
<b>4</b> :	XM 001566274 Reports Leishmania braziliensis MHOM/BR/75/M2904 heat-shock protein hsp70, putative (LbrM28_V2.2980) mRNA, partial cds gi 154340734 ref XM_001566274.1 [154340734]	Links
5:	XM_001566273 Reports Leishmania braziliensis MHOM/BR/75/M2904 heat-shock protein hsp70, putative (LbrM28_V2.2970) partial mRNA gi 154340732 ref XM_001566273.1 [154340732]	Links

#### 7. ...select those you want to use

I7:       AF291716       Reports         Leishmania braziliensis heat shock protein 70 (hsp70) gene, complete cds       gij9864198[gb]AF291716.1]AF291716[9864198]	Links	
DID T14005 Popula		

8. Click on Reports and select FASTA



9. Now you have your sequence in FASTA format

Leishmania braziliensis heat shock protein 70 (hsp70) gene, complete cds ygi]9864198[gb]AF291716.1[AF291716 Leishmania braziliensis heat shock protein 70 (hsp70) gene, complete cds traaccorcertrecertrecertrecertreceadegartrecaatecorceacca preactine construction of the construction	Format: GenBank FASTA Graphics More Formats	<u>Download</u> ▼ <u>Save</u> ▼ <u>Links</u> ▼
Customize View     Sequence Analysis Tools       > pgi   9864198   gb   AF291716.1   AF291716 Leishmania braziliensis heat shock protein 70 (hsp70) gene, complete cds     > BLAST Sequence       > TRACCCTCCTCCCCCTTCCCCTGCTTCACATAAAACCCTTGACACATGCAGGATTCACATAACCG SCACTTTCACCTCTCCCCTGCTTCACATAAAACCCTTGACACATGCAGCACCCCCTTACCCA ACTACTCCCCCCTGTTTGTATACGAGACTTCTCCACGTTCACATTACCACTACCCA ACTACTCCTCTCTATATTCCCTTCTCGGTTCAGGTTTTAAATGTTTCACCATACCTACC	GenBank: AF291716.1	Change Region Shown
<ul> <li>bgij9864198jgjA291716.1jA291716 Leishmania braziliensis heat shock protein 70</li> <li>bLAST Sequence</li> <li>blast Staccher Second State State</li></ul>	Leishmania braziliensis heat shock protein 70 (hsp70) gene, complete cds	Customize View
hsp70) gene, complete cds       , BLASI Sequence         TTAACCCTCCTTCCCCTTCCCTTCTCTCACAAAACCCTTTGACACATGCAGGATTCACAATACCG       , BLASI Sequence         CCCCCCCTTCCCCTTCCCTTCCCTTCCCTCACACTGCAGGATTCACAATACCG       , BLASI Sequence         CCCCCCCTTCCCCTTCCCCTTCCCCTCACCACTCCTCCACACTGCAGGATTCACAATACCG       , BLASI Sequence         CCCCCCCCTTTCCCTTCTCCGTTCCGTCCCACCCCTCCCCACCCCCTTACCCCCC       , BLASI Sequence         CCCCCCCCTTTCCCCTTCCCCCCACCCCCCCCCCCCCC		Sequence Analysis Tools
TTACCCTCTCCCCCTTTCCCCTGCTTCACTAAAACCCTTTGACACATGACAGGATTTCACATAACCG       + Pick Primers         SCACTTTCACTCTCCGGATATACGAACTTTCTCAAGGATTTCCAACTCCGACCACCCCTTACACCA       Recent Activity         TACACTTCCCCCCTGTTTGTACTGAAGGATCTTTCCCTCCC	>gi 9864198 gb AF291716.1 AF291716 Leishmania braziliensis heat shock protein 70 (hsp70) gene. complete cds	BLAST Sequence
TICGCCCATAACTCGCTTTCTATTTTCCTTTCTCGTTCAGTTTTAATTGTTCTAACTTTCACCTACT CTACTCCTCTCTTATTCACTACATAACTTTACCCTCCCCCC	TTAACCCTCCTTCCCCCTTTGCTTGACATAAAACCCTTTGACACATGCAGGATTTCACATAACCG CGCACTTTCACTCTACTCCAGATATACGAAGTTTCTGAAGGATTGTCAAGTCTCGACCACCCGTTACACA	Pick Primers
ITCCCTCATAACTCCCTTATTTCATTTTCCTTTTCAGTTTCAACTTCTAACTTCCACTACTACCTAC	TACATTACCCCCCTGTTTGTGTACGTGAAGGACTCTTTCCCTCACTCTCACTATATCCATCTACCCTA	Recent Activity
CCCCCCCTGTTTTGTTGACGTAAAAGGATTCTTTCCCCGCACTCTCCCGACTATATCCCATCTACCACCTACCCACCC	ITCGCTCATAACTCGCTTTCTATTTTCCTTTCTCGTTTCAGTTTTAAATTGTTTCTAACTTTCACCTACT	Recent Activity -
TATTCGCTTCATAAACCTCGCTTTCTAATTTTCCCTTTCTCGGTTTCAGTTTTAAATGTTTCTAA         TTTTCCACCTACTACAACTTCTCTTCTTTTTTCAACCAACAA	ACTACTCTCTCTTATTCACTACATAACTTTACACCTCTCCCCCACCCGTTTACCACATAACCATTACC	Turn Off Clear
Contraction of the second	CCCCCCTGTTTTGTTGTACGTTAAAAGGATTCTTTTCCCCTCACTTCTCCGACTATATCCCATCTACCC	
TTTTCCACCTACTTACAACTTCCTCTCTTTTTTCAACGACAAAACTTTTTACACCTCCTCTTTT CACTTCCCACGACGACGACGCCTGCGAGGGGGCTATTGGGATGACCAGGCACGACGACGACGCACCGCCTCTACGGG CSTGTGGCCGCTGCAGGACGCCGGAGGTCCCGCGAGGACGACGCACGC	IATTTCGCTTCATAAACCTCGCTTTTCTAATTTTCCCTTTCTCGGTTTCCAGTTTTAAATTGTTTCTTAA	🗏 Leishmania major Ini: 1578720291
CEGTGTGGCAGAACGAGCGCGGGAGATCATCGCGAACGACCAGGGCAACGGCCAGGCGCGTCGTACGTC CCTTCACGGACTCCGAGGCCTGTGTCGGCGAACGACCAGGGCGAGGAGCGCAACGAGGCAACGAGGCCAACGA CTGGCCCTTCAAGGTGACGACGACGGCGAAGGACGAACGA	CTTTTCCACCTACTTACAACTTCTCTCTCTTTATTTCAACGTACAATAACTTTTACACCTCCTTCTT	
Sectricacegaaccecetaaccacececetaaccacececetaaccacececetaaccacecetaaccacecetaaccacecetaaccacecetaaccacecetaaccacecetaaccacecetaaccacecetaaccacecetaaccacecetaaccacecetaaccacecetaaccacecetaaccacecetaaccacecetaaccacecetaaccacecetaceta	CACTTCTCTCAAAGATGACGTTCGAGGGTGCTATTGGTATTGATCTGGGCACGACGTACTCGTGCGTG	Leishmania brazil[gi:154340736]
Construction       Constructin       Constructin       C	CGTGTGGCAGAACGAGCGCGTGGAGATCATCGCGAACGACCAGGGCAACCGCACGACGCCGTCGTACGTC	0
CTGGCCCTTCAAGGTGACGACGACGACGACGACGACGCCGTGATCACGGTGCAGTTCCACGGCGAGGAG       Q Leishmania (31033)       Nucleotide         NacGACCTTCACCCCGGAGGAGGTGCTGCTGCTGCTGCAGAGTGCAGGCGAGGCGAGGCGACGCGAGGCGACGGAGGCGAGGGCACGGAGGGCACGGAGGGCGACGGAGGGCACGGCGAGGGCGACGGAGGGCACGGCGAGGGCACGGAGGGCACGGAGGGCACGGAGGGCACGGAGGGCACGGCGAGGGCACGGCGAGGGCGACGGAGGGCAAGGGCGACGGCACGGCGAGGCGACGGCGAGGCGACGGCGAGGGCGACGGCGAGGGCGACGGCGAGGGCGACGGCGAGGCGACGGCGAGGCGACGGCGAGGGCGACGGCGAGGCGACGGCGAGGCGACGGCGACGGCGAGGCGACGGCGAGGCGACGGCGAGGCGACGGCGAGGCGACGGCGAGGCGAGGCGACGGCGAGGCGACGGCGAGGCGACGGCGACGGCGAGGCGACGGCGAGGCGACGGCGACGGCGAGGCGACGGCGACGGCGAGGCGACGGCGACGGCGAGGCGACGGCGAGGCGACGGCGAGGCGACGGCGAGGCGAGGCGACGGCGAGGCGAGGCGAGGCGAGGGGGG	GCCTTCACGGACTCGGAGCGTCTGATCGGCGATGCCGCGAAGAACCAGGTGGCGATGAACCCGCACAACA	Q leishmania[Organism] AND (21)
ALGACCTTCACGCCGGAGGAGTGGGCTCGATGGTGCCTGCTGAAGATGAAGGCGAGGCGCAGGCGAAGGCGTACC TTGGCAAGCAGGCGAGGATGGGAGCTGGTGCTGCCTGCTGCTGCTGCAGGATGAGGCGAGGCGTACC SAAGGACGCCGGCACGATGGCGGCGCGGGGGGGCGCGCGCG	CGGTGTTCGACGCGAAGCGCCTGATTGGCCGCAAGTTCAACGACTCCGTTGTGCAGGCGGACATGAAGCA	O Laishmania (24022)
ITGECAAGCAGGTGAAGAAGAGGCCGTGGTGACGGTGCCTGCCT	CTGGCCCTTCAAGGTGACGACGAGGGTGACGACGAGCCCGTGATCACGGTGCAGTTCCACGGCGAGGAG	Celsninariia (31033) Nucleotide
AAAGGACGCCGGCACGATTGCGGGGCCTGGAGGTGCTGCGCATCATCAACGAGCCGACGGCCGCCGCCCCCCCC	AAGACCTTCACGCCGGAGGAGGTGAGCTCGATGGTGCTGCTGAAGATGAAGGAGACGGCGGAGGCGTACC	
SCGTACGGGCTGGACAAGGCGAAGACGCAAGAGGCGAACGTGCTGATCTTCGACCTTGGCGGCGGTA CGTTCGATGTGACGCTGCTGACGATCGACGGCGGATTCTTCGAGGCGAAGGCGACGAACGGCTACACCCA CGTCGGCGGCGAGGACTTCGACGACTCGGTGACCTTCTTCACCGGGAGGTGTCAAGGCGCAAGACAAG SGCAAGGACCTGTCGTCGACCTCGGTGCCCTTCGTCACCGCGGCGGCGCCAAGGCGCA SGCAAGGACCTGTCGTCGACGCGCCGCGCGCGCGCGCGCGC	TTGGCAAGCAGGTGAAGAAGGCCGTGGTGACGGTGCCTGCC	
CETTCGATGTGACGCTGCTGACGATCGACGGCGGTATCTTCGAGGTGAAGGCGACGAACGGCTACACCCA Full text in PMC CCTCGGCGGCGAGGACTTCGACAACCGCCTGGTGACCTTCTTCACCGAGGAGATCAAGCGCAAGAACAAG SGCAAGGACCTGTCGTCGACGACCGCGCGCGCGCGCGCGCG	GAAGGACGCCGGCACGATTGCGGGGCCTGGAGGTGCTGCGCATCATCAACGAGCCGACGGCTGCGGCCATC	All links from this record
CCCCGGCGGCGAGGACTTCGACAACCGCCTGGTGACCTTCTTCACCGAGGAGTTCAAGCGCAAGAACAAG SGCAAGGACCTGTCGTCGACGCCGCGCGCGCGCGCGCGCGC	GCGTACGGGCTGGACAAGGGCGACGACGGCAAGGAGCGCAACGTGCTGATCTTCGACCTTGGCGGCGGTA	
SSCAAGGACCTGTCGTCGAGCCACCGCGCGCGCGCGCCCCCGCGCGCG	CGTTCGATGTGACGCTGCTGACGATCGACGGCGGTATCTTCGAGGTGAAGGCGACGGACG	Full text in PMC
SOCAAGGACCIGICGICGAGCCACCGCGCGCGCGCGCGCGCG	CCTCGGCGGCGAGGACTTCGACAACCGCCTGGTGACCTTCTTCACCGAGGAGTTCAAGCGCAAGAACAAG	
CGCTGTCCGCCGCGACGCGACGACGACGACGACGACGCGCTGTTCGACAACGTGGACTTCCAGGCCAC	GGCAAGGACCTGTCGTCGAGCCACCGCGCGCGCGCCGCCCGC	Protein
	CGCTGTCCGCCGCGACGACGACGACCGACGACGCGCCGCTGTTCGACAACGTGGACTTCCAGGCCAC	Taxonomy

10. To save it click on Download as FASTA and save it in your sequence editor.

earch Nucleotide for Go Clear Limits Preview/Index History Clipboard Details			
Format: <u>GenBank</u> FASTA <u>Graphics</u> <u>More Formats</u> ▼	Download ▼ Save	▼ <u>Links</u> ▼	
GenBank: AF291716.1	GenBank	nown	
Leishmania braziliensis heat shock protein 70 (hsp70) gene, complete	GenBank(Full)		
cds	FASTA		
>qi 9864198 gb AF291716.1 AF291716 Leishmania braziliensis heat shock protein 70	ASN.1	s Tools	
(hsp70) gene, complete cds TraAcconcertrococcritocortrocatatataccortrigacacatocaggatticagatataccg	XML		
CGCACTTTCACTCTACTCCAGATATACGAAGTTTCTGAAGGATTGTCAAGTCTCGACCACCCGTTACACA	INSDSeq XML		
TACATTACCCCCCTGTTTGTGTACGTGAAGGACTCTTTCCCTCACTCTCTCACTATATCCATCTACCCTA TTCGCTCATAACTCGCTTTCTATTTTCCTTTCTCGTTTCAGTTTTAAATTGTTTCTAACTTTCACCTACT	TinySeq XML		
ACTACTCTCTCTCTTATTCACTACATAACTTTACACCTCTCCCCCC	Feature Table	<u>Turn Off</u>	Clear

11. You can also select more than one sequence

Display Summary V Show 20 V Sort By V Send to V	<b>~</b>
Page 1 of 2 Next	
☑ 20: L14604 Reports Links Leishmania amazonensis heat shock protein 70 (hsp70) mRNA, complete cds gi 293056 gb L14604.1 LEIHSP70C[293056]	
□ 19: L14601 Reports Links Leishmania amazonensis heat shock protein 70 (hsp70) gene (8), partial cds gi 293060 gb L14601.1 LEILHSPG8[293060]	
■ 18: L14605 Reports Links Leishmania amazonensis heat shock protein (hsp70) gene, 5' end of cds and 5' flanking region gi 293058 gb L14605.1 LEIHSP70G[293058]	
Image: Provide the state of the state o	

12. Display them in FASTA format

NCBI CCC AGAC	COREC OCT AG A GATCGGATC GO A TGG A GATCGGATC V T CACTAC CACCTAC TACAC	ALACINE STATE	ATAGE TEGATEGATET FAGEATGAETGATETA EOLIO CEAGE COLOCIO CEAG			My NCBI RA [Sian In] [Register]
PubMed Nucleotide	Protein	Genome	Structure	PMC	Taxonomy	OMIM Books
Search Nucleotide Y for	eishmania[Organism] AN	ID heat shock prot	ein 70[Ti Go C	lear		
Display FASTA Show 20	<ul> <li>Send to</li> </ul>					3
	Send to	Item 1	- 2 of 2			
	Text File					
1: L14604. Reports Leishmania am						Links
>gi 293056 gb L14604.1 LEIHSP7	OC Leishmania amaz	onensis heat	shock			O Next sequence
protein 70 (hsp70) mRNA, compl						
CTTTATTGGGTCCTAAACACGCACTCGCAC						
GTCTCTCTCGCTCTGCGCTCTATTACGTAA						
CTGCCGCAGAGATGACGTTCGACGGCGCCA						
GTGGCAGAACGACCGCGTGGAAATCATCGC TTCACGGACTCGGAGCGCCTGATCGGCGAT						
TGTTCGATGCGAAGCGCCTGATTGGTCGCA	00000111101110011001					
COOCTTON CONTON CONCERNMENT	AGIICAACGACIIGGIIG	COLORGICOGACAT	CARCOCACIO			

13. and send them to Text

😪 🏟 🕞 http://www.ncbi.nlm.nih.gov/sviewer/viewer	tas 🕶
	~
>gi 293056 gb L14604.1 LEIHSP70C Leishmania amazonensis heat shock protein 70 (hsp70) mRNA, complete cds	
CTTTATTGGGTCCTAAACACGCACTCGCACTCCAGCTGTCCGAAGAGAACACATACGCGCACAGGCACAC	
GTCTCTCTCGCTCTGCGCTCTATTACGTAACCCCTATAAACACCCCCCCC	
CTGCCGCAGAGATGACGTTCGACGGCGCCATCGGCATCGGCCACGGCGCGTCCGTGCGTG	
GTGGCAGAACGACCGCGTGGAAATCATCGCGAACGATCAGGGCAACCGCCGCGACGGCGACGGTCGTACGTTGCG	
TTCACGGACTCGGAGCGCCTGATCGGCGATGCCGCAAAGAACCAGGTGGCCATGAACCCGCACAACACGG	
TGTTCGATGCGAAGCGCCTGATTGGTCGCAAGTTCAACGACTTGGTTGTGCAGTCGGACATGAAGCACTG	
GCCGTTCAAGGTGACGACGACGACGACAAGCCCGTGATTTCGGTGCAGTACCGCGGCGAAGAGAAA	
ACCTTCACGCCGGAGAAGATCAGCTCGATGGTGCTGCTGAAGATGAAGGAGACGGCGGAGGCGTACCTGG	
GCAAGCAGGTGAAGAAGGCCGTGGTGACGGTGCCGGCGTACTTCAACGACTCGCAGCGACGAACGA	
GBACGCCGGCACGATTTCTGGGCTGGAGGTGTTGCGCATCATCAACGAGCCGACGGCGBCGGCCATCGCG	
TACGGCCTGGACAAGGCGACGACGACGGCAAGGAGCGCAACGTGCTGATCTTCGACCTTGGCGGCGGCACGT	
TCGATGTGACGCTGCTGACCATCGACGGCGACTCTTCGAGGTGAAGGCGACGACGGCGACACGCGACCT	
TGGGGGGGAGGACTTCGACAAACCGCCTCGTCACGTCTTCTCACCGAGGAGTTCAAGCGCAAGAACAAGAACAAGGGC	

14. Now you can copy and paste them in a text editor.

#### Protocol 5.2 Analysis of sequence chromatograms

# PurposeTo edit the chromatograms obtained from automated sequencers and<br/>resolve possible conflicts in order to obtain the final deduced sequence

#### A. INTRODUCTION

Output from automated sequencers is limited in size, and approaches the sequence from one direction only. It is necessary to edit the sequence chromatograms in order to extract trustworthy information, concatenate them into larger sequences, and resolve possible conflicts to obtain the final deduced sequence. To conclude, the new sequences can be submitted to the public domain sequence repositories (EBI / GenBank).

Sequences used: HSP70 and ITS1 chromatograms obtained from automated sequencing

Programs used: BioEdit (<u>www.mbio.ncsu.edu/BioEdit/BioEdit.html</u>)

Programs not used but excellent: MEGA (www. megasoftware.net)

**Sequences used:** HSP70 and ITS1 chromatograms obtained from automated sequencing.

**Programs used:** BioEdit (www.mbio.ncsu.edu/BioEdit/BioEdit.html)

Programs not used but excellent: MEGA (www.megasoftware.net)

Method: Editing chromatograms for HSP70 and ITS1 sequences by using BioEdit software (joint exercise)

· - · · · -

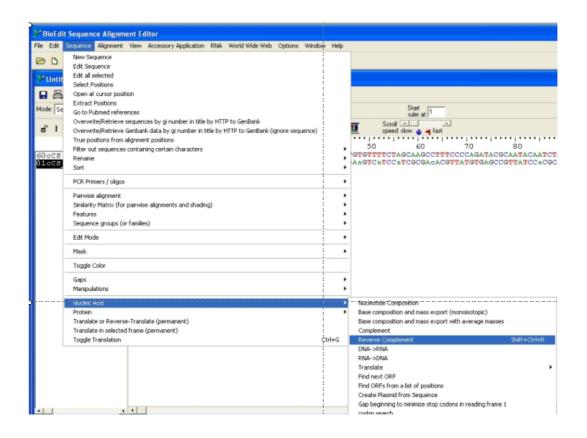
- Start the BioEdit software from FioEdit.exe 1.
- Go to File and New Alignment and Import your sequence alignment files: 2.

BioEdit Sequence Alignment Editor					
Fie Edit Sequence Alignment View Accessory Application RN	A World Wide Web O	Options Wi	ndow	Help	
New Alignment		Ctrl+N			
Open		Ctrl+O			
New from Clipboard					
New Text					
Open As Text					
Save		Ctrl+S		ne Start	
Save As				ne ruler at: 1	
Retrieve sequences from GenBank or GenPept				Scrol .	
Copy file name to clipboard				MI speed slow 🕁 🚽 fast	
Export			•		
Import			•	Sequence alignment file 70	80 90 10 GCAATACAATCTATATATGTATA
Import from Clipboard				From tab delimited file (eg Excel)	GCARTACAATCTATATATGTATA
Merge into Alignment based on a Reference Sequence			T		
Merge from Clipboard					
Append Alignment					
Go to GenBank					
Close					
Batch ABI to SCF trace file conversion					

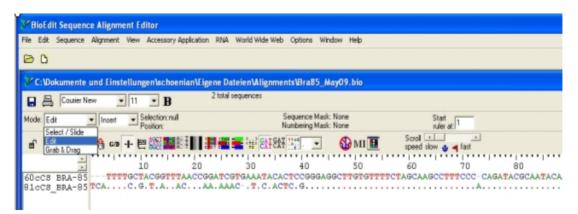
To see the chromatogram for a given sequence press Open. Check the quality of the 3. sequence:

1	BioEd	it Sequ	ence	Alignm	ent Edi	itor											
File	Edit	view	Zoom	Horizont	al Scale	Accesory	/ Applicati	on RNA	Window	Help							
2	b D																
34	Unti	tled															
E		Couri	er New	-	11	• B		2 total :	equences								
Mo	de: S	elect / SI	ide 💌			election: 0 osition:						ce Mask ng Mask					Start ruler at
E	ſΙ	DI	D	😚 сл	+- [편		£Ŧ		<b>e i tei</b>	AT CA	T Tres T : <sup>2</sup> c:		-	MI		Scroll speed s	low 🕁
						) STTTAA CCTTTT		<b>FCGTG</b> A		AC <b>T</b> C	4 CGGG	aGGC <mark>T</mark>	' <b>T</b> G <b>T</b> G'	50 FTTTC:	FAGC/	60 AAGCC	TTTC
	🎾 А	Bl Chro	matog	gram: C	: Woku	imente u	ınd Eins	tellunge	n\schoe	nian\	Eigene	Dateie	n\SEQ\	CSchwe	ynoch	nMAR2	5C\60c
			Sele	cted: nor	ne Sam	ple: 60cCS	_BRA-85_	LITSR	ïle: C:\Dok	umente	e und Ein	stelbinge:	n'schoen	ian\Eigene	Dateie:	n'SEQ\CS	chweync
		140 TTA.	ACA 1		150 GCG	TATAC	160 CAACA	. A A A A		170 CGT	тст	AC G (	180 3 C T T	тттт	гттт	190 ГGGC	GGC(
					ΛΛΛ 3, 232: 8		W	M					W		W	MA	MA
	•																

4. Compare the forward (Primer LITS for ITS1) and backward (primer L5.8S for ITS1) sequences. For this convert the backward sequence as follows: Go to Sequence, than to Nucleic Acid and than click on Reverse Complement.



- 5. Align the forward and the converted backward sequence by hand or go to Sequence and choose Pairwise alignment. Save the alignment obtained and close the file.
- 6. For editing go to File and Open your alignment. To check again with the chromatograms Open your sequence files (see point 3!). Use Mode for editing your sequences.



It is recommended to copy one of the sequences and to do all the editing there.

<mark>≫ BioEdit Seque</mark>	ence Alignm	ent Editor									
File Edit Sequenc	e Alignment	View Accesso	ory Application	RNA World Wid	le Web Options	Window Help					
🕞 🖪											
C: Wokument	te und Einst	ellungen\sch			gnments\Bra85	_May09.bio					
E 📇 Courier	r New 💌	11 <b>• E</b>	3	total sequences							
Mode: Edit	▼ Insert	Selection: Position:			Sequence M Numbering M			Start ruler at: 1			
f I D I	<u>D</u> 🔒 сл	++ 🗠 🎆		ter e	AT GAT TINE AT CAT THE	🚯 мі 🎛	Scroll speed sl	ow 🕁 ┥ fast	•		
	•		10	20	30	40	50	60	70	80	ا • • • • ا
	85 LITSR 85 L5-85					CTCCGGGAGGC CCGGGGAGGCT					
consensus H	BRA-85	TCATTTTC	CGATGATTA	CACCAAAAA	ACATACAACI	CCGGGGGAGGCT	TGTGTTTTC	AGCAAGCC	TTTCCCACAGA	TACGCAAT.	ACAA <mark>T</mark> O

7. After editing is finished, keep only the consensus sequence and delete the other two sequences by labeling them and pressing Edit and Cut. The consensus sequence will then be saved in Fasta format. This sequence can then be used for multiple alignments by either using BioEdit or MEGA softwares. In this training session we will use MEGA for multiple alignments.

#### Protocol 5.3 Sequence alignments, primer design and *in-silico* RFLP

Purpose To align sequences for identification of homologous positions

#### A. INTRODUCTION

Comparing sequences requires identification of homologous positions which is done by aligning them. Bases on such alignments, primers can be designed and RFLP experiments simulated *in-silico*.

Sequences used: Edited HSP70 and ITS1 sequences from previous workshop sessions

**Programs used:** MEGA (www. megasoftware.net)

Primer3 (online tool, http://frodo.wi.mit.edu)

Programs not used but excellent: GeneDoc (www.nrbsc,org/gfx/genedoc/)

BioEdit (www.mbio.ncsu.edu/BioEdit/BioEdit.html)

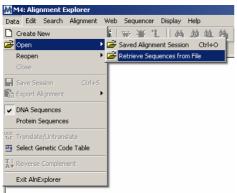
#### Method A: Aligning hsp70 nucleotide sequences (joint exercise)

- 8. Prepare a FASTA file containing all sequences you want to align. This file should include sequences retrieved by SRS and BLAST, as well as the consensus sequences from the analyzed chromatograms. Each sequence should have a concise but clear labeling: "GroupID\_accession\_short description". Include a species abbreviation in your description.
- 9. Make sure you have a copy of the exercise FASTA file "Rio\_Mega1.fsa".
- **10**. Start the MEGA software.
- **11**. Open the alignment explorer from the Alignment menu:



- **12**. Choose create a new alignment.
- **13**. Select Yes for nucleic acid alignment.

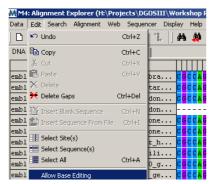
14. Retrieve sequences from the FASTA file "Rio\_Mega1.fsa" using the Data menu:



- **15.** In the Edit menu, choose <u>Select all</u>, or do this by the classical Windows mouse clicking method using the SHIFT key.
- 16. Align by ClustalW in the Alignment menu, using the default program options:

👫 M4: Alignment E	xplorer (H:\Projects\DGOSIII\'	Wor	ks	hop	) Ri	io\	te
Data Edit Search	Alignment Web Sequencer Dis	splay	/ 1	Help	)		
] 🗅 😅 🔚 🚏	W Align by ClustalW	4	ġ	<b>4</b>	ů	ġ,	đ
DNA Sequences T	📜 Mark/Unmark Site Ctrl+M						
	Align Marked Sites Ctrl+L	П	Π	Π	Π	Τ	Π
emb1 AF291716 #	Unmark All Sites	t	ta	ac	c	t	с
emb1 AY423867 A	🕺 Delete Gap-Only Sites	a	tg	aç	rc	jt	с
emb1 AY423868 A		a	tg	ac	a	tt	с
emb1 AY913842 A	t	gc	đđ	ra	tc	с	

- **17**. Wait a bit. Perhaps take a coffee.
- 18. Once aligned, switch off Base editing:

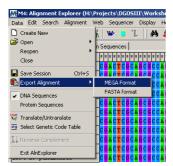


This is of crucial importance when starting to edit an alignment, as it prevents you from accidentally changing, inserting, or deleting bases.

- **19.** When not OK, modify the alignment manually by moving residues left (BACKSPACE) or right (SPACE).
- **20**. After finishing the alignment, add your own sequences to it. Do this from the Edit menu by allowing base editing and selecting the file option:

M4	: Alignment Explorer (I	H:\Projects\DG	OSIII\Workshop F
Data	Edit Search Alignment	: Web Sequen	cer Display Help
	🎦 Undo	Ctrl+Z	11. 🛛 🗛 🦊
DNA	唱 Copy	Ctrl+C	
	👗 Cut	Ctrl+X	
embl	🛃 Paste	⊂trl+V	bra ATTGGI
embl	imes Delete		tar ATTGGC
embl	Ӿ Delete Gaps	Ctrl+Del	don ATCGA
embl	🛅 Insert Blank Sequenc	e Ctrl+N	don
embl	🖄 Insert Sequence From	n File Ctrl+I	bra(2) <mark>acaag</mark> o
embl			tar(2) <mark>gaacgo</mark>
embl	Select Site(s)		tar(2) <mark>ggtga</mark> a
embl	Select Sequence(s)		don(2)
embl	Select All	Ctrl+A	r_h ATCGGC
embl	✓ Allow Base Editing		ili
embl	X52314 X52314_Lei	shmania_hsp7	O_g <mark>ATCGGC</mark>

- **21.** Use only sequences between the primer pair Hsp70sen (5' GACGGTGCCTGCCTACTTCAA 3') and Hsp70ant (5' CCGCCCATGCTCTGGTACATC 3').
- **22**. Save your alignment from the Data menu.
- **23.** Export your alignment in MEGA format, input a title and indicate that these are coding sequences:



**24.** Done. You are now ready to start aligning your own sequences in the individual group exercise.

#### Additional remarks:

- 1. Sometimes during aligning you will get an error message that some sequence are too divergent. Just ignore this.
- 2. The biggest alignment problem is working with partial sequences. ClustalW does not deal with this well. There are two ways of preventing this to some extent: either manual editing after the aligning procedure, or clipping your sequences prior to aligning. The latter is possible by using the features in the GenBank or EMBL entries, or by looking for particular motives that should appear at the beginning or the end of your sequence, such as PCR primers:

M4: Alignment Explorer (C:\Do	cuments and Settings\gvdauwera\My Documer
Data Edit Search Alignment We	eb Sequencer Display Help
📄 😅   🚧 Find Motif 💦 🖸	trl+F 🔃 🐍 🛛 🗛 🦊 🦓 🦓 🗠 🗈
DNA Segue 🌺 Find Next	F3 ces
	t+F3
LH2182 L	F4 BCCACGATTGCTGGCCTGGAGGTGCT
sidon PEI V Highlight Motif	BGCACGATTGCTGGCCTGGAGGTGCT
LCA08cl2 L. peruviana	C C C C C C C C C C C C C C C C C C C
geauw_CA08c12_per	C G C C G G C A C G A T T G C T G G C C T G G A G G T G C T
LH2439 L. peruviana	CGCCGGCACGATTGCTGGCCTGGAGGTGCT
sidon_PER126_1_per	C G C C G G C A C G A TT G C T G G C C T G G A G G T G C T
LH2864 L. peruviana	C G C C G G C A C G A TT G C T G G C C T G G A G G T G C T
sidon_PER204_1_per	C G C C G G C A C G A TT G C T G G C C T G G A G G T G C T
M5210 L. naiffi	C C C C G C A C G A T T G C T G G C C T G G A G G T G C T

- 3. When working with data base retrieved sequences, you will usually have sequences in your alignment that do not match the others. These sequences have probably "escaped" your selection, and are not what you want. In such case double check the entry description in the data base you retrieved it from. Alternatively, this may be a partial sequence, part of a sequence not present in the other entries, or a reverse complement.
- 4. You can translate your nucleic acids into amino acids to aid in aligning in case you work with coding sequences. If your reading frame does not start from position 1, you have to select the corresponding columns first. Careful: alignment gaps are causing a frame shift in your translation, unless they are a multitude of 3.

#### Exercises in individual groups:

 Align your retrieved and analyzed HSP70 sequences with the ones in the above exercise alignment. Remove any nucleotides outside the region of interest, and remove all PCR primer sequences by site selection and cutting:

_	Alignment Explorer (C:\De			-	-	_	_	-	au	we	ra\	M	/ D	00	u	ne	nt	s\
Data	Edit Search Alignment We	eb Sequer	cer	Di	spla	зy	н	elp										
D	🖍 Undo	Ctrl+Z	1		]	<b>1</b> 4	4	Ņ	Ņ		¥,		1	2	[	٥		X
DNA	Сору	Ctrl+C																
	👗 Cut	Ctrl+X	* * 7	*	* *	± 1	*	* *	*	ż	* *	* *	*	ż	*	*:		×
	🔁 Paste	Ctrl+V	302	G	CG	clo	A	GG	cla	A	cic	ala	G	G	lc.	G	: c	G
	🗙 Delete		B C A	G	CG	СС	A	GG	СС	A	:  G	A A	G	G	l C	G	: c	G
LEM6	₩ Delete Gaps	Ctrl+Del	3 C A	G	CG	СС	A	GG	СG	A	: G	ΑA	G	G	c	G	:c	G
ITMA	🕒 Insert Blank Sequence	Ctrl+N	3 C 4	G	C G	СС	A	GG	СС	A	G	ΑÀ	G	G <mark>1</mark>	c	G	: c	G
M417	🖄 Insert Sequence From File	Ctrl+I	3 C Z	G	CG	СС	A	GG	СС	À	G	АÀ	G	G	LC	G	c c	G
geau	Select Site(s)		9 C A	G	CG	СС	A	GG	CG	À	C	A P	G	G	1C	G	c c	G
LH23			B C A	G	CG	СС	A	GG	СĢ	A	C	A A	G	G	۱C	G	c c	G
sido	Select Sequence(s)		G C A	G	CG	СС	A	GG	СС	A	CG	A A	G	G	C	G	сc	G
M403	Select All	Ctrl+A	5 C 4	G	C G	сс	A	GG	СС	A	2 G	АÅ	G	G	c	G	c c	G
sido	✓ Allow Base Editing		3 C Z	G	C G	сс	A	GG	СС	A	2 G	АÀ	G	G	c	G	:c	G
LS94	L. panamensis	CGACTC	Ğ C A	G	CG	СС	A	GG	CG	À	2  G	АÀ	G	G	LC	G	c c	G
geau	r_LS94_pan	CGACTC	G C A	G	CG	СС	A	GG	CG	A	CG	ΑA	G	G	C	G	: c	G

- Make a de novo alignment from your gathered and analyzed ITS1 sequences. When exporting into MEGA format, remember that these are not protein coding sequences. Use only sequences between primer pair LITSR (5' CTGGATCATTTTCCGATG 3') and L5.8S (5' TGATACCACTTATCGCACTT 3').
- **3**. Use these alignments in subsequent primer design exercises.

#### Method B: PCR primers design using Primer3 software (joint exercise)

A primer is a short synthetic oligonucleotide which is used in many molecular techniques from <u>PCR</u> to <u>DNA sequencing</u>. A pair of primers is used in most PCR variants and these are designed to have a sequence which is the reverse complement of a region of template or target DNA to which we wish the primer to anneal. When designing primers for PCR it is often necessary to make predictions about these primers, for example melting temperature (Tm) and propensity to form dimers with themselves or other primers in the reaction.

Some aspects should be taken into account when designing PCR primers:

- 1. primers should be 15-30 bases in length.
- 2. base composition should be 40-60% (G+C)
- primers should end (3') in a G or C, or CG or GC: this prevents "breathing" of ends and increases efficiency of priming (<u>Note</u>: three or more Cs or Gs at the 3'- ends ofprimers may promote mispriming at G or C-rich sequences, because of stability of annealing, and should be avoided)
- 4. Tms between 55-70°C are preferred. Ideally, both primers should anneal at the same temperature. The annealing temperature (Ta) will ne dependent upon the primer with the lowest Tm. A simple formula to estimate the Tm of a DNA molecule is Tm= 2(A+T) + 4 (G+C); but you may be aware that this is only orientative, as different factors such as salt concentration may change the Tm value.
- 5. 3'-ends of primers should not be complementary, otherwise primer-dimers will be created preferentially to any other product.
- primer self-complementarity (ability to form 2<sup>ary</sup> structures such as hairpins) should be avoided.

Following these rules you can design your own set of PCR primers. Nevertheless there are different software tools which can be helpful for this purpose. In this training course you will deal with Primer3 software.

#### Primer3 (v. 0.4.0)

## 1. Open http://frodo.wi.mit.edu/ and go to Primer3plus interface

Primer3 Input 0.4.0 (primer)	3-web/htdocs/input-040.	htm) - Windows Intern	et Explorer	- 7 ×
🚱 🗸 🖉 http://frodo.wi.mit.edu/			🖌 🗲 🗙 primer3	• ٩
Y! • ∅• 🔤 🖌	🖁 Baja el IE7 🔹 🖉 🗔 ד 🎍 ד 🏻 🛱	Guardar en My Web 🔹 Entrar 💌	🖂 Correo Yahoo! 🔹 🚳 Mi Ya	hoo! >>
😭 🕸 🌈 Primer3 Input 0.4.0 (primer3-we	b/htdocs/in	6	🔹 📓 🔹 🖶 🝷 🔂 Página	🕶 🎯 Herramientas 👻 🎽
Primer3 (v. 0.4.0) Pick prim	ers from a DNA sequence.	Primer3plus interface Vore primer		Primer3 Home           FAQ/Wiki
Paste source sequence below (5'->3', string undesirable sequence (vector, ALUs, LINE	· · · · · · · · · · · · · · · · · · ·		ignored). FASTA format ok.	Please N-out
✓ Pick left primer, or use left primer below:	Pick hybridization probe (interroligo), or use oligo below:		: primer, or use right primer b pposite strand):	elow
Pick Primers Reset Form				

2. Be sure that you are on Task: Detection and paste your sequence (FASTA format) or Upload the File

🖉 Prime	r3Plus - Windows	Internet Explor	er			- 2 2
00-	+ 🖪 http://www.bioinform	natics.nl/cgi-bin/primer3p	ılus/primer3plus.cgi		🗸 🛃 🗙 primer3	P -
<b>Y!</b>	🖉 🔹 🚽 🔶 Busc	ar 🔹 🏉 Baja el IE7 🔹	🖉 🗔 र 🐠 र   🛱 Guar	dar en My Web 🔻 Entrar	💌 🖂 Correo Yahoo! 🔹 🚳 Mi Yah	100! ×
😤 🏟 🔒	🖳 Primer3Plus				🟠 🔹 🔝 👘 🖶 🔹 🔂 Página '	🕶 💮 Herramientas 👻 🎽
Prin	ner3Plus			Primer3Manager	Help	
	ers from a DNA sequence			About	Source Code	
Task: D		ct primer pairs to detect th uded/excluded regions car	ie given template sequence. 1 be specified.	Optionally targets and	Pick Primers Reset Form	n
Main	General Settings	Advanced Settings	Internal Oligo	Penalty Weights	Sequence Quality	
Sequence 1	Id:	٦				
	ce sequence below	Or upload sequence f	ile:		pload File	<b>a</b>
Mark sele	ected region: <> [] {}	Clear		Save Sec	quence	
Excluded F	Regions: <		>			
Targets:	[		]			
Included R	Partion: 1		<b>,</b>			×
					😜 Internet	🔍 100% 🔹 🛒
🤚 Inici	io 🔁 TRAINING	🖉 Primer3Plu 🛛	🖬 Primers De 🛛 🕎 Ge	nBank e ES Escrito	rio 🎽 Búsqueda en el escritorio	🔎 🔇 🥵 🎦 17:34

3. We will use the *L. braziliensis* hsp70 sequence AF291716, which has been obtained in the previous exercise. Identify the sequence (Sequence ID) as Lbra AF291716 hsp70 PCR

C Primer	3Plus - Windows	Internet Explorer					_ F 🗙
<b>OO</b> -	A http://www.bioinform	natics.nl/cgi-bin/primer3plus/	/primer3plus.cgi		🗸 🛃 🗙 prin	ner3	<b>P</b> •
Y! · Ø	🕶 🔷 🔶 Busca	ar 🔹 🏉 Baja el IE7 🔹 🖉	🔄 • 🔶 • 🛛 🛱 Gua	rdar en My Web 🔻 Entra	r 🔻 🖂 Correo Yahoo! 🔹	🧆 Mi Yahoo!	**
🚖 🎄 🗖	Primer3Plus				🏠 • 🔊 - 🖶 •	🔂 Página 🔻 🍈 H	leṟramientas 🕶 🎽
L			·				~
Task: De		ct primer pairs to detect the gi uded/excluded regions can be		Optionally targets and	Pick Primers	Reset Form	
Main	General Settings	Advanced Settings	Internal Oligo	Penalty Weights	Sequence Quality		
	Lbra AF291716 hsp70 F	Or upload sequence file:		Examinar	Upload File		
cds TTAACCCTC CTGAAGGAT TATTCGCTC ACACCTCTC ACACCTATTT TTTATTTCA CGTGGGCGC ATGGATGCCG ATGAAGCAC	CTTCCCCTTTCCCTTGC TGTCAAGTCTCGACCACC ATAACTCGCTTTCTATTT CCCCCACCCGTTTACCACA GCTTCATAAACTTTGC GCTCCATAAACTTTAC GCTACGAACACCAGCGCGC GCGACGACCAGGTGGCGA TGGCCCTCAAGGTGACG GCTGCTGAAGATGAAGGA	F291716 Leishmania bu TTCACATAAAACCCTTTGAC CGTTACACCATACATTTACCCC TCCTTTCTCGTTTCAGTTTT TAACCATTACCCCCCCCTGT TTACTAATTTCCCCTTCTTC GGAGATCATCGGAAACGACC GGAGGGCGAACGACCAGCCC GACGGCCGAGGGCGACCAGCCT Clear	ACATGCAGGATTTCAC CCTGTTTGTGTACGTG AAATTGTTTCTAACTT TTTGTTGTACGTTAAA 3GTTTCCAGTTTTAAA TCAAAGATGACGCACCGCA AGGCAACCGCACCGC	ATAACCOCCCC ACTTTCAC AAGGACTCTTTCCCTCAC TCACCTACTACTACTCTC AGGATTCTTTTCCCCCCA TTGTTTCTAACTTTTCC GCCGTCGTACTATTGGTATT GCCGTCGTACCTCCCCTT TGATTGGCCCCAAGTTCA.	ICTACTCCAGATATACGAA ICTCTCACTATATCCATC ICTCTTATTCACTACATAA ICTCTCACGACITATTCCC ICTACTGCACGACITATTCCC ICTACTGCACCGACITACT ICTCGCCGCACGACITCT ICTCGCCGCCGCGCGCGCGCGCGCC ICTCCACGCCGCGGCGGCG ICTCTCAACGCCGCGGCGGCG ICTCTCCACGCCGCGCGCGCC ICTCCACGCCGCGCGCGCCCCC ICTCCACGCCGCGCGCGCCCCCCCCCCCCCCCCCCCCCCC	AGTTT TACCC CACTTT CATCT TTCTC TCGTG SATCG CGGAC TTGAG	

4. AF291716 sequence has 2566 bp, by alignment analysis of the sequences previously obtained we "now" that region on which we are interested is between the bases 400 and 1900. Thus in the Included Region text box we will write 400,1500, which means that we are interested on primers addressed to the region starting at base 400 and finishing at 1900 (approx. 1500 bp size).

C Prime	er3Plus - Windov	ws Internet Explorer					
00	- 🕰 http://www.bioinf	ormatics.nl/cgi-bin/primer3plus/	primer3plus.cgi		🗸 🛃 prim	ier3	<b>P</b> •
<b>Y!</b> -	🖉 - 🔛 🖌 🖌	uscar 🝷 🏉 Baja el IE7 🔹 🖉	🕞 • 🔮 • 🗍 📮 Gua	rdar en My Web 🔻 Entra	ar 🔹 🖂 Correo Yahoo! 🔹	🚳 Mi Yahoo!	**
🚖 🏟	🕰 Primer3Plus				🔄 • 🖻 • 🖶 •	🕑 Página 🔻 🍈 H	e <u>r</u> ramientas 🕶
Main	General Settings	s Advanced Settings	Internal Oligo	Penalty Weights	Sequence Quality		~
Sequence	Id: Lbra AF291716 hsp	70 P					
Paste sour	rce sequence below	Or upload sequence file:		Examinar	Upload File		
gi 9864 cds	198 gb AF291716.1	AF291716 Leishmania bra	aziliensis heat :	shock protein 70 (	hsp70) gene, comple	te 🔺	
TTAACCC		IGCTTCACATAAAACCCTTTGACA				GTTT	
TATTCGC	TCATAACTCGCTTTCTAT		AATTGTTTCTAACTT	TCACCTACTACTACTCTC	TCTCTTATTCACTACATAA	CTTT	
		CTTTTCTAATTTTCCCTTTCTCC	GTTTCCAGTTTTAAA'		ACCTACTTACAACTTCTCT GATCTGGGCACGACGTACT	TCTC	
		GTGGAGATCATCGCGAACGACCA	GGGCAACCGCACGAC	GCCGTCGTACGTCGCCTT			
ATGAAGC	ACTGGCCCTTCAAGGTGA	CGACGAAGGGTGACGACAAGCCC GGAGGCGGCGGAGGCGTACCTTGC	GTGATCACGGTGCAG	TTCCACGGCGAGGAGAAG	ACCTTCACGCCGGAGGAGG	TGAG	
			CARGCAGGIGHGHA			CHOC	
Mark sel	ected region: <> [] {	[] Clear		Save S	equence		
Excluded	Regions: <		>				
Targets:	]		]				
Included I	Region: {	400,1500	}				

- 5. Now we will click on the General Settings tab and:
  - delete Product Size Ranges
  - set Primer Size Min: 15 and Max:30
  - set Primer Tm Min: 55 Max: 70 Opt: 60
  - set Primer GC% Min: 40 Max: 60
  - Let the other settings as default by the system

Main Gen	eral Settings	Advanced Settings	Internal Oligo	Penalty Weights	Sequence Quality	
oduct Size Range	<u> </u>					
rimer Size	<u>Min:</u> 15	Opt 20	Max: 30			
rimer Tm	Min: 55.0	Opt: 60.0	Max: 70.0	Max Tm Differe	ence: 100.0	
rimer GC%	Min: 40.0	Opt	Max: 60.0	Fix the 5 pr	rime end of the primer	
oncentration of m	onovalent cations:	50.0 <u>An</u>	nealing Oligo Concent	ration: 50.0		
oncentration of di	valent cations:	0.0 <u>Co</u>	ncentration of dNTPs	0.0		
spriming/Repeat I	Library: NONE	~				
	al settings here: D	efault 🗸 🗸	A	utton to load the selected	1	

- 6. Click on the Advanced Settings tab and:
  - set Number tor Return (number of primers pairs) to 10
  - delete the GC Clamp text box
  - Mark Use Product Size Input and ignore Product Size Range
  - set Product Size Min: 1000 Max: 1500
  - let the other settings as default

C Primer3Plus - Window	s Internet Explore	r .					- 6 🛛
GO - Attp://www.bioinfo	rmatics.nl/cgi-bin/primer3plu	s/primer3plus.cgi			🗸 🗲 🗙 prim	er3	P •
Y! - ⊘+Bus	scar 🔻 🏉 Baja el IE7 🔹 🖉	🖞 🗔 🔻 🌰 🕶 🕴 🛱 Guard	dar en My Web 🔻	Entrar 🔻 🖂	Correo Yahoo! 🔹	🎯 Mi Yahoo!	**
😭 🕸 🚨 Primer3Plus				<u>ن</u>	S - 🖶 - 🛛	🕑 Página 🔻 🍈 I	Herramientas 🔻 🂙
			_				^
Primer3Plus			Primer3Ma	iager	<u>Help</u>		
pick primers from a DNA sequen	ce		About		Source Code	2	
	lect primer pairs to detect the s		Optionally targets a	nd Pi	k Primers	ResetForm	
in in	cluded/excluded regions can b	e specified.					
Main General Settings	Advanced Settings	Internal Oligo	Penalty Weigh	its Seq	www.guality		
Max Poly-X:	5 <u>Ta</u>	ble of thermodynamic p	arameters:	Breslauer e	t al. 1986 💌		
Max #N's:	0 <u>Sa</u>	lt correction formula:		Schildkraut	and Lifson 1965	<b>~</b>	
<u>Number To Return:</u>	10 <u>CC</u>	<del>3</del> Clamp:					
Max Self Complementarity:	8.00 <u>M</u>	ax 3' Self Complementa	<u>urity:</u>	3.00			
	Ma	ax 3' Stability:		9.0			
Max Repeat Mispriming:	12.00 <u>Pa</u>	ir Max Repeat Misprim	ing:	24.00			
Max Template Mispriming:	12.00 <u>Pa</u>	ir Max Template Mispr	iming:	24.00			
Left Primer Acronym:	F Int	ernal Oligo Acronym:		IN			
Right Primer Acronym:	R Pri	mer Name Spacer:		_			
Product Tm	Min:	Opt:	Max:				

Use Product Size	e Input and ignore Prod		ge Warni	ng: slow and expensive!	
Product Size	Mi	n: 1000	Opt:	Max: 1500	
✓ Liberal Base	-		libraries as consensus	Use Lowerca	0
Sequencing					
<u>Lead</u>	Bp:	50	Spacing	Bp:	500
Accuracy	Bp:	20	Interval	Bp:	250
Pick Reverse Primer	<u>s</u>				

## 7. Click on Pick Primers.

You will obtain detailed information for the best 10 primer pairs selected by Primer3.

Carefully check which pair is most suitable for your objectives and keep in mind that you have to test their performance on the bench!!

✓ Left Primer 1:	Lbra AF291716 hsp70	PCR gi 9864198 gb AF				
Sequence:	GGTATTGATCTG	GGCACGAC				
Start: 596	Length: 20 bp	Tm: 60.3 °C	GC: 55.0 %	ANY: 4.0	SELF: 1.0	
Right Primer 1:	Lbra AF291716 hsp7	0 PCR gi 9864198 gb AF				
Sequence:	CTCCGTCTGCTT	GCTCTTTC				
Start: 1750	Length: 20 bp	Tm: 60.3 °C	GC: 55.0 %	ANY: 2.0	SELF: 0.0	
Product Size: 1155	bp	Pair Any: 4.0	Pair End: 1.0			

#### Protocol 5.4 Phylogeny

**Purpose** To construct and interprete phylogenetic trees

#### A. INTRODUCTION

Reconstructing evolution is the primary goal of any phylogenetic analysis. Methods are numerous, and interpretation of resulting dendrograms is not trivial. This session aims at construction and interpretation of phylogenies based on the previously built HSP70 and ITS1 alignments, by using the basic and most common algorithms. Trees obtained for these two targets will be compared.

Sequences used: HSP70 and ITS1 alignments from previous workshop sessions

Programs used: MEGA (www. megasoftware.net)

Programs not used but excellent: PHYLIP

(evolution.genetics.washington.edu/phylip.html)

#### Method: Aligning hsp70 nucleotide sequences (joint exercise)

- 1. Start MEGA 4.0
- 2. Open your HSP70 alignment in MEGA format from the File menu.
- 3. Explore the tabs

4.

M4: Sequence Data Explor	rer											
Data Display Highlight Statis	stic	s	Hel	P								
🕞 😼 😻 👯 Color C	v	Pi	S		0	2	4		U	UC L	Phe	
<u>.                                    </u>						_						
	С	G	A	С	т	С	G	С	A	G	С	G
✓ 029-ZAV L. guyanensis ✓ sidon ECGuy3 1 guy	C C	G	A	C C	T T	C C	G G	C C	A	G	C C	G

- 5. to familiarize yourself with data viewing, defining groups and reading frames.
- 6. Define the two *Leishmania* subgenera *L. (Leishmania)* and *L. (Viannia)* and add the respective sequences to these groups.
- 7. Make a sequence selection of the complete sequences, unchecking the boxes from the partial sequences.
- 8. Explore the options in the Display menu

Data	Display	Highlight	Statistics	Help									
8	Shou	w Only Sele	cted Seque	nces		2	4		υų	C • Pł			
_	Use	Identical Sy	/mbol		E			-			=		
029-:		r Cells				С	Т	С	G	С	А	G	С
🖌 sidor	Sort	Sequences				С	т	С	G	С	А	G	С
🖌 LBM	CA Dore	ore Input (	, ~~		A	С	т	С	G	С	А	G	С
🖌 ITM#	Resi	ore input (	Jruer		A	С	т	С	G	С	А	G	С
✔ M41	Shou	w Sequence	e Names		A	С	т	С	G	С	А	G	С
🗸 geau	Show	« Group Na	mes		A	С	т	С	G	С	А	G	С
LH2					-	С	т	С	G	С	А	G	С
✓ sidor	Cha	nge Font			A	С	т	С	G	С	А	G	с

- 9. Identify parsimony informative sites after theoretical introduction on maximum parsimony.
- 10. Close the data viewer.
- 11. From the Phylogeny menu, select maximum parsimony trees:

	MM	EGA 4									
	File	Data	Distances	Phylogeny	Pattern	Selection	Alignme	ent	Windows	Help	
		🔊 I	1 😼 👽	Constru	ct Phyloge	hylogeny		🕩 👯 Neighbo		-Joining (NJ)	
					ap Test of	Phylogeny		• 3	∑ð; Minimum	Evolution (ME)	1
Tutorial on How to Use		1 Interior branch resc or Phylogeny			١	🔤 Maximum	Parsimony (MP).				
	Click	me to a	activate a da	🖽 Relative	D-1- T			. 6	E UPGMA		
	Citing	MEG	A in publica	E% Relative	Rate les	ts		1			
	<u>Go to</u>	the M	EGA web p	Display Saved Tree Session							
				Display	Newick Tre	ees from Fil	e				

12. Use the following options and Compute:

otide (Coding) geny reconstruction
geny reconstruction
num Parsimony
trap (100 replicates; seed=64864)
nini Branch-and-bound .
lete Deletion
nd+3rd+Noncoding

- 13. Wait a bit...
- 14. And some more...
- 15. Have a coffee...
- 16. Have another one ...
- 17. Go powder your nose ...
- 18. Be patient...
- 19. Go out for lunch...
- 20. Stop and abort the calculation.
- 21. Try again with the following options:

Option	Selection
Data Type	Nucleotide (Coding)
Analysis	Phylogeny reconstruction
Tree Inference	
->Method	Maximum Parsimony
->Phylogeny Test and options	None
->Search Options	CNI (level = 3) with initial tree by Random addition (50 reps)
Include Sites	-
->Gaps/Missing Data	Complete Deletion
->Codon Positions	1st+2nd+3rd+Noncoding

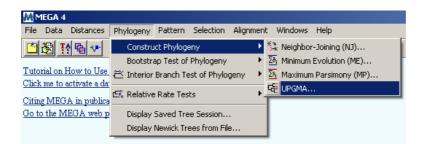
- 22. Look at the different output trees and compute a consensus. Play around with the different view options. Close the Tree explorer.
- 23. From the Distance menu, select Compute pairwise.
- 24. Compute with the following options:

Option	Selection
Data Type	Nucleotide (Coding)
Analysis	Pairwise distance calculation
->Compute	Distances only
Include Sites	
->Gaps/Missing Data	Complete Deletion
->Codon Positions	1st+2nd+3rd+Noncoding
Substitution Model	
->Model	Nucleotide: Number of differences
->Substitutions to Include	d: Transitions + Transversions
->Pattern among Lineages	Same (Homogeneous)
->Rates among sites	Uniform rates

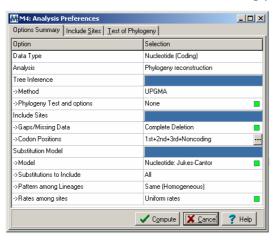
25. Repeat the calculation with the following options:

Option	Selection
Data Type	Nucleotide (Coding)
Analysis	Pairwise distance calculation
->Compute	Distances only
Include Sites	
->Gaps/Missing Data	Complete Deletion
->Codon Positions	1st+2nd+3rd+Noncoding
Substitution Model	
->Model	Nucleotide: Jukes-Cantor
->Substitutions to Include	All
->Pattern among Lineages	Same (Homogeneous)
->Rates among sites	Uniform rates

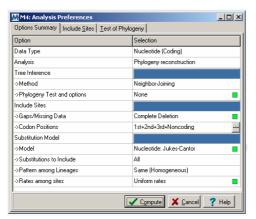
26. From the Phylogeny menu, select the following option:



27. Calculate a UPGMA tree with the following parameters:



- 28. Observe the equal distance of all taxa to the root.
- 29. Calculate a Neighbor-Joining tree:



- 30. Observe the unequal distance to the root of all taxa.
- 31. Re-root the tree on a branch:



32. Calculate a Minimum Evolution tree:

M4: Analysis Preferences	
Options Summary Include <u>S</u> ites <u>M</u> E	Tree Options I lest of Phylogeny
Option	Selection
Data Type	Nucleotide (Coding)
Analysis	Phylogeny reconstruction
Tree Inference	
->Method	Minimum Evolution
->Phylogeny Test and options	None
->Search Options	CNI (level = 2) with initial tree = NJ MaxTrees = 50
Include Sites	
->Gaps/Missing Data	Complete Deletion
->Codon Positions	1st+2nd+3rd+Noncoding
Substitution Model	
->Model	Nucleotide: Jukes-Cantor
->Substitutions to Include	All
->Pattern among Lineages	Same (Homogeneous)
->Rates among sites	Uniform rates

- 33. As for Maximum Parsimony, different topologies are obtained, and consensus trees can be built.
- 34. Build a Neighbor-Joining tree with the bootstrap option (you can pick a different seed number):

f Phylogeny
Selection
Nucleotide (Coding)
Phylogeny reconstruction
Neighbor-Joining
Bootstrap (100 replicates, seed=32156)
Complete Deletion
1st+2nd+3rd+Noncoding
Nucleotide: Jukes-Cantor
All
Same (Homogeneous)
Uniform rates

35. Re-root the tree and observe what happens with the bootstrap values.

#### Additional remarks:

- **5**. A lot of different tree building options are available in MEGA, using different substitution models, tree construction algorithms, and confidence tests.
- **6**. Trees can be built using different site selections, such as synonymous versus nonsynonymous, with or without gaps, different codon positions, and also on the basis of protein sequences.
- **7**. The better the chosen model describes the actual evolutionary mechanisms, the more reliable your phylogeny can be. It is, however, difficult to go back in time.

- **8**. Beware of over-interpretation of confidence levels such as obtained with bootstrap: they only tell you something about the consistency in your data set with the method used, they don't tell you whether you correctly captured evolution.
- **9**. The starting point of any analysis is a decent alignment.
- 10. Phylogenies start from the assumption that evolution is divergent, and works through a series of bifurcations from a common ancestor by accumulation of mutations. Unfortunately, evolution does not always treat us kindly, and the history of an organisms genome includes also horizontal gene transfers, duplications, DNA transpositions, and filtering by natural selection. It is essential to realize that the phylogeny of a piece of DNA or protein is not necessarily identical to the phylogeny of the organism it is derived from.

## Exercises in individual groups:

- **4.** Using your HSP70 alignment, build trees from protein sequences and synonymous sites only. Compare these with the trees from the joint exercise.
- 5. Build phylogenies from the ITS1 alignment, and compare these with the HSP70 phylogeny. In this alignment, different gap treating may have a big impact as they are numerous. As these sequences are non-coding, protein alignment is not possible to aid in identification of homologous sites.
- **6**. From both HSP70 and ITS1 phylogenies, determine the species and if possible subspecies of the unknown sequences you received for the *in silico* analysis.

## 6. Multilocus sequence typing (MLST)

Protocol 6.1 Analysis of MLST data

 Purpose
 To type Leishmania species using the sequences of multiple housekeeping genes.

#### A. Introduction:

Multilocus sequence typing (MLST) is used primarily to type bacteria in the format of sequencing about 500bp of an average of 10 single copy housekeeping genes (Maiden et al 1998; Spratt 1999). You can find more information and databases for current MLST systems at http://www.mlst.net/. The same methodology has been adapted to some pathogenic fungal species and now for Leishmania (Mauricio et al 2006; Zemanova et al 2006). In the system being developed we have so far five genes that code for enzymes often used for enzyme electrophoresis typing, and which are fully sequenced (from 1000 to 2000 bp). Genes are amplified by PCR using primers external to the coding region and sequenced using internal primers. This methodology has several advantages over other methods, as well as drawbacks. Compared with enzyme electrophoresis, for example, it has a greatly increased discriminatory power, it can be applied to biological samples instead of cultures, it does not require reference strains and it is portable (comparable between labs and through databases). The last two features are also an advantage over PCR-RFLPs. MLST is less sensitive and has less discriminatory power than microsatellite typing, but it is more easily applicable across different species of Leishmania and to unknown samples. Its analysis is also potentially more straightforward. MLST has so far been used to identify genetic groups in the L. donovani complex and recombination.

The MLST system for the sub-genus *L.* (*Viannia*) is on its final stages of development. However, 5 genes have already been fully tested and chosen to integrate the fugure 10-15 system. These are: *MPI*, *ACON*, *NH2*, *MDH* and *G6PDH*. Other genes being tested are: *ASAT*, *NH2*, *C20070*, *C320290*, *GPI*, *NH1*, *ICD*, *PGM*, *6PG*, *ME* and *HK*. The final systems and conditions will be published shortly.

Technically, MLST requires knowledge of PCR and DNA sequencing, which have been covered elsewhere.

Here, participants will practice MLST analysis on 3 gene sequences that have been published for the *L. donovani* complex, which includes *L. infantum* (syn. *L. chagasi*).

#### **B. MATERIALS:**

Samples:Edited chromatograms for 10 selected strains of the *L. donovani* complex.For PCR conditions and primers for the full 10 gene MLST system for the *L. donovani* complex, see references Mauricio et al (2006) and Zemanova el al (2007).

#### Equipment and software:

Computer with internet access BioEdit (www.mbio.ncsu.edu/BioEdit/BioEdit.html) MEGA (<u>www.megasoftware.net</u>) SplitsTree4 (<u>http://www.splitstree.org/</u>) PHASE (http://stephenslab.uchicago.edu/software.html) Text editor program (Word, for example)

Spreadsheet program (Excel, for example)

#### C. METHODS

All programs will be available on the Desktop, as well as a folder with the initial data files.

# a) Assemble consensus sequence for each gene and for each sample from partial sequences obtained from internal primers.

<u>Step 1</u> – Create alignment file - Open BioEdit Sequence Alignment Editor. From the File menu, open New Alignment (or Control+N). Save it in your chosen name (Note: keep saving after each main step).

<u>Step 2</u> – Import relevant raw sequences - From the File menu, open Import, then Sequence Alignment File. Select all the electropherograms for this alignment. You may need to select All files or Abi formats

Step 3 – Clean sequences - Remove stretches of NNNN from the beginning and the end.

<u>Step 4</u> – Assemble sequences - Select sequences by clicking on their names on the left hand side. From the Accessory Application menu, choose CAP Contig Assembly Program. In the window click Run Application button (you can change the settings first). You may need to press Enter once the program is finished (it reads SHOW on the command prompt window). If the alignment is successful you should obtain a new file (window) with the sequences aligned and an additional sequence called "Contig-0". Rename the consensus sequence with a short easily identifiable name.

Alternative: if you are not able to obtain a single contig sequence with this method, import a reference sequence for the entire gene sequence (from GeneBank or GeneDB or previous alignments), select the reference strain and one of the raw sequences then from the Accessory Application menu, choose ClustalW Multiple Alignment. You may need to reverse some sequences prior to alignment (Select sequence name, go to the Sequence menu, choose Nucleic Acid, then Reverse Complement; or Shift+Control+R). Repeat the process for all raw sequences. You can copy the aligned raw sequence

from one window to the other with Control+F8, then paste with Control+F9. Once all aligned sequences are imported to the alignment file, you can create a consensus sequence by copying and pasting one (Select sequence name, then Control+C, then Control+V) and then adding the missing regions from the other aligned sequences.

<u>Step 5</u> – Check sequences for anomalies: gaps, misalignments, incorrect multibase codes (a glitch does not reverse multibase codes in CAP). If necessary go back to the chromatograms.

Note: Instructions on how to edit raw sequences are given elsewhere.

# b) Align consensus sequences from all samples for each gene and against sequences retrieved from GeneBank/EMBL

<u>Step 1</u> - Create alignment file - Open BioEdit Sequence Alignment Editor. From the File menu, open New Alignment (or Control+N). Save it in your chosen name (Note: keep saving after each main step).

<u>Step 2</u> – Import relevant consensus sequences. – One way is to copy them from each individual alignment file, as they are finished by copy the aligned raw sequence from one window to the other with Control+F8, then paste with Control+F9. Alternatively you can copy a sequence from the file of origin by going to the Edit menu, then Copy Sequences to Clipboard (FASTA format), and on the destination file by going to the File menu, then Import from Clipboard.

<u>Step 3</u> – Align sequences – Select the names for all sequences, then from the Accessory Application menu, choose ClustalW Multiple Alignment.

<u>Step 4</u> – Select coding region (remove primer sequences) – For the majority of targets, the PCR primers are outside the coding region, so by selecting the coding region their sequences are automatically eliminated. From the Edit menu, choose Search then Find Next ORF. The open reading frame (ORF, or coding region) will be highlighted in black. You can vertically select the regions outside and press the Delete button. If you need to select primer sequences, you can go to the Edit menu, choose Search then Find (Control+F).

#### c) Concatenate the sequences for all samples for all genes

<u>Step 1</u> – All alignment files for each target should have the same samples and in the same order.

<u>Step 2</u> - Save one of the alignment files with a new name. Go to the File menu, then choose Append Alignment. Choose the alignment file for the target you want to add.

<u>Step 3</u> – Repeat procedure for each target at a time until all targets have been added.

#### d) Identify SNPs

Step 1 – Make sure all alignment files from BioEdit were saved in FASTA format.

Step 2 – Open MEGA, then from the File menu, choose Open Data (F5). Select the file.

<u>Step 3</u> – Convert FASTA file to MEGA format – From the Utilities Menu on the window, choose Convert to MEGA format (Control+M). Press OK on the next two windows. You should now have a new tab with the alignment in MEGA format. Save the file. Close windows.

<u>Step 4</u> – Open MEGA format file – From the File menu, choose Open Data (F5). Select the file. On the following windows, click OK/Yes for default. It should be: 1) Nucleotide sequences, 2) Protein Coding

sequence data, 3) Standard Genetic Code. You should now have a window with an alignment view.

<u>Step 5</u> – Identify SNPs – From the Highlight menu, choose Variable sites (or the V button on the top toolbar).

<u>Step 6</u> – Create SNP only file – From the Data menu, choose Export Data. Make sure to select Only Highlighted Sites (at the bottom of the window). It is also advisable to choose the option of writing site numbers for each site, so that you have a record of the location of each SNP. You can select different file types.

# e) Determine phase of heterozygous strains, using the program PHASE (Stephens et al 2001)

This program implements a Bayesian statistical method for reconstructing haplotypes from population genotype data. You should be aware that it provides a probability result rather than the actual phase for sequencing or microsatellite data, so you should be very careful when interpreting results. However, it is very useful if physical linkage between SNPs or microsatellites is impossible to determine.

<u>Step 1</u> – Open an MS-DOS window. Go to the Windows Start Menu, then choose Run, type COMMAND in the prompt and press OK. Alternatively, also from the Start Menu, choose Command Prompt (probably in the Programs, then Accessories menus.

<u>Step 2</u> – Your PHASE program should be saved in the PHASE directory in your C:\ drive. To get there, you may need to write "cd.." at the prompt until you only see C:\. Then type "cd PHASE". You should obtain C:\PHASE\

<u>Step 3</u> – Prepare the input file. Open the file you prepared with SNPs only from your sequence data for one of the gene targets. Prepare an input file in the following format:

General description: NumberOfIndividuals NumberOfLoci P Position(1) Position(2) Position(NumberOfLoci) LocusType(1) LocusType(2) LocusType(NumberOfLoci) ID(1) Genotype(1) ID(2) Genotype(2) ID(NumberOfIndividuals)] Genotype(NumberOfIndividua ls)	Example: 3 5 P 300 1313 1500 2023 5635 MSSSM #1 12 1 0 1 3 11 0 1 0 3 #2 12 1 1 1 2 12 0 0 0 3 #3 -1 ? 0 0 2 -1 ? 1 1 13	LocusType(i): (a) S for a biallelic (SNP) locus, or biallelic site in sequence data. (b) M for microsatellite, or other multi- allelic locus (eg tri-allelic SNP, or HLA allele). The default is that this denotes a microsatellite locus with stepwise mutation mechanism. Genotype (i) for the ith individual. Given on two rows: one allele on the 1st row, and one on the 2nd row. It does not matter which allele is entered on each row. For biallelic loci, any two characters (e.g. A/C, G/T, 0/1) can be used to represent the two alleles, and they do not need to be separated by a space. Missing alleles at SNP loci should be entered as ?. For multiallelic loci a
		positive integer must be used for each allele (number of repeats at microsatellite loci), and data for each locus should be separated by a space. Missing alleles at multiallelic loci should be represented by -1.

In this case, your data should all be coded as S (for a biallelic locus) and you should code it as A, C, G or T.

<u>Step 4</u> – Run PHASE. At the DOS prompt, type PHASE <filename.inp> <filename.out> <number of iterations> <thinning into interval> <burn in>

For example for the gene ASAT you may prepare an ASAT.inp file, then type:

PHASE ASAT.inp ASAT.out 100 1 100 (these 3 last numbers are the default parameters, so you can miss them out)

As it runs, it keeps you informed, until you obtain a new PHASE prompt.

<u>Step 5</u> – Results. You should obtain a summary results file, with the name that you specified.

In the output file, you should obtain a recreation of the input file with the estimated phase of the alleles. Uncertain phase are indicated by (), wheras uncertain genotypes are indicated by [], with p (for phase) and q (for genotypes) at the default value of 0.9. The list of probabilities for each uncertain phase call are listed at the end of the output file.

You also obtain 6 other output files with more detailed results.

<u>Step 6</u> – Evaluate results. You can check the .monitor output file to examine the goodness-of-fit and the .freqs output file to check that the estimates accross each run are consistent. You can test the consistency of your results by increasing the number of iterations or the thinning interval. If the different runs keep giving different results, you can select the results from the run with the highest <u>average</u> value for the goodness of fit. For the purpose of this exercise, you can try another run with a higher number of iterations and compare with the results from the first run.

When you are happy with the results you can use the estimated phase to code your diplotype data as haplotypes for the next analyses, particularly for bifurcating trees.

#### f) Code diplotypes

<u>Step 1</u> – Go to the website http://linux.mlst.net/nrdb/nrdb.htm. This opens the online program NRBD that selects non-redundant sequences.

<u>Step 2</u> - Paste full alignment in FASTA or MEGA format. Click the Submit button. You should obtain an alignment of unique sequences with concatenated titles.

<u>Step 3</u> - You should assign a number or letter for each unique sequence. Heterozygous sequences should be given a separate code. Where phase can be determined for heterozygous strains you can assign two codes.

<u>Step 4</u> – Prepare a table listing each diplotype per sample and per target.

<u>Step 5</u> – Assign an MLST code for each unique combined diplotype.

## g) Conduct network analysis of sequences on the program SplitsTree4 (Huson & Bryant, 2006)

<u>Step 1</u>. Make sure you have a file with aligned non-redundant sequences. SplitsTree4 accepts several formats (FASTA, Nexus, PHYLIP, Clustal, for example, as well as distance files).

<u>Step 2</u>. Open SplitsTree4. From the File menu, select Open. Choose your file and click Open. The program automatically produces a network. At the bottom of the window you should find information about the network produced. The default is likely to be a SplitsTree.

<u>Step 3</u> – Choose distance measure. From the Distances menu, you can choose different statistics. The default will be uncorrected P. Other distances incorporate factors to produce more realistic models. However, the K2P (Kimura-2-Parameter) is very robust and widely used. Try this one first. Keep the default settings shown on the new window and click the Apply button.

<u>Step 4</u> – Choose Network building method. From the Networks menu, choose Neighbor-Net if not already selected, or try another method for comparison. Keep default settings and click the Apply button.

## h) Manipulate and analyze networks and trees. Explore different analyses and presentations.

<u>Step 1</u>. Produce bifurcating trees. From the menu Trees, choose a tree building method. Neighbor-Joining (NJ) is recommended.

<u>Step 2</u>. Modify presentation of the trees and networks. From the Draw menu select a presentation style. For example, Phylogram for the NJ tree.

<u>Step 3</u>. Modify view settings for the trees and networks. From the View menu select, for example, Rotate Right and then Left. You can also use the arrows keys on your keyboard: Right and Left for rotating, Down for zooming in and Up for zooming out. You can also click on branches, which become highlighted in red. Then click on a node for that branch to move it around. If you click on a branch that forms part of a split, you'll notice that all splits will be highlighted. You can move splits and sequence names to make the network easier to view.

<u>Step 4</u>. Remove sequences from tree/network. You may want to focus on a smaller group. From the menu Data, select Filter Taxa. On the left hand side, a box lists shown taxa. Select taxa you want to remove and click the button Hide in the centre. The sequence should appear on the right hand box. Click the button Apply.

Note: you can save trees and networks by going to the menu File, then selecting Export Image.

#### i) Analyze networks for evidence of recombination.

<u>Step 1</u>. Notice presence of splits. A large number of splits suggests recombination. Either within the entire group or within smaller groups.

<u>Step 2</u>. Notice where taxa appear. Taxa at the end of long branches suggest absence of recombination, whereas taxa sitting on intermediate branches are putative hybrids. Try to identify instances of each of these cases in your dataset.

<u>Step 3</u> – Test for recombination. From the Analysis menu, chose Conduct Phi Test of Recombination. You can do the analysis for the entire dataset and for subsets.

#### **REFERENCES:**

Huson D.H. & Bryant D. (2006) Application of Phylogenetic Networks in Evolutionary Studies, *Mol. Biol. Evol.*, **23**:254-267,

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant,

D.A., Feavers, I.M., Achtman, M., & Spratt, B.G. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. USA*, **95**, 3140-3145.

Mauricio IL, Yeo M, Baghaei M, Doto D, Pratlong F, Zemanova E, Dedet J-P, Lukes J, Miles MA (2006) Towards multilocus sequence typing in the *Leishmania donovani* complex: resolving genotypes and haplotypes for five polymorphic metabolic enzymes (ASAT, GPI, NH1, NH2, PGD). *Int. J. Parasitol.* **36**:757-69.

Spratt, B.G.Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the Internet. Curr. Opin. Microbiol. **2**, 312-316, 1999.

Stephens, M., Smith, N., & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Human Genetics*, **68**, 978-989

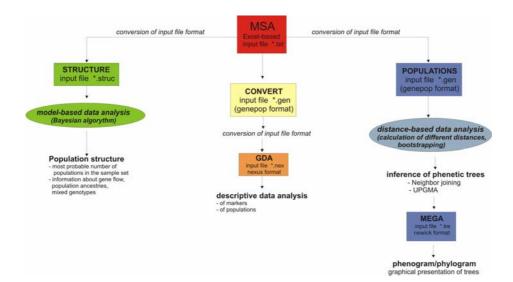
Zemanova E, Jirku M, Mauricio IL, Horak A, Miles MA & Lukes J. (2006) Analysis of genetic polymorphism of the *Leishmania donovani* complex in five metabolic enzymes. *Int. J. Parasitol.* **37**: 149-160

## 7. Analysis of MLMT data

Protocol 7.1	Population genetic analysis of Leishmania strains based on MLMT data
Purpose	To construct phylogenetic trees based on microsatellite distance measures, to assign individuals to population by using a Bayesian statistics-based approach, to estimate the genetic isolation of the populations defined and to perform a descriptive analysis for the populations and the microsatellite loci.

## A. Introduction:

Microsatellite profiles consisting of the repeat numbers for the different microsatellite loci that have been assembled for every strain under study will analysed by distance based methods and by models based on Bayesian statistics. Microsatellite genetic distances based on the "Chord distance" or "proportion of shared alleles" measures will be calculated using the program POPULATIONS and distance trees will be constructed using MEGA4. The Bayesian statistics-based method implemented in STRUCTURE uses patterns of allele frequencies for identification of distinct subpopulations and determines fractions of the genotype for each strain that belong to each subpopulation. It was shown to accurately infer individual ancestries and to provide information on population relationships and history. Fstatistics which allows for testing for the genetic variation among populations is calculated by MSA software and descriptive analysis (mean number of alleles, observed and expected heterozygosity) of microsatellite loci and of the different populations obtained is performed by GDA.



# Method A. Preparation of data and output formats using the MSA software package.

**MSA (Microsatellite Analyzer)** provides a variety of data and format outputs which are useful for analysis of microsatellite data.

Prepare an excel file with your data as following:

Two format column							
2			2*	2	2	2	Repeat size
			92*	92	62	62	Length of flanking region
			Lm2TG*	Lm2TG	TubCA	TubCA	Locus names
DON1	d	1	110*	110	82	82	Fragment
DON39	d	1	110	110	82	82	size
DON45	d	1	110	110	82	82	]
DON28	d	1	110	110	82	82	1
DON53	d	1	110	110	82	82	1
DON56	d	1	110	110	82	82	1

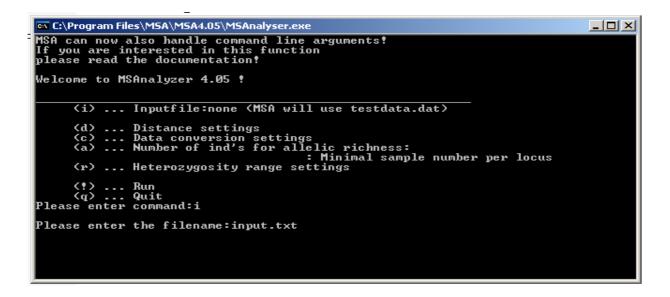
<u>\*Hint:</u>

Marker Lm2TG contains a dinucleotide repeat (2), a the flanking region of the following has 92 bp. If the fragment size is 110 which equals to ((repeat numbers x 2) + flanking region); the repeat number in this example is 9. d=outbred individual (h=inbred), 1=group number, here we consider all individuals as belonging to one group. Missing data are indicated by empty cells, -1, nd, etc.

Convert the excel file into text file format (\*.txt) before starting MSA.

Copy your input file into the folder of the MSA program, then press the MSA starting

icon MSAnalyser.exe to get the following starting screen:



Type the command (i) and enter the name of your input file as shown. Run (!), then close the command line window and go to the MSA folder where the output folder named

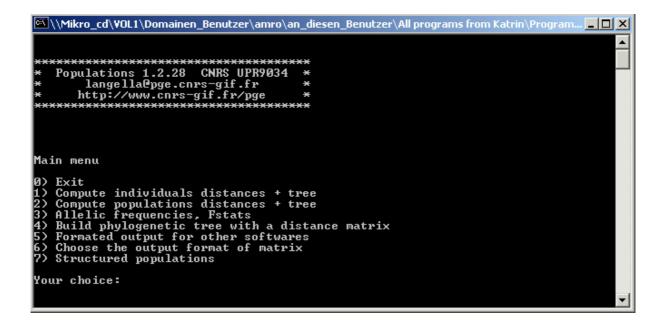
is found This folder which contains many **input files** for other programs and results **as allele counts and allele frequencies**.

#### Method B. Construction of distance matrices using POPULATIONS

Open the MSA folder input.txt\_MSAresult00 and select the following files:

🚞 Formats&Data 🔜 🚬 🗒 Genepop.gen

To start the program POPULATIONS open the program folder and copy the genepop.gen file to it. Open the command line window using populations.exe :



type (1) for computing individual distances, press enter. Type the name of the input file (in this example genepop.gen), press enter to get to this screen:

Mikro\_cd/VOL1\Domainen\_Benutzer\amro\an\_diesen\_Benutzer\All programs from Katrin\Program...
X
1) Compute individuals distances + tree
2) Compute populations distances + tree
2) Allelic frequencies. Pstats
4) Build phylogenetic tree with a distance matrix
5) Formated output for other softwares
6) Choose the output format of matrix
7) Structured populations
Your choice:
1
Name of input file (Populations or Genepop format) ?
genepop.gen
8) back
1) Get distance matrix between individuals
2) Phylogenetic tree of individuals with bootstraps on locus
4) Choose locus to compute distances (default: all)
Your choice:

Type (2) to calculate the phylogenetic tree of individuals without bootstrap, or (3) for the phylogenic tree of individuals with bootstraps on locus. Enter and choose on the following screen (3) for Chord distance Cavalli-Sforza and Edwards, Dc, or (5) for "proportion of shared alleles" distance measure DAS (also called Dps). Here we select the Dc measure. Enter and select either (1) UPGMA or (2) Neighbor Joining for tree calculations in the next screen. We will calculate a neighbor-joining tree in this session. If the bootstrap option is used it is needed to type the number bootstraps wanted. In this example we selected 100.

Type an output file name. The tree calculation process will take about 15 seconds depending on to your sample size.

Close the command line screen and go to the program folder of **POPULATIONS** where the output file can be found. Open this file using the program **EDITOR** and delete the lines before the brackets as shown in the following screen:

🌌 ahmad+mohammad - Editor 📃	
Datei Bearbeiten Format ?	
[108 inidividuals, 108 populations\DAS, 100 bootstraps on locus\#G 'POP_1' 'red' '48'\#G 'POP_2' 'gre 'brown2' '29'\#G 'POP_38' 'nawy blue' '30'\#G 'POP_39' 'goldenrod' '31'\#G 'POP_40' 'tomato' '26'\#G 'PO P_76' 'brown1' '103'\#G 'POP_77' 'aquamarine' '104'\#G 'POP_78' 'orange' '105'\#G 'POP_79' 'brown2'	)P_(
((((((((((((((((((((((((((((((((((((((	0N
-10,DN133:-3.72464e-10)6:1.08635e-10,DN159:-1.08635e-10)6:3.25906e-11,BD26:-3.25906e-11)6:	

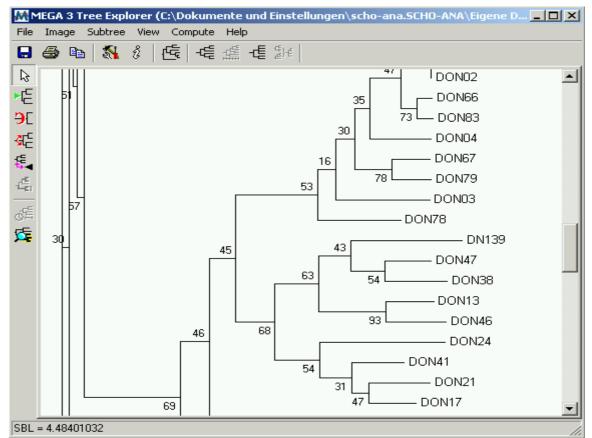
Then save the file as text file and switch to the program **MEGA 4** in order to draw your tree.

## Method C: Drawing distance trees using MEGA 4

Open <u>MEGA 4</u> and select 'Phylogeny', then select the second option 'Display Newick Trees from File':

Molecular Evolutionary Genetics Analysis, Version 3.1	
File Phylogeny Alignment Windows Help	
Display Saved Tree Session	
Display Newick Trees from File	
Tutonal on How to Use WIEGA	
<u>Click me to activate a data file</u>	
Citing MEGA in publications	
Go to the MEGA web page	
TATCH	
ATGCTACGTA GATGCTT	
15:08:39	
י ההחודה כי די	b

Go to the folder of **POPULATIONS** and select your outputfile. The tree will appear as follows:



MEGA 4 allows for rooting the trees and changing the format, colors, appearance etc. by pressing the small icons at the top or at the left side of the screen. For further graphical applications and changes you can export the tree as graphic in \*.emf (enhanced metafile) format.

### Method D: Inference of population structure using STRUCTURE

The program **STRUCTURE 2.1** is used for studying the nature and extent of genetic variation within and between populations.

To obtain the **STRUCTURE** input file open the MSA folder again and got to the folder input.txt\_MSAresult00. There you will find a folder Formats&Data from where you can select the \*.struct file which should look as follows:

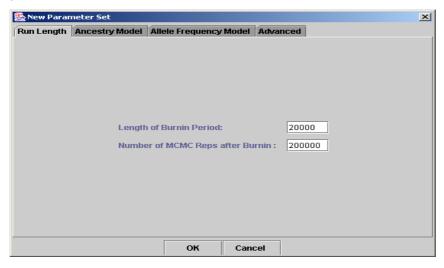
Date         Bearbeiten         Format         Ansicht         2           DON-01000         110         82         67         92         74         102         117         95         103         80         90           DON-01000         110         82         67         92         74         102         115         95         103         80         90           DON-39000         110         82         67         92         74         102         115         95         103         80         90           DON-39000         110         82         67         92         74         102         115         95         103         80         90           DON-40000         110         82         67         92         74         102         115         95         103         80         90           DON-45000         110         82         67         92         74         102         115         95         103         80         90           DON-51000         110         82         67         92         74         102         115         95         103         80         90         DON-5200
DON-01000          110         82         67         92         74         102         117         95         103         80         90           DON-39000         110         82         67         92         74         102         115         95         103         80         90           DON-39000         110         82         67         92         74         102         115         95         103         80         90           DON-40000         110         82         67         92         74         102         115         95         103         80         90           DON-40000         110         82         67         92         74         102         115         95         103         80         90           DON-45000         110         82         67         92         74         102         115         95         103         80         90           DON-51000         110         82         67         92         74         102         115         95         103         80         90           DON-52000         110         82         67         92         74         102<

Open **STRUCTURE 2.1**, select '**File**' and then '**New Project**'. Enter the name of your new project and select your input file. Click '**Next**' and enter the number of individuals, the ploidy (**2**), the number of loci and the code for missing values (-9) as shown on this screen:

🌺 Step 2 of 4 - Project Wizard	X
Step 2 of 4: Information of input data set	
Number of individuals:	67
Ploidy of data:	2
Number of loci:	15
Missing data value:	-9
Show data file for	rmat
< <back next="">&gt;</back>	Cancel

Click 'Next'. In the '*step 3 screen'* don't add anything. Click 'Next' again to open the '*step 4 screen'* where you should select Individual ID for each individual, then press Finish to get to a new screen and there on Proceed.

After this the small screens will disappear and a big table will show up on the main screen of the program. There click on '**Parameter set**' on the upper tool bar and select '**new parameter set**':



Type in the length of burn-in (usually between 10000-20000) and the number of MCMC repetitions after burn-in (usually between 100000-200000). Choose also the appropriate 'Ancestry model', usually 'admixture model'. All other values are default values.

Press on **OK**, and type the new parameter set (20000-200000) on the following small screen:

🎘 Inpu	£		×
୭	Please name th	ie new parameter set	
ă	20000-200000		
	ок	Abbrechen	

Press **OK**, go back to the main tool bar, select '**Project**' and then '**Start Job**'. Define the numbers of K and of iterations. K represents the number of populations that will be simulated and tested. It depends on the nature and origin of organisms tested and usually varies between 1 and 10. For a first analysis use one iteration.

If you wish to calculate delta K (see below) you have to run at least 10 iterations per each K.

Structure Scheduler	×
Select Simulations to Run	
20000-200000	
Set K from 1 to 10	
Number of Iterations: 1	
Start Cancel	

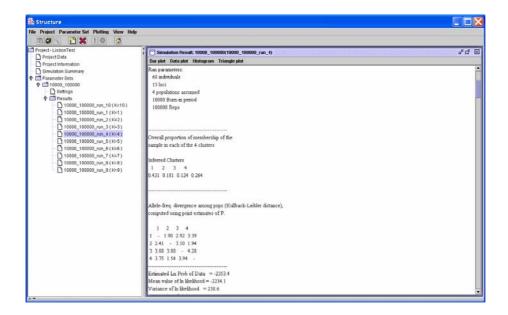
K represents the number of populations that will be simulated and tested. It depends on the nature and origin of organisms tested and usually varies between 1 and 10. For a first analysis use one iteration.

If you wish to calculate delta K,s ee below, you have to run at least 10 iterations per K.

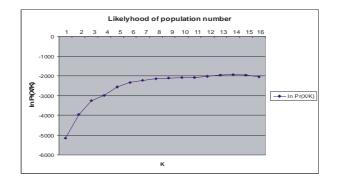
Press '**Start**', the job may take a long time depending on your data size and the number of iterations.

Press 'Start', the job may take a long time depending on your data size and the number of iterations.

To show **STRUCTURE** results for each tested K value click on the result folder for the respective K value in the left part of the screen.



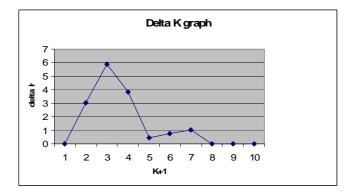
In this example the results for K = 4 are shown, including the values for the estimated ln probability and the mean value of ln likelihood. The latter are used for each K to draw the "likelihood of population number graph" by **EXCEL**. When the derived Gaussian graph reaches a plateau, the value of K captures the major population structure in the data set.



The number of population can be obtained more precisely by calculating the delta ( $\Delta$ ) K (first derivative of K) values according to *Evanno et al. 2005*. For this 10 iterations are needed for each K to calculate Mean and Standard deviation values.  $\Delta K$  is estimated as the mean of the absolute values of L''(K) averaged over 10 runs divided by the standard deviation of L(K), DK = m(|L''(K)|)/s[L(K)], which expands to

 $\Delta K = m(|L(K + 1) - 2 L(K) + L(K - 1)|)/s[L(K)].$ 

For our example (K=4) the following graph will be obtained:

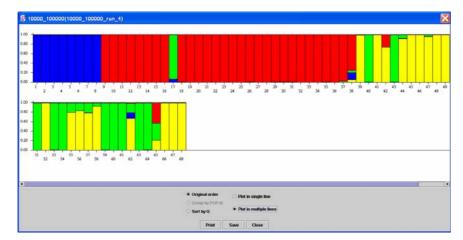


The maximum of the graph shows the most likely number of populations in the data set.

To see the **barplots** for the data analysed click on **`bar plot**' and then on **`show'**.

§ 10000_100000	0(10000_100000_run_4)		Þ
.00 - 1.60 - 1.40 - 1.20 -			_
	Original order	Plot in single line	
	<ul> <li>Group by POP Id</li> <li>Sort by Q</li> </ul>	O Plot in multiple lines	
	Print	Save Close	

(A) Barplots for K = 4 using order according to Q (population assignment) plotted as single lines.



If you want to see the barplots in input order press on the respective key.

(b) Barplots for K = 4 using order according to Q (population assignment) plotted as multiple lines with strain input numbers – assignment of individual strains is possible (see order of numbers in the input file!).

To see the summary of the **STRUCTURE** run click on 'view' and then 'simulation summary':

Presenter     File       Distriction formation       Distriction       Distriction	pect Data		ry of Project LisbonTest										್ರೆ
Parameter Sen		File											
00000_100000_un_110(0010)                10000_10000_un_110(0010)               10000_10000_un_210(00100)               10000_10000_un_210(00100)             10000_10000_un_210(00100)             10000_10000_un_210(00100)             10000_10000_un_210(00100)             10000_10000_un_210(00100)             10000_10000_un_210(00100)             10000_10000_un_210(00100)             10000_10000_un_210(00100)             10000_10000_un_210(00100)             10000_10000_un_210(00100)             10000_10000_un_210(00100)             10000_10000_un_210(0000_un_210(0000)             10000_10000_un_210(0000_un_210(0000)             10000_10000_un_210(0000)             10000_10000_un_210(0000)             10000_10000_un_210(0000)            10000_100000_un_210(0000_un_21							ammary of	Simulations					
■ serings         ■ serings         ■ 1000_1_00000_1.un_1 (0+10)           ■ 1000_1_00000_1.un_1 (0+10)         ■ 1000_1_00000_1.un_2 (0+2)           ■ 1000_1_00000_1.un_2 (0+2)         ■ 1000_1.un_2 (0+2)           ■ 1000_1_00000_1.un_2 (0+2)         ■ 1000_1.un_2 (0+2)           ■ 10000_1.0000_1.un_2 (0+2)         ■ 10000_1.0000_1.un_2 (0+2)           ■ 10000_1.00000_1.un_2 (0+2)         ■ 10000_1.0000_1.un_2 (0+2)           ■ 10000_1.0000_1.un_2 (0+2)         ■ 10000_1.0000_1.un_2 (0+2)           ■ 10000_1.00000_1.un_2 (0+2)         ■ 1000				I IC	Ln P(D)	VarlLnP(D)]	.01	Fst_1	Fs1_2	Fit.3	Fit_4	Fst_5	Fat
<ul> <li></li></ul>		10000_100	10000_100000_run_1	1		43.1	+			+		+	+
10000_100000_un_10 (1010)         10000_100000_un_2 (1010)         10000		10000_100	10000_100000_run_2	2						+		+	+
□ 10000_100000_um_(14(k+10)) □ 10000_10000_um_(14(k+10)) □ 10000_10000_um_(14(k+10)) □ 10000	Results	10000_100	10000_100000_run_3	3								+	-
D 10000_100000_unr,1 (161) 10000_100000_unr,2 (161) 10000000_unr,2 (161) 10000_100000_unr,2 (161) 10000_100000_unr,2 (161) 10000_100000_unr,2 (161) 10000_100000_unr,2 (161) 10000_100000_unr,2 (161) 10000_100000_unr,2 (161) 10000_100000_unr,2 (161) 10000_unr,2 (161) 10000_unr	10000_100000_run_10(K=10)	10000_100	10000_100000_run_4							0.5346		-	
☐ 10000_100000_unu_2(16+2) 10000_100000_unu_2(16+2) 10000_100000_unu_2(16+2) 10000_100000_unu_2(16+2) 10000_100000_unu_2(16+2) 10000_100000_unu_2(16+2) 10000_100000_unu_2(16+2) 10000_100000_unu_2(16+2) 10000_100000_unu_2(16+2)	T 10000 100000 run 1 (H-1)	10000_100	10000_100000_run_5	- 2-									0.4562
□ 10000_100000_mm_3 (1+02) □ 10000_100000_mm_4 (1+02) □ 10000_100000_mm_6 (1+02) □ 100000_100000_mm_6 (1+02) □ 10000_100000_mm_6 (1+02) □ 10000_10000_mm_6 (1+02) □		10000 100	10000 100000 nm 7	2									0.1422
□ 10000_100000_um_4 (1962)           □ 10000_100000_um_4 (1962)           □ 10000_100000_um_4 (1962)           □ 10000_10000_um_4 (1962)           □ 10000_100000_um_4 (1962)           □ 10000_100000_um_6 (1962)		10000 100	10000 100000 run 8	18									0.4301
D 10000_100000_mm_6 10x63 D 10000_100000_mm_6 10x63 D 10000_100000_mm_6 10x75 D 10000_100000_mm_9 (1xx9)		10000_100.	10000_100000_run_9	9	-1695.8	423.0	0.0255	0.5003	0.3119	0.5578	0.1767		0.4295
D 10000_100080_NUL_6 10:09 ) D 10000_100080_NUL_6 10:07 ) D 10000_100080_NUL_6 10:09 ) D 10000_100080_NUL_9 10:09 )		10000_100	10000_100000_run_10	10	-1669.3	482.1	0.0259	0.4806	0.1919	0.3525	0.2608	0.4978	0.5586
· · · · · · · · · · · · · · · · · · ·										_	1		

## Method E. Measurement of Fst values using MSA

*F*statistics allows for testing the genetic variation among populations.

Make an input file by using the **same MSA input file** that has been applied before. The populations as inferred by STRUCTURE should be arranged in the first column of the input file. Indicate the population number in the third column. Save this file as Text file (\*.txt). The input file will look as follows:

<u>D</u> atei <u>B</u> earbeiten	Format	<u>A</u> nsicht	2												
Indial d Indial d Ind	111111111111111112222222223444	2 92 110 110 110 110 110 110 110 110 110 11	2 92 1m2TG 110 110 110 110 110 110 110 110 110 11	2 52 TubCA 82 82 82 82 82 82 82 82 82 82	2 52 TubCA 82 82 82 82 82 82 82 82 82 82	2 55 167 67 67 67 67 67 67 67 67 67 67 67 67 73 3 73 73 71 73 73 71 71 71	2 5 5 67 67 67 67 67 67 67 67 67 67 67 67 67	2 70 B 992 992 992 992 992 992 992 992 992 992	2 70 8 92 92 92 92 92 92 92 92 92 92 92 92 92	2 62 C 74 74 74 74 74 74 74 74 74 74 74 74 74	2 62 C 74 74 74 74 74 74 74 74 74 74 74 74 74	2 68 E 96 96 96 96 96 96 96 96 96 96 96 96 96	2 68 E 96 96 96 96 96 96 96 96 96 96 96 96 96	2 53 F 117 1115 1115 1115 1115 1115 1115 1115	

Copy this input text file to the folder of the MSA program.

Open the <u>MSA</u>nalyser.exe, type (i) and type your input file name in the command line of file name:

```
MC:\Program Files\MSA\MSAnalyser.exe

MSA can now also handle command line arguments?

If you are interested in this function

please read the documentation?

Welcome to MSAnalyzer 4.00 ?

(i) ... Inputfile:none (MSA will use testdata.dat)

(d) ... Distance settings

(c) ... Distance settings

(c) ... Heterozygosity range settings

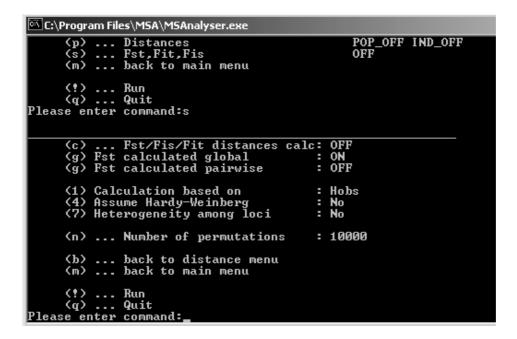
(r) ... Heterozygosity range settings

(?) ... Run

(q) ... Quit

Please enter the filename:MSA_Nabila_ok_61_popsnwithhybrids.txt_
```

Enter, type (d) in the command line and enter:



Type (s) in the command line to activate **(ON)** the Fst, Fit, Fis option, and enter. Then type (c) in the command line and enter, then type (g) and enter again.

Type (b) and press enter, to see that Fst, Fit, Fis are ON. Now type (!) in the command line and press enter to run the program. It will take approximately 30 seconds.

After that close the window and open the output folder, which is deposited after analysis in the MSA program folder as \*.txt\_MSAresult00.

Open the output folder and find the F-Statistic folder:

Now open the 'FST\_WC84-pValue...' data file:

FIS_WC84.txt	1 KB	Textdatei	16.07.2007 11:22
FIT_WC84.txt	1 KB	Textdatei	16.07.2007 11:22
FST_WC84.txt	1 KB	Textdatei	16.07.2007 11:22
💐 FST_WC84-pValu	1 KB	Microsoft Excel-Arb	16.07.2007 11:22
🕷 GlobFst.xls	2 KB	Microsoft Excel-Arb	16.07.2007 11:22

The *F*st-values and *p*-values as shown in the following screen:

	A	В	С	D
1	Fst-values:			
2	pop1MON24	0.000000	0.276766	0.393683
3	pop2MON1	0.276766	0.000000	0.340363
4	pop3MON1	0.393683	0.340363	0.000000
5				
6				
7	P-values:			
8	pop1MON24	0.000000	0.000100	0.000100
9	pop2MON1	0.000300	0.000000	0.000100
10	pop3MON1	0.000300	0.000300	0.000000
11				

The p-value indicates the statistical significance of *F*st- value (p<0.05: significant, p>0.05: not significant).

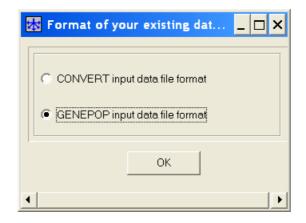
Guideline for interpretation of  $F_{st}$  - values according to Wright (1978):

- < 0.05 little genetic differentiation
- 0.05-0.15 moderate genetic differentiation
- 0.15-0.25 great genetic differentiation
- > 0.25 very great genetic differentiation

## Method F: Descriptive statistics analysis by using GDA

GDA software allows calculation of mean number of alleles (MNA), observed heterozygosity  $(H_o)$  and expected heterozygosity  $(H_e)$ 

As input file use the Genepop.gen file already generated by **MSA** analysis for calculation of *F*st values. This file is located inside the folder of Formats and Data. Convert the Genepop.gen file into the GDA input file which is a nexus file (\*.nex). For this copy of the Genepop.gen file into the folder of the **CONVERT131** program. Open the **CONVERT.exe**, click on file, after that on Load data file, select GENEPOP input data file format and click on **OK**.



Select your file from the **CONVERT131** folder and click on **open** to see the following screen:

Population no.	Last individual in pop	Population name	
1	pop1MON24	Pop001	
2	pop2MON1	Pop002	
3	pop3MON1	Pop003	

Click on OK to open the following screen. Select GDA (nexus format).

Choose desired format for file conversion
GDA (nexus format: *.nex)
© GENEPOP
O ARLEQUIN (*.arp)
C POPGENE (*.dat)
MICROSAT (*.mst)
O PHYLIP (allele frequency infile)
C STRUCTURE (*.str)
O Produce table of allele frequencies
ΟΚ

GDA can also be used to create input files for other programs as shown above.

Click on OK and then on Save in the following screen:

Name of next	us file to be created (input for G	ida)			? ×
Spe <u>i</u> chern	CONVERT131	• +	£	r 🖽	
<ul> <li>■ Nabilapop:</li> <li>■ paperIndia</li> </ul>	sK3nohybrids.nex				
📋 paperIndia	a_DON_IN+SD.nex				
Datei <u>n</u> ame:	NabilapopsK3withhybrids.nex			<u>S</u> peich	hern
Datei <u>t</u> yp:	Nexus files (*.nex)		•	Abbred	chen

A message of successful conversion will be received!

Select the file converted to nexus format from the **CONVERT131** folder and copy it (\*.nex) into the folder of the **GDA** program. Open the file:

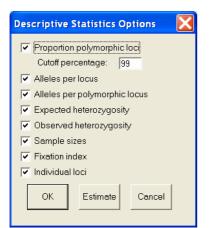
Example of a nexus file:

🛃 NabilapopsK3with	-	nex - Eo	litor									
Datei Bearbeiten For	mat ?											
#nexus												
begin gdadata; dimensions np format missin locusallelela 1 Lm2TG, 2 TubCA, 3 Lm4TA, 4 B, 5 C, 6 E, 7 F, 8 G, 9 P, 10 Q, 11 R, 12 CS20, 13 7031, 14 7039 ;	[!MS/ ops= g=? se bels	A-Out 3   2para	put fro nloci= tor=/;	om fi 14;	ile: MS	A_Nal	bila_ok_	_61_p	opsnohy	/bric	ls.txt	
matrix												
Pop001 : pop1MoN24 pop1MoN24 pop1MoN24 pop1MoN24 pop1MoN24 pop1MoN24 pop1MoN24	7/ 9/ 6/ 8/ 9/	7 9 8 9 9 9	1/ 1/ 1/ 1/	1 1 4 1 1	7/ 7/ 2/ 10/ 7/ 6/	7 7 10 9 10	3/ 3/ 2/ 3/ 3/	wwwww	4/ 4/ 1/ 4/ 4/	4 4 4 4 4	6/ 6/ 25/ 5/	000000

Cave: marker names should have more than one letter otherwise they have to be edited, e.g by adding a second letter!

### The GDA nexus file can also be created manually!

<u>Now run GDA</u> by opening the gda.exe, click on the file menu and after that on open. Select the desired file and open. Go to the file, click on Log to save the file and type a file name. Then go to Descr and option and choose the values to be calculated:



- P proportion of polymorphic loci
- n- number of individuals
- A- alleles per locus (mean number of alleles-MNA)
- Ap- alleles per polymorphic locus
- He- expected heterozygosity (genetic diversity)
- Ho- observed heterozygosity
- f- fixation index

Click on Estimate to start the calculation and go back to the file menu to make the file unlog.

The results of descriptive statistics are shown on the following screen:

Results per population for all I
----------------------------------

Genetic Data	Analysis Hierarchy F-sta	ate Diet Diear	Help		
	nijerarchy r-s <u>u</u>	aus Di <u>s</u> i <u>D</u> isey	Ūdh		
scriptive s	tatistics (by	population)	:		
Population	n	Р	A	Ap	
Pop001 Pop002 Pop003 Pop004 Pop005	8.000000 8.000000 20.933333 18.000000 13.000000	0.066667 0.533333 0.933333 0.933333 1.000000	1.133333 1.866667 4.133333 4.200000 6.466667	3.000000 2.625000 4.357143 4.428571 6.466667	
Mean Population	13.586667 He	0.693333 Ho	3.560000 f	4.175476	
Pop001 Pop002 Pop003 Pop004 Pop005	0.022778 0.253333 0.530347 0.436296 0.707897	0.008333 0.050000 0.265079 0.044444 0.184615	0.650000 0.813333 0.506413 0.900754 0.746999		
Mean	0.390130	0.110495	0.723135		
					Ready
					,
vecuite Rec:	all Delete				Abo

#### Descriptive data of the microsatellite loci:

iptive s		tats Dist Disec	Deb		
	tatistics (b	y locus):			
Locus	n	Р	A	Ар	
Lm2TG	68.000000	1.000000	17.000000	17,000000	
TubCA	68.000000	1.000000	8.000000	8.000000	
Lm4TA	68.000000	1.000000	9.000000	9.000000	
B1	68.000000	1.000000	9.000000	9.000000	
C1	68,000000	1.000000	4.000000	4.000000	
E1	68,000000	1.000000	18,000000	18.000000	
F1	68.000000	1.000000	17.000000	17.000000	
G1	68.000000	1.000000	12.000000	12.000000	
P1	68.000000	1.000000	14.000000	14.000000	
01	68.000000	1.000000	4.000000	4.000000	
R1	68.000000	1.000000	8.000000	0.000000	
CS19	68.000000	1.000000	7.000000	7.000000	
CS20	67.000000	1.000000	12.000000	12.000000	
7031	68.000000	1.000000	11.000000	11.000000	
7039	68.000000	1.000000	14.000000	14.000000	
A11	67.933333	1.000000	10.933333	10.933333	
Locus	He	Ho	f		
F	0.799129	0.102941	0.872015		
Lm2TG					
TubCA Lm4TA	0.749891 0.834532	0.044118 0.058824	0.941578		
Lm41A B1	0.804575	0.264706	0.672638		
C1	0.676362	0.191176	0.718851		
E1	0.889325	0.191176	0.786284		
F1	0.890959	0.132353	0.852387		
G1	0.868736	0.132353	0.848607		
P1	0.774292	0.220588	0.716622		
01	0.639216	0.117647	0.817065		
R1	0.799564	0.029412	0.963478		
CS19	0.783987	0.014706	0.981379		
CS20	0.854337	0.230806	0.721996		
7031	0.802070	0.147059	0.817761		
7039	0.879303	0.147059	0.833788		
A11	0.803085	0.135528	0.832284		
				Ready	

## **References:**

Manuals for the softwares used

Evanno, G., S. Regnaut, J. Goudet, Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study, Mol. Ecol. 14 (2005) 2611-2620