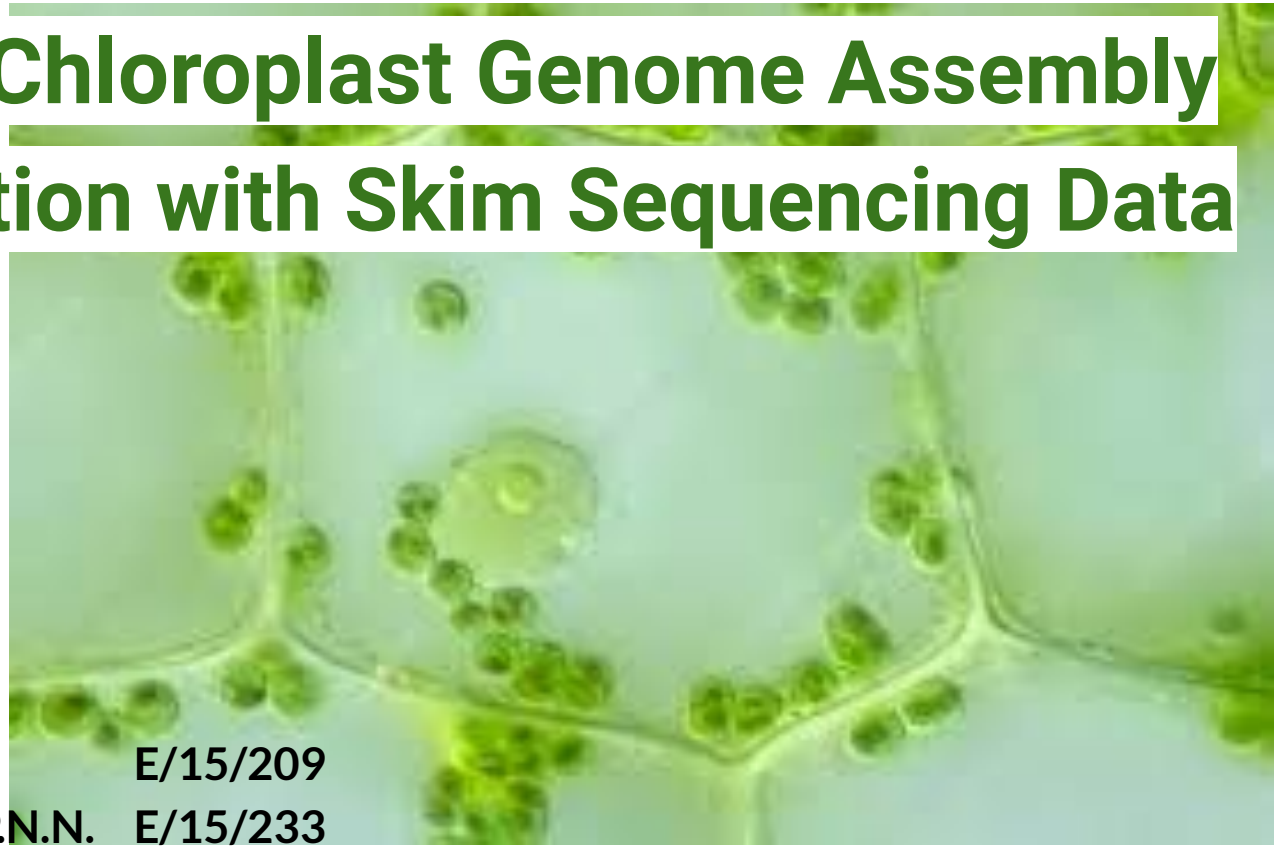
Optimizing Chloroplast Genome Assembly and Annotation with Skim Sequencing Data



Group Members:

1. Madhushani H. K. E/15/209
2. Muthucumarana P.N.N. E/15/233
3. Sankalpana W.A.P.C. E/15/325

Why this project?

Identifying the plant species is really important.

Analyse Sequences

- How they vary?
- Where they vary?
- Is the same species?

What are the technologies for assembly?

Biologists don’t know which one gives the best.

Why Chloroplast Genome?

- Specific to plant cells
- Useful in phylogenetic and evolutionary studies
- Evolve comparatively faster than nuclear genomes
- Easy to sequence and assemble

- How to resolve complex evolutionary relationships?
- How to identify plant species?

?

Tested Datasets

- Arabidopsis Thaliana (17Mb)
- Cinnamon (40Gb)
- Oryza Zativa (14Mb)

Tested Assembly Tools

- GetOrganelle
- Fast-Plast
- NovoPlasty

Why these three tools?

Tested Parameters

- Assembly time
- Memory Usage
- CPU utilization
- Genome coverage
- Accuracy

Server Specification

Why the server specification is important?

- Have to work with different servers
- Effect of number of threads in the tool

Tested servers

- Aiken server
- Tesla server

Aiken Server	Tesla Server
● High performance server ● Computing power is better ● Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz ● Number of Cores - 8 ● Number of threads - 16 ● 256 GB of RAM ● Ubuntu 18.04.5 LTS	● A GPU Workstation ● Can do the calculations parallely ● Intel(R) Core(TM) i7-6700K CPU @ 4.00GHz ● Number of Cores - 4 ● Number of threads - 8 ● 32 GB Memory ● Ubuntu 14.04.5 LTS

Results

DeNOVO Tools

Results

Data Set	Assembly Tool	Aiken Server			Tesla Server			Accuracy	
		Run Time	Memory Usage	CPU Usage	Run Time	Memory Usage	CPU Usage	Genome fraction	Missassemblies
Arabidopsis Thaliana (17 Mb)	Get-Organelle	4m 5s	0.276	6	3m 26s	0.26	12.0	82.998	1
	NOVO Plasty	23m 33s	0.100	99.40	28m 29s	0.10	102.0	100	0
	Fast-Plast	9m 48s	0.100	99.6	18m 33s	0.1	60.7	100	0
Cinnamon (20 Gb)	Get-Organelle	469m 54s	1.100	99.8	405m 31	1.1	99.9	100	0
	NOVO Plasty	45m 32s	0.600	99.60	28m 29s	0.1	102.0	100	0
	Fast-Plast	92m 35s	3.900	0.7	468m 48s	0.2	100.0	80.575	0

Usage of Threads - GetOrganelle

Data Set	Assembly Tool	Num of threads	Aiken Server			Tesla Server			Accuracy	
			Run Time	Memory Usage	CPU Usage	Run Time	Memory Usage	CPU Usage	Genome fraction	Missassemblies
Arabidopsis Thaliana	Get-Organelle	1	4m 5s	0.276	6	3m 26s	0.260	12	100	0
		2	4m 36s	0.277	6	4m 55s	0.261	12	100	0
		3	3m 16s	0.276	5	2m 58s	0.260	14	100	0
		4	2m 2s	0.100	3	2m 5s	0.300	18	100	0
		5	2m 49s	0.100	5	3m 25s	0.200	21	100	0
		10	4m 5s	0.276	6	4m 59s	0.260	13	100	0

Reference-guided de novo assembly approach

- combination of DeNovo approach and Reference mapping approach

- Effective and powerful

Reference-guided de novo assembly approach using NovoPlasty pipeline.

Results for the assembly of the Oryza sativa chloroplast (dataset SRR1328237)

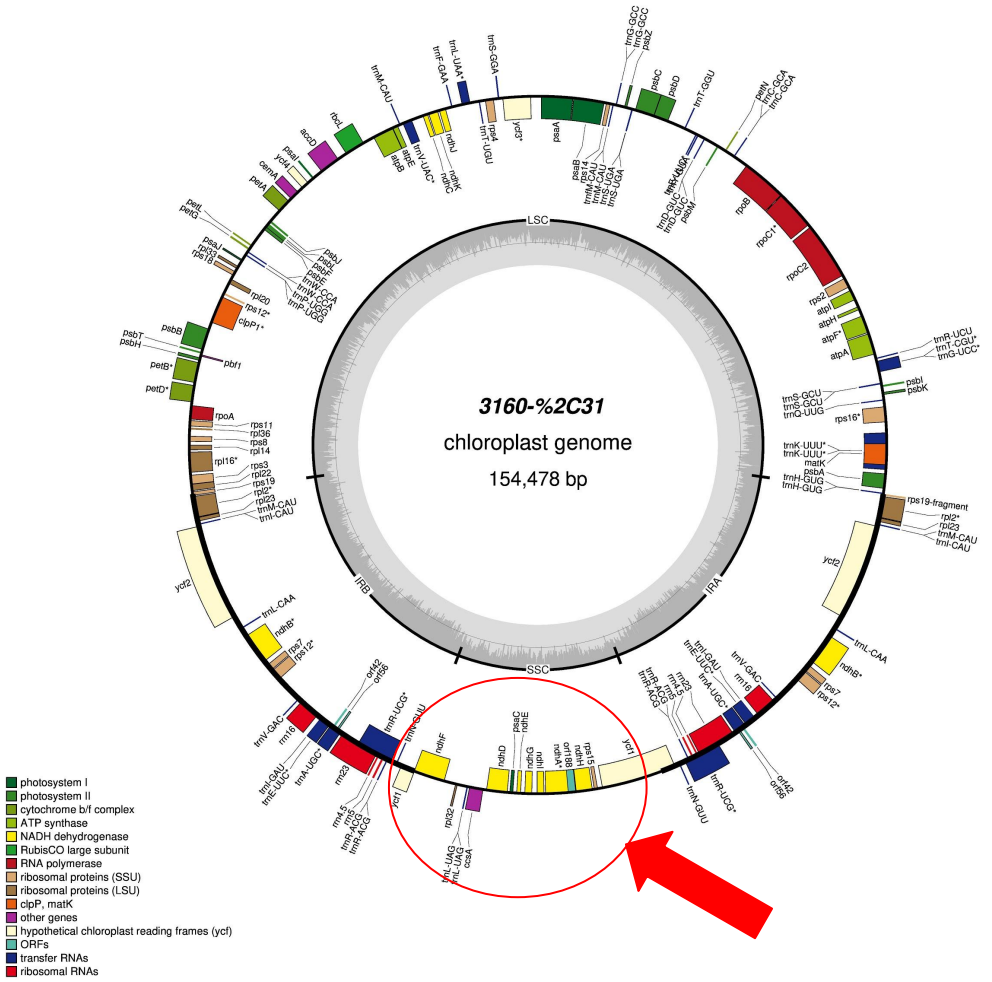
Performance	Without Ref	With Ref
Duration (min)	3m39.511s	3m10.811s
System+user time (min)	3m39.460s + 0m0.360s	0m0.694s + 3m8.376s
Memory %	0.10%	0.10%
CPU%	99.70%	99.01%
Total contigs	3	1
Average insert size	480 bp	480 bp
Total reads	279568	279568
Aligned reads	273234	268650
Assembled reads	248700	243944
Organelle genome %	97.73 %	96.09 %
Average organelle coverage	307	302

Genome Annotation

Result of GetOrganalle for

Arabidopsis Thaliana

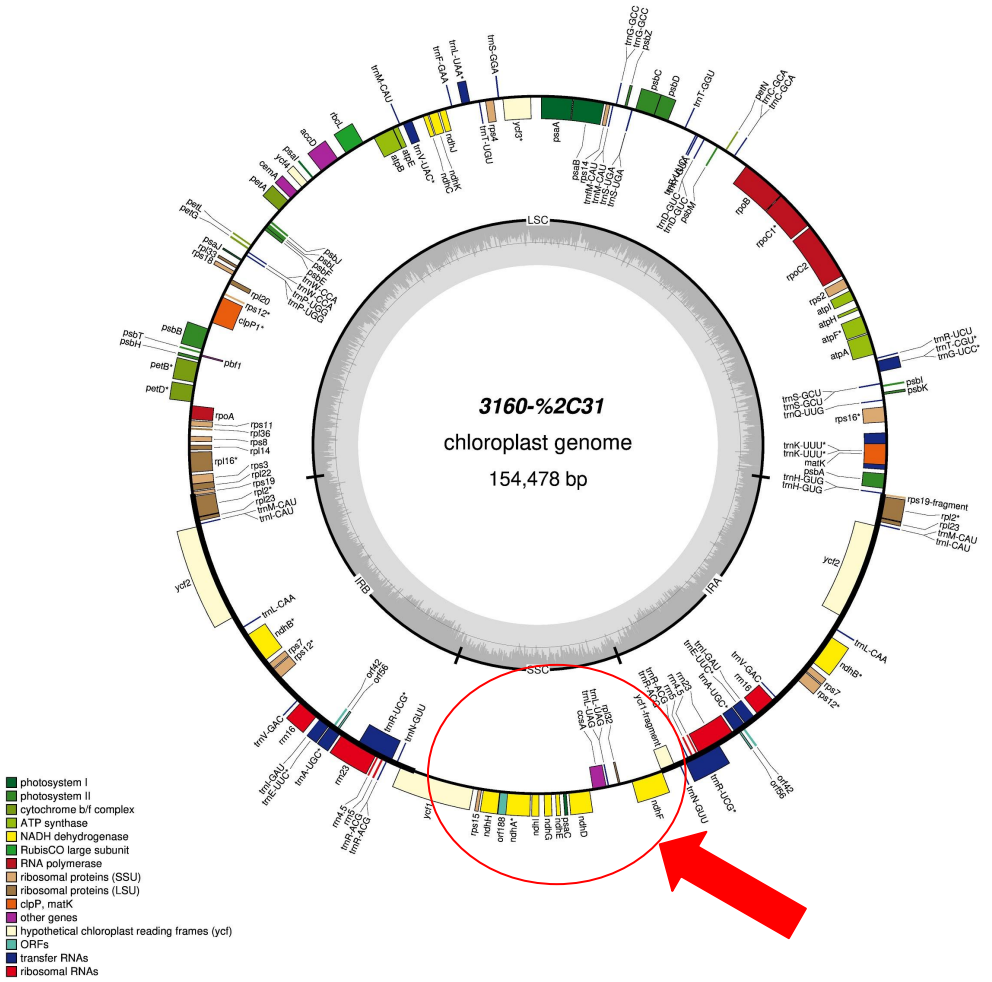
Option 1



Result of GetOrganalle for

Arabidopsis Thaliana

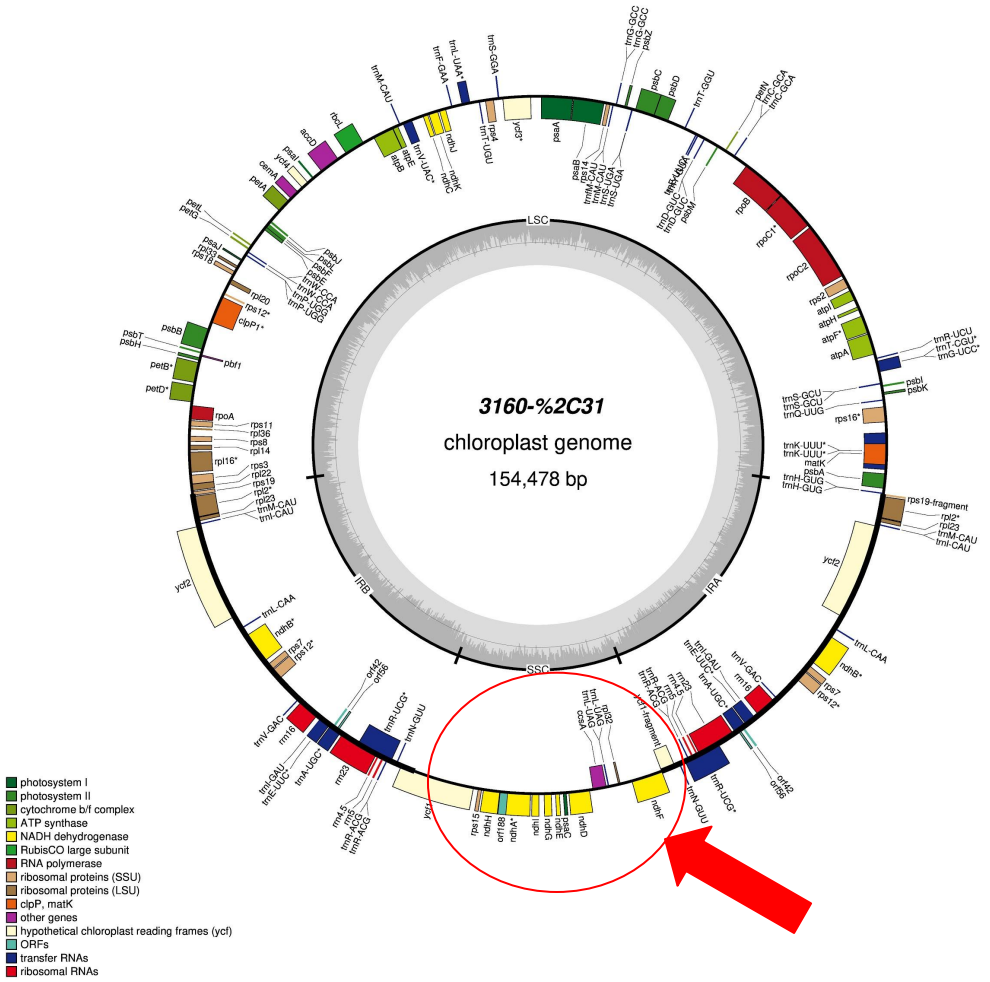
Option 2



Result of GetOrganalle for

Cinnamon

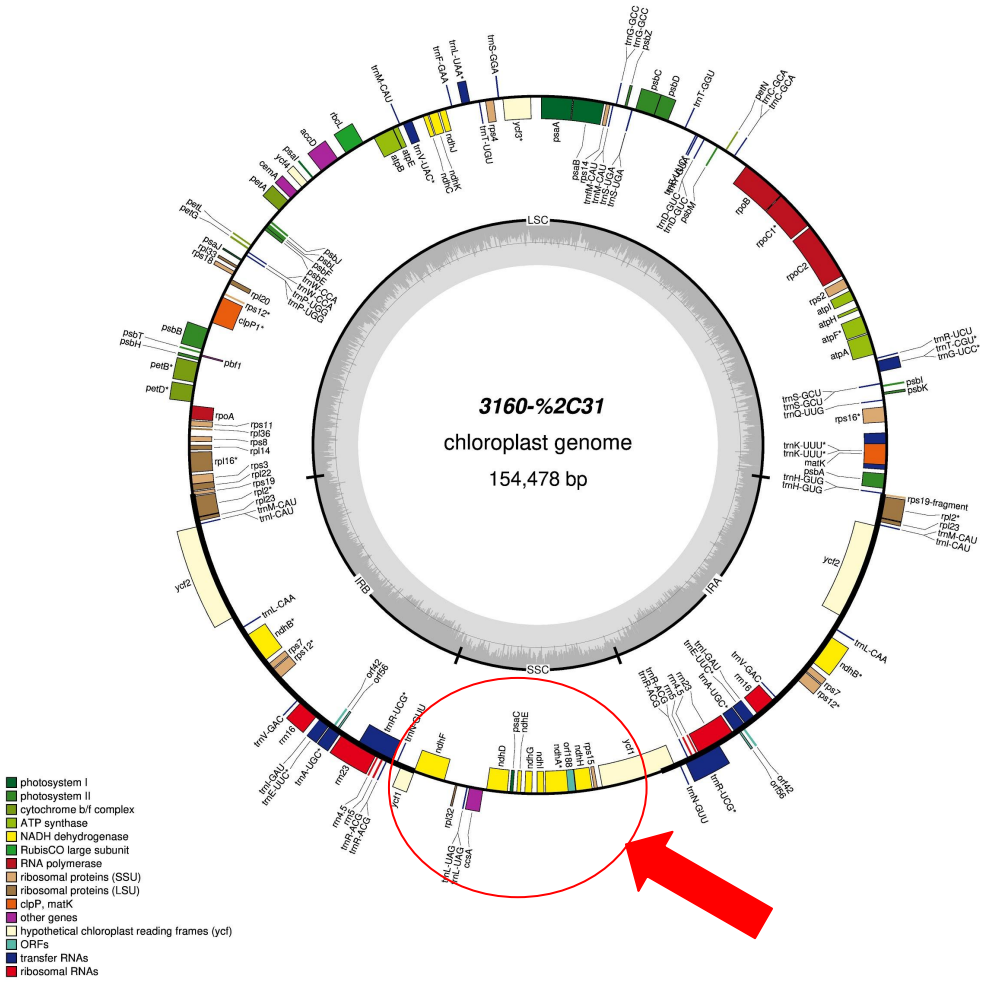
Option 1



Result of GetOrganalle for

Cinnamon

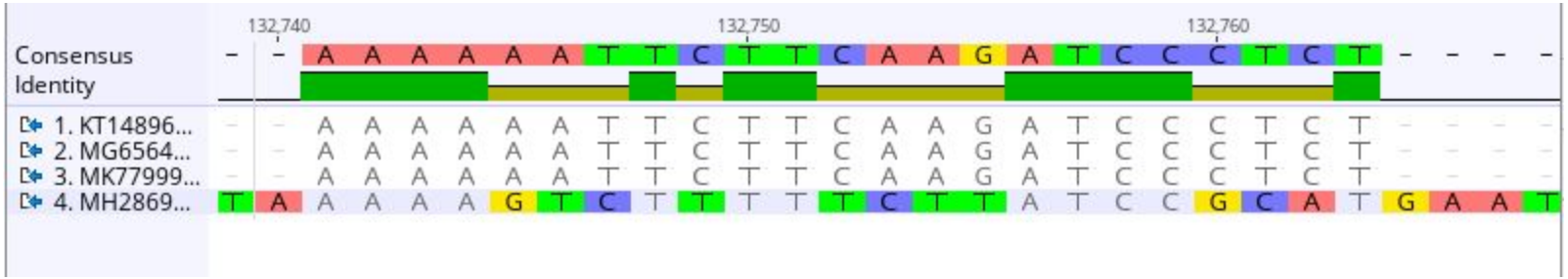
Option 2



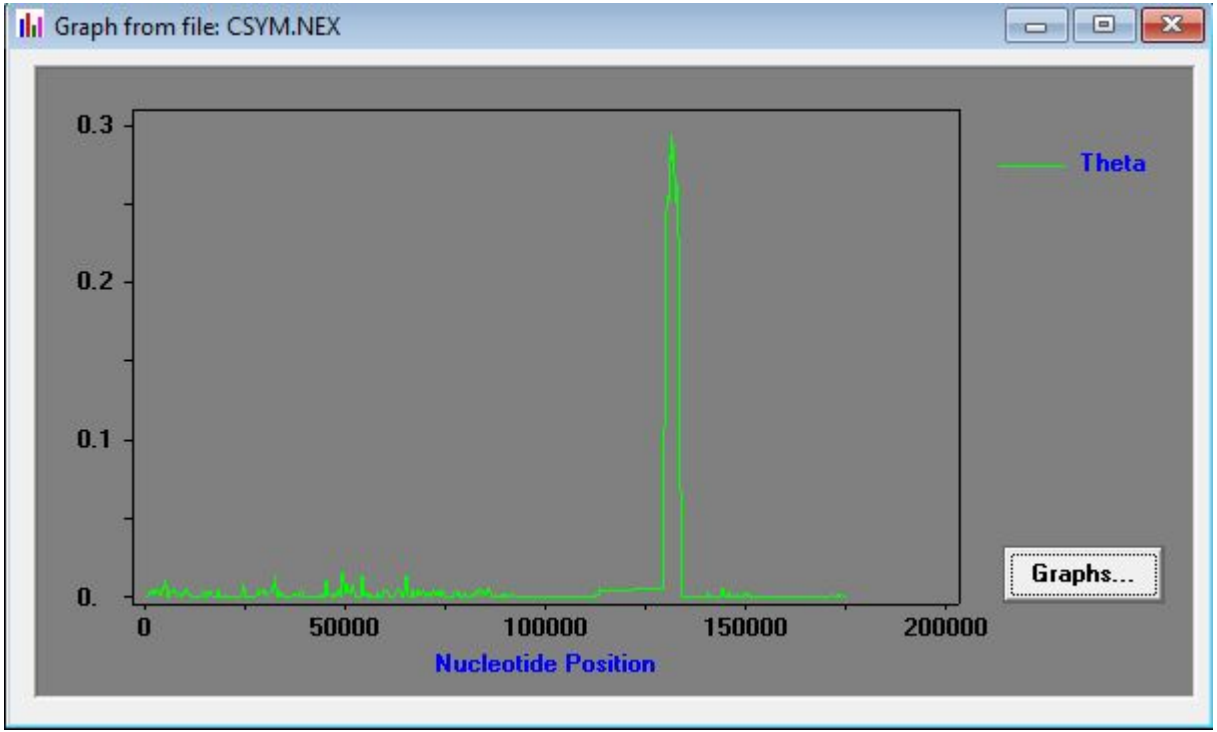
Usage of Genome Annotation

- Aquilaria Crassna (MK779998)
- Aquilaria Sinensis (KT148967)
- Aquilaria Yunnanensis (MG656407)
- Aquilaria Malaccensis (MH286934)

Variations of the base pairs of Aquilaria species



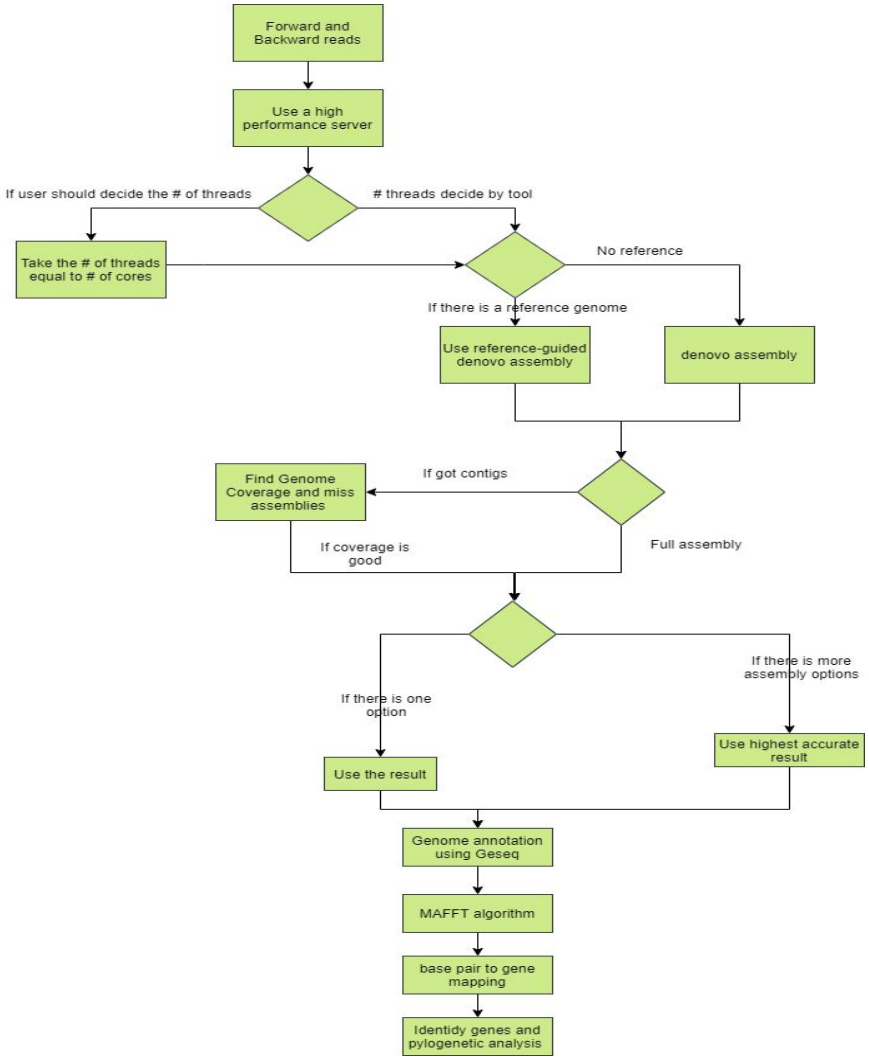
Varying regions of Aquilaria Species



Comparison between the methods

Tool	Usage	Speed	Accuracy
GetOrganelle	Default	Fast	Not much as NOVO-Plasty
Fast-Plast	Second option	Not much fast as GetOrganellla	Not much as GetOrganelle
NOVO-Plasty	Third option	Not much fast as Fast-Plast	Highest

Recommended Pipeline



Thank You !

Q & A