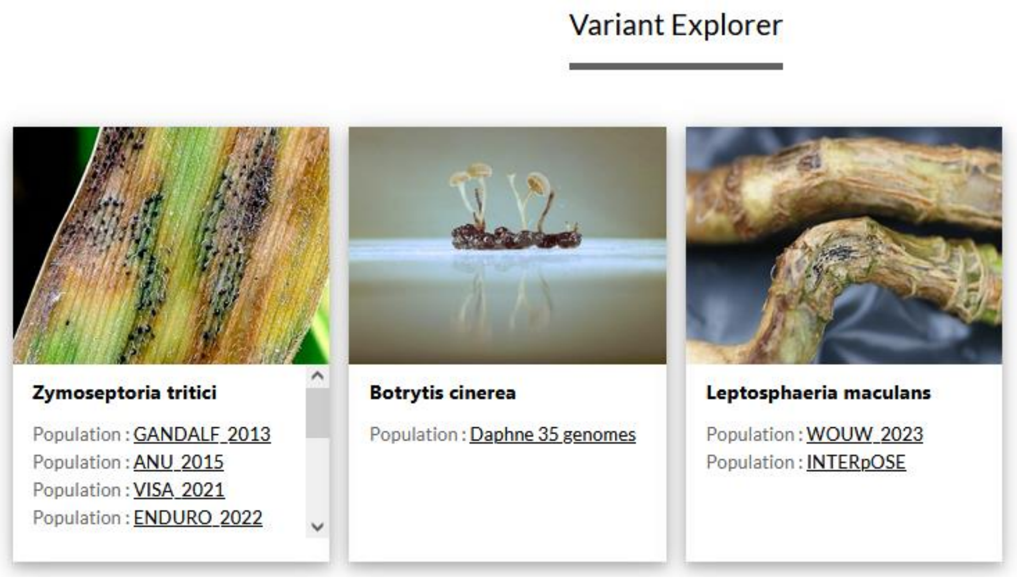
Variant Explorer



Variant Explorer

Zymoseptoria tritici

Population : GANDALF_2013
Population : ANU_2015
Population : VISA_2021
Population : ENDURO_2022

Botrytis cinerea

Population : Daphne 35 genomes

Leptosphaeria maculans

Population : WOUW_2023
Population : INTERpOSE

https://bioinfo.bioger.inrae.fr/portal/variant-explorer/

Adeline Simon & Nicolas Lapalu

Variant Explorer

Reference genome

(Assembly OK + Gene annotation OK)

Population

(Illumina / short-reads sequencing)

- 1- How to compare the sequence of a genomic region in reference genome & in population ? (SNP, InDels)
- 2- How to retrieve the sequences of reference proteins in population ?

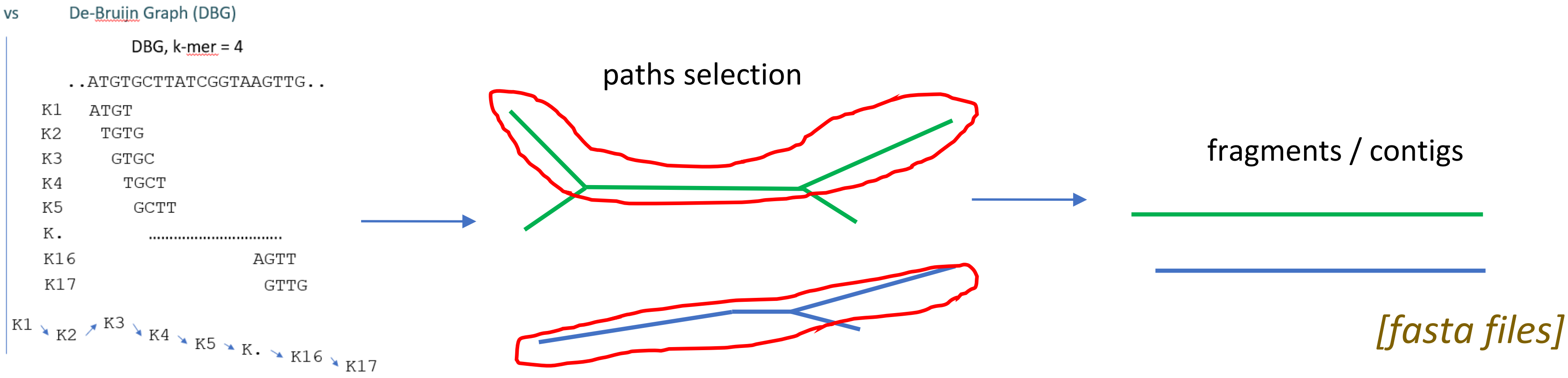
1- genomic region

Variant Explorer

1A/ Assembly based - preliminary analyses (bioinfobioger)

Population → **De novo assembly**

(Illumina / short-reads sequencing) (Spades / Velvet, SOAP denovo)



vs De-Bruijn Graph (DBG)

DBG, k-mer = 4

```
..ATGTGCTTATCGGTAAGTTG..
K1   ATGT
K2    TGTG
K3     GTGC
K4      TGCT
K5       GCTT
K.        ..............................
K16                AGTT
K17                 GTTG

K1 ↘ K2 ↗ K3 ↘ K4 ↘ K5 ↘ K. ↘ K16 ↘ K17
```

paths selection

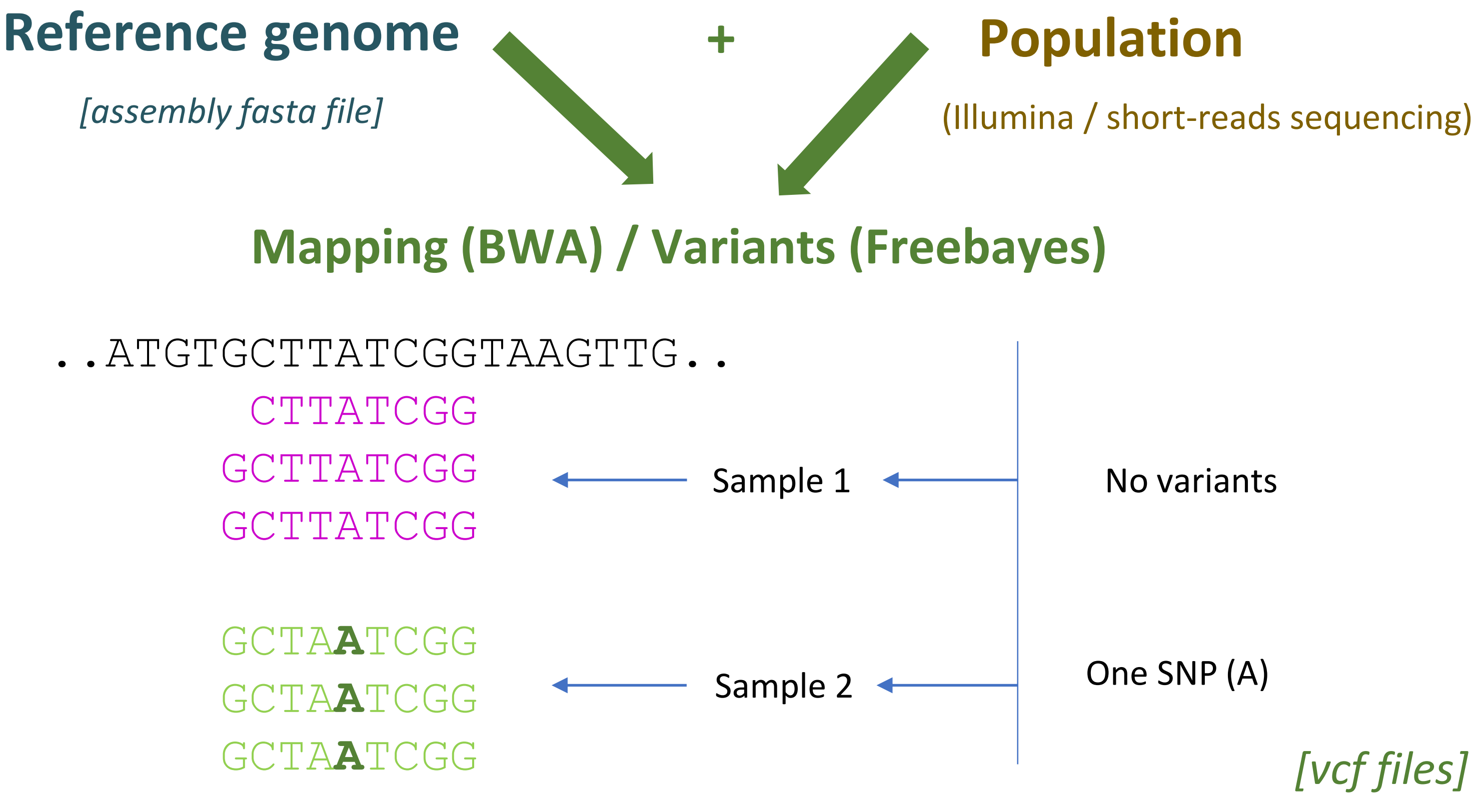
fragments / contigs

[fasta files]

Warning: Genome assembly from short-reads provides fragmented assembly. Many contigs. Possible bias in specific regions like TE-rich.

Variant Explorer

1B/ Variant calling - preliminary analyses (bioinfobioger)



Reference genome

[assembly fasta file]

\+

Population

(Illumina / short-reads sequencing)

Mapping (BWA) / Variants (Freebayes)

..ATGTGCTTATCGGTAAGTTG..

CTTATCGG
GCTTATCGG
GCTTATCGG

Sample 1

No variants

GCTA**A**TCGG
GCTA**A**TCGG
GCTA**A**TCGG

Sample 2

One SNP (A)

[vcf files]

Warning: No mapping = No data. Specific regions higher than 10/20pb are not detectable (short-reads size sequence = limiting factor)

Variant Explorer

1B/ Variant calling - preliminary analyses (bioinfobioger)

Reference genome + **Population**

[assembly fasta file]

(Illumina / short-reads sequencing)

Mapping (BWA) / Variants (Freebayes)

[vcf files]

List of all analyzed positions
= variant + non-variant sites

```
TGTGCTTATCGGTAAGT
```

```
Ref pos3 T => A
Ref pos7 T => A
```

« Pseudo-assembly »

[fasta files]

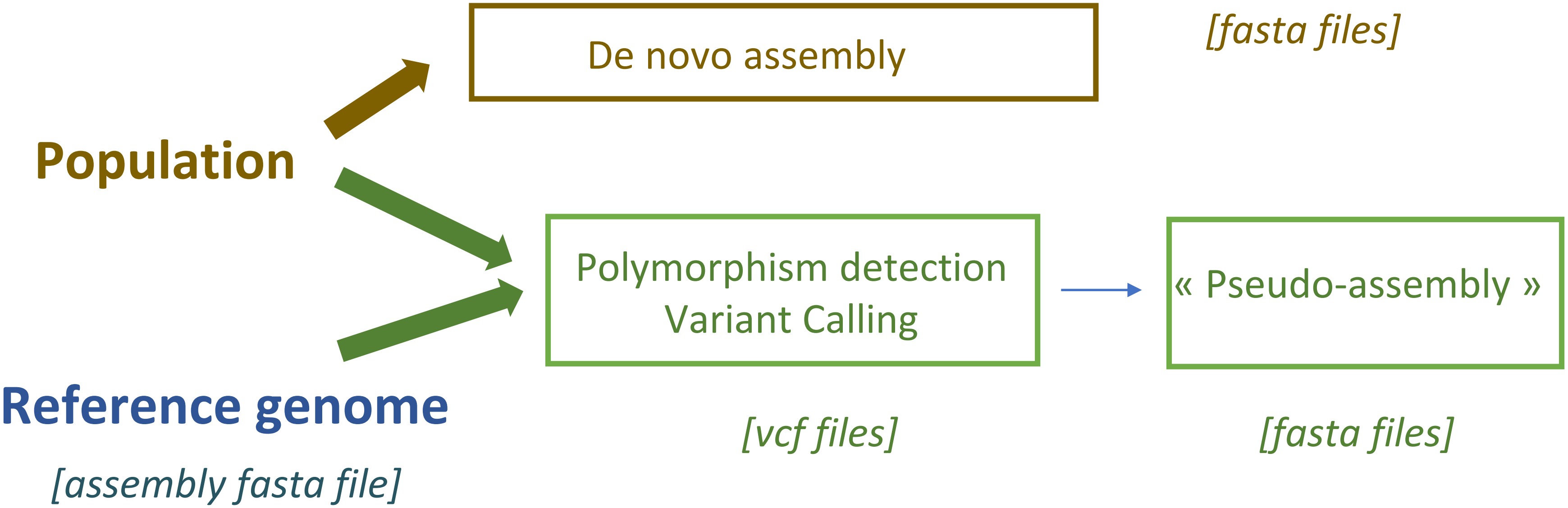
Inferred regions in fasta files from variant and non-variant sites

```
TGTGCTTATCGGTAAGT
```

AGTGCT**A**ATCGGTAAGT

1- genomic region

Variant Explorer



De novo assembly

[fasta files]

Population

Polymorphism detection
Variant Calling

« Pseudo-assembly »

Reference genome

[assembly fasta file]

[vcf files]

[fasta files]

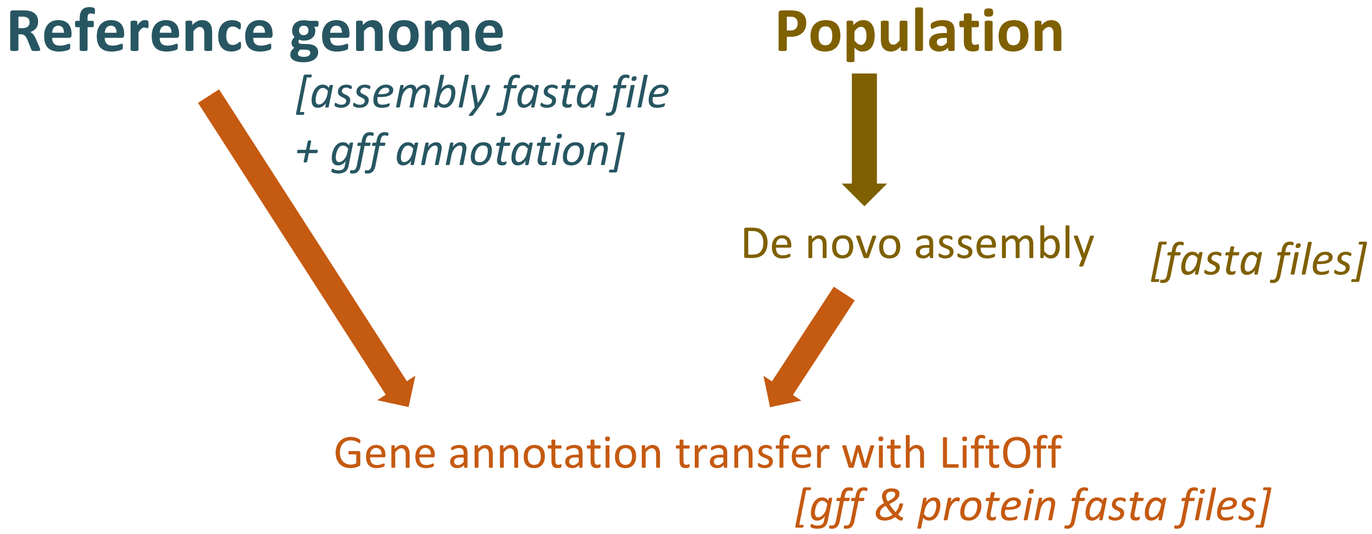
- <u>How can you explore these files with Variant Explorer ?</u>

Blast a region of interest in reference genome against population assemblies (evalue < e 10^{-10} & identity > 90 %)

2- proteins

Variant Explorer

Protein based - Preliminary analyses (bioinfobioger)



Reference genome

[assembly fasta file + gff annotation]

Population

De novo assembly *[fasta files]*

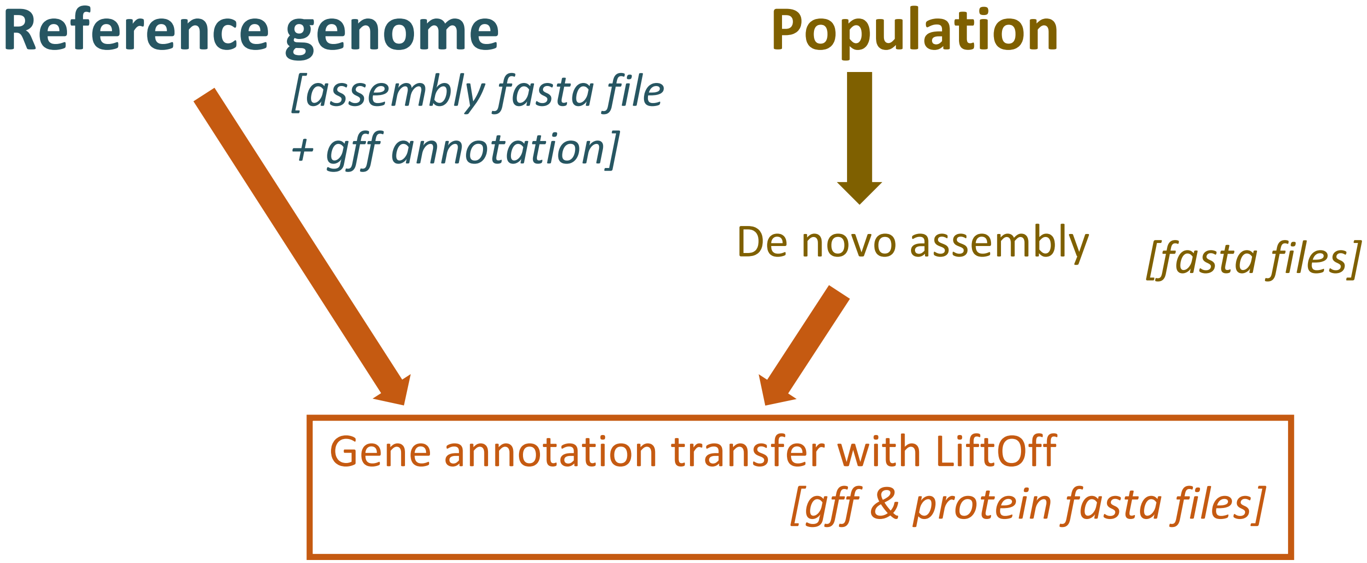
Gene annotation transfer with LiftOff

[gff & protein fasta files]

Warning: Comparison at protein level -> inferred from reference annotation, possible problems if bad annotations.

2- proteins

Variant Explorer



Reference genome

[assembly fasta file
+ gff annotation]

Population

De novo assembly

[fasta files]

Gene annotation transfer with LiftOff
[gff & protein fasta files]

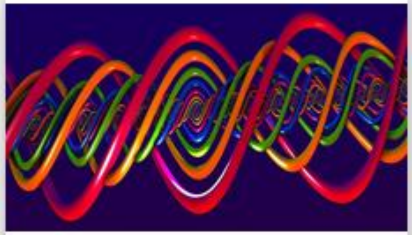
➢ How can you explore these files with Variant Explorer ?

Query the pre-computed population protein sequences, based on reference genome protein ids

Display the blast result (evalue < e 10^{-10})

Access to the data

Variant Explorer



Genomes & populations

https://bioinfo.bioger.inrae.fr/portal/genome-portal/

Description of populations

Botrytis

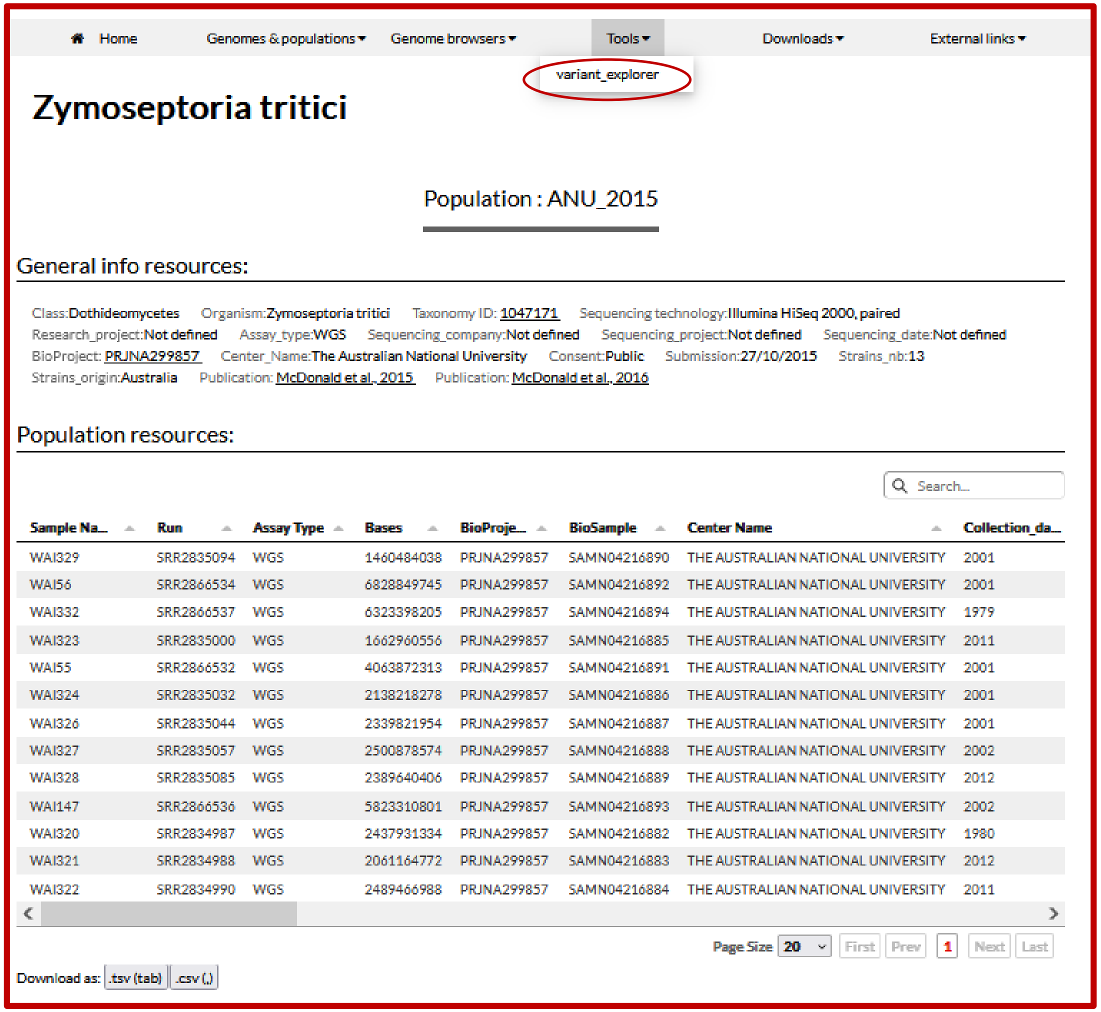
Genome : Botrytis private
Genome : SI3
Genome : Vv3
Mutants : Botrytis mutants
Population : Daphne 35 genomes

Leptosphaeria

Genome : INV13_269
Genome : JN3
Genome : Lepto private
Genome : NZ-T4
Population : Gandalf Lepto
Population : INTERpOSE

Zymoseptoria

Population : ANU_2015
Population : APOGE_2022
Population : BASTAFUN_2021
Population : ENDURO_2022
Population : ETH_2016
Population : GANDALF_2013



Zymoseptoria tritici

Population : ANU_2015

General info resources:

Class:Dothideomycetes Organism:Zymoseptoria tritici Taxonomy ID: 1047171 Sequencing technology:Illumina HiSeq 2000, paired Research_project:Not defined Assay_type:WGS Sequencing_company:Not defined Sequencing_project:Not defined Sequencing_date:Not defined BioProject: PRJNA299857 Center_Name:The Australian National University Consent:Public Submission:27/10/2015 Strains_nb:13 Strains_origin:Australia Publication: McDonald et al., 2015 Publication: McDonald et al., 2016

Population resources:

Sample Na...	Run	Assay Type	Bases	BioProje...	BioSample	Center Name	Collection_da...
WAI329	SRR2835094	WGS	1460484038	PRJNA299857	SAMN04216890	THE AUSTRALIAN NATIONAL UNIVERSITY	2001
WAI56	SRR2866534	WGS	6828849745	PRJNA299857	SAMN04216892	THE AUSTRALIAN NATIONAL UNIVERSITY	2001
WAI332	SRR2866537	WGS	6323398205	PRJNA299857	SAMN04216894	THE AUSTRALIAN NATIONAL UNIVERSITY	1979
WAI323	SRR2835000	WGS	1662960556	PRJNA299857	SAMN04216885	THE AUSTRALIAN NATIONAL UNIVERSITY	2011
WAI55	SRR2866532	WGS	4063872313	PRJNA299857	SAMN04216891	THE AUSTRALIAN NATIONAL UNIVERSITY	2001
WAI324	SRR2835032	WGS	2138218278	PRJNA299857	SAMN04216886	THE AUSTRALIAN NATIONAL UNIVERSITY	2001
WAI326	SRR2835044	WGS	2339821954	PRJNA299857	SAMN04216887	THE AUSTRALIAN NATIONAL UNIVERSITY	2001
WAI327	SRR2835057	WGS	2500878574	PRJNA299857	SAMN04216888	THE AUSTRALIAN NATIONAL UNIVERSITY	2002
WAI328	SRR2835085	WGS	2389640406	PRJNA299857	SAMN04216889	THE AUSTRALIAN NATIONAL UNIVERSITY	2012
WAI147	SRR2866536	WGS	5823310801	PRJNA299857	SAMN04216893	THE AUSTRALIAN NATIONAL UNIVERSITY	2002
WAI320	SRR2834987	WGS	2437931334	PRJNA299857	SAMN04216882	THE AUSTRALIAN NATIONAL UNIVERSITY	1980
WAI321	SRR2834988	WGS	2061164772	PRJNA299857	SAMN04216883	THE AUSTRALIAN NATIONAL UNIVERSITY	2012
WAI322	SRR2834990	WGS	2489466988	PRJNA299857	SAMN04216884	THE AUSTRALIAN NATIONAL UNIVERSITY	2011

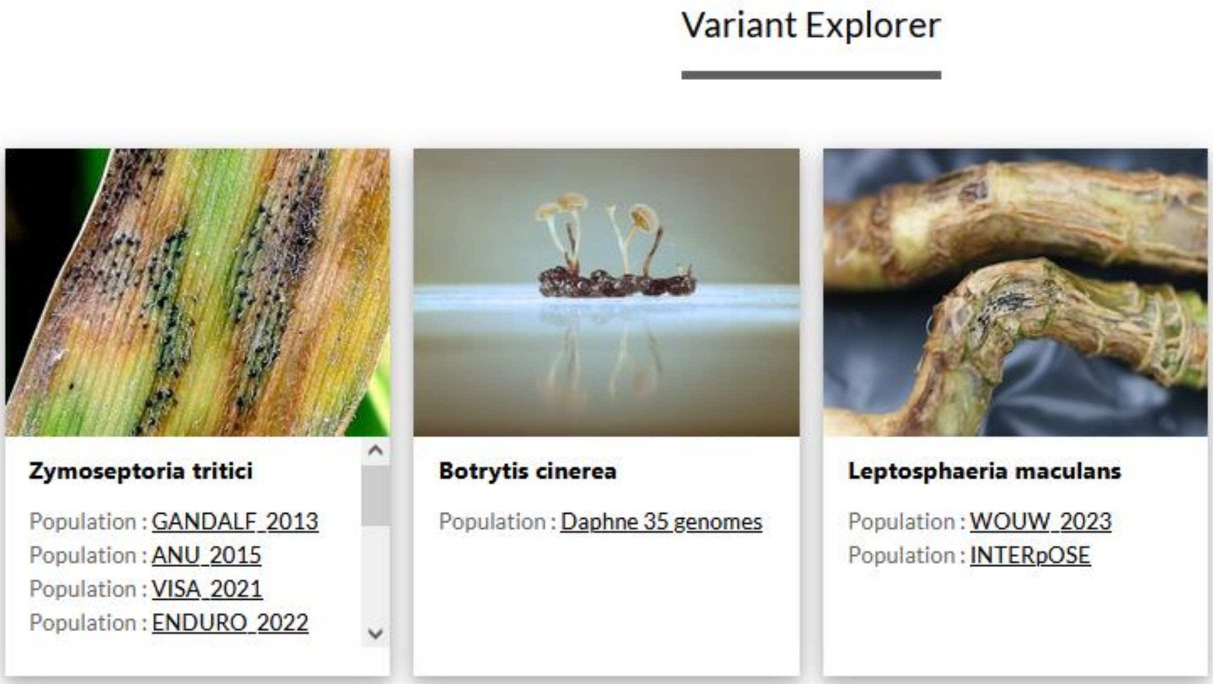
Access to the tool

Variant Explorer

Tools

https://bioinfo.bioger.inrae.fr/portal/variant-explorer/

Direct access to Variant Explorer tool



Variant Explorer

Zymoseptoria tritici

Population : GANDALF_2013
Population : ANU_2015
Population : VISA_2021
Population : ENDURO_2022

Botrytis cinerea

Population : Daphne 35 genomes

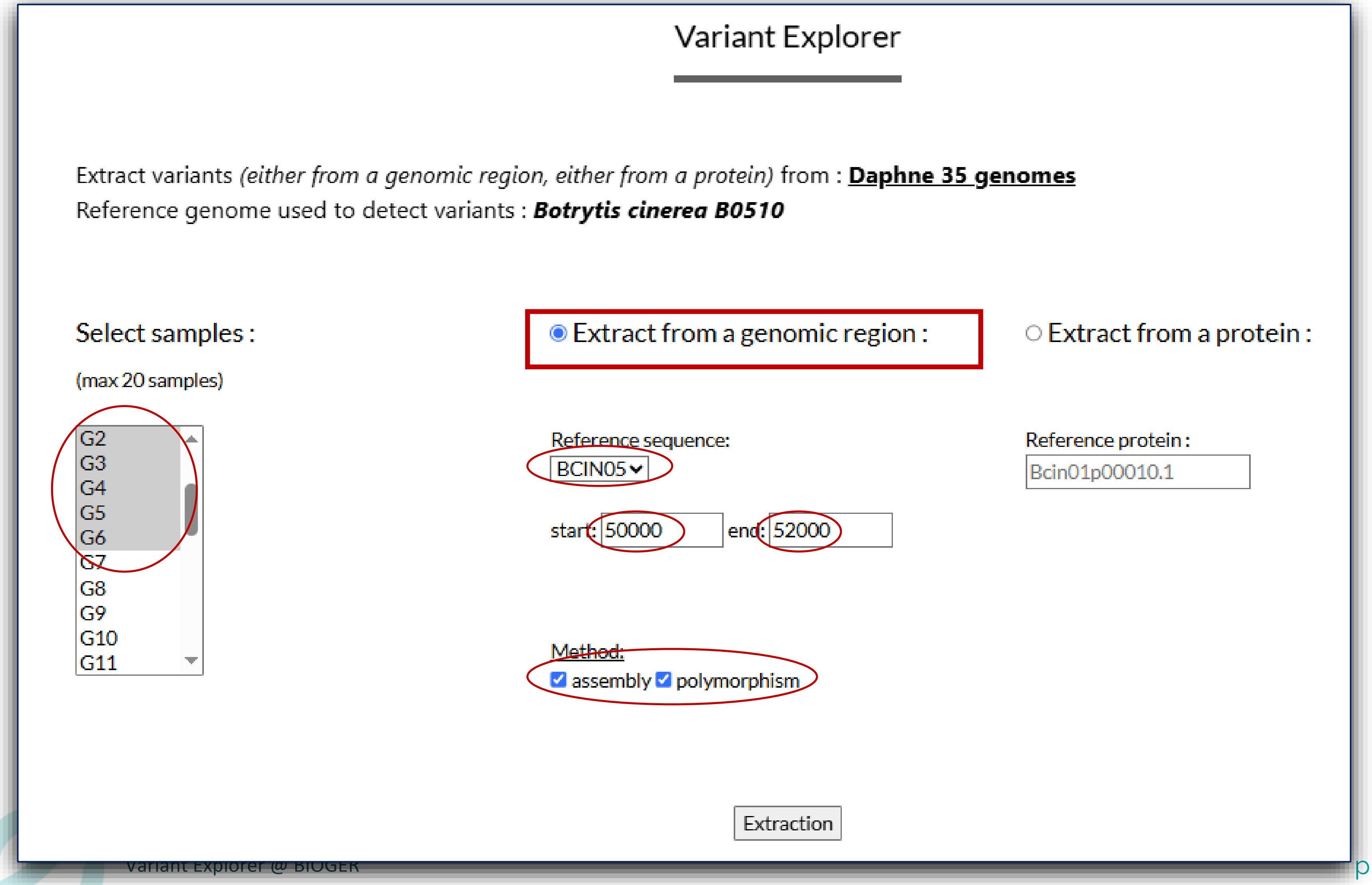
Leptosphaeria maculans

Population : WOUW_2023
Population : INTERpOSE

1- genomic region

Variant Explorer

Compare the sequence of a genomic region
in reference genome & in population



Variant Explorer

Extract variants *(either from a genomic region, either from a protein)* from : **Daphne 35 genomes**

Reference genome used to detect variants : ***Botrytis cinerea B0510***

Select samples :

(max 20 samples)

G2
G3
G4
G5
G6
G7
G8
G9
G10
G11

● Extract from a genomic region :

Reference sequence: BCIN05

start: 50000 end: 52000

Method:

☑ assembly ☑ polymorphism

○ Extract from a protein :

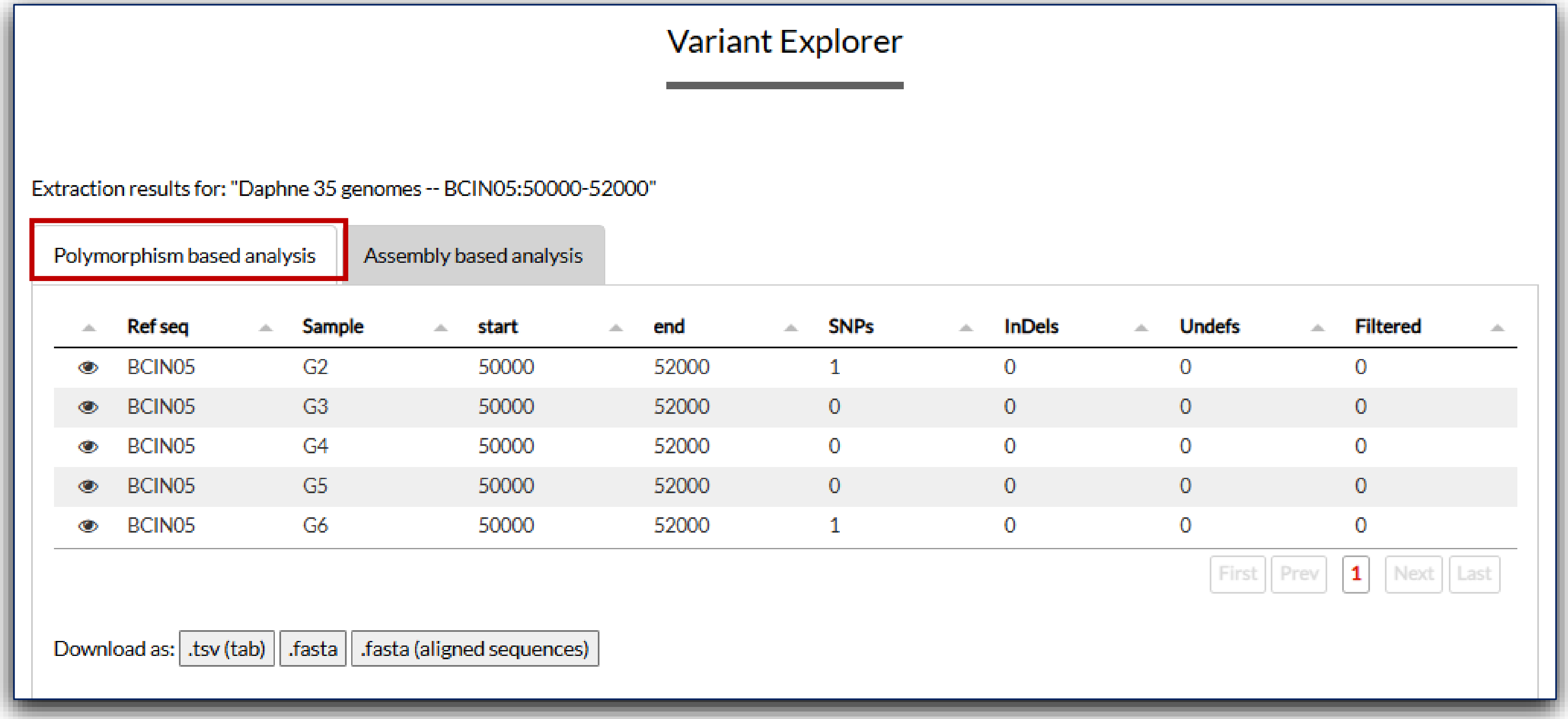
Reference protein : Bcin01p00010.1

1- genomic region

Variant Explorer

Compare the sequence of a genomic region
in reference genome & in population

Polymorphism based analysis



Variant Explorer

Extraction results for: "Daphne 35 genomes -- BCIN05:50000-52000"

	Ref seq	Sample	start	end	SNPs	InDels	Undefs	Filtered
👁	BCIN05	G2	50000	52000	1	0	0	0
👁	BCIN05	G3	50000	52000	0	0	0	0
👁	BCIN05	G4	50000	52000	0	0	0	0
👁	BCIN05	G5	50000	52000	0	0	0	0
👁	BCIN05	G6	50000	52000	1	0	0	0

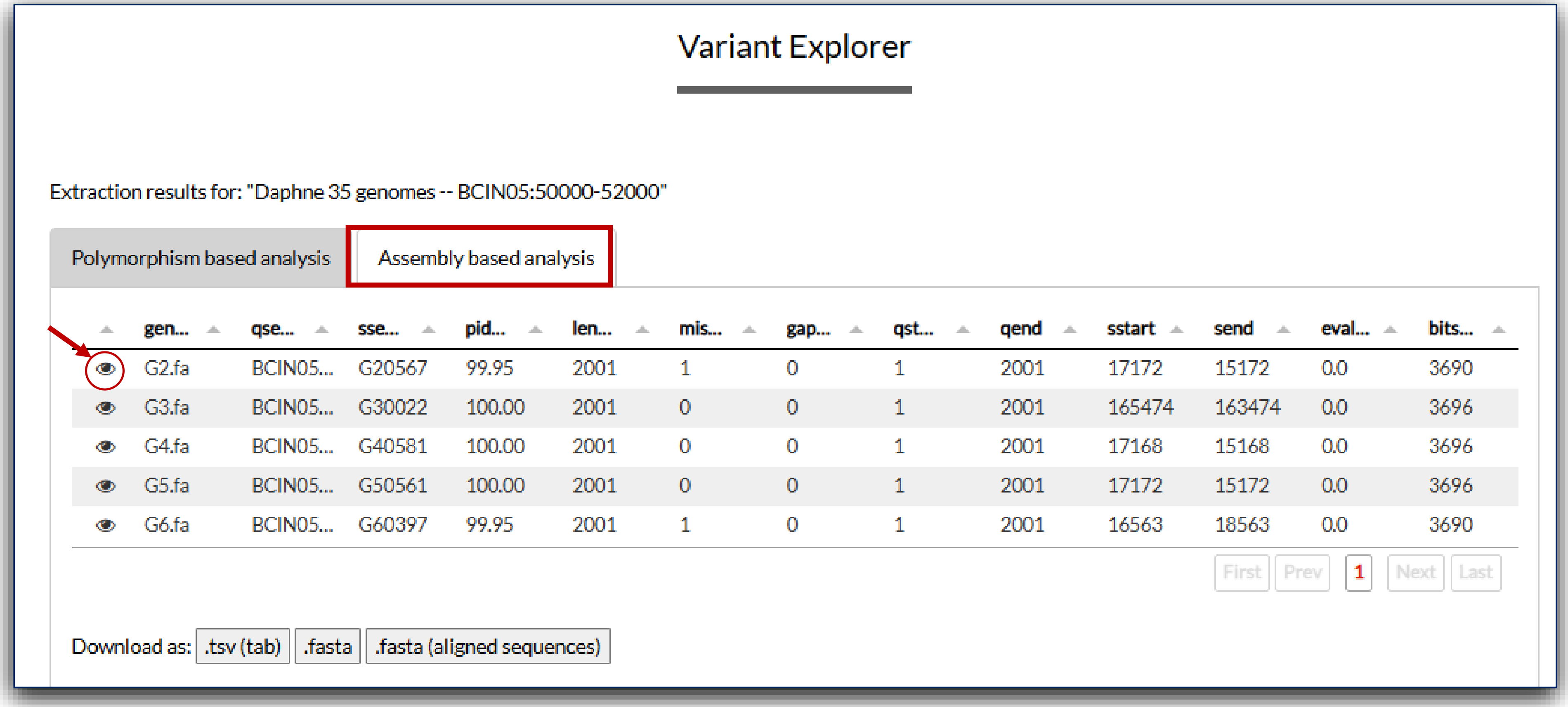
Download as: .tsv (tab) .fasta .fasta (aligned sequences)

1- genomic region

Variant Explorer

Compare the sequence of a genomic region
in reference genome & in population

Assembly based analysis



Extraction results for: "Daphne 35 genomes -- BCIN05:50000-52000"

	gen...	qse...	sse...	pid...	len...	mis...	gap...	qst...	qend	sstart	send	eval...	bits...
👁	G2.fa	BCIN05...	G20567	99.95	2001	1	0	1	2001	17172	15172	0.0	3690
👁	G3.fa	BCIN05...	G30022	100.00	2001	0	0	1	2001	165474	163474	0.0	3696
👁	G4.fa	BCIN05...	G40581	100.00	2001	0	0	1	2001	17168	15168	0.0	3696
👁	G5.fa	BCIN05...	G50561	100.00	2001	0	0	1	2001	17172	15172	0.0	3696
👁	G6.fa	BCIN05...	G60397	99.95	2001	1	0	1	2001	16563	18563	0.0	3690

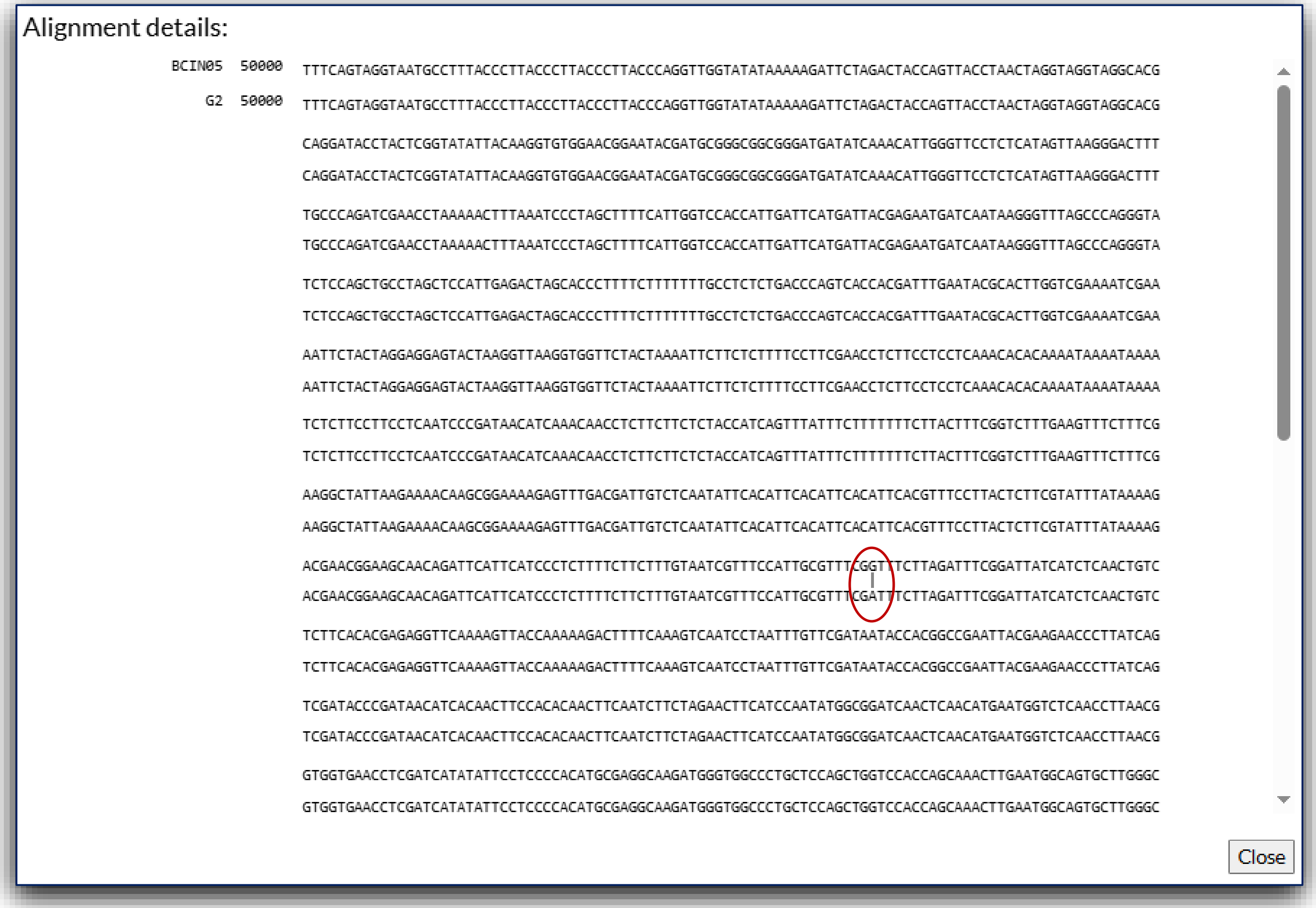
Download as: .tsv (tab) .fasta .fasta (aligned sequences)

1- genomic region

Variant Explorer

Compare the sequence of a genomic region in reference genome & in population

Alignment details



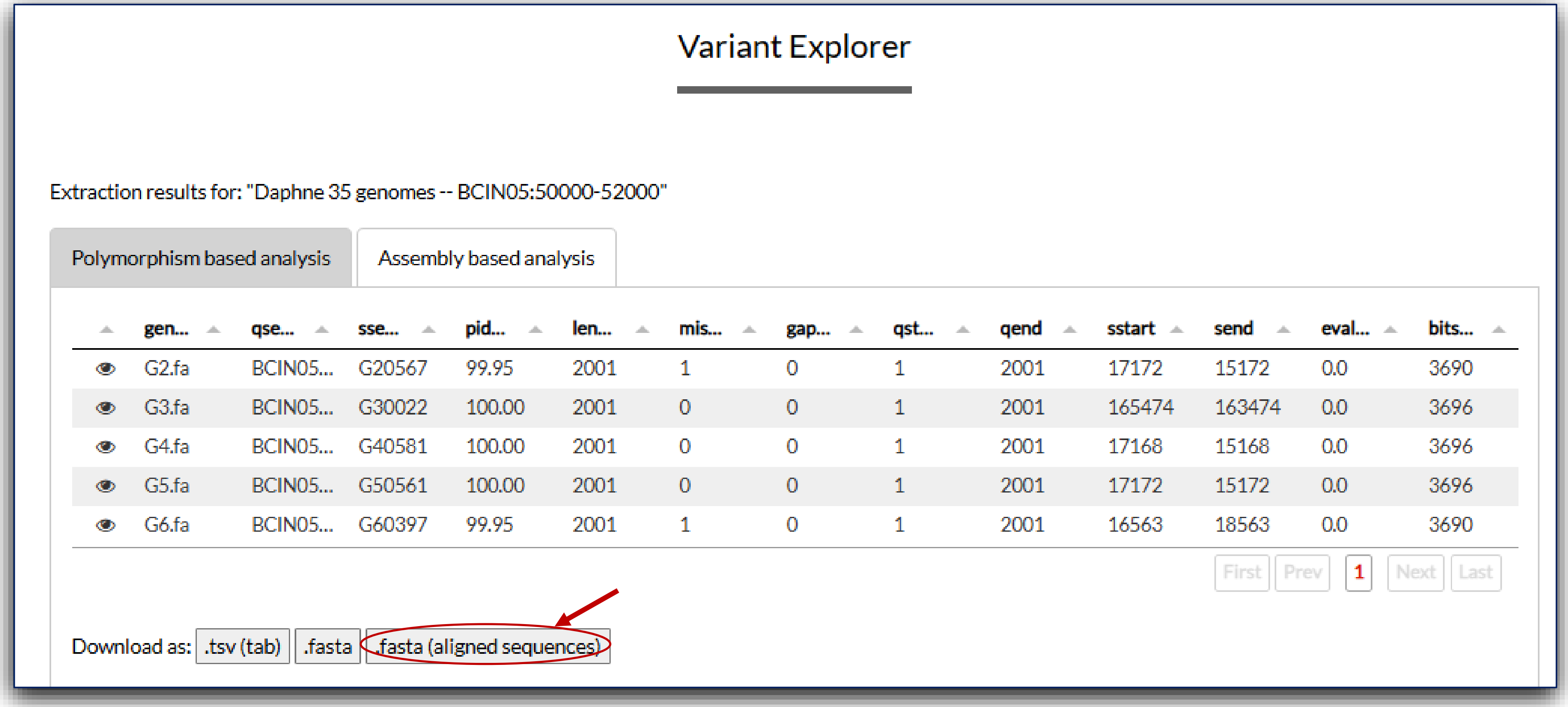
Alignment details:

```
BCIN05  50000  TTTCAGTAGGTAATGCCTTTACCCTTACCCTTACCCTTACCCAGGTTGGTATATAAAAAGATTCTAGACTACCAGTTACCTAACTAGGTAGGTAGGCACG
    G2  50000  TTTCAGTAGGTAATGCCTTTACCCTTACCCTTACCCTTACCCAGGTTGGTATATAAAAAGATTCTAGACTACCAGTTACCTAACTAGGTAGGTAGGCACG

               CAGGATACCTACTCGGTATATTACAAGGTGTGGAACGGAATACGATGCGGGCGGCGGGATGATATCAAACATTGGGTTCCTCTCATAGTTAAGGGACTTT
               CAGGATACCTACTCGGTATATTACAAGGTGTGGAACGGAATACGATGCGGGCGGCGGGATGATATCAAACATTGGGTTCCTCTCATAGTTAAGGGACTTT

               TGCCCAGATCGAACCTAAAAACTTTAAATCCCTAGCTTTTCATTGGTCCACCATTGATTCATGATTACGAGAATGATCAATAAGGGTTTAGCCCAGGGTA
               TGCCCAGATCGAACCTAAAAACTTTAAATCCCTAGCTTTTCATTGGTCCACCATTGATTCATGATTACGAGAATGATCAATAAGGGTTTAGCCCAGGGTA

               TCTCCAGCTGCCTAGCTCCATTGAGACTAGCACCCTTTTCTTTTTTTGCCTCTCTGACCCAGTCACCACGATTTGAATACGCACTTGGTCGAAAATCGAA
               TCTCCAGCTGCCTAGCTCCATTGAGACTAGCACCCTTTTCTTTTTTTGCCTCTCTGACCCAGTCACCACGATTTGAATACGCACTTGGTCGAAAATCGAA

               AATTCTACTAGGAGGAGTACTAAGGTTAAGGTGGTTCTACTAAAATTCTTCTCTTTTCCTTCGAACCTCTTCCTCCTCAAACACACAAAATAAAATAAAA
               AATTCTACTAGGAGGAGTACTAAGGTTAAGGTGGTTCTACTAAAATTCTTCTCTTTTCCTTCGAACCTCTTCCTCCTCAAACACACAAAATAAAATAAAA

               TCTCTTCCTTCCTCAATCCCGATAACATCAAACAACCTCTTCTTCTCTACCATCAGTTTATTTCTTTTTTTCTTACTTTCGGTCTTTGAAGTTTCTTTCG
               TCTCTTCCTTCCTCAATCCCGATAACATCAAACAACCTCTTCTTCTCTACCATCAGTTTATTTCTTTTTTTCTTACTTTCGGTCTTTGAAGTTTCTTTCG

               AAGGCTATTAAGAAAACAAGCGGAAAAGAGTTTGACGATTGTCTCAATATTCACATTCACATTCACATTCACGTTTCCTTACTCTTCGTATTTATAAAAG
               AAGGCTATTAAGAAAACAAGCGGAAAAGAGTTTGACGATTGTCTCAATATTCACATTCACATTCACATTCACGTTTCCTTACTCTTCGTATTTATAAAAG

               ACGAACGGAAGCAACAGATTCATTCATCCCTCTTTTCTTCTTTGTAATCGTTTCCATTGCGTTTCGGTTCTTAGATTTCGGATTATCATCTCAACTGTC
                                                                                 |
               ACGAACGGAAGCAACAGATTCATTCATCCCTCTTTTCTTCTTTGTAATCGTTTCCATTGCGTTTCGATTCTTAGATTTCGGATTATCATCTCAACTGTC

               TCTTCACACGAGAGGTTCAAAAGTTACCAAAAAGACTTTTCAAAGTCAATCCTAATTTGTTCGATAATACCACGGCCGAATTACGAAGAACCCTTATCAG
               TCTTCACACGAGAGGTTCAAAAGTTACCAAAAAGACTTTTCAAAGTCAATCCTAATTTGTTCGATAATACCACGGCCGAATTACGAAGAACCCTTATCAG

               TCGATACCCGATAACATCACAACTTCCACACAACTTCAATCTTCTAGAACTTCATCCAATATGGCGGATCAACTCAACATGAATGGTCTCAACCTTAACG
               TCGATACCCGATAACATCACAACTTCCACACAACTTCAATCTTCTAGAACTTCATCCAATATGGCGGATCAACTCAACATGAATGGTCTCAACCTTAACG

               GTGGTGAACCTCGATCATATATTCCTCCCCACATGCGAGGCAAGATGGGTGGCCCTGCTCCAGCTGGTCCACCAGCAAACTTGAATGGCAGTGCTTGGGC
               GTGGTGAACCTCGATCATATATTCCTCCCCACATGCGAGGCAAGATGGGTGGCCCTGCTCCAGCTGGTCCACCAGCAAACTTGAATGGCAGTGCTTGGGC
```

1- genomic region

Variant Explorer

Compare the sequence of a genomic region
in reference genome & in population



Variant Explorer

Extraction results for: "Daphne 35 genomes -- BCIN05:50000-52000"

	gen...	qse...	sse...	pid...	len...	mis...	gap...	qst...	qend	sstart	send	eval...	bits...
👁	G2.fa	BCIN05...	G20567	99.95	2001	1	0	1	2001	17172	15172	0.0	3690
👁	G3.fa	BCIN05...	G30022	100.00	2001	0	0	1	2001	165474	163474	0.0	3696
👁	G4.fa	BCIN05...	G40581	100.00	2001	0	0	1	2001	17168	15168	0.0	3696
👁	G5.fa	BCIN05...	G50561	100.00	2001	0	0	1	2001	17172	15172	0.0	3696
👁	G6.fa	BCIN05...	G60397	99.95	2001	1	0	1	2001	16563	18563	0.0	3690

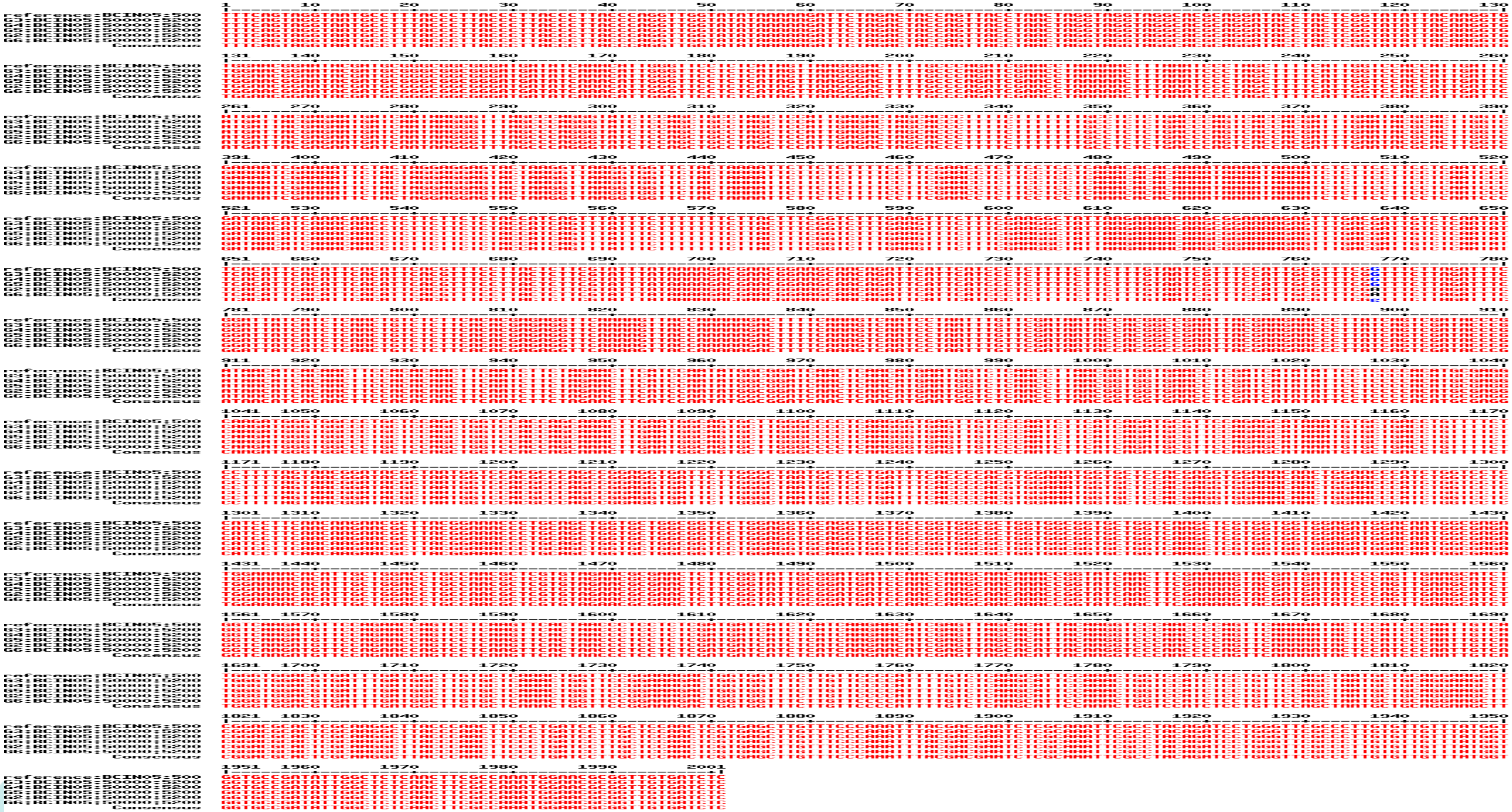
1- genomic region

Variant Explorer

Compare the sequence of a genomic region
in reference genome & in population

Fasta aligned sequences

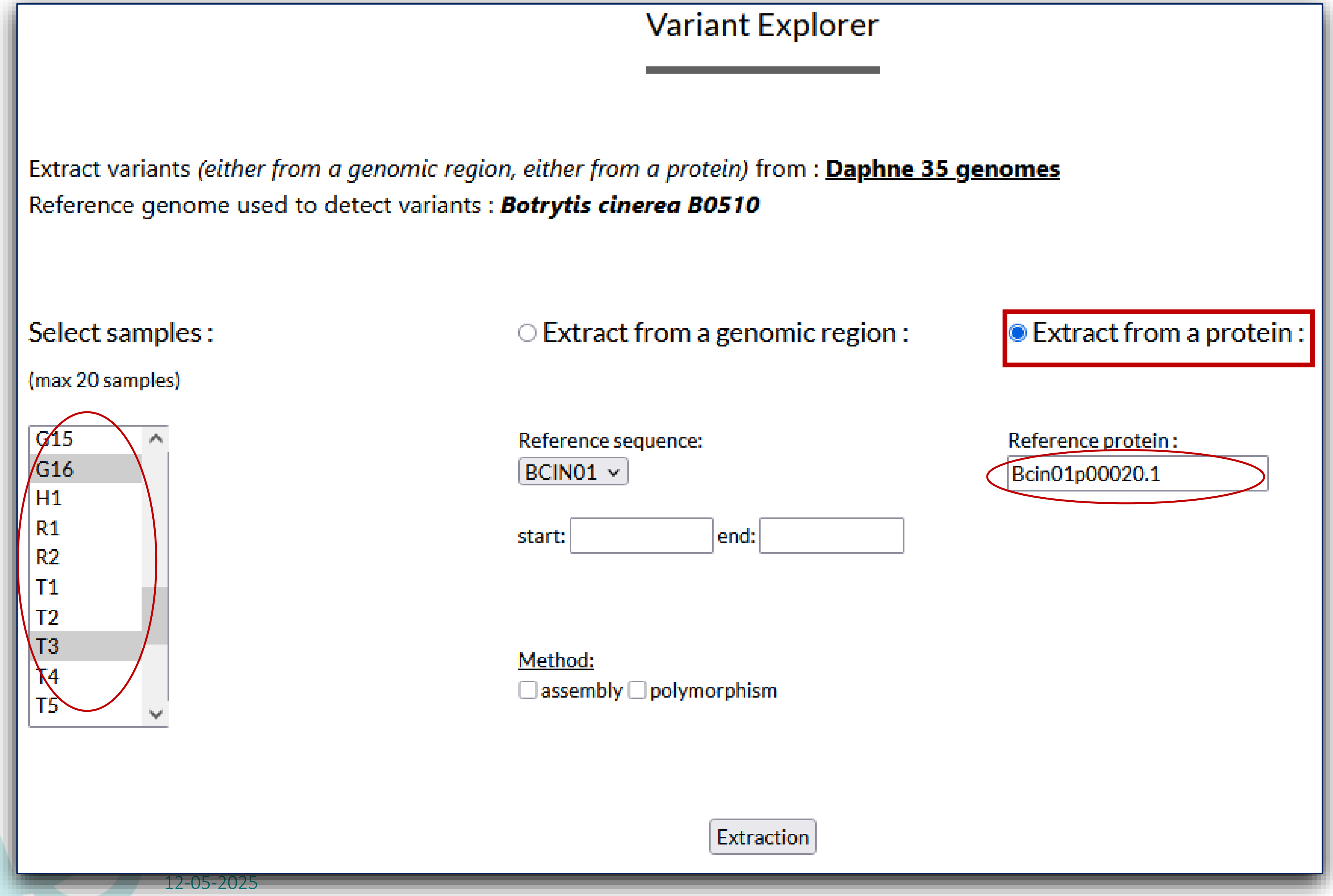
Visualised with http://multalin.toulouse.inra.fr/multalin/cgi-bin/multalin.pl



2- proteins

Variant Explorer

Retrieve the sequences of reference proteins in population



Variant Explorer

Extract variants *(either from a genomic region, either from a protein)* from : **Daphne 35 genomes**

Reference genome used to detect variants : ***Botrytis cinerea B0510***

Select samples :

(max 20 samples)

- G15
- G16
- H1
- R1
- R2
- T1
- T2
- T3
- T4
- T5

○ Extract from a genomic region :

Reference sequence: BCIN01

start: end:

Method:

☐ assembly ☐ polymorphism

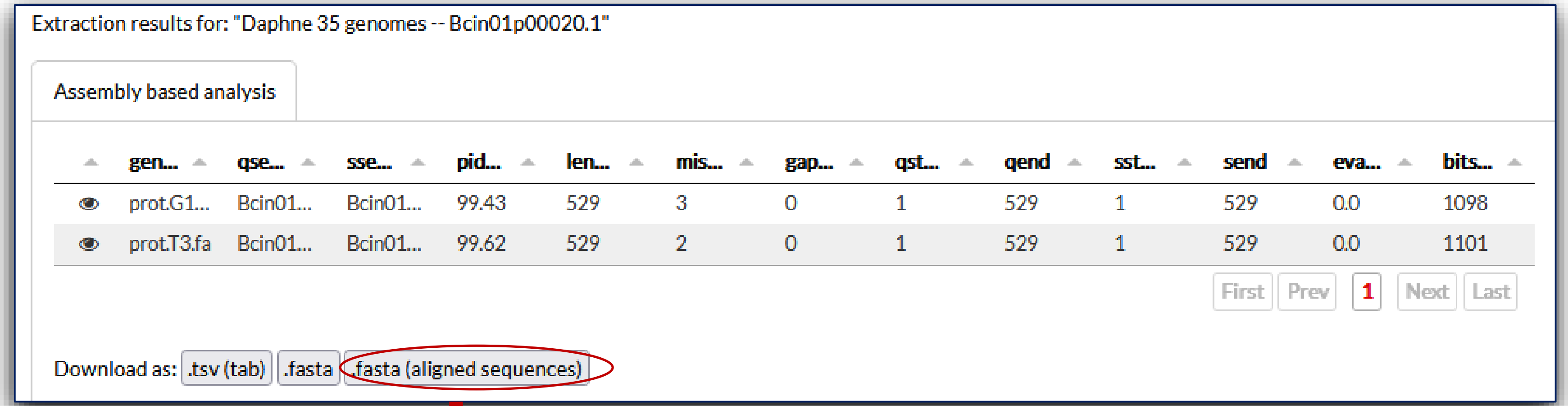
● Extract from a protein :

Reference protein : Bcin01p00020.1

2- proteins

Variant Explorer

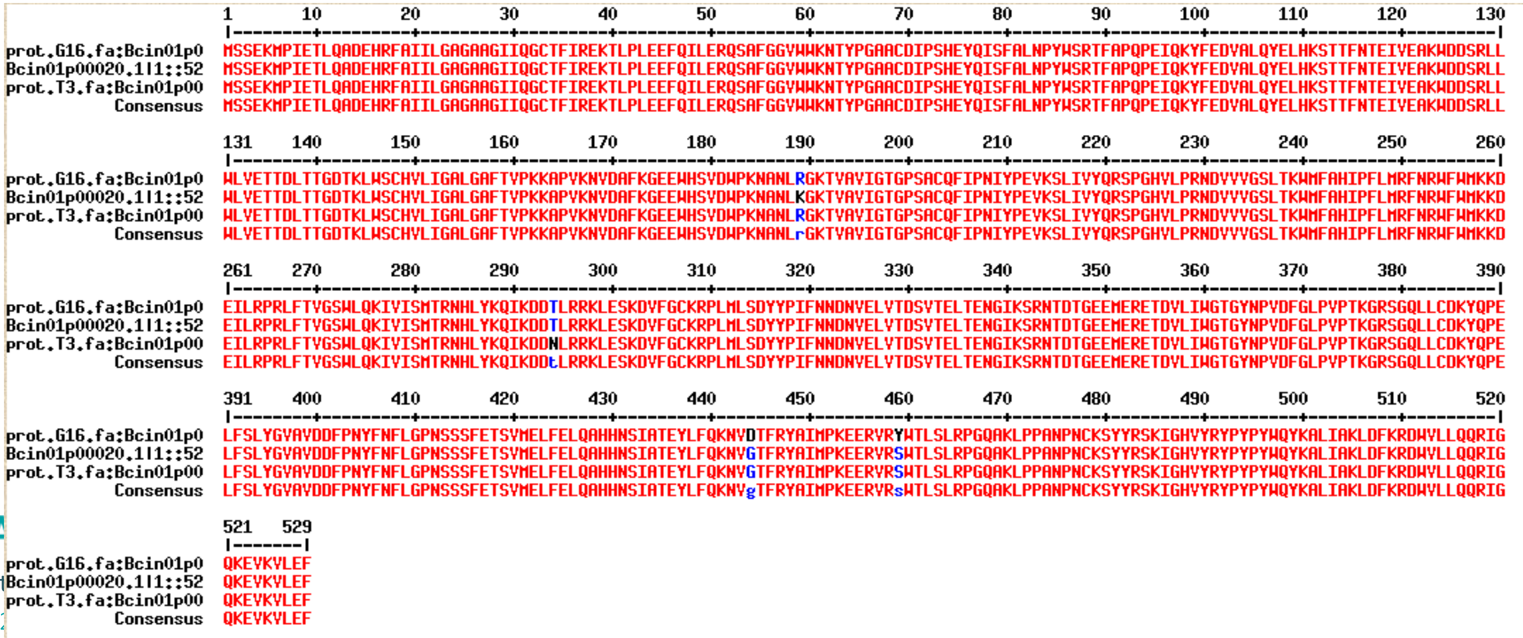
Retrieve the sequences of reference proteins in population



Extraction results for: "Daphne 35 genomes -- Bcin01p00020.1"

	gen...	qse...	sse...	pid...	len...	mis...	gap...	qst...	qend	sst...	send	eva...	bits...
👁	prot.G1...	Bcin01...	Bcin01...	99.43	529	3	0	1	529	1	529	0.0	1098
👁	prot.T3.fa	Bcin01...	Bcin01...	99.62	529	2	0	1	529	1	529	0.0	1101

Visualised with http://multalin.toulouse.inra.fr/multalin/cgi-bin/multalin.pl



```
                      1       10        20        30        40        50        60        70        80        90       100       110       120       130
                      |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
prot.G16.fa:Bcin01p0  MSSEKMPIETLQADEHRFAIILGAGAAGIIQGCTFIREKTLPLEEFQILERQSAFGGVWWKNTYPGAACDIPSHEYQISFALNPYWSRTFAPQPEIQKYFEDVALQYELHKSTTFNTEIVEAKWDDSRLL
Bcin01p00020.1|1::52  MSSEKMPIETLQADEHRFAIILGAGAAGIIQGCTFIREKTLPLEEFQILERQSAFGGVWWKNTYPGAACDIPSHEYQISFALNPYWSRTFAPQPEIQKYFEDVALQYELHKSTTFNTEIVEAKWDDSRLL
prot.T3.fa:Bcin01p00  MSSEKMPIETLQADEHRFAIILGAGAAGIIQGCTFIREKTLPLEEFQILERQSAFGGVWWKNTYPGAACDIPSHEYQISFALNPYWSRTFAPQPEIQKYFEDVALQYELHKSTTFNTEIVEAKWDDSRLL
           Consensus  MSSEKMPIETLQADEHRFAIILGAGAAGIIQGCTFIREKTLPLEEFQILERQSAFGGVWWKNTYPGAACDIPSHEYQISFALNPYWSRTFAPQPEIQKYFEDVALQYELHKSTTFNTEIVEAKWDDSRLL

                      131    140       150       160       170       180       190       200       210       220       230       240       250       260
                      |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
prot.G16.fa:Bcin01p0  WLVETTDLTTGDTKLWSCHVLIGALGAFTVPKKAPVKNVDAFKGEEWHSVDWPKNANLRGKTVAVIGTGPSACQFIPNIYPEVKSLIVYQRSPGHVLPRNDVVVGSLTKWMFAHIPFLMRFNRWFWMKKD
Bcin01p00020.1|1::52  WLVETTDLTTGDTKLWSCHVLIGALGAFTVPKKAPVKNVDAFKGEEWHSVDWPKNANLKGKTVAVIGTGPSACQFIPNIYPEVKSLIVYQRSPGHVLPRNDVVVGSLTKWMFAHIPFLMRFNRWFWMKKD
prot.T3.fa:Bcin01p00  WLVETTDLTTGDTKLWSCHVLIGALGAFTVPKKAPVKNVDAFKGEEWHSVDWPKNANLRGKTVAVIGTGPSACQFIPNIYPEVKSLIVYQRSPGHVLPRNDVVVGSLTKWMFAHIPFLMRFNRWFWMKKD
           Consensus  WLVETTDLTTGDTKLWSCHVLIGALGAFTVPKKAPVKNVDAFKGEEWHSVDWPKNANLrGKTVAVIGTGPSACQFIPNIYPEVKSLIVYQRSPGHVLPRNDVVVGSLTKWMFAHIPFLMRFNRWFWMKKD

                      261    270       280       290       300       310       320       330       340       350       360       370       380       390
                      |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
prot.G16.fa:Bcin01p0  EILRPRLFTVGSWLQKIVISMTRNHLYKQIKDDTLRRKLESKDVFGCKRPLMLSDYYPIFNNDNVELVTDSVTELTENGIKSRNTDTGEEMERETDVLIWGTGYNPVDFGLPVPTKGRSGQLLCDKYQPE
Bcin01p00020.1|1::52  EILRPRLFTVGSWLQKIVISMTRNHLYKQIKDDTLRRKLESKDVFGCKRPLMLSDYYPIFNNDNVELVTDSVTELTENGIKSRNTDTGEEMERETDVLIWGTGYNPVDFGLPVPTKGRSGQLLCDKYQPE
prot.T3.fa:Bcin01p00  EILRPRLFTVGSWLQKIVISMTRNHLYKQIKDDNLRRKLESKDVFGCKRPLMLSDYYPIFNNDNVELVTDSVTELTENGIKSRNTDTGEEMERETDVLIWGTGYNPVDFGLPVPTKGRSGQLLCDKYQPE
           Consensus  EILRPRLFTVGSWLQKIVISMTRNHLYKQIKDDtLRRKLESKDVFGCKRPLMLSDYYPIFNNDNVELVTDSVTELTENGIKSRNTDTGEEMERETDVLIWGTGYNPVDFGLPVPTKGRSGQLLCDKYQPE

                      391    400       410       420       430       440       450       460       470       480       490       500       510       520
                      |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
prot.G16.fa:Bcin01p0  LFSLYGVAVDDFPNYFNFLGPNSSSFETSVMELFELQAHHNSIATEYLFQKNVDTFRYAIMPKEERVRYWTLSLRPGQAKLPPANPNCKSYYRSKIGHVYRYPYPYWQYKALIAKLDFKRDWVLLQQRIG
Bcin01p00020.1|1::52  LFSLYGVAVDDFPNYFNFLGPNSSSFETSVMELFELQAHHNSIATEYLFQKNVGTFRYAIMPKEERVRSWTLSLRPGQAKLPPANPNCKSYYRSKIGHVYRYPYPYWQYKALIAKLDFKRDWVLLQQRIG
prot.T3.fa:Bcin01p00  LFSLYGVAVDDFPNYFNFLGPNSSSFETSVMELFELQAHHNSIATEYLFQKNVGTFRYAIMPKEERVRSWTLSLRPGQAKLPPANPNCKSYYRSKIGHVYRYPYPYWQYKALIAKLDFKRDWVLLQQRIG
           Consensus  LFSLYGVAVDDFPNYFNFLGPNSSSFETSVMELFELQAHHNSIATEYLFQKNVgTFRYAIMPKEERVRsWTLSLRPGQAKLPPANPNCKSYYRSKIGHVYRYPYPYWQYKALIAKLDFKRDWVLLQQRIG

                      521   529
                      |-------|
prot.G16.fa:Bcin01p0  QKEVKVLEF
Bcin01p00020.1|1::52  QKEVKVLEF
prot.T3.fa:Bcin01p00  QKEVKVLEF
           Consensus  QKEVKVLEF
```

* Graphs in bioinformatics - assembly

Overlap Layout Consensus (OLC) vs De-Bruijn Graph (DBG)

OLC, overlap cutoff = 5

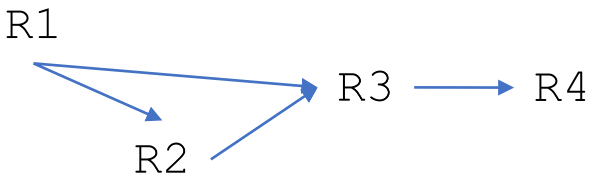
..ATGTGCTTATCGGTAAGTTG..

R1 ATG**TGCTTATC**

R2 **TGCTTATC**GGT

R3 GCTTAT**CGGTAA**

R4 **CGGTAA**GTTG



DBG, k-mer = 4

..ATGTGCTTATCGGTAAGTTG..

K1 ATGT

K2 TGTG

K3 GTGC

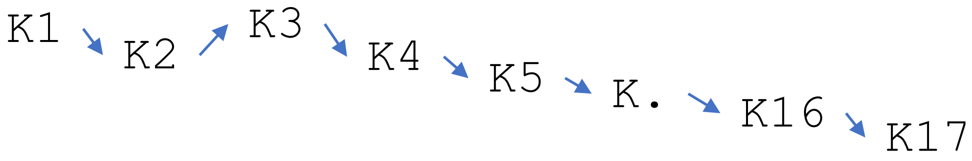
K4 TGCT

K5 GCTT

K. ....

K16 AGTT

K17 GTTG

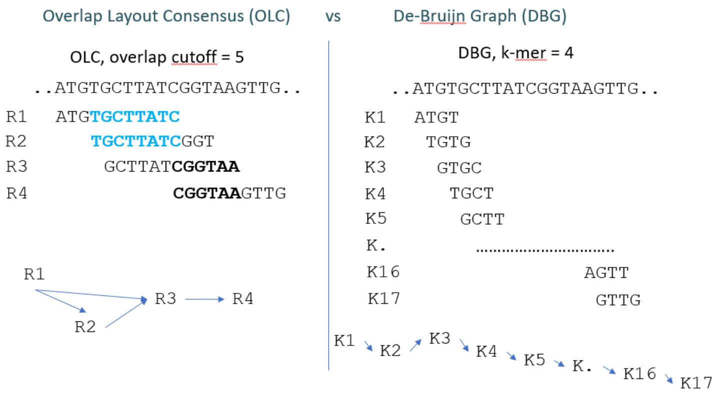


Sequencing errors and sequence repetitiveness !!!

Zhenyu Li, et al, Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph, *Briefings in Functional Genomics*, Volume 11, Issue 1, January 2012, Pages 25–37, https://doi.org/10.1093/bfgp/elr035

* Graphs in bioinformatics - assembly

From reads to scaffolds



Overlap Layout Consensus (OLC) vs De-Bruijn Graph (DBG)

OLC, overlap cutoff = 5

..ATGTGCTTATCGGTAAGTTG..

R1 ATG**TGCTTATC**
R2 **TGCTTATC**GGT
R3 GCTTAT**CGGTAA**
R4 **CGGTAA**GTTG

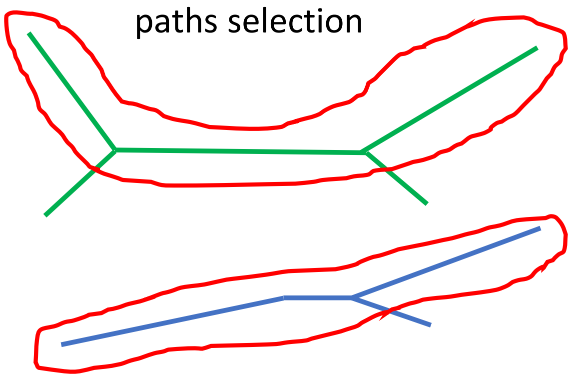
R1 → R2 → R3 → R4
R1 → R3

DBG, k-mer = 4

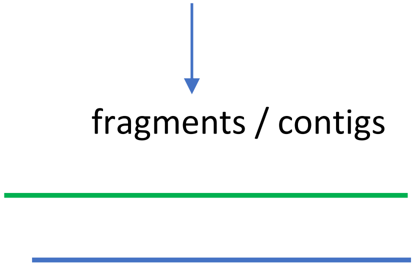
..ATGTGCTTATCGGTAAGTTG..

K1 ATGT
K2 TGTG
K3 GTGC
K4 TGCT
K5 GCTT
K. ………………………….
K16 AGTT
K17 GTTG

K1 → K2 → K3 → K4 → K5 → K. → K16 → K17

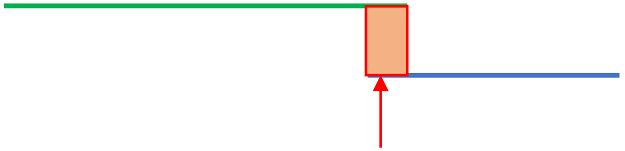


paths selection



fragments / contigs

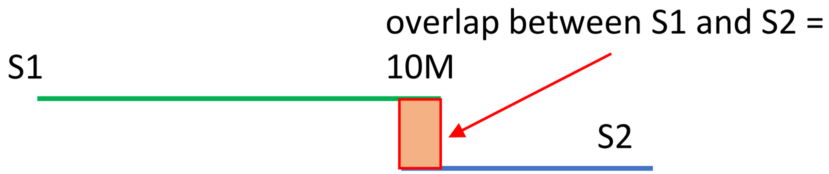
scaffolds / contigs



overlap between contigs, allows scaffolding / chaining

Origin of the pangenome graphs, the GFA Format

GFA (Graphical Fragment Assembly) format (main format to store pangenome graph) comes from Genome Assembly methods



GFA 1.0:

```
S   S1   TAAAGCGCCTAAAAGGGGGTTAACGGAGCTAAACAGACCCGAACTAACTATATATACCCCTT
S   S2   TATACCCCTTCTTCTCTTTATCCTTTAGTACAAGGATAAGAGGCATCGGT
L   S1   +   S2 +   10M
P   contig1   S1+,S2+   *
```