Genostar

Microbial bioinformatics

Bioinformatics for NGS data analysis

Genostar

Microbial bioinformatics

ABS4NGS

June 22-23, 2015

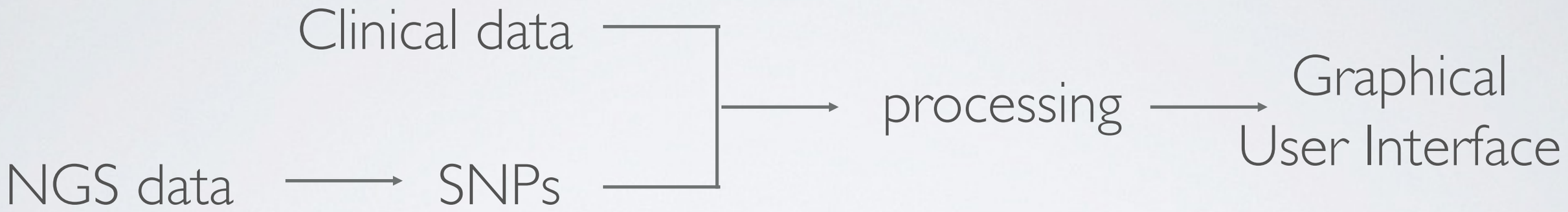
François Rechenmann, CEO

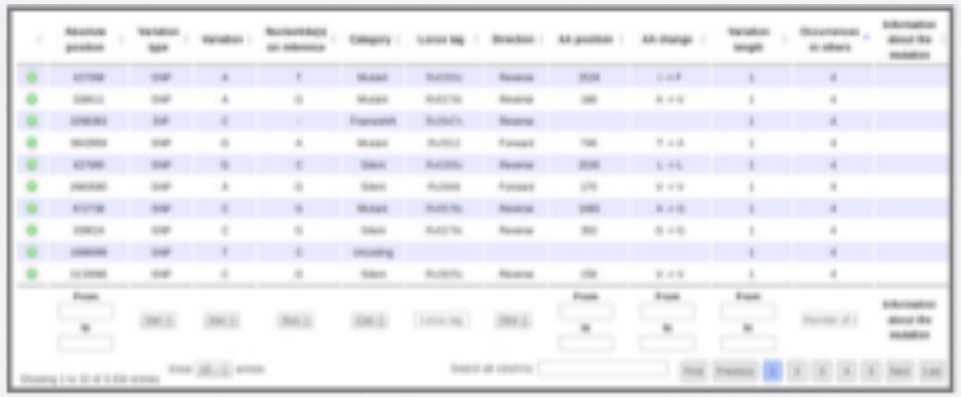
INVESTISSEMENTS D'AVENIR

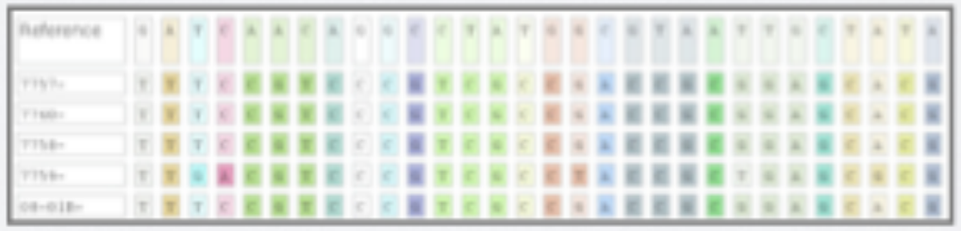
Genostar
Bioinformatics solutions

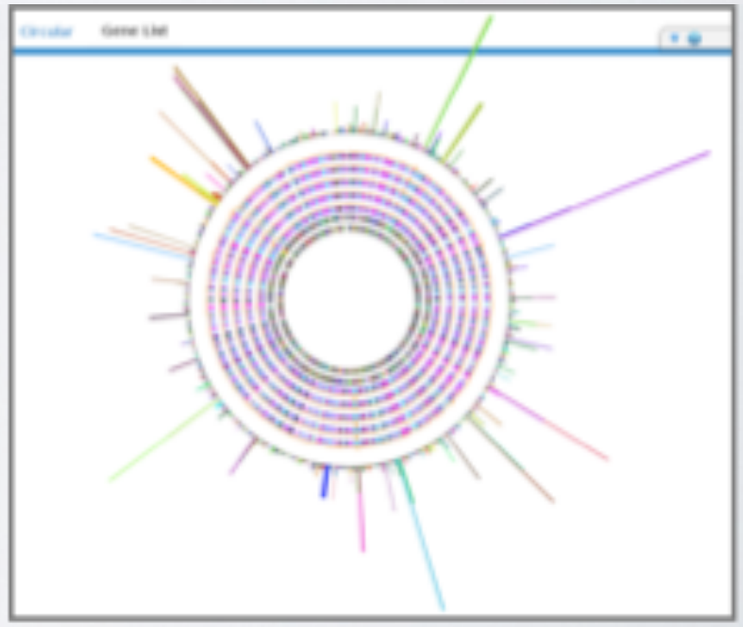
VARIANT

PRESENTATION





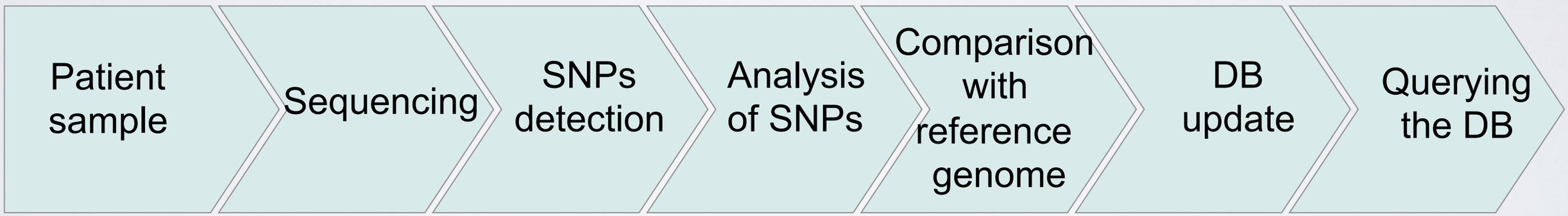




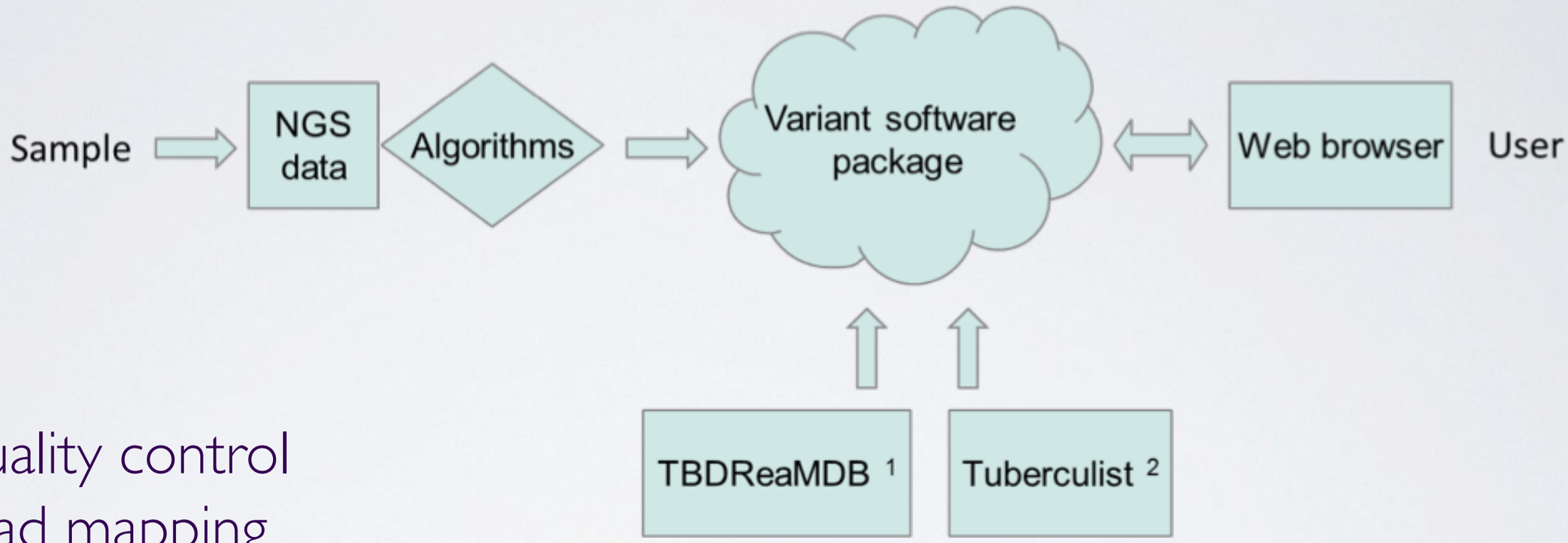
Genostar
Bioinformatics solutions

VARIANT

WORKFLOW



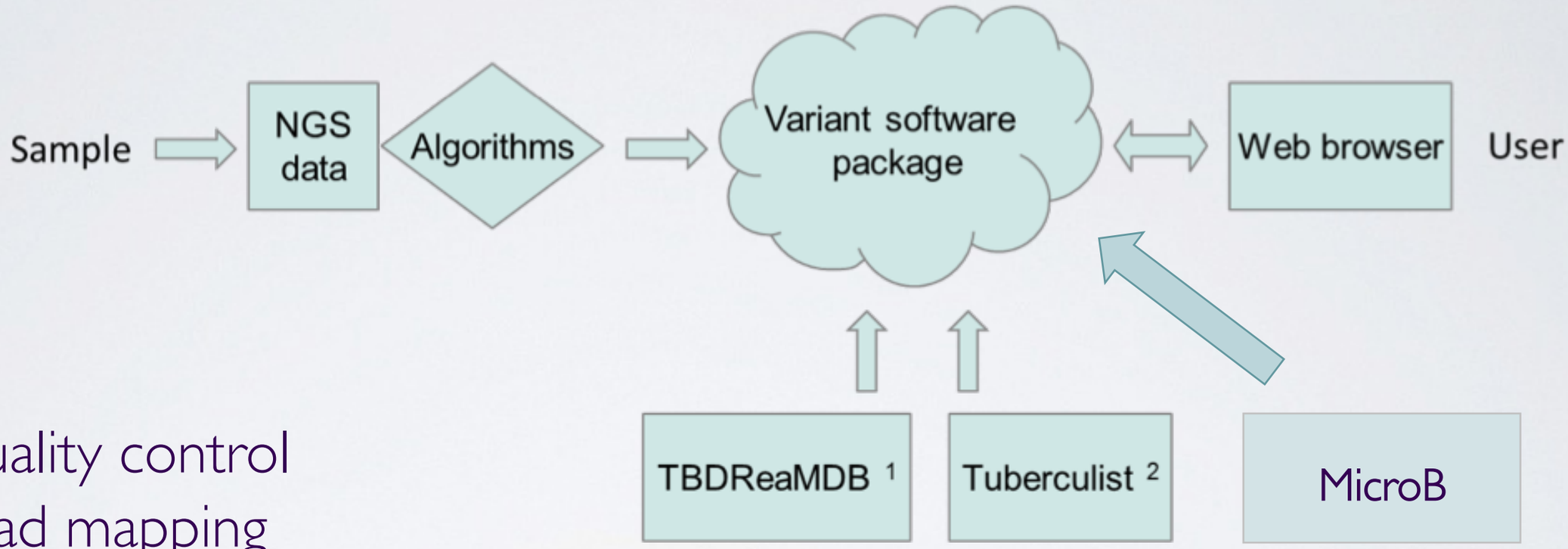
VARIANT



- Quality control
- Read mapping
- Variant calling
- SNP characterization
  (mutant, silent, intergenic)
- SNP filtering

- Reference database
- Meta-analysis methods
- Graphical User interface

VARIANT

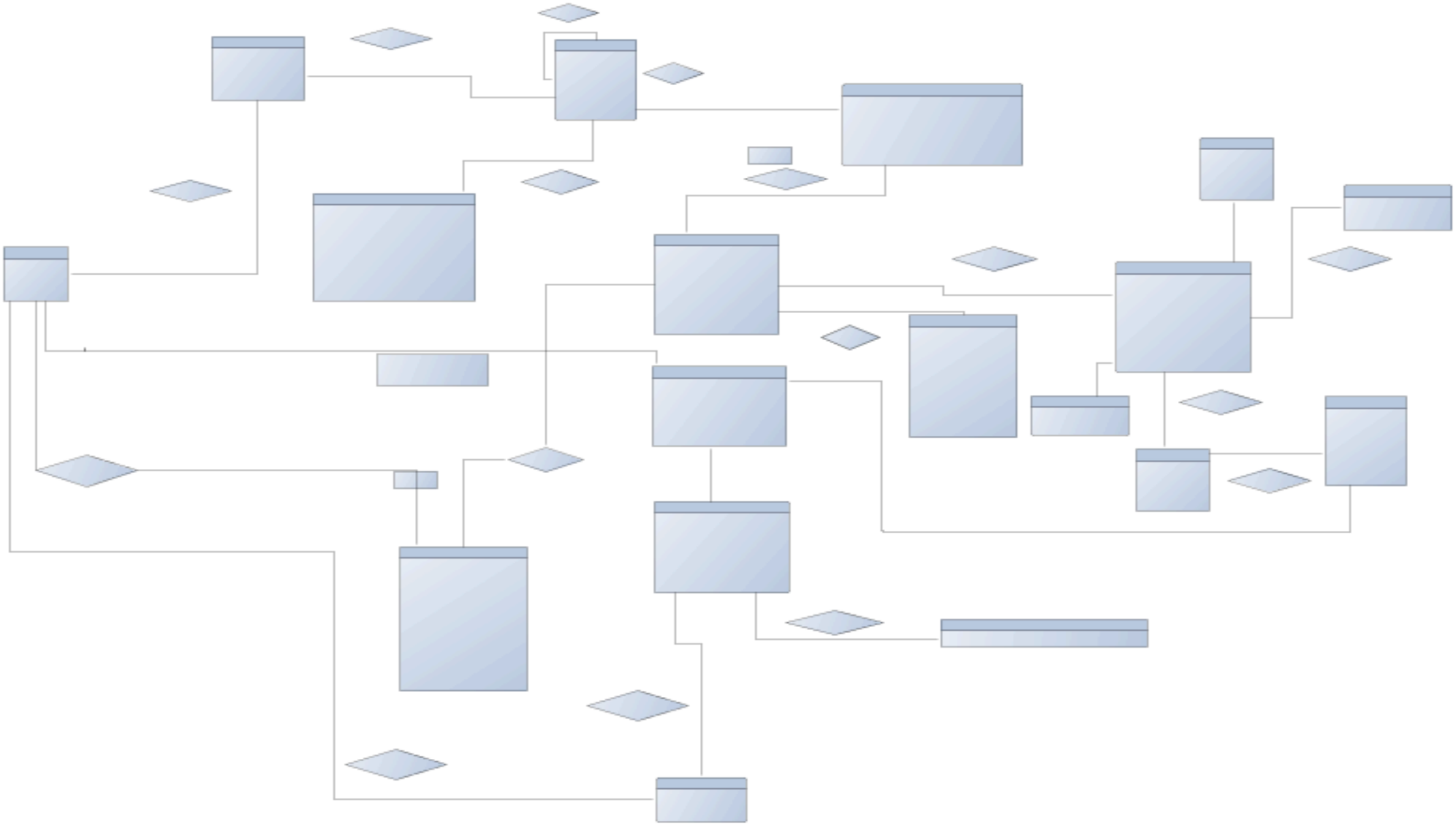


- Quality control
- Read mapping
- Variant calling
- SNP characterization
  (mutant, silent, intergenic)
- SNP filtering

- Reference database
- Meta-analysis methods
- Graphical User interface

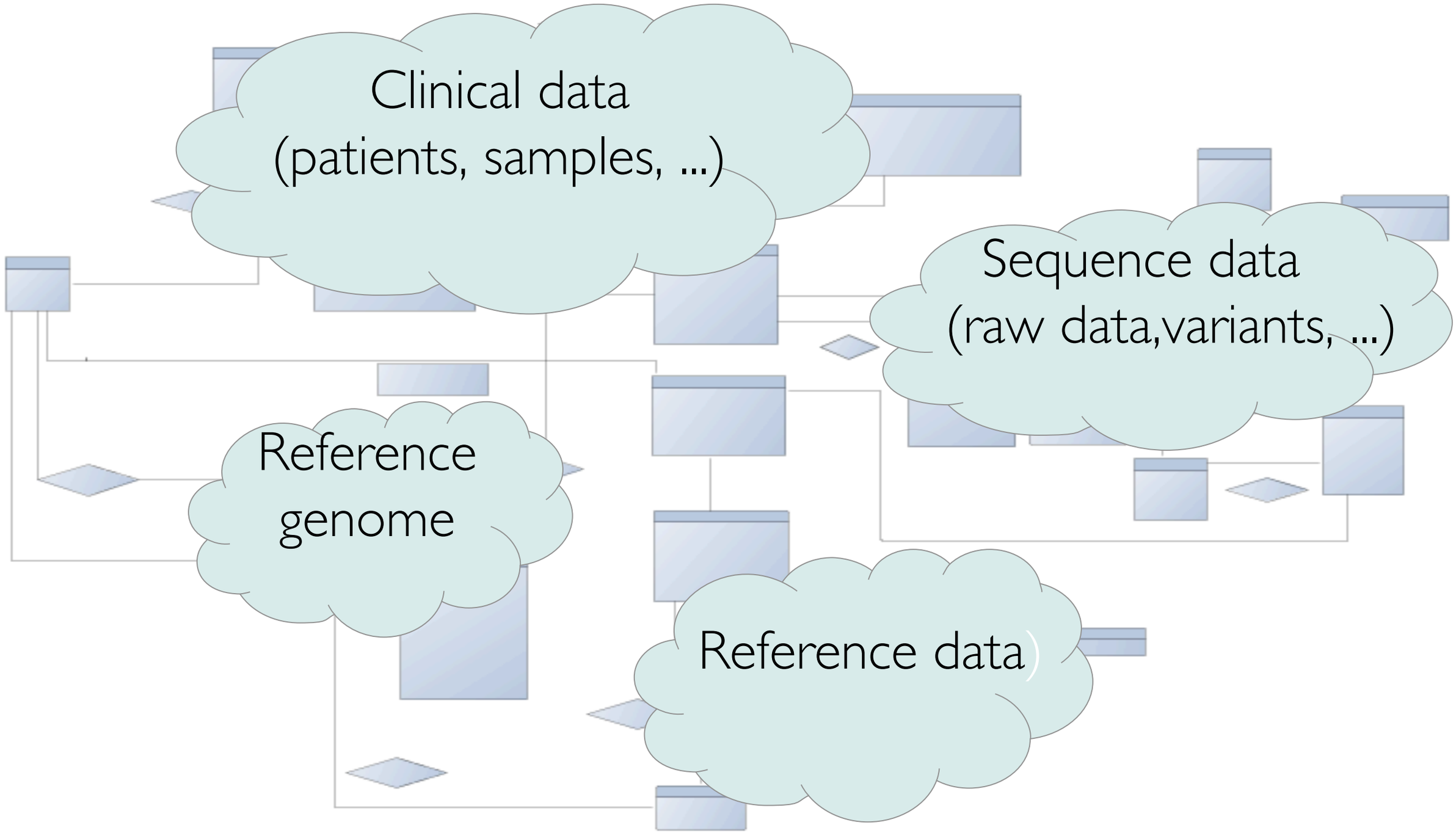
Genostar
Bioinformatics solutions

VARIANT DATABASE



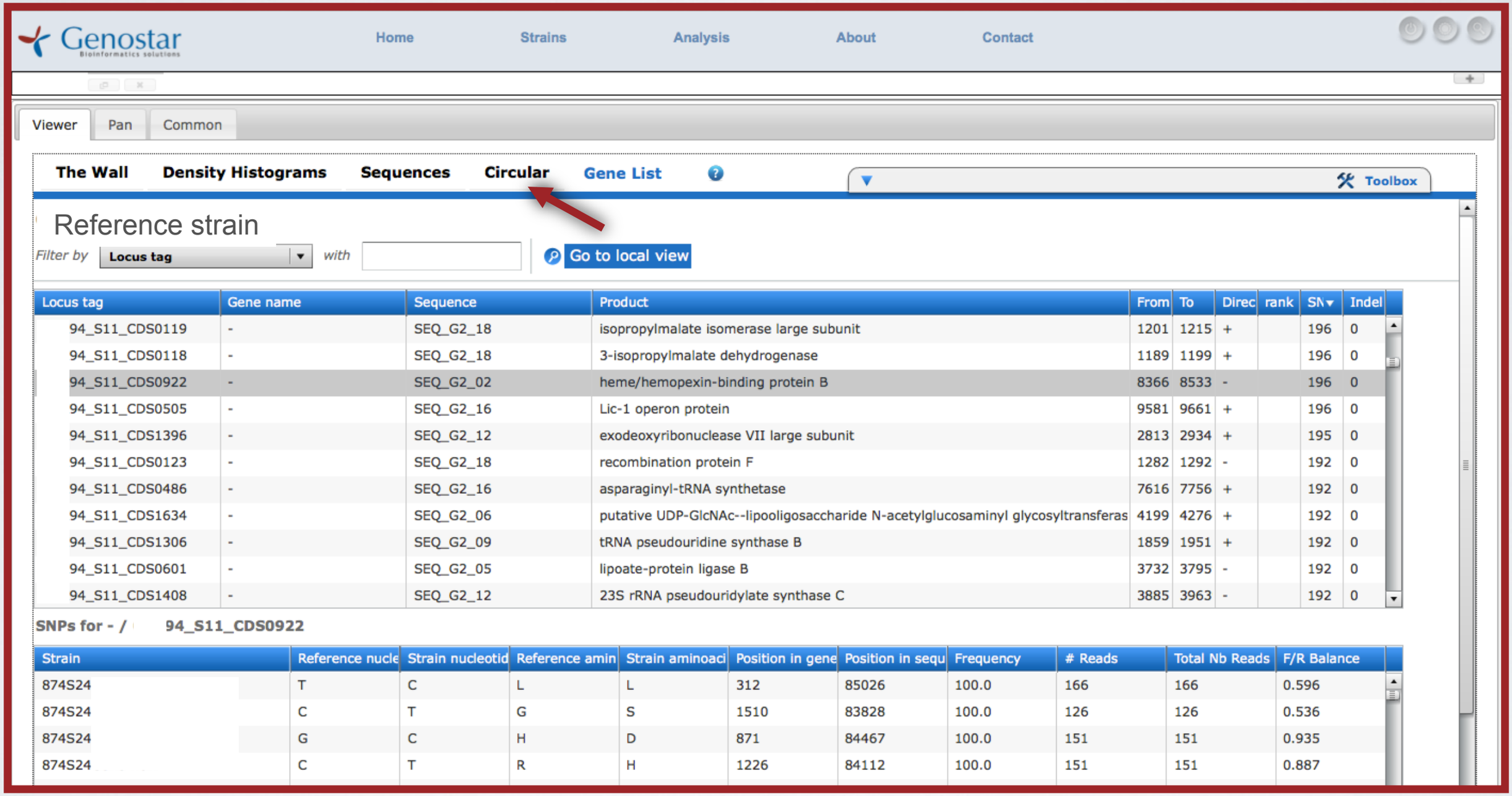
Genostar
Bioinformatics solutions

VARIANT DATABASE



Genostar
Bioinformatics solutions

VARIANT



Locus tag	Gene name	Sequence	Product	From	To	Direc	rank	SN	Indel
94_S11_CDS0119	-	SEQ_G2_18	isopropylmalate isomerase large subunit	1201	1215	+		196	0
94_S11_CDS0118	-	SEQ_G2_18	3-isopropylmalate dehydrogenase	1189	1199	+		196	0
94_S11_CDS0922	-	SEQ_G2_02	heme/hemopexin-binding protein B	8366	8533	-		196	0
94_S11_CDS0505	-	SEQ_G2_16	Lic-1 operon protein	9581	9661	+		196	0
94_S11_CDS1396	-	SEQ_G2_12	exodeoxyribonuclease VII large subunit	2813	2934	+		195	0
94_S11_CDS0123	-	SEQ_G2_18	recombination protein F	1282	1292	-		192	0
94_S11_CDS0486	-	SEQ_G2_16	asparaginyl-tRNA synthetase	7616	7756	+		192	0
94_S11_CDS1634	-	SEQ_G2_06	putative UDP-GlcNAc--lipooligosaccharide N-acetylglucosaminyl glycosyltransferas	4199	4276	+		192	0
94_S11_CDS1306	-	SEQ_G2_09	tRNA pseudouridine synthase B	1859	1951	+		192	0
94_S11_CDS0601	-	SEQ_G2_05	lipoate-protein ligase B	3732	3795	-		192	0
94_S11_CDS1408	-	SEQ_G2_12	23S rRNA pseudouridylate synthase C	3885	3963	-		192	0

SNPs for - / 94_S11_CDS0922

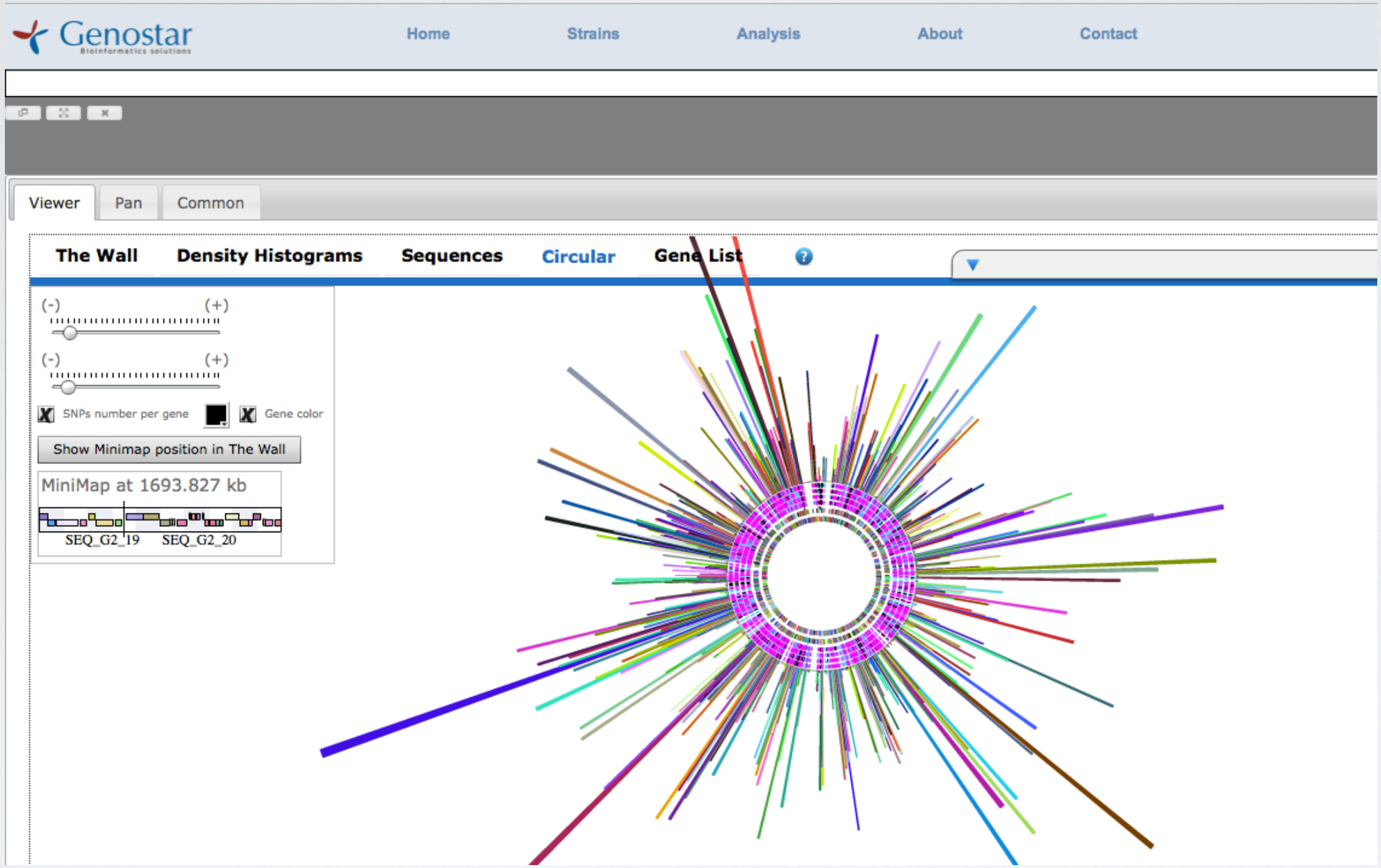
Strain	Reference nucle	Strain nucleotid	Reference amin	Strain aminoaci	Position in gene	Position in sequ	Frequency	# Reads	Total Nb Reads	F/R Balance
874S24	T	C	L	L	312	85026	100.0	166	166	0.596
874S24	C	T	G	S	1510	83828	100.0	126	126	0.536
874S24	G	C	H	D	871	84467	100.0	151	151	0.935
874S24	C	T	R	H	1226	84112	100.0	151	151	0.887

Reference genome Str_0094

Compared strains : Str_0095, Str_0102, Str_0103, Str_0874 and Str_0876

Genostar
Bioinformatics solutions

VARIANT

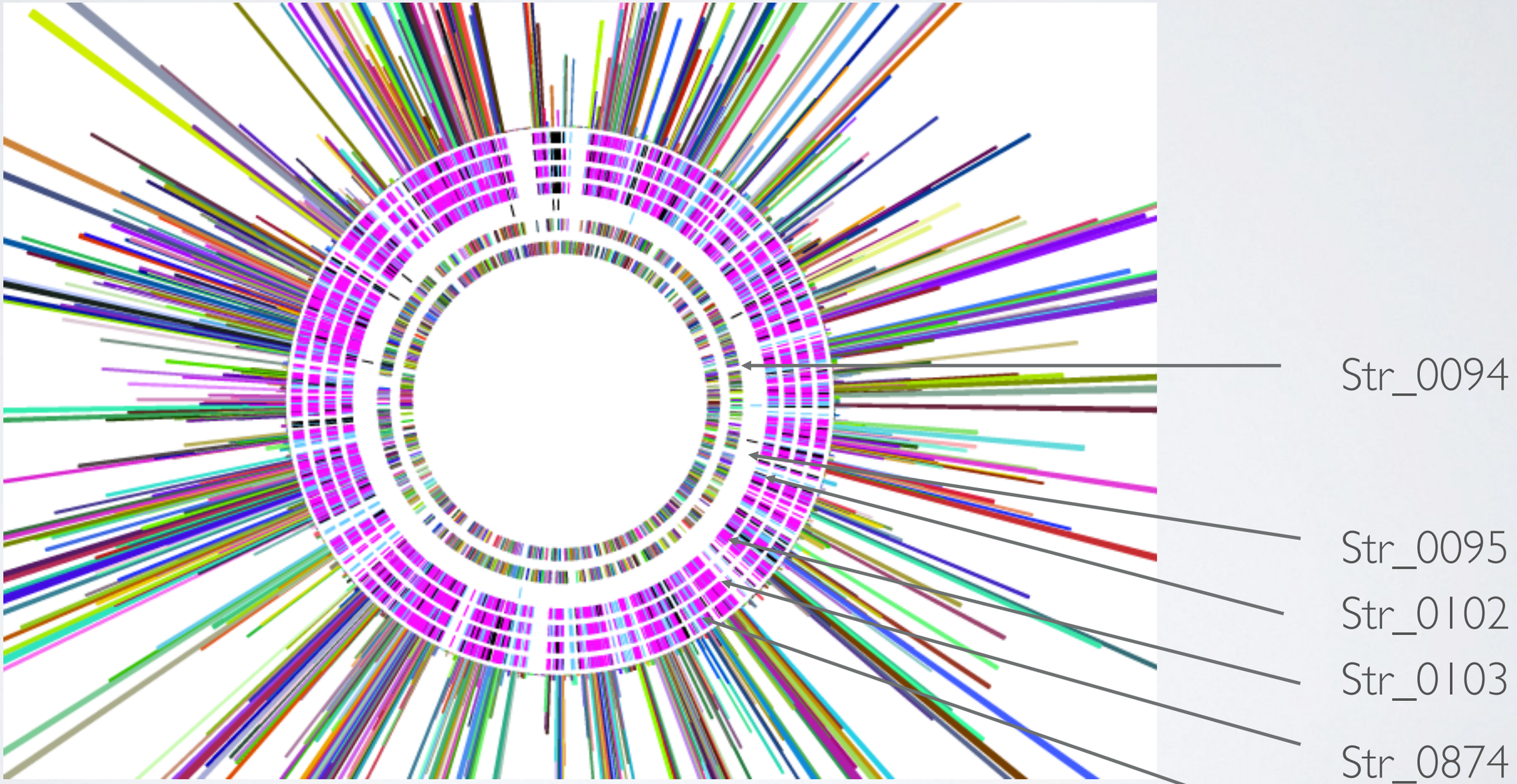


Positions and variant type (intergenic, silent, mutant)

Mutation hot spot: synchronisation with the wall

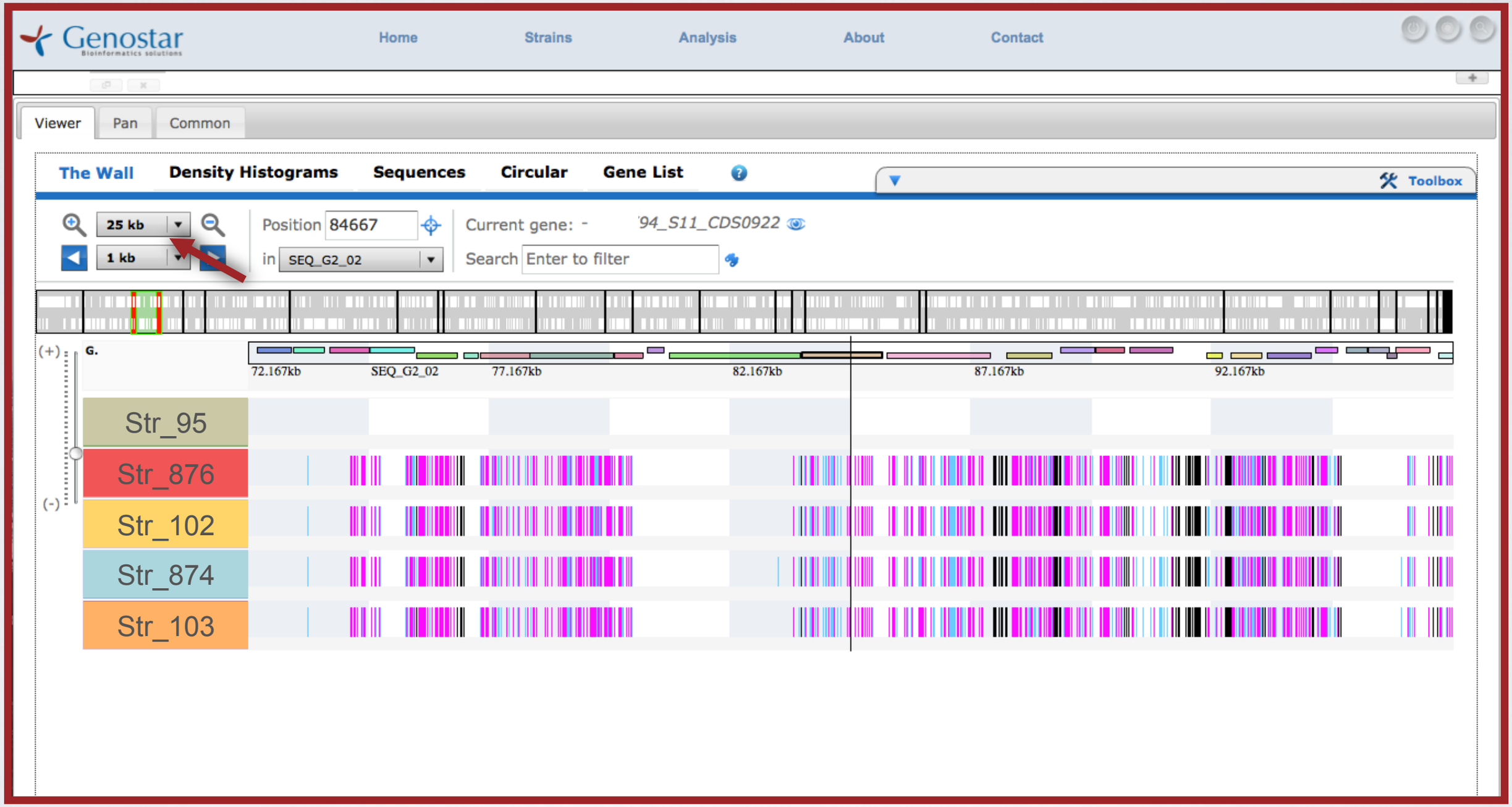
Genostar
Bioinformatics solutions

VARIANT



Genostar
Bioinformatics solutions

VARIANT

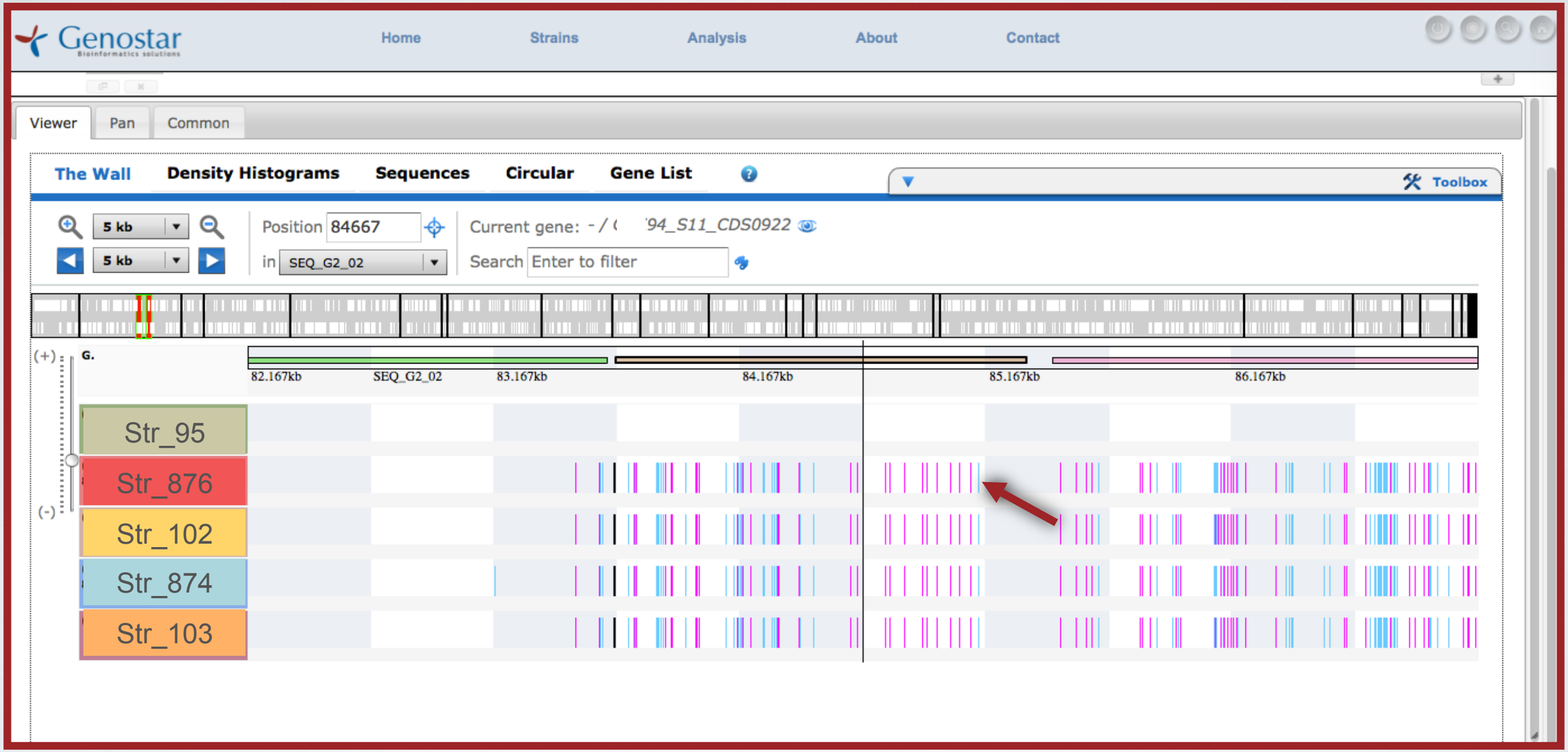


SNPs on genomic map

Hot spot : Str_94_S11_CDS_0922 (heme/hemopexine-binding protein B)

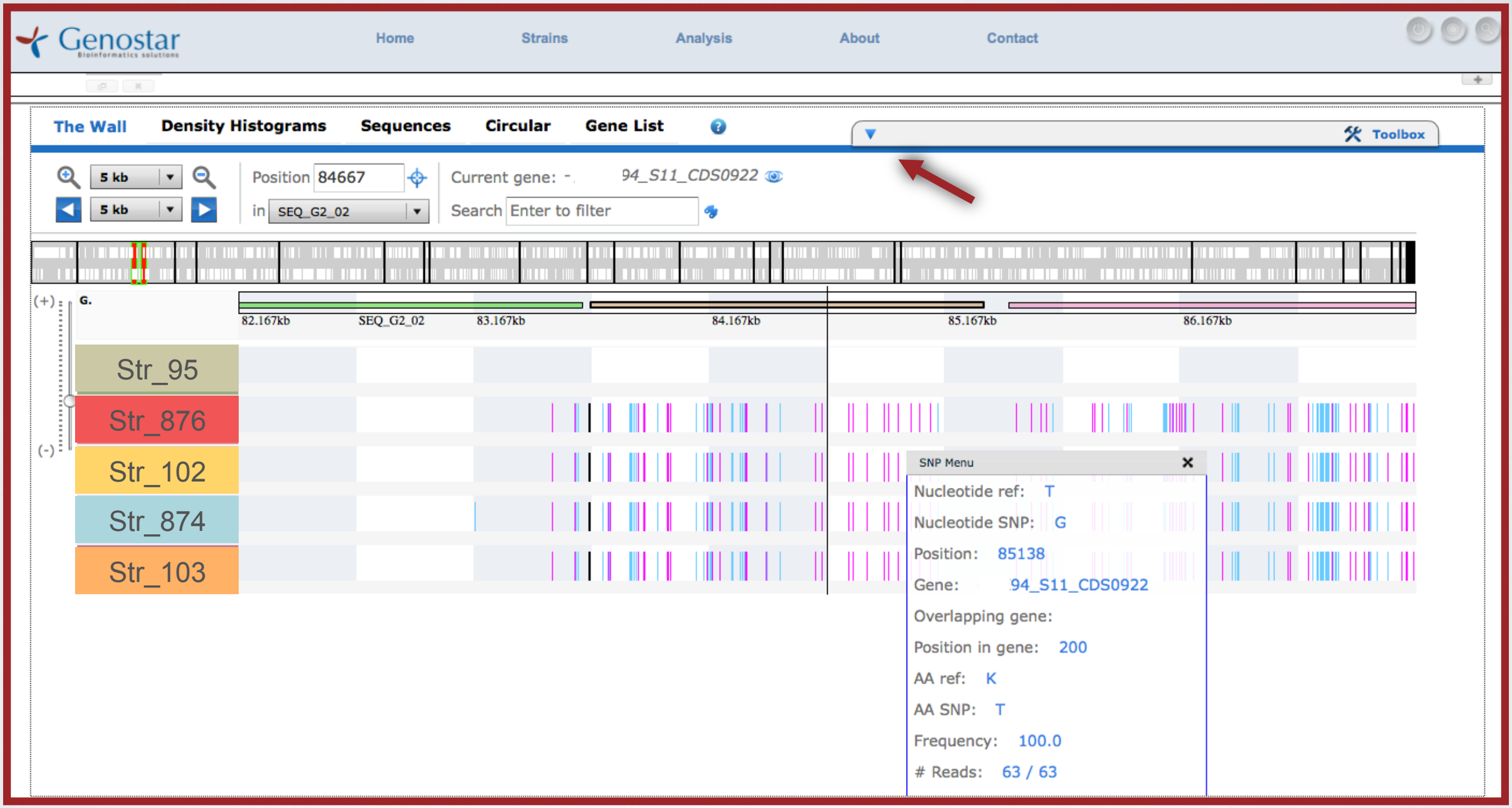
Genostar
Bioinformatics solutions

VARIANT



Zoom on gene Str_94_S11_CDS_0922 (heme/hemopexine-binding protein B)

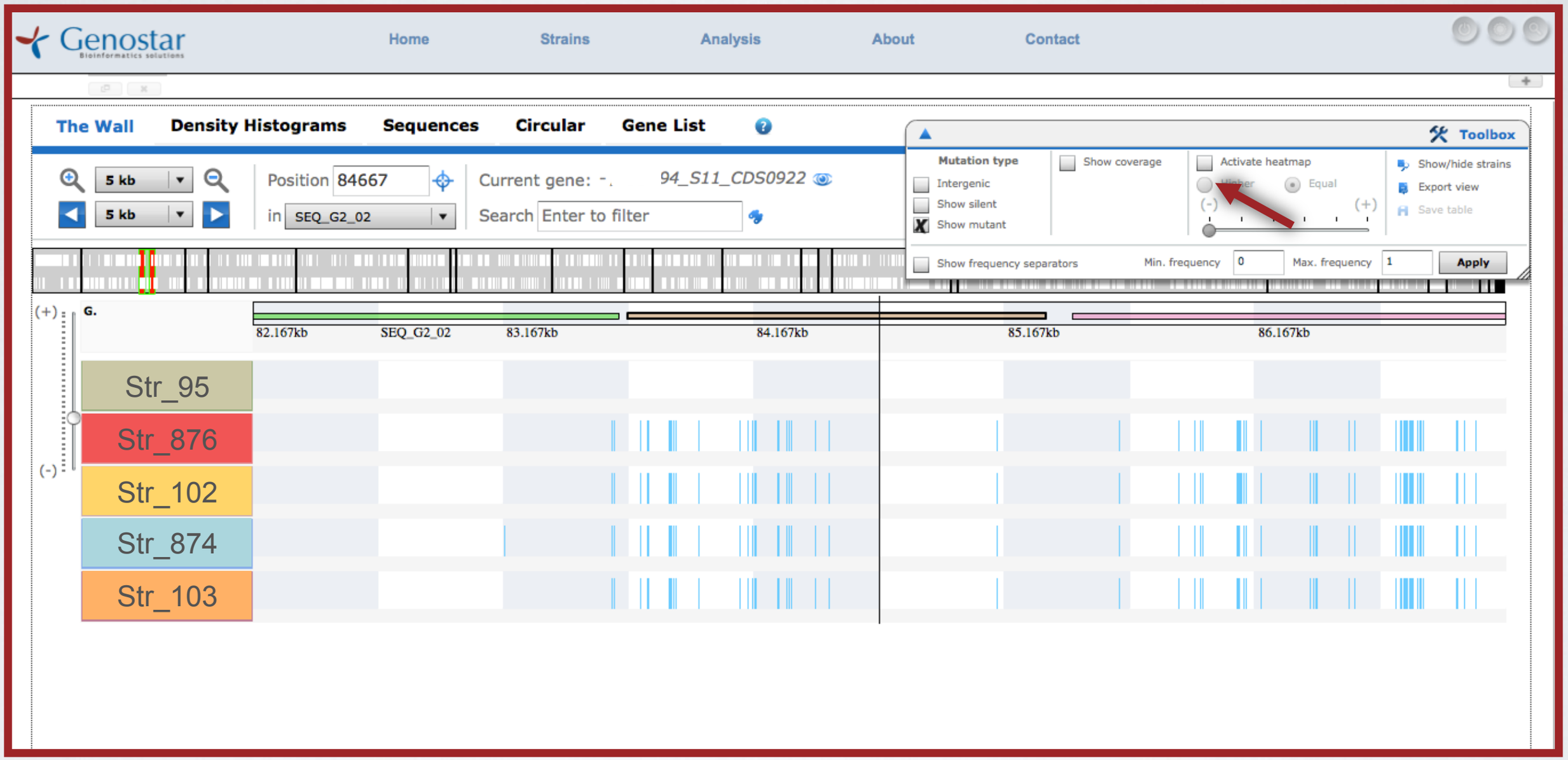
VARIANT



Clic on a particular SNP

Genostar
Bioinformatics solutions

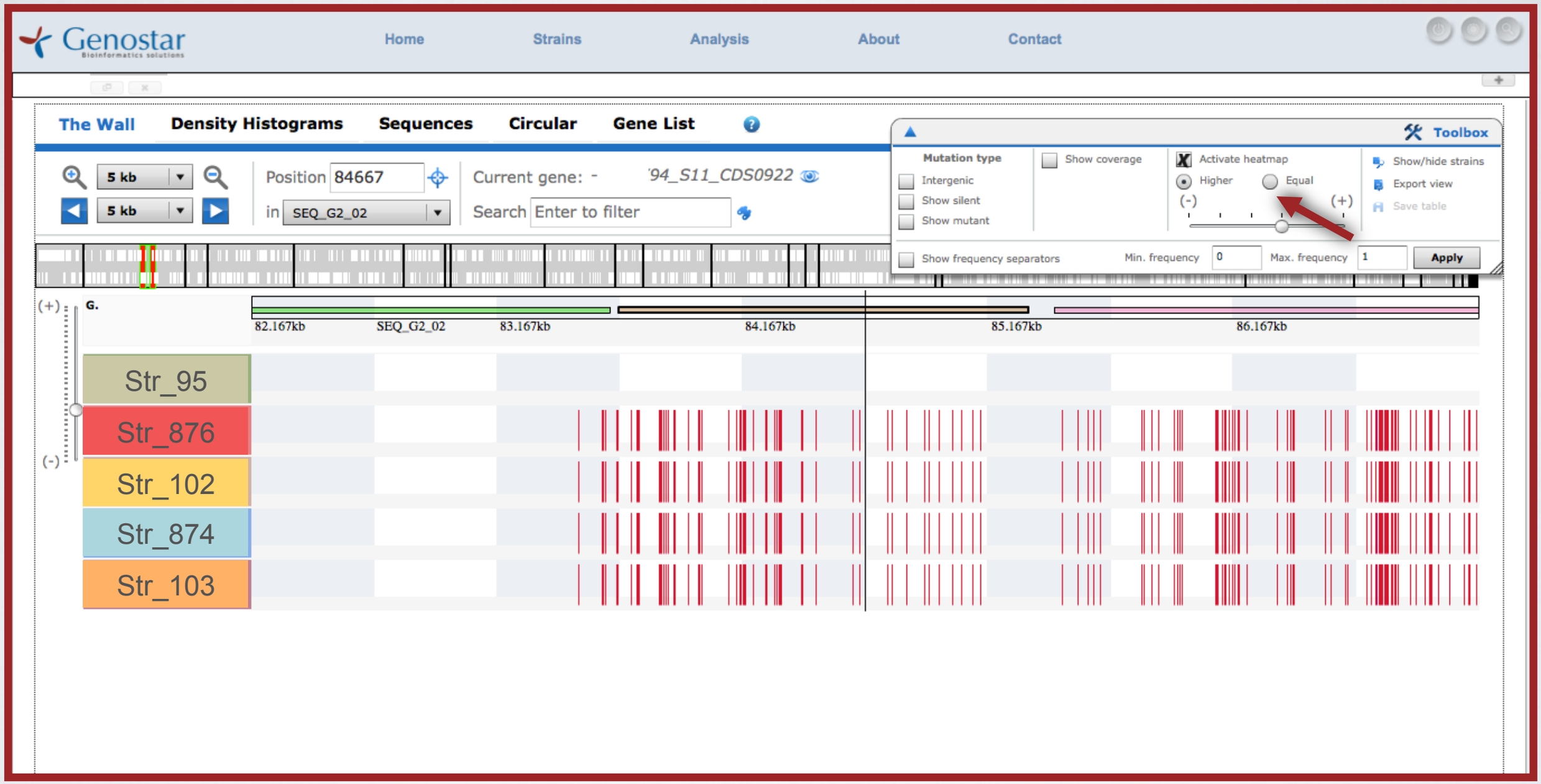
VARIANT



Filtering SNPs: show only mutants

Genostar
Bioinformatics solutions

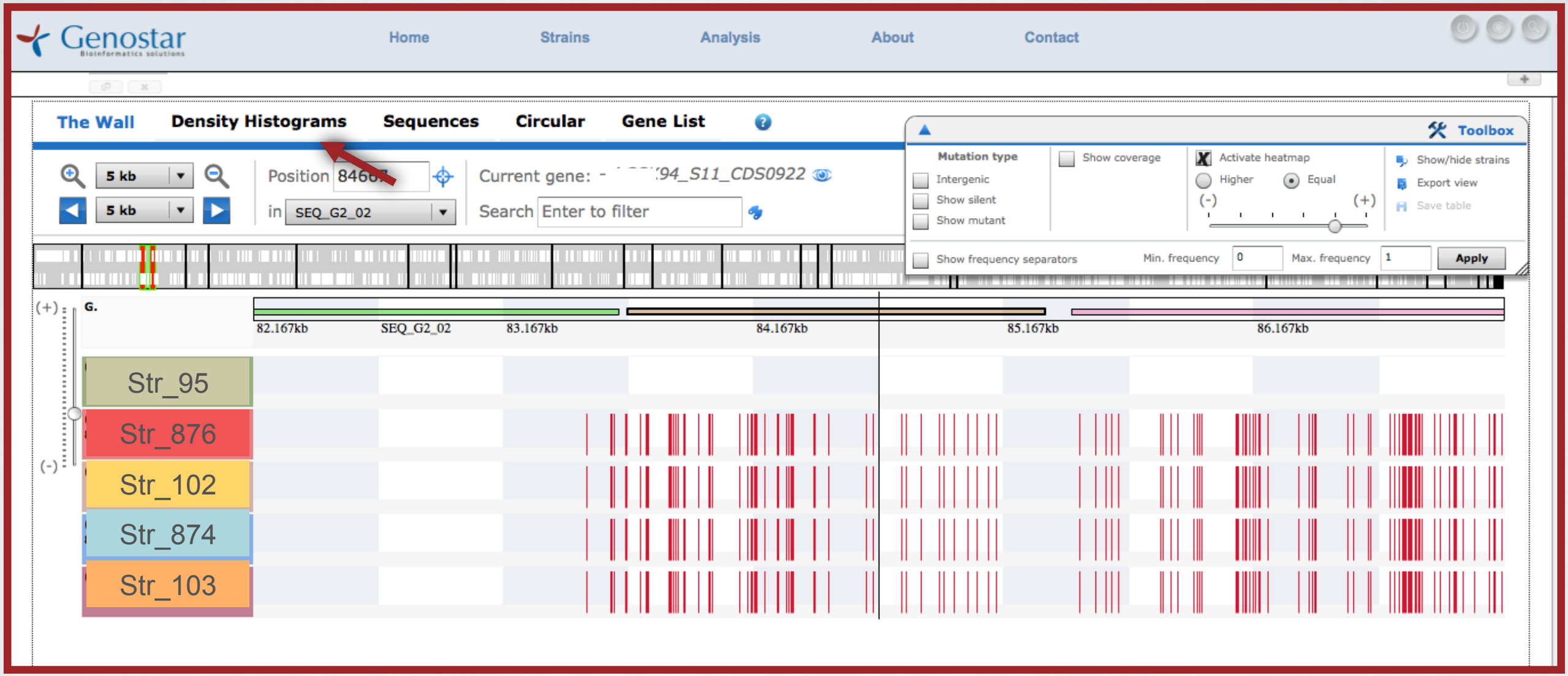
VARIANT



Comparative analysis: SNPs present in at least 4 strains

Genostar
Bioinformatics solutions

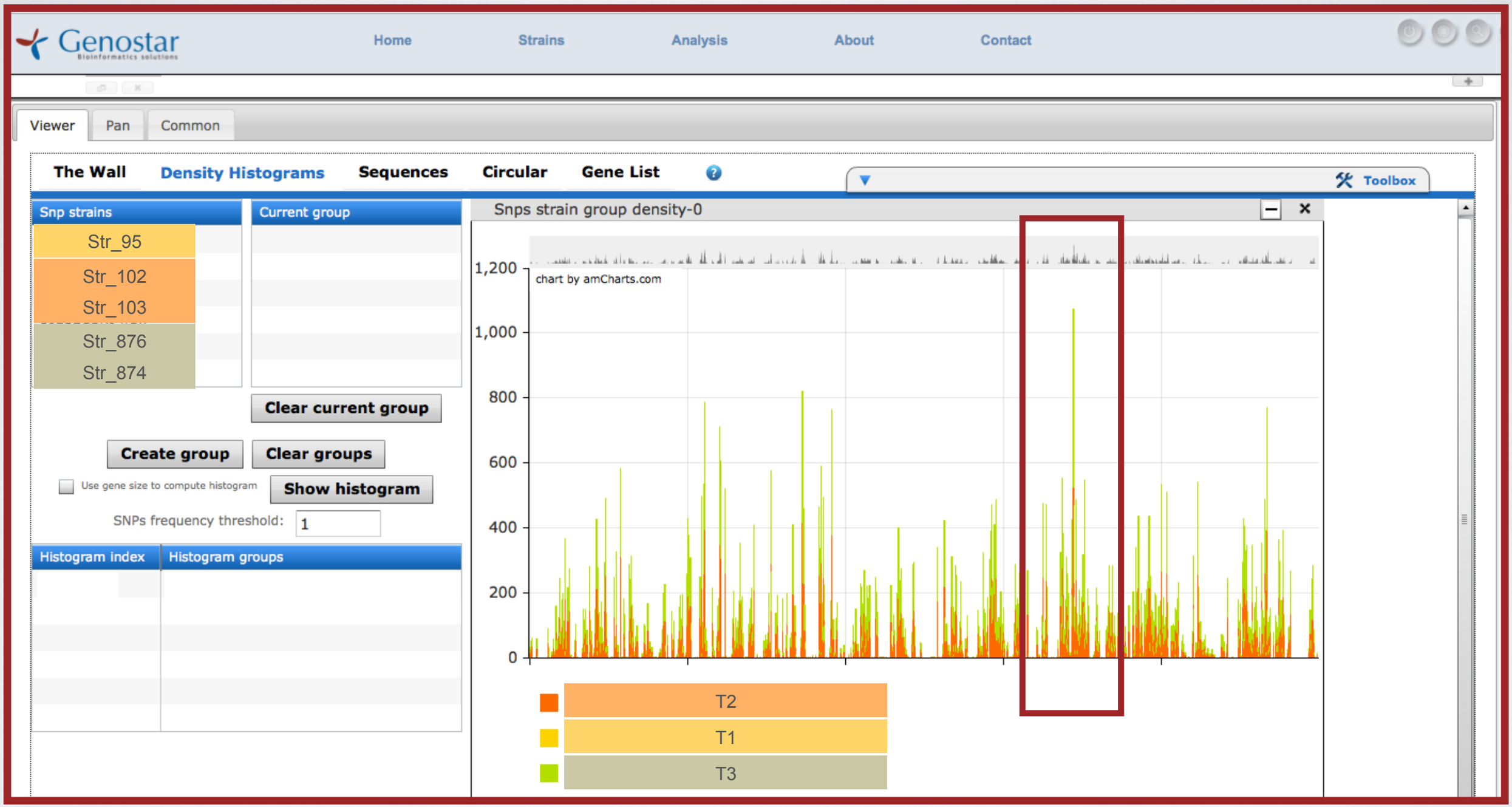
VARIANT



Comparative analysis: SNPs present in exactly 4 strains

Genostar
Bioinformatics solutions

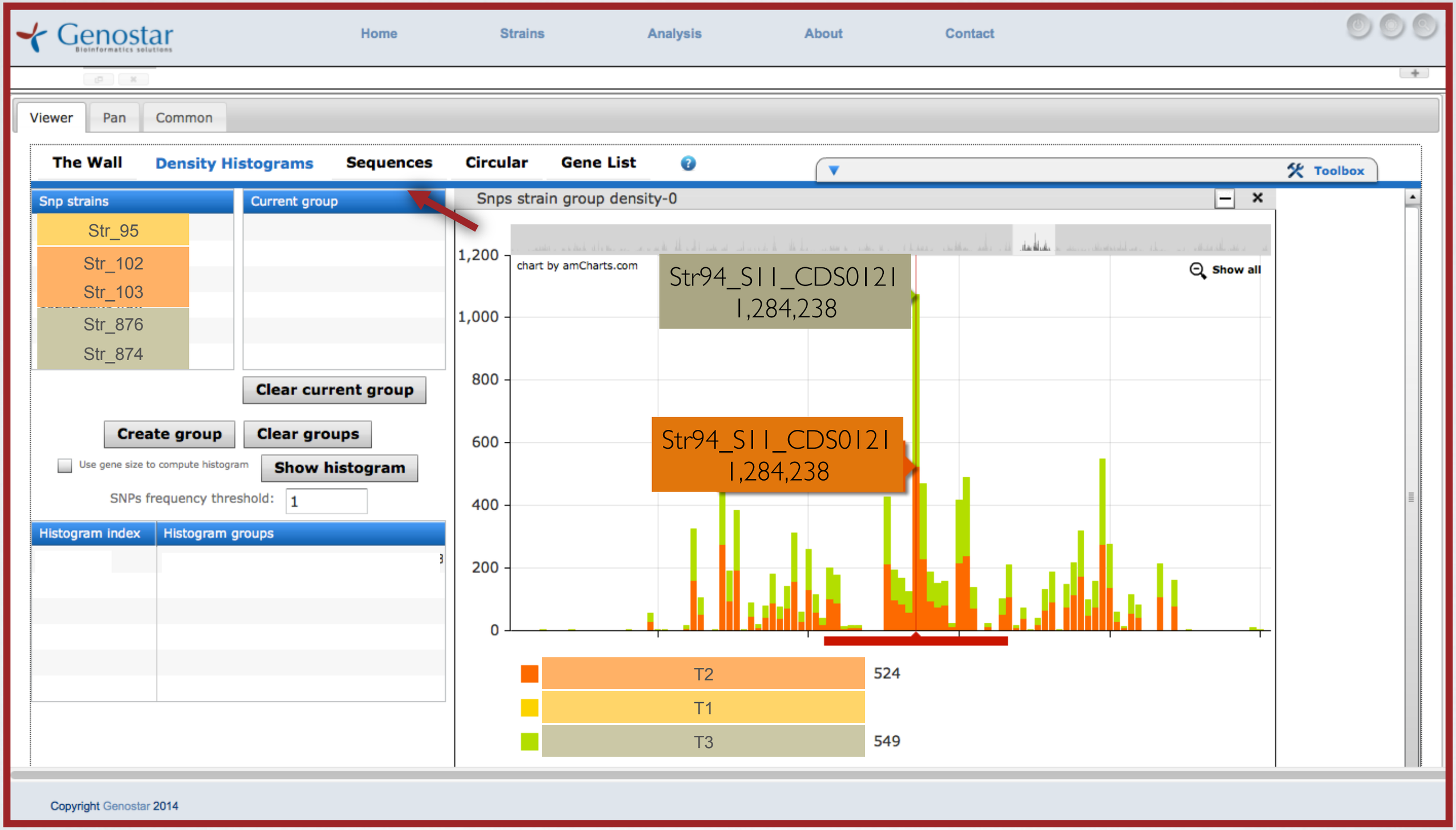
VARIANT



Comparative analysis : one group per time point
Number of SNPs at each genome position (Str_0094)

Genostar
Bioinformatics solutions

VARIANT



524 and 549 SNPs respectively for T2 and T3 at position 1 284 238

SNP filters

- generic filters: by coverage, reverse/forward balance, SNP frequency
  → *in progress*
- specific filters: genome analysis: location of regions with assembly problems
  → *needs expertise*

Meta-analysis methods

- Links between: patient strain SNPs - strain phenotypic traits - patient symptoms and history
- Clustering

VARIANT

VARIANT is still a prototype

VARIANT is a generic framework

- Integration of new NGS analysis methods
- Extension of the database model
- Customizable
    - Other microorganisms
    - Example : *M. leprae*