Секвенирование, сборка и аннотация SNP у штаммов *Saccharomyces cerevisiae* Петергофской Генетической Коллекции

Радченко Элина Александровна

Научный руководитель:
Добрынин П. В.
Центр геномной биоинформатики
им. Ф.Г. Добржанского

Институт Биоинформатики

Институт Биоинформатики,
13.09.14.

Sequencing, assembly and SNP annotation of *Saccharomyces cerevisiae* strains from Peterhof Genetic Collection

Elina Radchenko

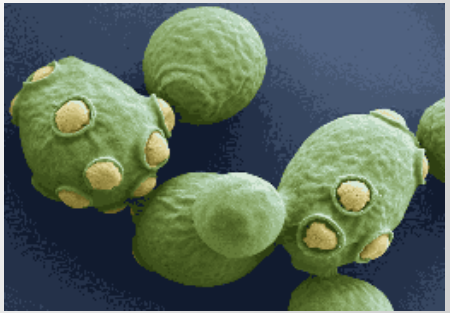
Scientific advisor:
Pavel Dobrynin
Theodosius Dobzhansky Center for Genome Bioinformatics

Bioinformatics Institute

Bioinformatics Institute,
13.09.14.

Yeast *Saccharomyces cerevisiae*

- Kingdom Fungi, Phylum Ascomycota
- Stable in haploid and diploid forms
- Model unicellular object of genetics, one of the most studied eukaryotic organisms
- Widely used in industry
- The first eukaryotic organism sequenced (1996 y.) (S288C strain)
- n = 16 chromosomes
- 12 million base pairs
- 6607 open reading frames
- 5059 verified genes encoding proteins
- Resequencing of S288C showed 498 SNPs compared to first-sequenced S288C (Liti et al., 2009)



Saccharomyces Genome Database (www.yeastgenome.org)

- 32 sequenced genomes other than S288C (contigs)
- Coverage from 2X to 378X
- Contig length from 683 bp to 376113 bp
- N50 from 1 kb to 800 kb (mean 150 kb)
- Number of contigs from 31 to 12493

Strains of Peterhof Genetic Collection

XII industrial race of baker's yeast

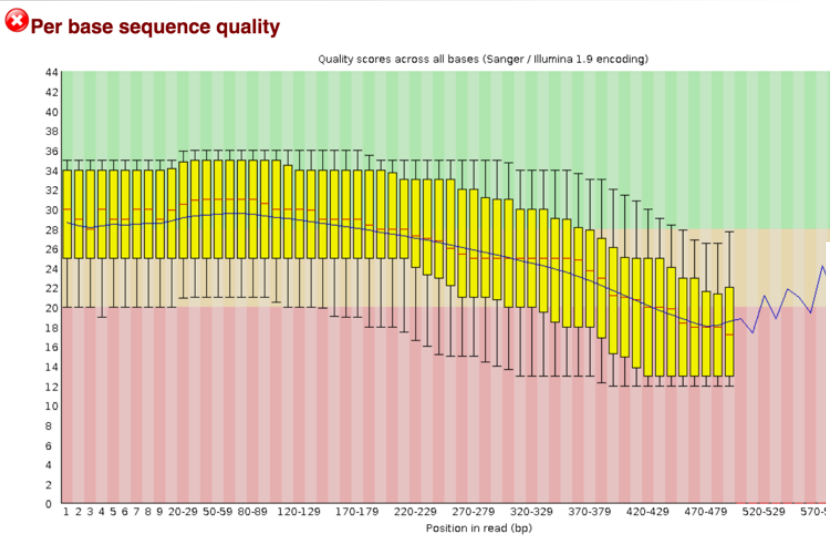
- 15V-P4 (ancestor)
- 25-25-2V-P3982 (two clones of one strain)
- 1B-D1606

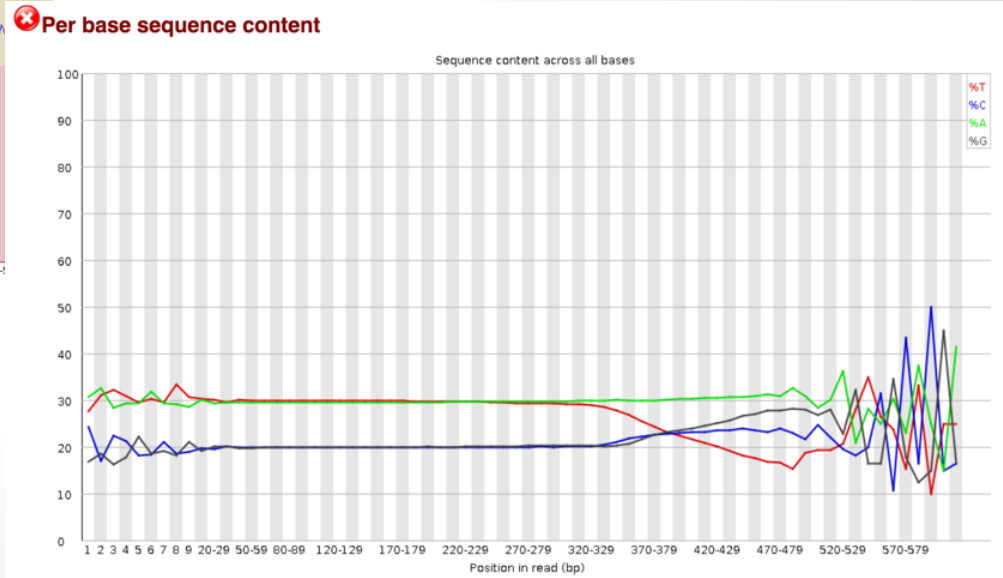
Sequencing:

- Ion Torrent PGM
- unpaired reads

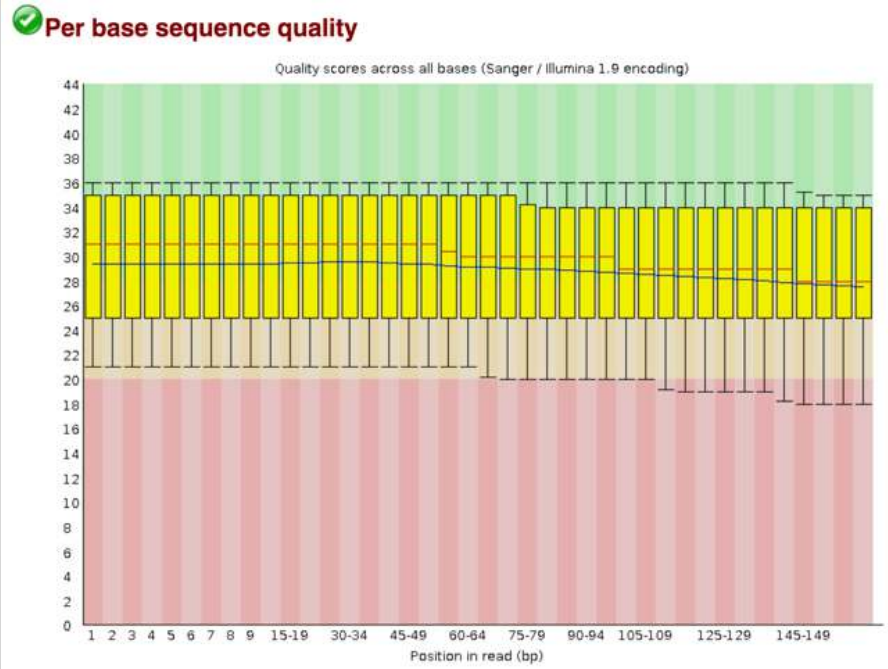
Sequencing was performed in Research Resource Center for Molecular and Cell Technologies SPbSU (Центр "Развитие Молекулярных и Клеточных технологий" СПбГУ)

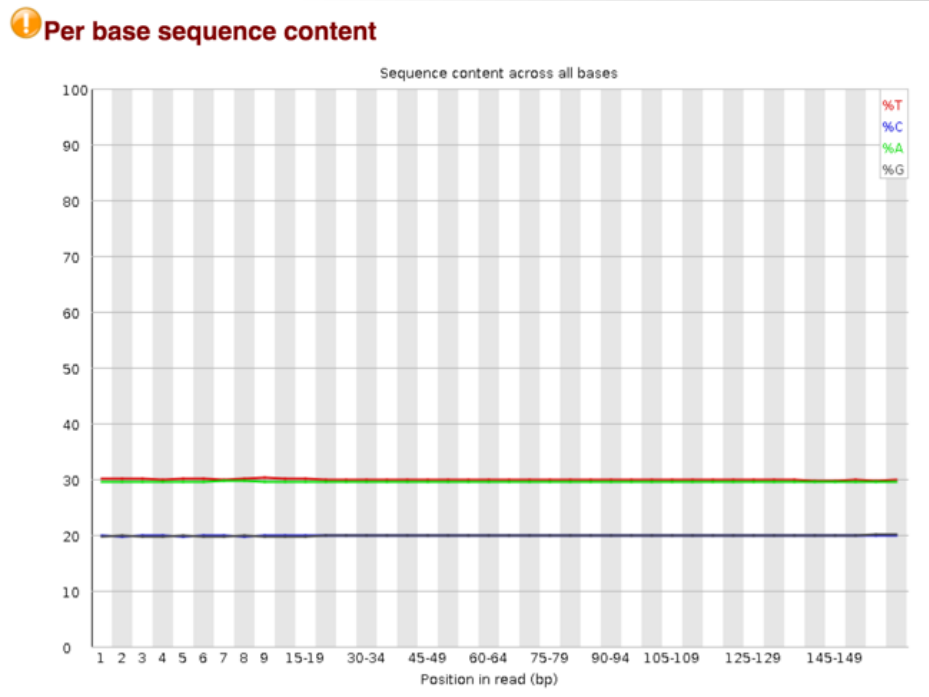
FASTQC before trimming



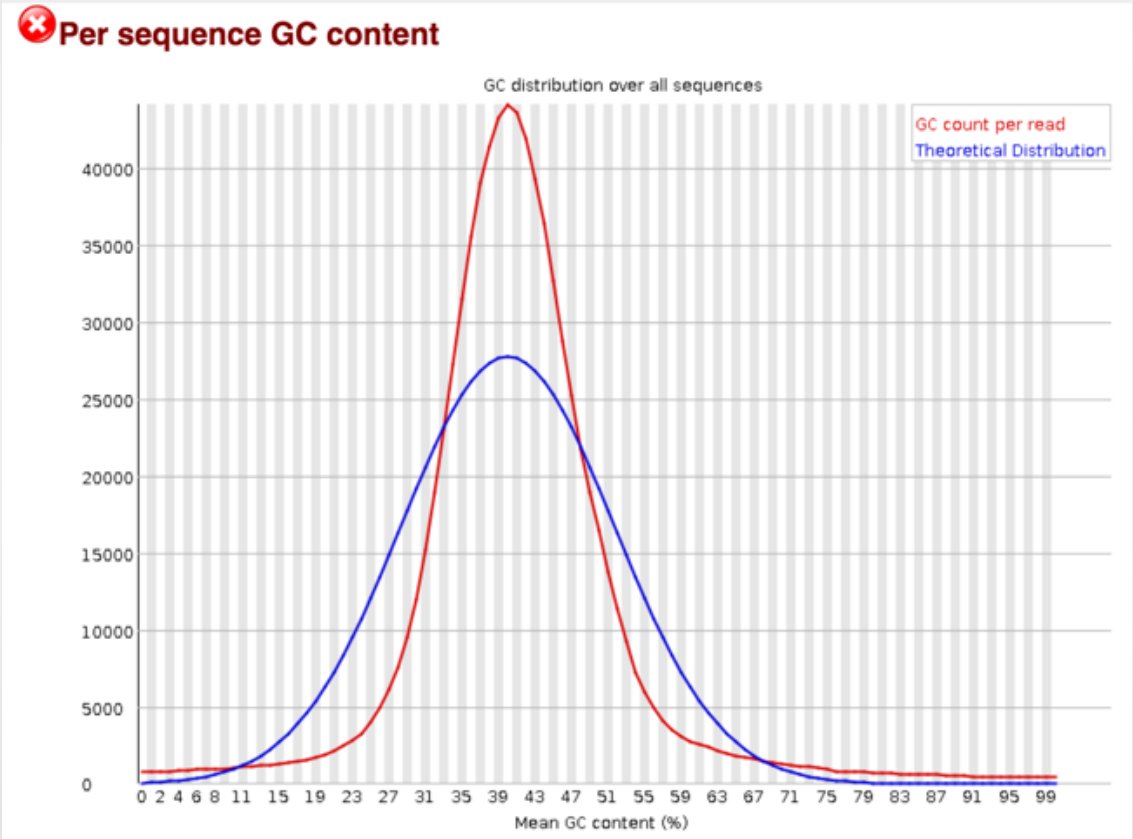


FASTQC after trimming (-f 40 -l 200)





FASTQC after trimming



SPAdes -> QUAST

Reference genome: 12157,105 bp, G+C content: 38.15 %

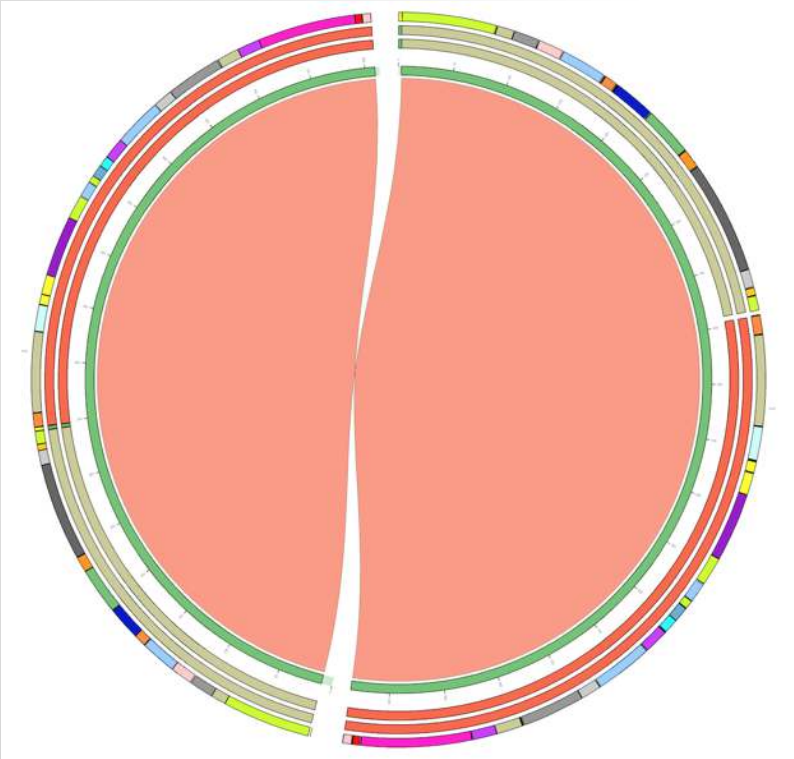
	15V-P4	1B-D1606	25-25-2V-P3982	
• **#contigs:**	3,422	3,367	3,352	3,182
• **total length:**	7410,996	8194,436	7227,077	9200,176
• **N50:**	2,479	2,872	2,407	3,546
• **genome fraction:**	43.308	52.936	42.677	60.942
• **#predicted genes:**	4,107	4,328	3,973	4,626
• **GC (%):**	39.25	39.08	39.25	38.81

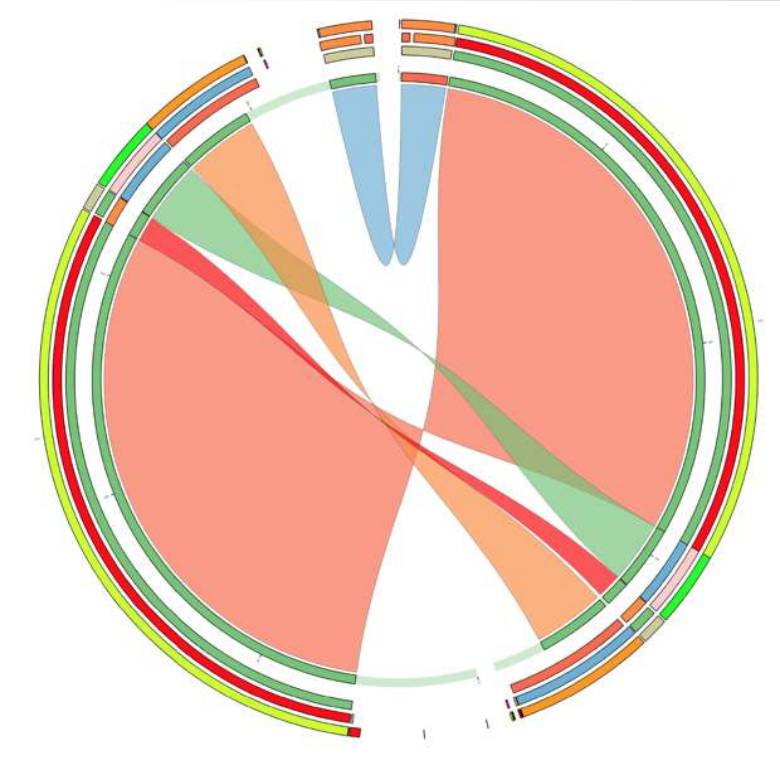
Chromosomer -> QUAST

Reference genome: 12157,105 bp, G+C content: 38.15 %

	15V-P4	1B-D1606	25-25-2B-P3982	
• **#contigs:**	17	17	17	17
• **total length:**	17477,651	13708,082	13903,149	14140,664
• **N50:**	1340,252	1062,545	1076,249	1099,137
• **genome fraction:**	77.015	80.912	75.301	84.316
#predicted genes:	3,635	4,255	4,087	4,229
• **GC (%):**	38.270	38.25	38.270	38.160

Sibelia
per chromosome analysis





SNP annotation

Align: Bowtie2

Convert and sort files (.sam -> .bam -> sorted .bam): samtools

SNP calling: Samtools -> Bcftools

Filtering: Vcftools

SNP annotation: SnpEff

SNPs: analysed strain to reference

15V-P4:	44,723
1B-D1606:	24,857
25-25-2V-P3982 I:	32,823
25-25-2V-P3982 II:	35,191

Checked for already known mutations:

ADE1, *LYS9*, *LYS2*, *HIS7* etc

Summary

- 4 PGC (Peterhof Genetic Collection) strains genomes were sequenced on ABI IonTorrent platform
- sequencing quality was assessed using FastQC
- reads were assembled to contigs with SPAdes
- contigs were assembled to 17 scaffolds (chromosomes) with Chromosomer
- assemblies quality was assessed and compared with one of other strains from SGD
- SNPs in these 4 strains were annotated

Thank you for attention!
Questions?