

Перетасовка экзонов в геномах дрожжей: ошибка сборки или биологическая реальность

Выполнил: Раменский Дмитрий

Руководители: Олег Тарасов, Полина Дроздова
Кафедра генетики и биотехнологии СПбГУ

Предпосылки:

Ген *SNC1* (YAL030W) : Рецептор везикулярной мембраны (v-SNARE), участвует в слиянии транспортных пузырьков аппарата Гольджи с плазматической мембраной. Содержит два экзона и один интрон.

У части штаммов экзоны в аннотированных генах переставлены:

1 75

MSSSTPFDPYALSEHDEERPQNVQSKSRTAELQAEIDDTVGIMRDNINKVAERGERLTSIEDKADNLAVSAQGFKR
-----EIDDTVGIMRDNINKVAERGERLTSIEDKADNLAVSAQGFKR

GANRVRKAMWYKDLKMKMCLALVIIILLVVIIVPIAVHFSR-----
GANRVRKAMWYKDLKMKMCLALVIIILLVVIIVPIAVHFSRMSSSTPFDPYALSEHDEERPQNVQSKSRTAELQA

76 150

Задачи проекта:

1. Получить названия и последовательности всех генов *Saccharomyces cerevisiae* (возможно, и других видов *Saccharomyces* из базы yeastgenome.org), содержащих интроны.
2. Построить выравнивания по каждому из таких генов.
3. Проанализировать, есть ли другие гены (помимо *SNC1*) с перестановкой экзонов.
4. Пересобрать последовательности всех генов, в которых будут обнаружены перестановки экзонов, на основе первичных данных секвенирования.
5. Оценить возможные биологические последствия перестановки экзонов.

Задачи проекта:

1. Получить названия и последовательности всех генов *Saccharomyces cerevisiae* (возможно, и других видов *Saccharomyces* из базы yeastgenome.org), содержащих интроны.

В результате:

- С помощью поиска по базе данных <http://yeastmine.yeastgenome.org/> получили 344 нуклеотидные последовательности *Saccharomyces cerevisiae*, содержащие интроны
- Из них 74 кодируют тРНК, мРНК и митохондриальные белки - их убираем из рассмотрения

Задачи проекта:

2. Построить выравнивания по каждому из таких генов.

В результате:

Для построения выравниваний сгенерировали список ссылок на локусы интересующих нас генов в базе <http://yeastgenome.org/> и использовали встроенную процедуру выравнивания между штаммами (Strain Alignment)

Задачи проекта:

3. Проанализировать, есть ли другие гены (помимо *SNC1*) с перестановкой экзонов.

В результате:

ORF Systematic Name	ORF Standard Name	ORF Name	ORF Length
YAL030W	<i>SNC1</i>	Suppressor of the Null allele of CAP	467
YBL050W	<i>SEC17</i>	SECretory	995
YBR255C-A			457
YCR097W	<i>HMRA1</i>	Hidden Mat Right A	487
YDL075W	<i>RPL31A</i>	Ribosomal Protein of the Large subunit	763
YER117W	<i>RPL23B</i>	Ribosomal Protein of the Large subunit	885
YFR045W			1002
YJL189W	<i>RPL39</i>	Ribosomal Protein of the Large subunit	542

Продолжение:

ORF Systematic Name	ORF Standard Name	ORF Name	ORF Length
YKL006C-A	<i>SFT1</i>	Suppressor of sed Five Ts	435
YLR329W	<i>REC102</i>	RECombination	892
YLR406C	<i>RPL31B</i>	Ribosomal Protein of the Large subunit	691
YLR445W	<i>GMC2</i>	Grand Meiotic recombination Cluster	649
YML034W	<i>SRC1</i>	Spliced mRNA and Cell cycle regulated gene	2631
YML036W	<i>CGI121</i>	homolog of human CGI-121	652
YNL004W	<i>HRB1</i>	Hypothetical RNA-Binding protein	1707
YNL038W	<i>GPI15</i>	GlycosylPhosphatidylinositol anchor biosynthesis	764

Задачи проекта

- **Оценить выравнивание белков с выявленными перестановками на геном**
- 4. Пересобрать последовательности всех генов, в которых будут обнаружены перестановки экзонов, на основе первичных данных секвенирования.**

В результате:

Для выравнивания использовали программу Exonerate, позволяющую производить выравнивание белковых последовательностей на геном с учетом интронов

В результате:

```
1 : MetAlaArgAspIleThrPheLeuThrValPheLeuGluSerCysGlyAlaValAsnAsn :      20
  |||
  MetAlaArgAspIleThrPheLeuThrValPheLeuGluSerCysGlyAlaValAsnAsn
155447 : ATGGCAAGAGATATCACATTTTTGACCGTATTTTTAGAAAGTTGTGGCGCTGTAAATA : 155390
. . .

40 : ProGluSerThrAspSerAsnSerLeuTyrIleProLeuLeuProPro{G} >>>> Ta :      55
  |||
  GluSerThrAspSerAsnSerLeuTyrIleProLeuLeuProPro{G}+-
155332 : GAATCAACCGACTCTAATTCATTATATATTCCACTGCTACCACCT{G}ga..... : 155282

56 : rget Intron 1 >>>> {ly}MetLeuLysIleLysLeuAsnPheLysMetAsnAsp :      67
    97 bp           {||}! :!!
                    ++{ly}LysValLysIleLysLeuAsnPheLysMetAsnAsp
155281 : .....ag{GA}AAAGTGAAGATTAACTGAATTTTAAAATGAACG : 155152
```

Задачи проекта



- **Оценить выравнивание белков с выявленными перестановками на геном**
- 4. Пересобрать последовательности всех генов, в которых будут обнаружены перестановки экзонов, на основе первичных данных секвенирования.**

В результате:

← → ↻ 🏠 sra.dnanexus.com/experiments/SRX050623

DNAnexus POWERED Sequence Read Archive +

 **Experiment SRX050623** [DOWNLOAD](#)

Title	W303
Alias	W303
Design description	no content
References	no content
Organisms	Saccharomyces cerevisiae
Submitter	Stanford University School of Medicine
Instrument	Illumina Genome Analyzer II
Library name	W303
Library strategy	WGS
Library source	GENOMIC
Library selection	RANDOM
Library layout	PAIRED
Library construction protocol	no content
Related objects	 1  1  1
Submissions	SRA030835
Links	no content

← → ↻ 🏠 www.ncbi.nlm.nih.gov/sra/SRX050623

NCBI Resources How To

SRA SRA Advanced

Full ▾

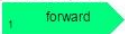

SRX050623: W303
1 ILLUMINA (Illumina Genome Analyzer II) run: 8.8M spots, 633M bases, 370.3Mb downloads

Submitted by: STANFORD UNIVERSITY SCHOOL OF MEDICINE

Study: Rapid Evolution of Ethanol Tolerance in Yeast by Selection in a Turbidostat
• [SRP006150](#) • [All experiments](#) • [All runs](#)
[show Abstract](#)

Sample: *S. cerevisiae* Ancestral Strain
[SAMN00253867](#) • [SRS183192](#) • [All experiments](#) • [All runs](#)
Organism: *Saccharomyces cerevisiae*

Library:
Name: W303
Instrument: Illumina Genome Analyzer II
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED

Spot descriptor:
 1 forward  37 reverse

Runs: 1 run, 8.8M spots, 633M bases, 370.3Mb

Run	# of Spots	# of Bases	Size	Published
SRR154334	8,791,974	633M	370.3Mb	2012-03-28

ID: 59786

You are here: NCBI > DNA & RNA > Sequence Read Archive (SRA)

В результате:

The image displays two side-by-side screenshots of the FastQC software interface. Each window shows a list of quality control metrics on the left and a table of 'Basic sequence stats' on the right. The left window shows results for 'SRR154334_1.fastq' and 'SRR154334_2.fastq', while the right window shows results for 'SRR154334_1.fastq' and 'paired_1.fq'.

Left Window (SRR154334_1.fastq, SRR154334_2.fastq):

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Measure	Value
Filename	SRR154334_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	8791974
Filtered Sequences	0
Sequence length	36
%GC	40

Right Window (SRR154334_1.fastq, paired_1.fq):

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Measure	Value
Filename	paired_1.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	8338055
Filtered Sequences	0
Sequence length	26
%GC	38

В результате:

