# Processing data of BLAST, clustering and ordering

**Student:** Elena Bushmanova, Bioinformatics Institute, SPbSU

**Advisor:** Pavel Dobrynin, Dobzhansky Center

Saint-Petersburg,
2013

# Goal:

Creating a program, which allows

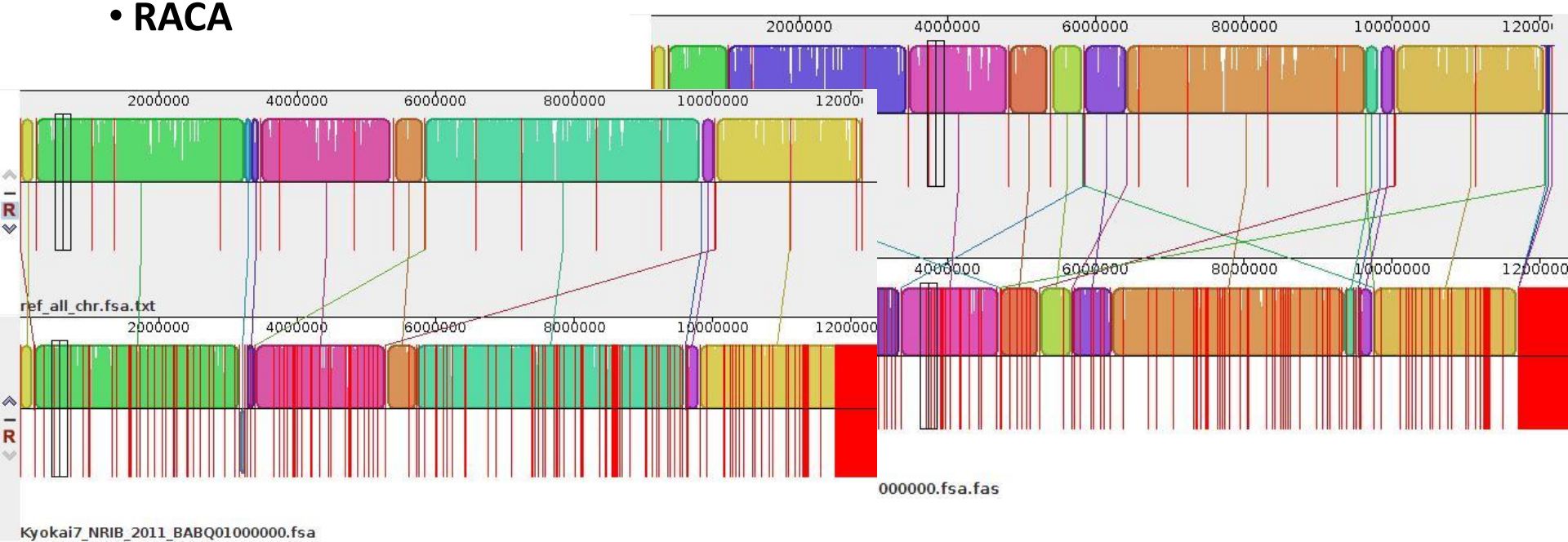
- to form single units of overlapping hits
- to rank the received units according to the coordinates on the chromosome
- and get some statistics

# Problem:

To increase of the accuracy and efficiency of assembling of the genome on the basis of the reference.

# Other tools:

- SyMAP
- ABACAS
- Arachne
- MAUVE
- LASTZ
- RACA



**SyMAP** – very long time to analyze only one pair of genomes. But this is only one program that currently used for reference assisted assembly of large genomes;

**ABACAS** – very good program, but working only with bacteria or other small (20mbp) genomes;

**Arachne** – we tried to adapt this assembler for Sanger sequences to assemble contigs but failed;

**MAUVE** – not trivial output, only small genome, work slowly;

**LASTZ** – great tool for alignment of large genomes, but working not very fast, and also quite complicated pipeline for genome assembly ;

**RACA** – in theory one of the best ref assisted assembler, but still only in theory (no publications that used it).

# Algorithm

- Processing:

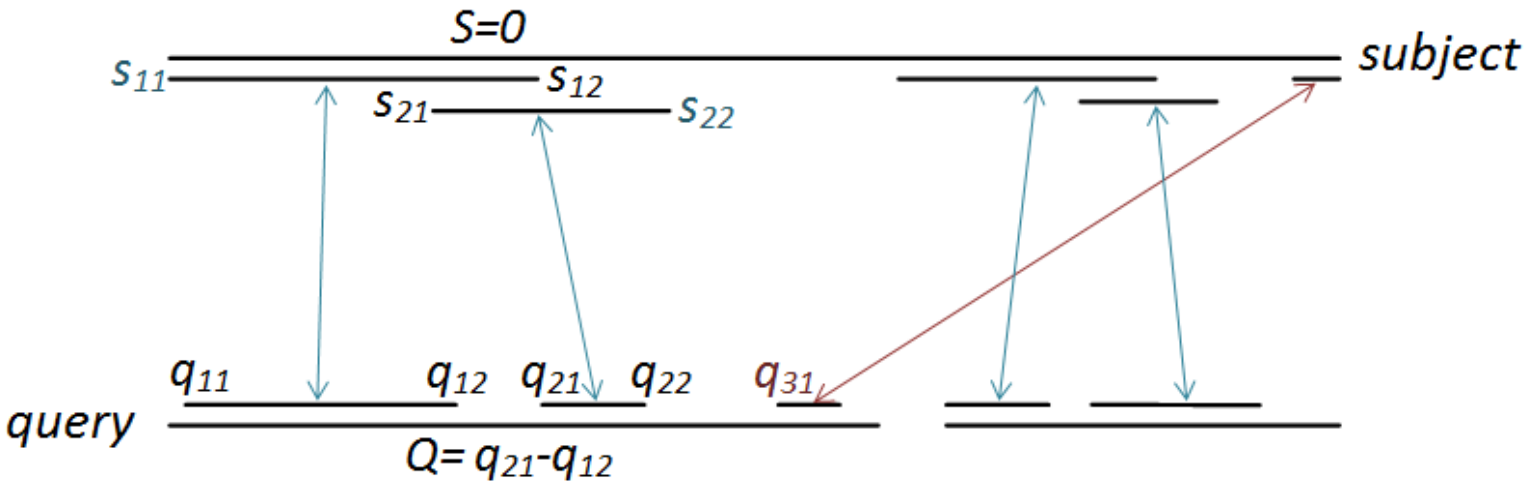
makeblastdb -in ref_all_chr.fsa.txt -dbtype nucl -title yeast -out yeast

blastn -query Kyokai7_NRIB_2011_BABQ01000000.fsa -db yeast -outfmt 6 -out data.txt

- Selection
- Sorting
- Checking:
- Splicing
- Statistics

| $S_{l1}$ | $S_{n1}$ | $S_{m1}$ |
|---|---|---|

$$Q - 1000 \leq S \leq Q + 1000$$



# Data

## Input data:

**yeast_blastn.txt:** input list of fragments

query_id, subject_id, identity, alignment_length, mismatches, gap_opens, q.start, q.end, s.start, s.end, e_value, bit_score

## Output data:

**outputList.txt:** output list of fragments

**unusedList.txt:** list of duplicate fragments

**In_statistics.txt, Out_statistics.txt:**

- total number of fragments and percent of unique fragments;
- average, median, maximum and minimum count of the same id fragments;
- similar characteristics of the number of matches.

**In_statisticsMatch.txt, Out_statisticsMatch.txt;**

**In_statisticsLen_match.txt, Out_statisticsLen_match.txt.**

# Some statistics

| | IN | OUT |
|---|---|---|
| all_count | 35909 | 21697 |
| per_unique | 0.07 | 0.29 |
| match_middle | 4 | 2 |
| match_median | 2 | 1 |
| match_min | 1 | 1 |
| match_max | 133 | 111 |
| len_match_middle | 900 | 1162 |
| len_match_median | 606 | 607 |
| len_match_min | 200 | 191 |
| len_match_max | 108589 | 223617 |

# Results

1. To develop and test an approach for data processing in a simple tabular output format of blast;
2. Test program on small genomes;
3. Test program on large genomes.

# Future

1. Compare results from different programs for reference assisted assembly;
2. Start working with large genomes;
3. Adapt program to work with different phylogeny groups (eukaryote, vertebrates, mammals, birds etc).

# References

1. W. James Kent, Robert Baertsch, Angie Hinrichs, Webb Miller, and David Haussler **«Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes»**. - PNAS, 2003

2. Anna I. Rissman, Bob Mau, Bryan S. Biehl, Aaron E. Darling, Jeremy D. Glasner and Nicole T. Perna **«Reordering contigs of draft genomes using the Mauve Aligner»**. - BIOINFORMATICS, 2009

3. Serafim Batzoglou **«The many faces of sequence alignment»**. - BRIEFINGS IN BIOINFORMATICS, 2004

4. Sante Gnerre, Eric S Lander, Kerstin Lindblad-Toh and David B Jaffe **«Assisted assembly: how to improve a de novo genome assembly by using related species»**. - Genome Biology, 2009

5. Mikhail Kolmogorov, Son Pham et al **«Complete Assembly of Common Bacterial Genomes Using Only Short Reads»**

# Thanks!