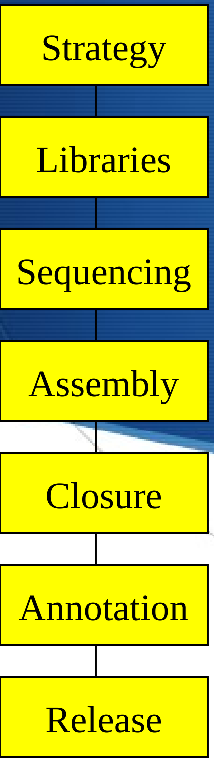
Whole Genome Sequencing, Assembly and Annotation



Dr. G P S Raghava (FASc, FNASc),
Head Bioinformatics Centre,
Institute of Microbial Technology,
Chandigarh, INDIA

Introduction

- Whole genome sequencing provide information about coding and noncoding part of genome.
- To fetch out important pathways.
- For evolutionary studies and species comparison.
- For more effective personalized medicine (why a drug works for person X and not for Y).
- Disease-susceptibility prediction based on gene sequence variation.

History of Sequencing

- Allan Maxam and Walter Gilbert developed an important method of DNA sequencing in 1976-1977.
- This method of chemical modification of DNA was technically complex and fallen out of flavor due to the use of extensive hazardous chemicals, and difficulties with scale-up.

History of Sequencing

- Sanger and his team developed the chain-termination method of DNA sequencing in 1977.
- Only be used for fairly short strands (100 to 1000 base pairs) and longer sequences must be subdivided into smaller fragments.
- After this, these small fragments subsequently re-assembled to give the overall sequence

History of Sequencing

- Shotgun sequencing has been developed for sequencing of large fragments of DNA in 1979.
- DNA is broken up randomly into numerous small segments, which are sequenced using the chain termination method and then short reads have been produced.
- Shotgun sequencing was the initiative for full genome sequencing.

WHOLE GENOME SEQUENCING

- Information about coding and non coding part of an organism.
- To find out important pathways in microbes.
- For evolutionary study and species comparison.
- For more effective personalized medicine (why a drug works for person X and not for Y).
- Identification of important secondary metabolite pathways (*e.g.* in plants).
- Disease-susceptibility prediction based on gene sequence variation.

NEXT GENERATION SEQUENCING

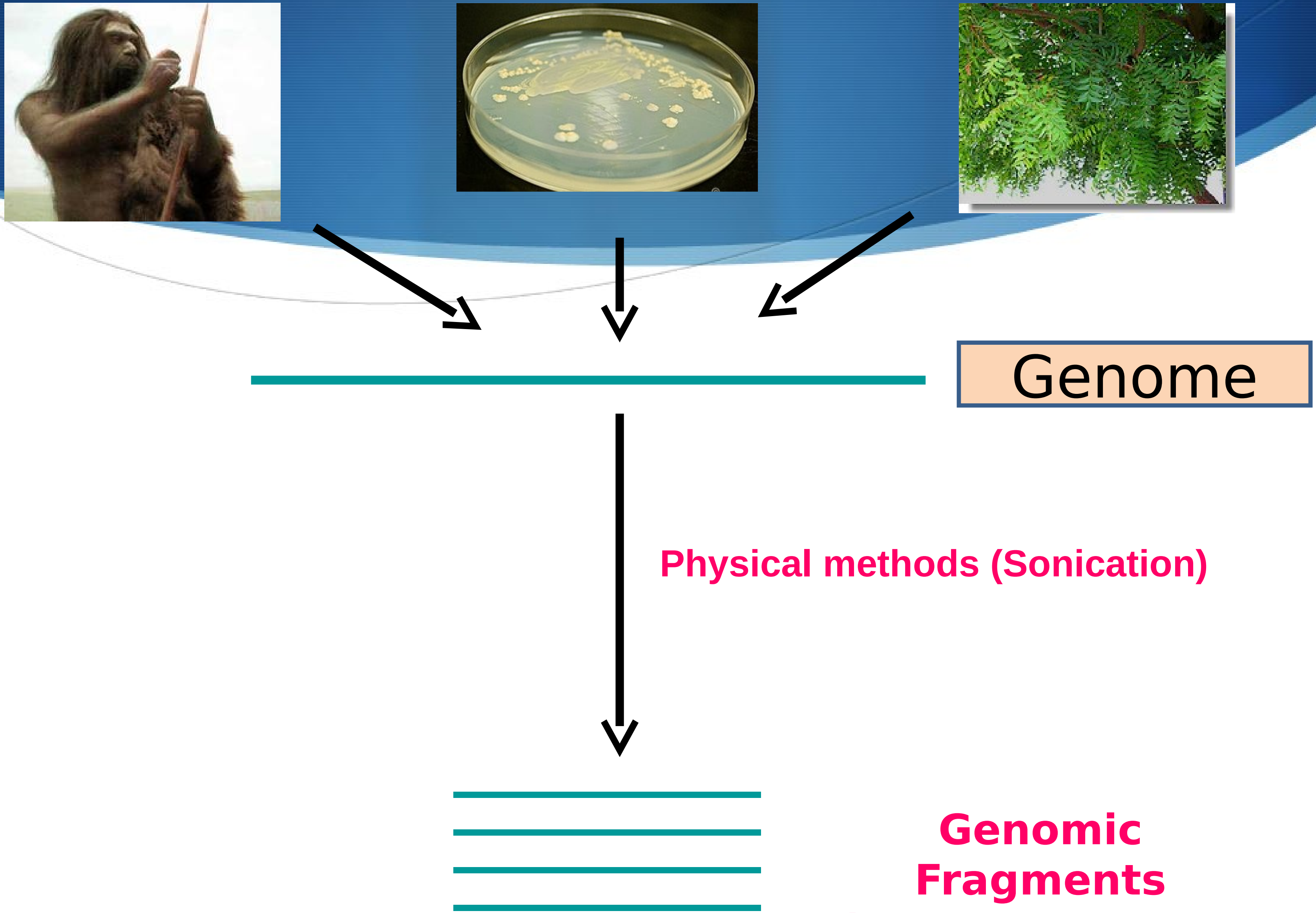
- Sequence full genome of an organism in a few days at a very low cost.
- Produce high throughput data in form of short reads.

Illumina

ABI’s Solid

Roche’s 454 FLX

Ion torrent



Genome

Physical methods (Sonication)

Genomic Fragments (200 nt or 400 nt or 1kb)

Single end sequencing

Genomic Fragments (200 nt or 400 nt or 1kb)

Low cost & Less time

Paired end sequencing

454 FLX

Ion torrent

ABI’s Solid

Illumina

Short Reads

CHALLENGES

Removal of artifacts in short reads ??

Genome assembly of short reads ??

Several assemblers available, which is best ??

Annotation and validation of assembled genome ??

Recent techniques

- High throughput sequencing also called Next Generation Sequencing (NGS) have the capacity to sequence full genomes.
- These technologies Includes Roche’s 454 GS FLX, Illumina’s Solexa technology, ABI’s SOLiD technology and Ion torrent technology.

Next Generation Sequencing

Technique	**Ion torrent**	**Roche's 454**	**Illumina**	**ABI's SOLiD**
Data (Mb per run)	100	100	600	700
Time per run	1.5 Hrs	7 Hrs	9 Days	9 Days
Read length	200 bp	400 bp	150 bp	75 bp
Cost per Mb	5 $	84.39 $	0.03 $	0.04 $

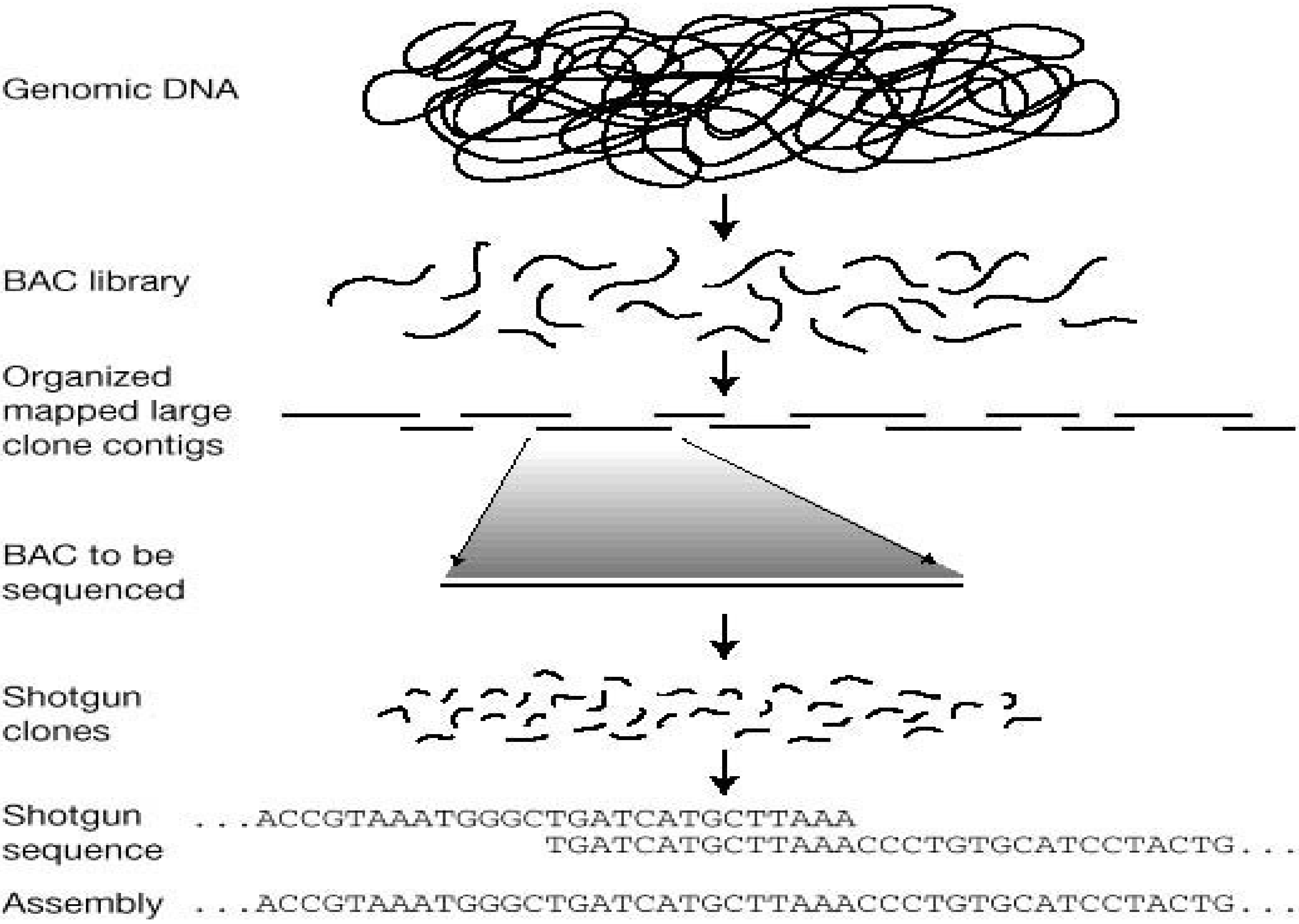
History of genome sequencing

- Bacteriophage fX174, was the first genome to be sequenced, a viral genome with only 5,368 base pairs (bp).
- First bacterial genome sequenced was *Haemophilus influenza*.
- The first nearly complete human genomes sequenced were J. Craig Venter's, James Watson's, a Han Chinese, a Yoruban from Nigeria, a female leukemia patient, and Seong-Jin Kim.
- As of June 2012, there are 69 nearly complete human genomes publicly available.

Challenges of genome sequencing

- Data produce in form of short reads, which have to be assembled correctly in large contigs and chromosomes.
- Short reads produced have low quality bases and vector/adaptor contaminations.
- Several genome assemblers are available but we have to check the performance of them to search for best one.

Hierarchical shotgun sequencing



Genomic DNA

BAC library

Organized mapped large clone contigs

BAC to be sequenced

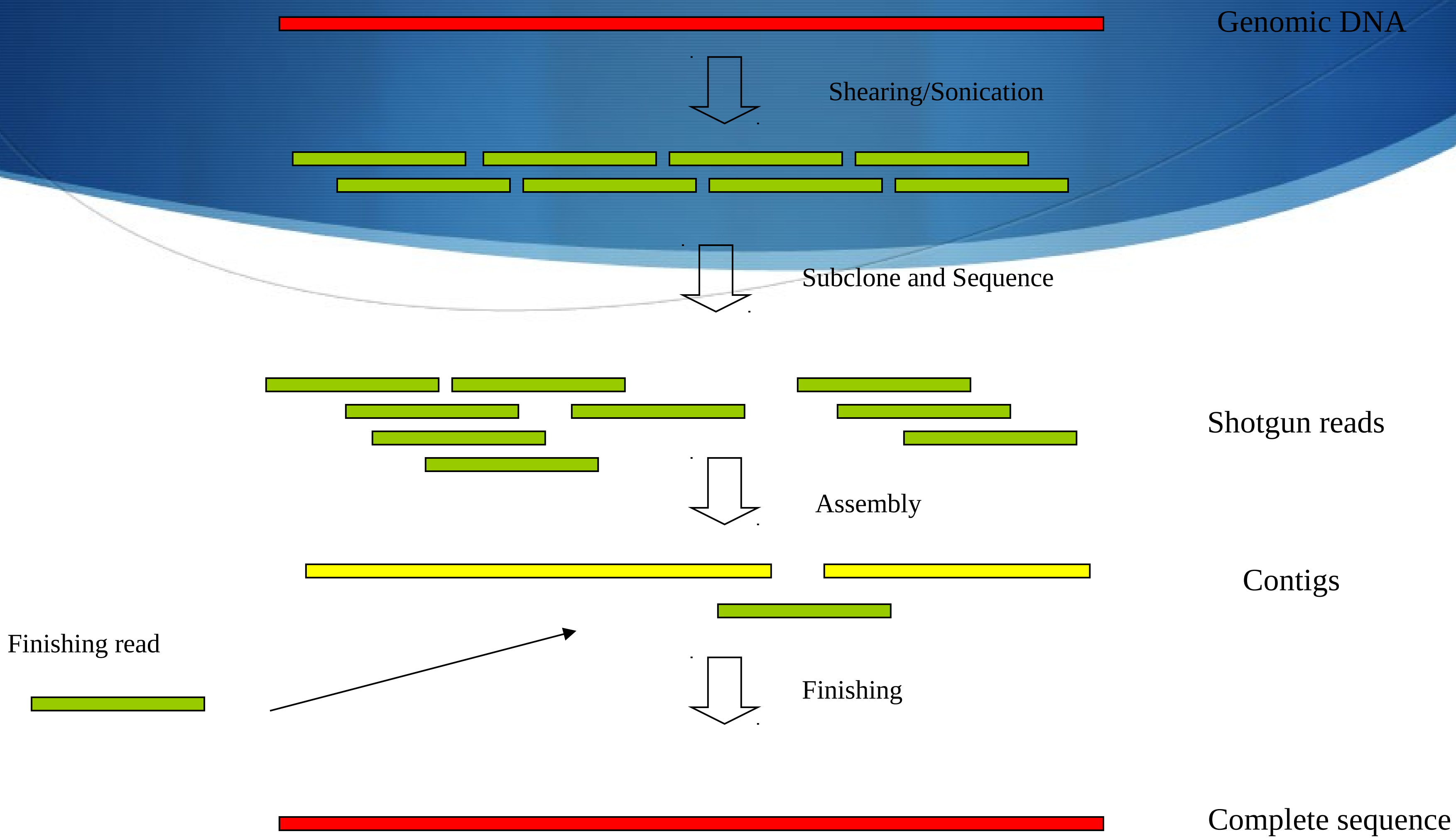
Shotgun clones

Shotgun sequence

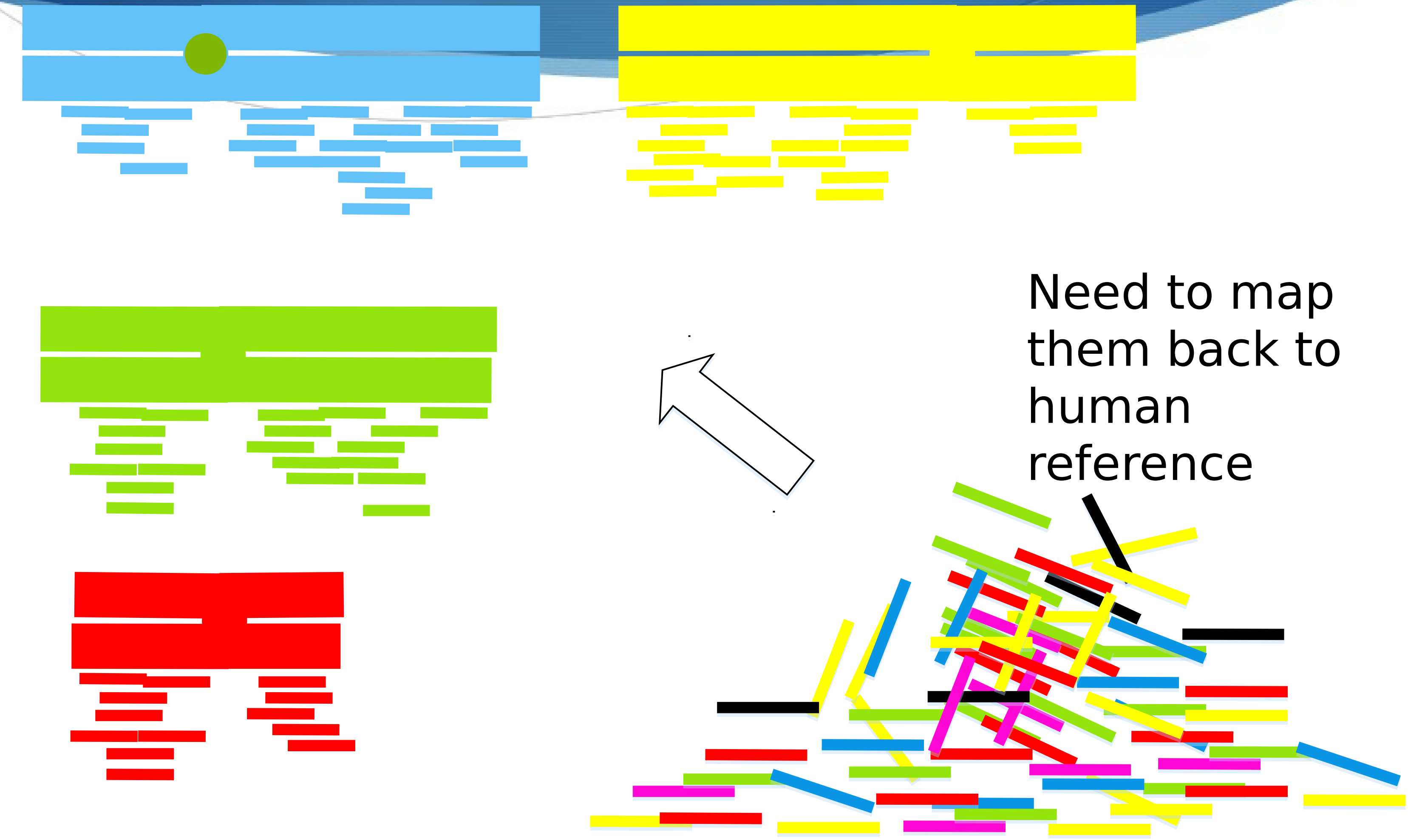
```
...ACCGTAAATGGGCTGATCATGCTTAAA
                    TGATCATGCTTAAACCCTGTGCATCCTACTG...
```

Assembly

```
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...
```


Genomic DNA
Shearing/Sonication
Subclone and Sequence
Shotgun reads
Assembly
Contigs
Finishing read
Finishing
Complete sequence

Short read alignment

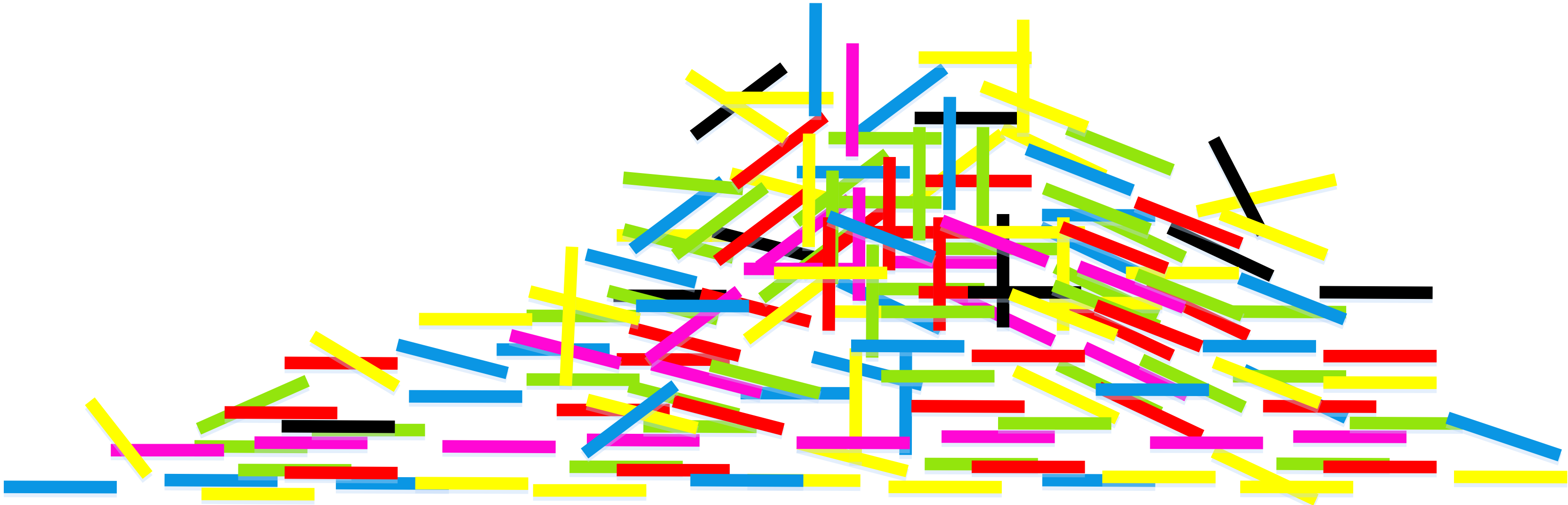


Need to map them back to human reference

Short read alignment

Sequencing machine

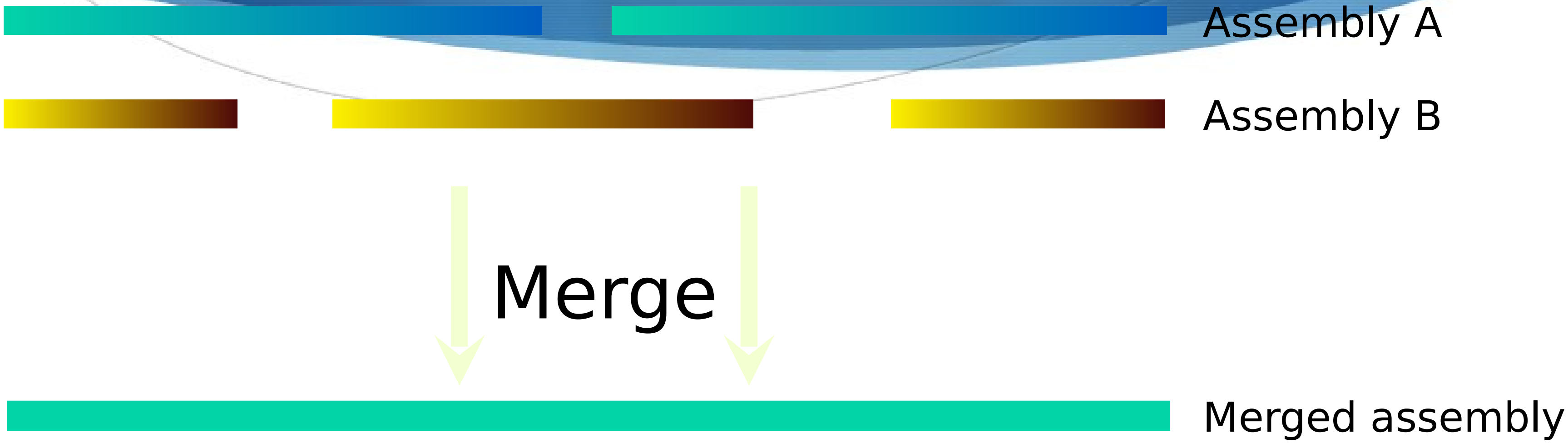
And you get MANY of them



De novo assembly strategies

- SSAKE
  - Warren et al., 2007
  - Uses DNA prefix tree to find k-mer matches
- Edena
  - Hernandez et al., 2008
  - overlap-layout algorithm adapted for short reads
- Velvet
  - Zerbino and Birney, 2008
  - Uses DeBruijn graph algorithm plus error correction

Comparative assembly using multiple genomes



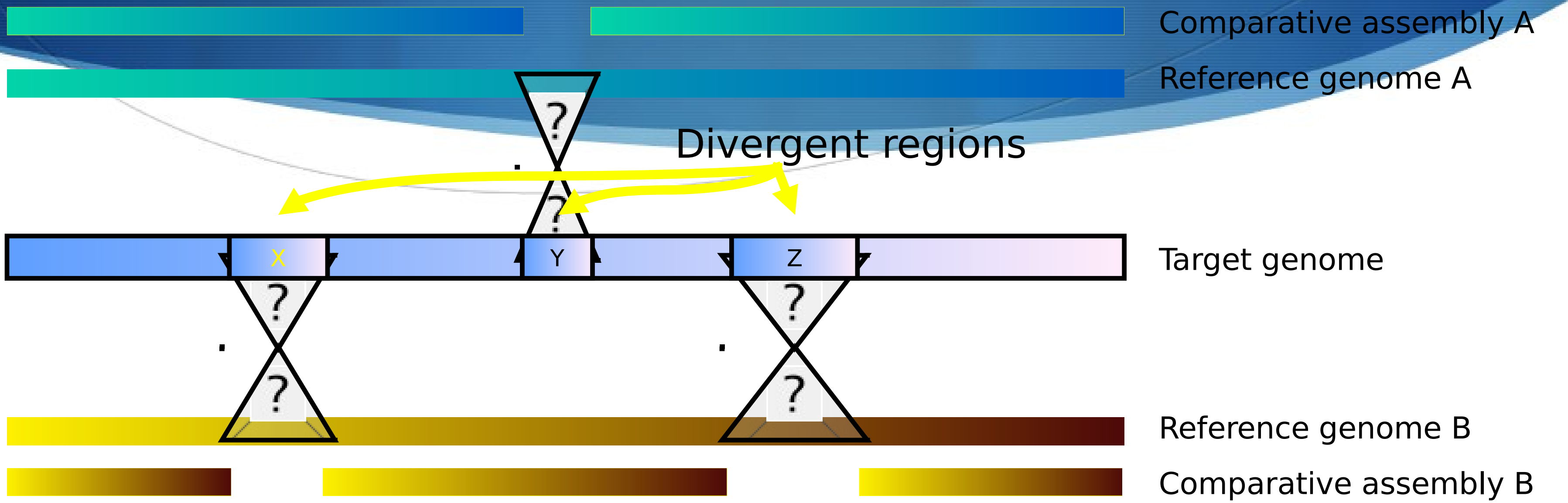
Assembly A

Assembly B

Merge

Merged assembly

Comparative assembly using multiple genomes



Comparative assembly A

Reference genome A

Divergent regions

Target genome

Reference genome B

Comparative assembly B

Genome annotation

- A process of attaching biological information to sequences (contigs or chromosomes).
- Consists of two main steps: -

A. Identifying elements on genome a process called gene prediction (*Structural annotation*) .

B. Attaching biological information to these elements (*Functional annotation*).

Genome annotation

- *Structural annotation*
- ORFs and their localisation
- Gene structure
- Coding regions
- Location of regulatory motifs
- *Functional annotation*
- Biochemical function
- Biological function
- Involved regulation and interactions
- Expression

Genome annotation

- Can be done manually (require human expertise) or with automated pipelines.
- Pipelines available :-
  - PGAAP (NCBI)
  - RAST server
  - IMG-ER,
  - ISGA
  - MAKER (for eukaryotes).

Genome annotation tools at IMTECH

- **Protein Structure prediction servers**
- **Servers for predicting function of proteins**
- **Servers for designing epitope based vaccine**
- **Genome annotation**
- **Molecular Interactions & Modifications**
- **Designing of Therapeutic Molecules**
- **Computer Aided Drug Design**

http://www.imtech.res.in/raghava/

Genome submission to NCBI (GenBank)

- NCBI (GenBank) accepts both complete and incomplete genomes (contigs produced after genome assembly).
- Bacterial genome submission instructions available at http://www.ncbi.nlm.nih.gov/genbank/genomesubmit/ .
- Eukaryotic genome submission instructions availble at
- http://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission/

Publications

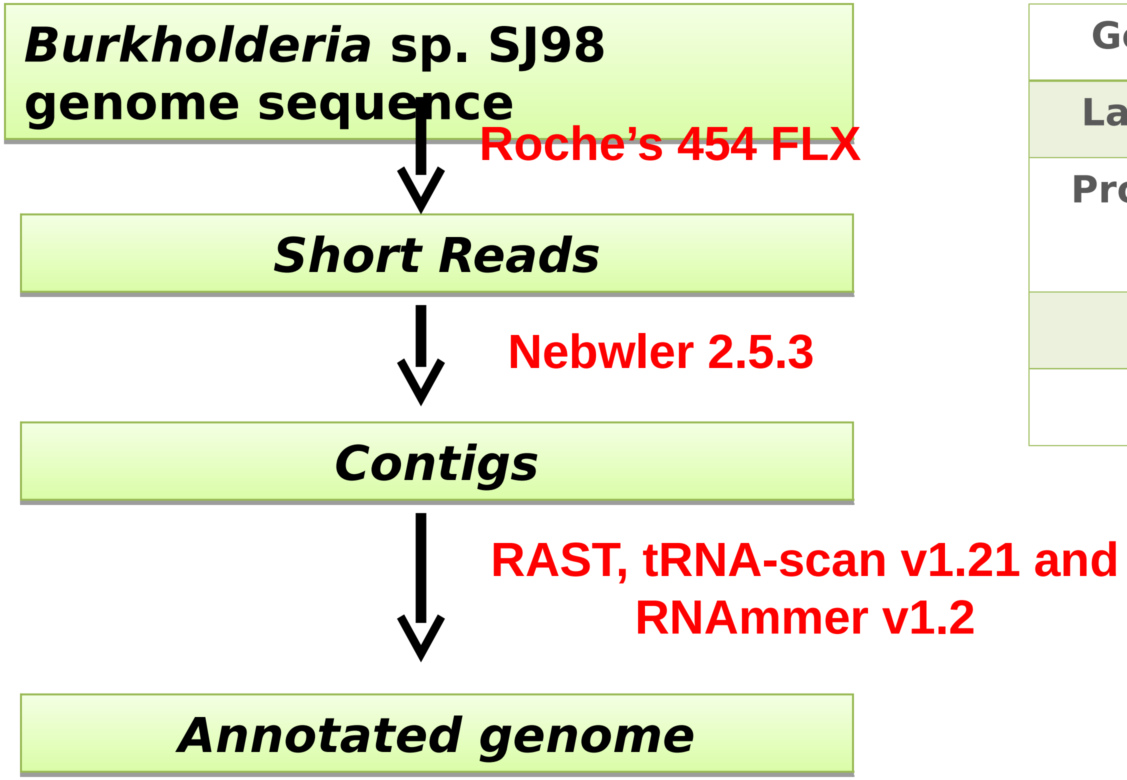
- Whole genome assembly and annotation of microbes with preliminary analysis can be published in reputed journals like Journal of Bacteriology (http://jb.asm.org/) and Eukaryotic cell (http://ec.asm.org/).
- Other journals are Genome Biology, Genome Reaserch and Nature Biotechnology(according to the analysis done).

Genome assembly and annotation done at IMTECH

- ***Burkholderia sp.* SJ98 (Kumar *et al.* 2012).**
- ***Debaryomyces hansenii* MTCC 234 (Kumar *et al.* 2012).**
- ***Imtechella halotolerans* K1^{T} (Kumar *et al.* 2012).**
- ***Marinilabilia salmonicolor* JCM 21150^{T} (Kumar *et al.* 2012).**
- ***Rhodococcus imtechensis sp.* RKJ300 (Vikram *et al.* 2012).**
- ***Rhodosporidium toruloides* MTCC 457 (Kumar *et al.* 2012).**

Burkholderia sp. SJ98

- Degrade a number of aromatic compounds, e.g., p nitrophenol, o-nitrobenzoate, p-nitrobenzoate, and 4-nitrocatechol (Pandey G, *et. al*. 2002), 2-chloro-4-nitrophenol(Pandey J, *et al.* 2011), and 3-methyl-4-nitrophenol (Bhushan B, *et. al*. 2000).



***Burkholderia* sp. SJ98 genome sequence**

Roche's 454 FLX

Short Reads

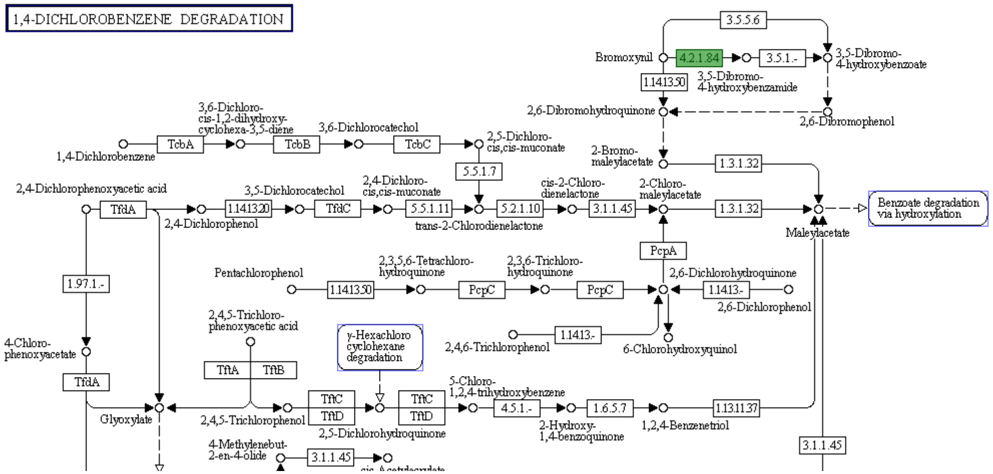
Nebwler 2.5.3

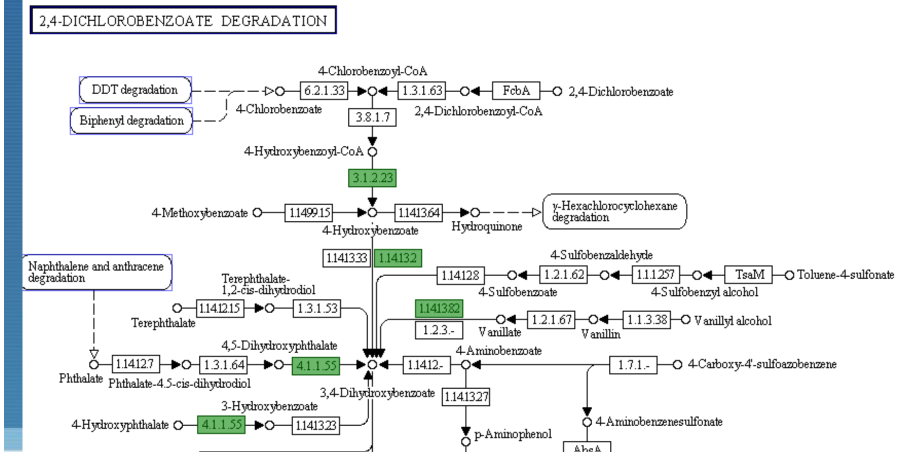
Contigs

RAST, tRNA-scan v1.21 and RNAmmer v1.2

Annotated genome

Genome size	**7.89-Mb**
Large contigs	**79**
Protein coding genes	**7,364**
rRNAs	**3**
tRNAs	**51**



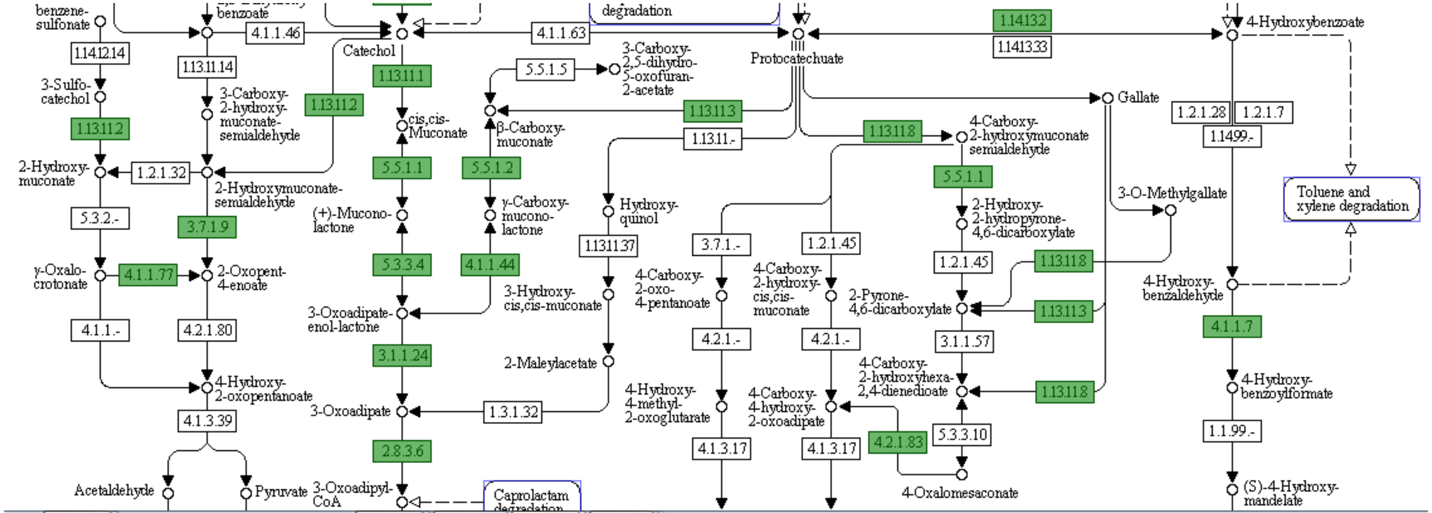


Journal of Bacteriology

Genome Sequence of the Nitroaromatic Compound-Degrading Bacterium Burkholderia sp. Strain SJ98

Shailesh Kumar, Surendra Vikram and Gajendra Pal Singh Raghava

***J. Bacteriol.* 2012, 194(12):3286. DOI: 10.1128/JB.00497-12.**



Azadirachta indica (Neem) Genome and transcriptome assembly and annotation

Dr. Prof. Siddhartha Roy (Director), IICB, Kolkata

Dr. Rupak K. bhadra , IICB, Kolkata

Dr. G P S Raghava, IMTECH, Chandigarh

Dr. Saikat Chakrabarti, IICB, Kolkata

Dr. Prabodh Trivedi, NBRI, Lucknow

Dr. Sumit Bag, NBRI, Lucknow

Dr. Mehar Asif, NBRI, Lucknow

Dr. Sridhar Sivasubbu, IGIB, New Delhi,

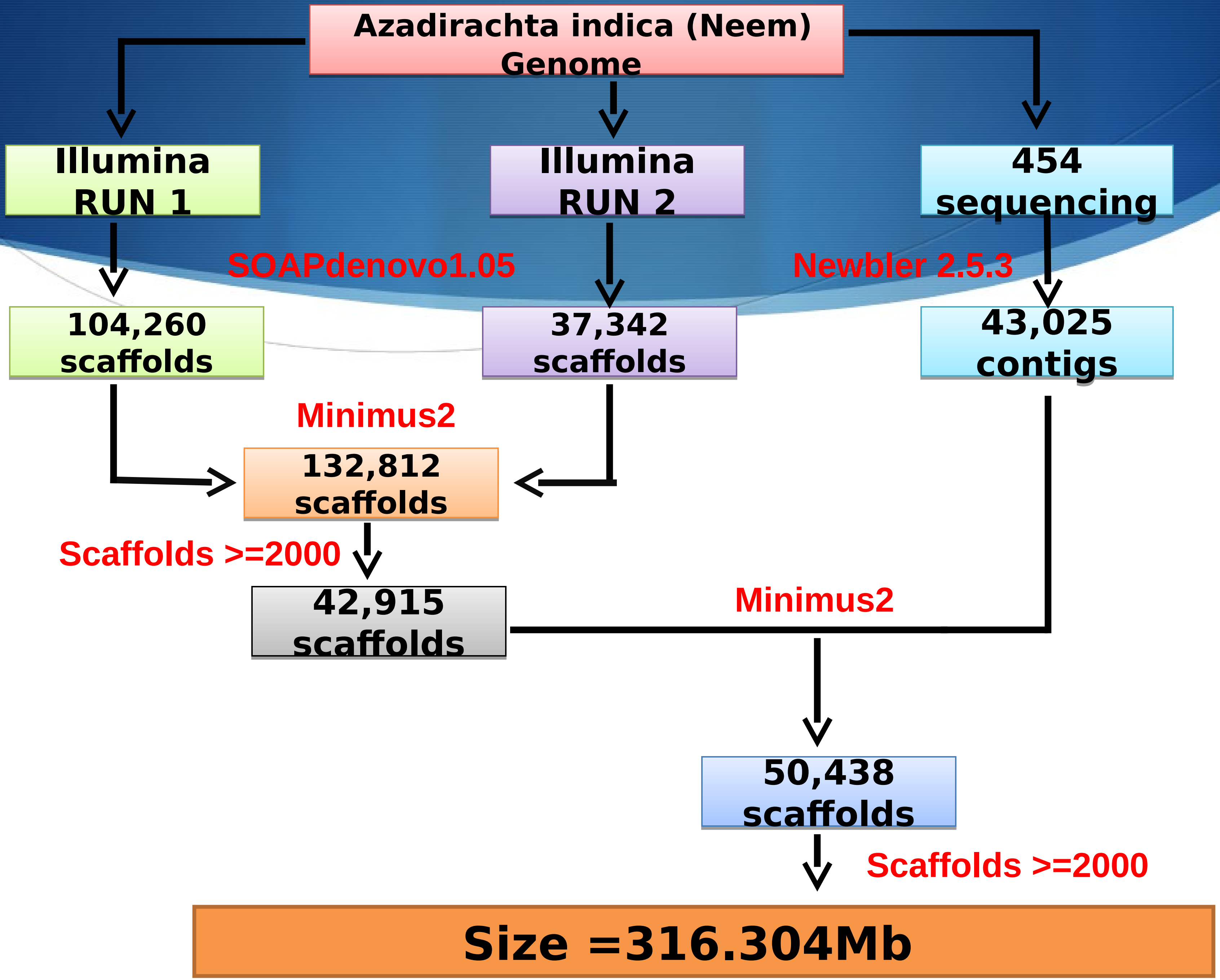
Dr. Vinod Scaria, IGIB, New Delhi

Azadirachta indica (Neem)

Each part of the neem tree has some medicinal property and is thus commercially exploitable.

Neem Genome and transcriptome sequencing

Genome sequecing	Illumina and Roche’s 454
Transcriptome sequencing	Roche’s 454
Genome assembly	SOAPdenovo and Newbler
Transcriptome assembly	Newbler
Gene Prediction	FGENESH and Augustus
Annotation	BLAST2GO and manualy
Repeatmasking	Repeatmasking
Transcripts mapping to Genome	BLAST programe



Azadirachta indica (Neem) Genome

Illumina RUN 1

Illumina RUN 2

454 sequencing

SOAPdenovo1.05

Newbler 2.5.3

104,260 scaffolds

37,342 scaffolds

43,025 contigs

Minimus2

132,812 scaffolds

Scaffolds >=2000

42,915 scaffolds

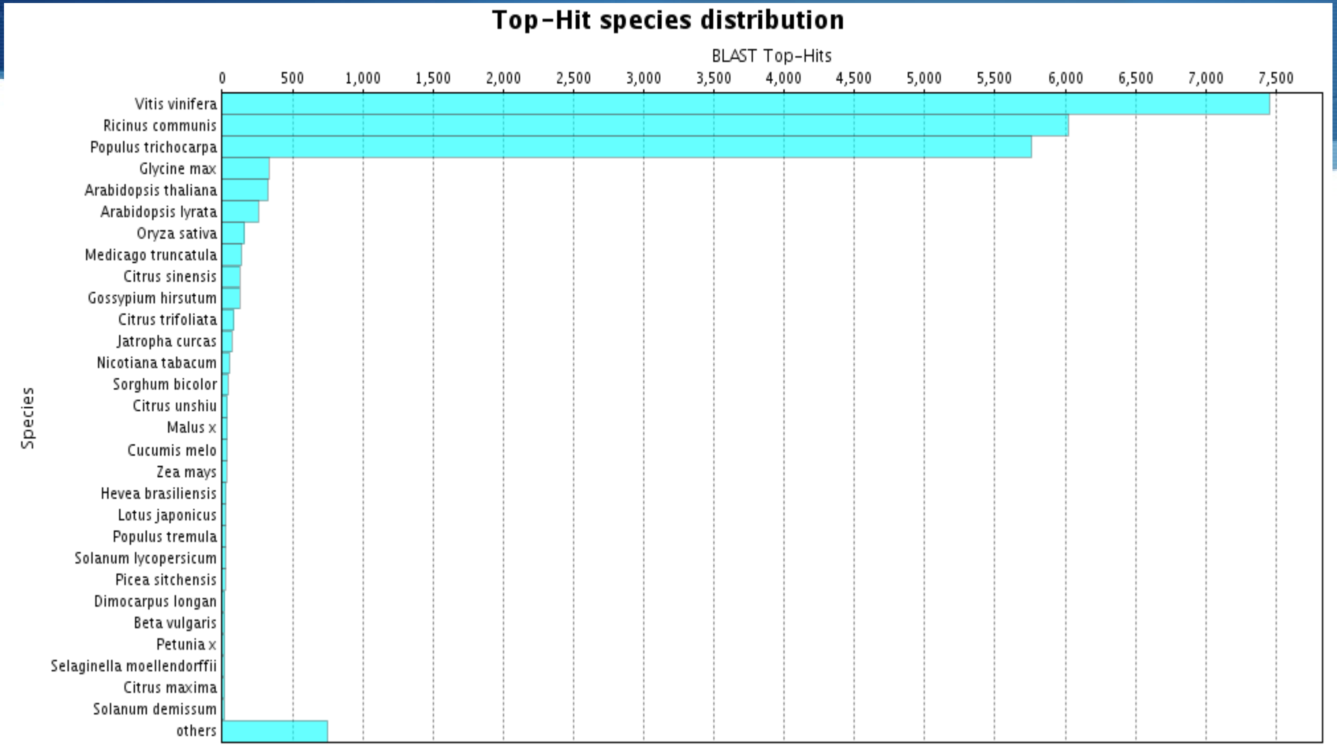
Minimus2

50,438 scaffolds

Scaffolds >=2000

Size =316.304Mb

BLAST2GO annotation



1

Vitis vinifera
487Mb

3

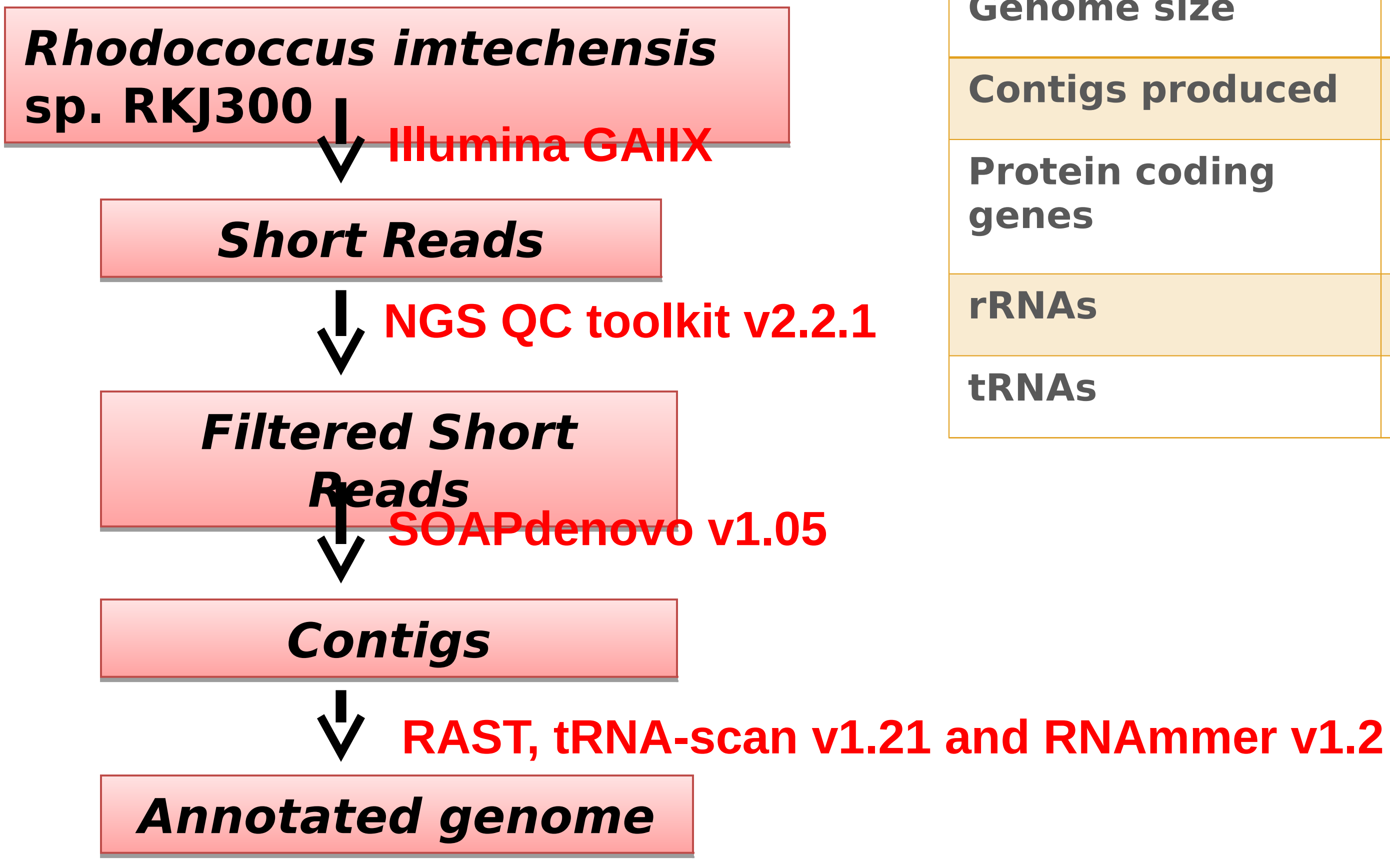
Ricinus communis
352Mb

Populus trichocarpa
485Mb

2

Rhodococcus imtechensis RJ300

- Strain RKJ300 is capable of utilizing 4 nitrophenol, 2-chloro-4-nitrophenol, and 2, 4-dinitrophenol as sole sources of carbon and energy(Ghosh A, et al. *2010*).



***Rhodococcus imtechensis* sp. RKJ300**

Illumina GAIIX

Short Reads

NGS QC toolkit v2.2.1

Filtered Short Reads

SOAPdenovo v1.05

Contigs

RAST, tRNA-scan v1.21 and RNAmmer v1.2

Annotated genome

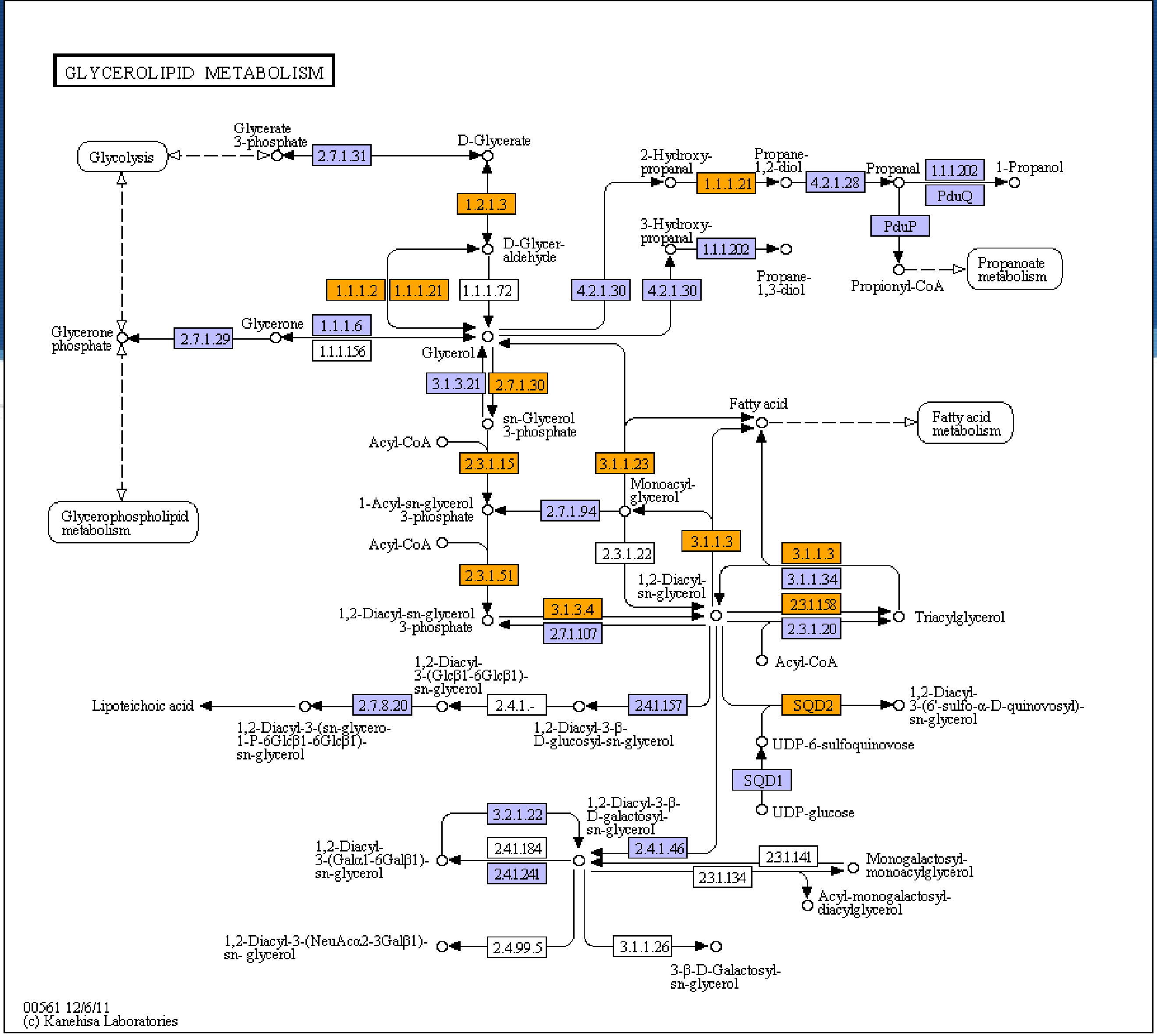
Genome size	8.231-Mb
Contigs produced	178
Protein coding genes	8,059
rRNAs	5
tRNAs	49

Rhodosporidium toruloides

MTCC 457

It can accumulate lipids to a higher level (~75% of dry weight under certain conditions) than most other oleaginous yeasts and fungi (Ageitos, J. M. *et. al*.).

R. toruloides offers many opportunities for being developed as an additional yeast model and synthetic biology platform to *Saccharomyces cerevisiae*.

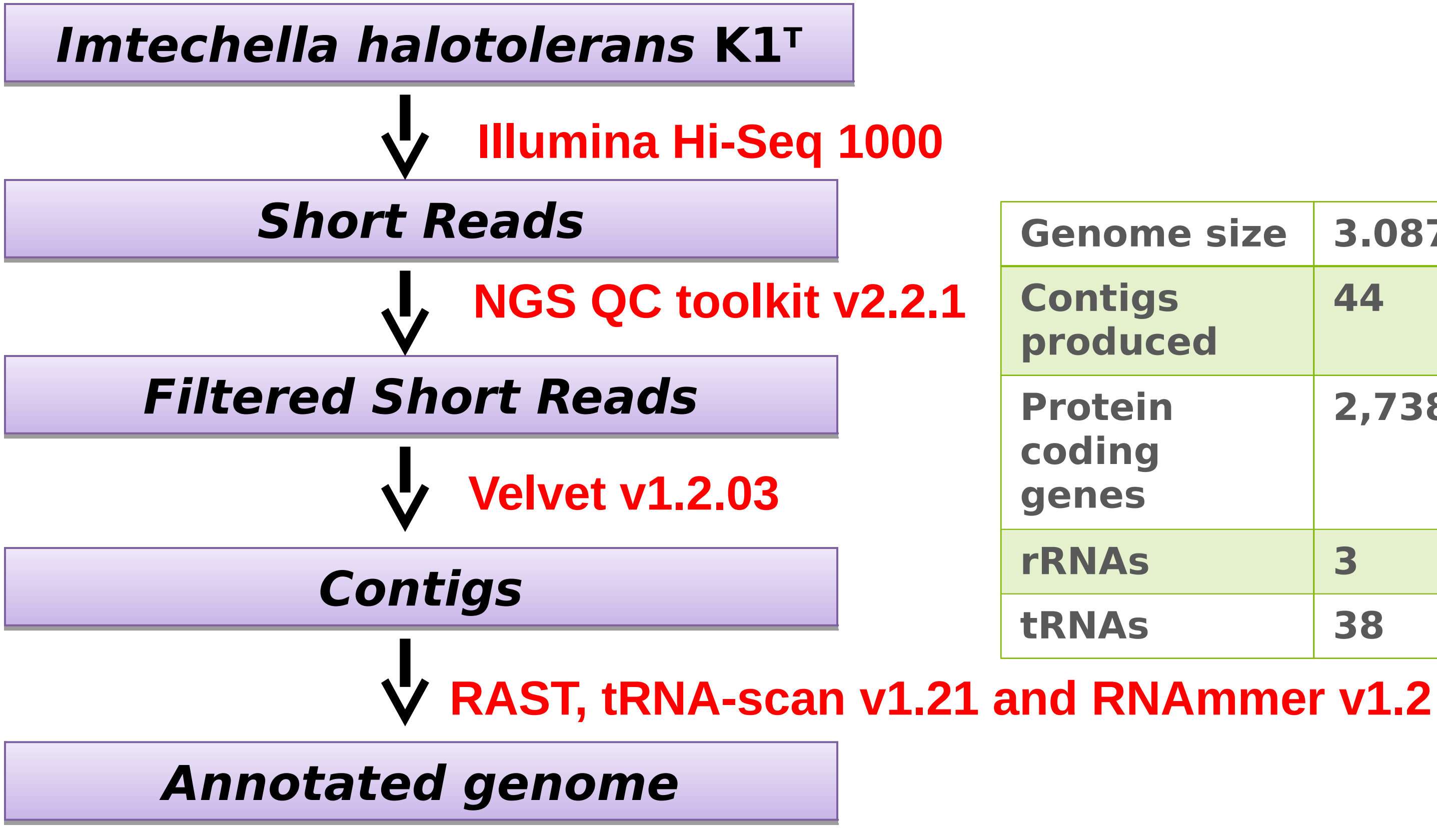


KEGG Pathways (www.genome.jp/kegg/pathways.html)

Kumar, S., Kushwaha, H., Bachawat, A.K., Raghava G.P.S. and Ganesan, K. Genome sequence of the oleaginous red yeast Rhodosporidium toruloides MTCC 457. Eukaryotic Cell (In Press).

Imtechella halotolerans K1^T

Strain K1T is known to possess various enzymatic activities, such as lipase, ϒ-glutamyl transferase,glycine arylamidase, and Glu-Gly-Arg-arylamidase (Vikram S et. al. 2012).



***Imtechella halotolerans* K1^T**

Illumina Hi-Seq 1000

Short Reads

NGS QC toolkit v2.2.1

Filtered Short Reads

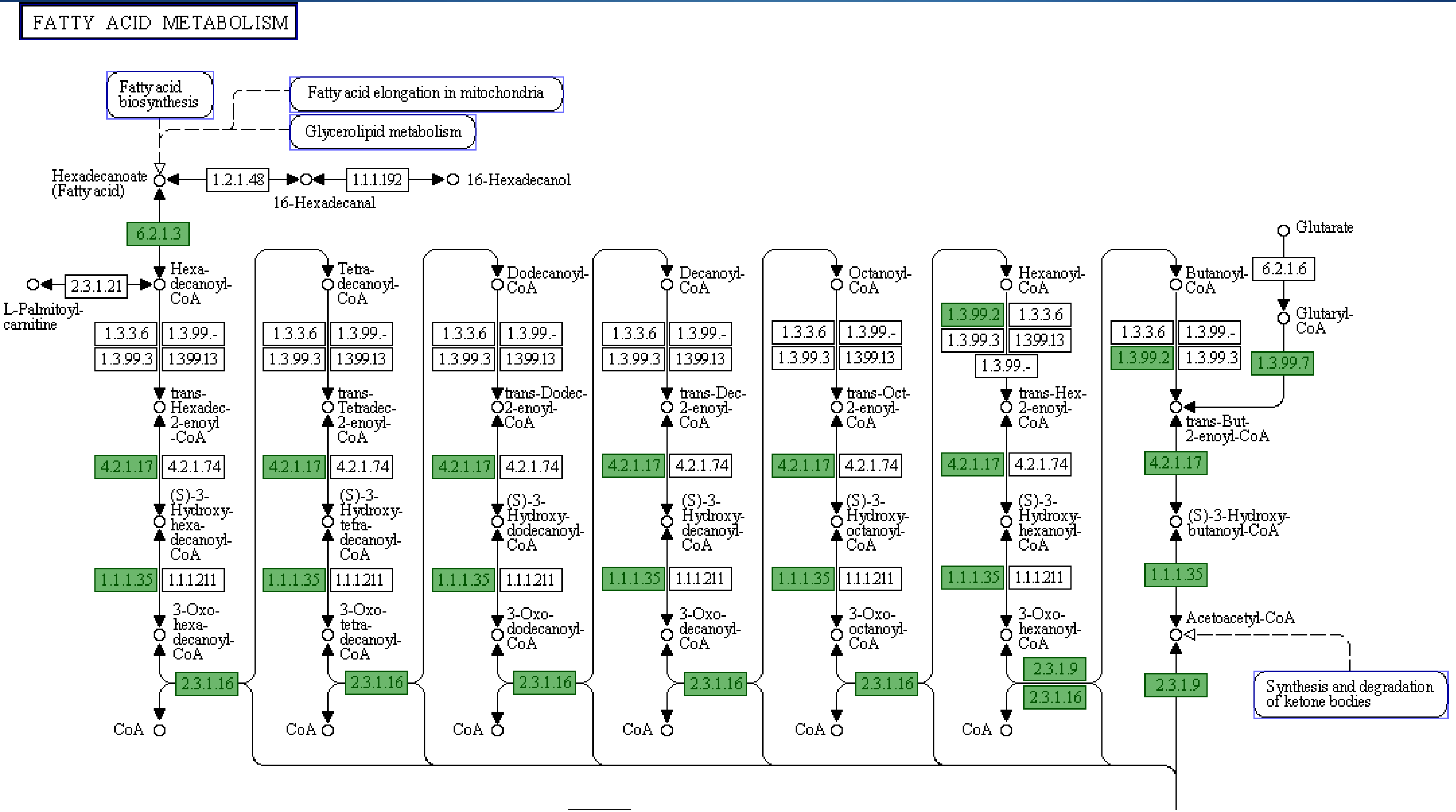
Velvet v1.2.03

Contigs

RAST, tRNA-scan v1.21 and RNAmmer v1.2

Annotated genome

Genome size	3.087-Mb
Contigs produced	44
Protein coding genes	2,738
rRNAs	3
tRNAs	38



Journal of Bacteriology

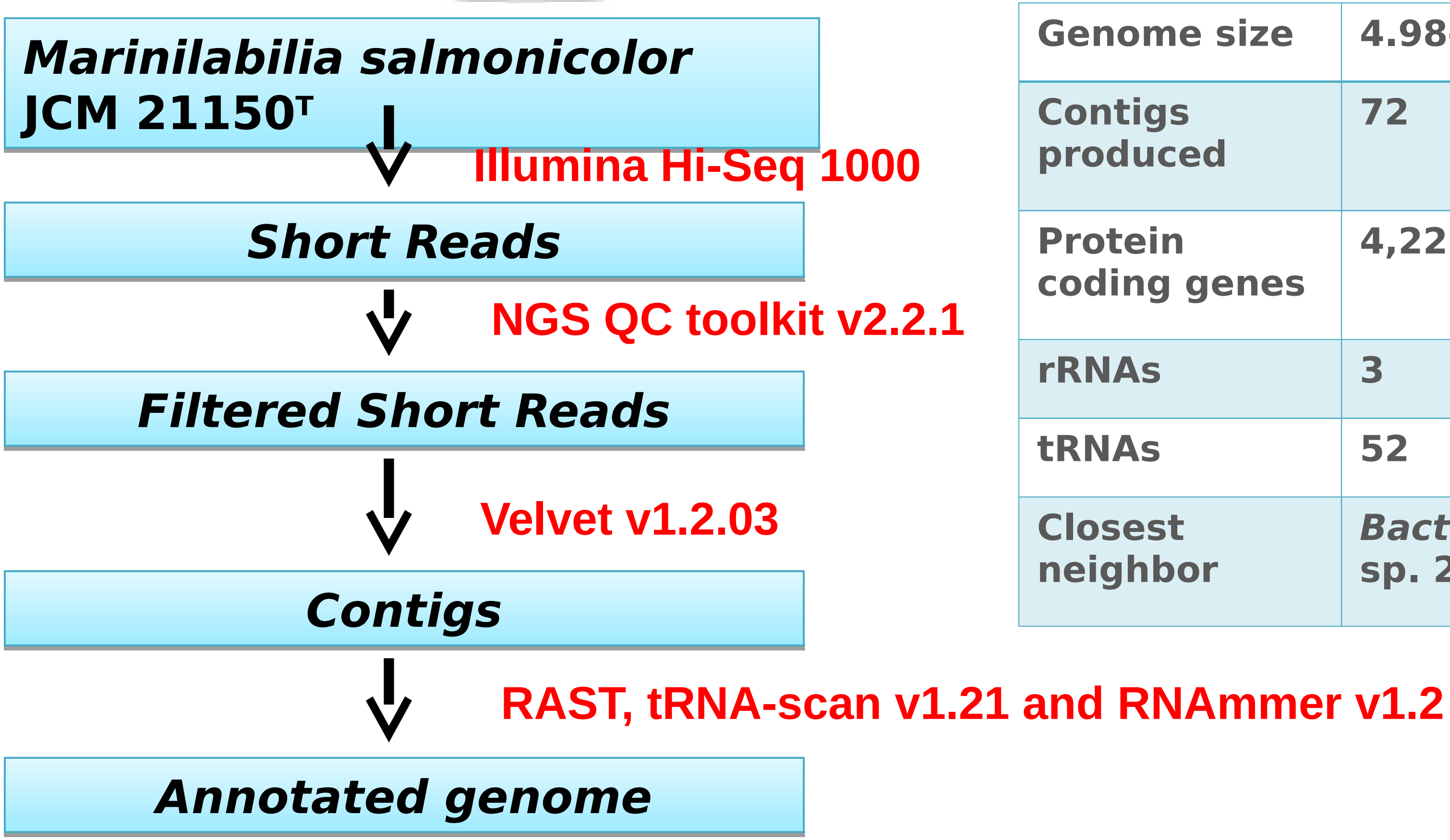
Genome Sequence of the Halotolerant Bacterium Imtechella halotolerans K1 T

Shailesh Kumar, Surendra Vikram, Srikrishna Subramanian, Gajendra Pal Singh Raghava and Anil Kumar Pinnaka

J. Bacteriol. 2012, 194(14):3731. DOI: 10.1128/JB.00506-12.

Marinilabilia salmonicolor JCM21150^T

The strain is capable of gelatin liquefaction. All the strains of the genus Marinilabilia were reported to decompose various biomacromolecules (Muller HE *et. al.* 1996).



***Marinilabilia salmonicolor* JCM 21150^T**

Illumina Hi-Seq 1000

Short Reads

NGS QC toolkit v2.2.1

Filtered Short Reads

Velvet v1.2.03

Contigs

RAST, tRNA-scan v1.21 and RNAmmer v1.2

Annotated genome

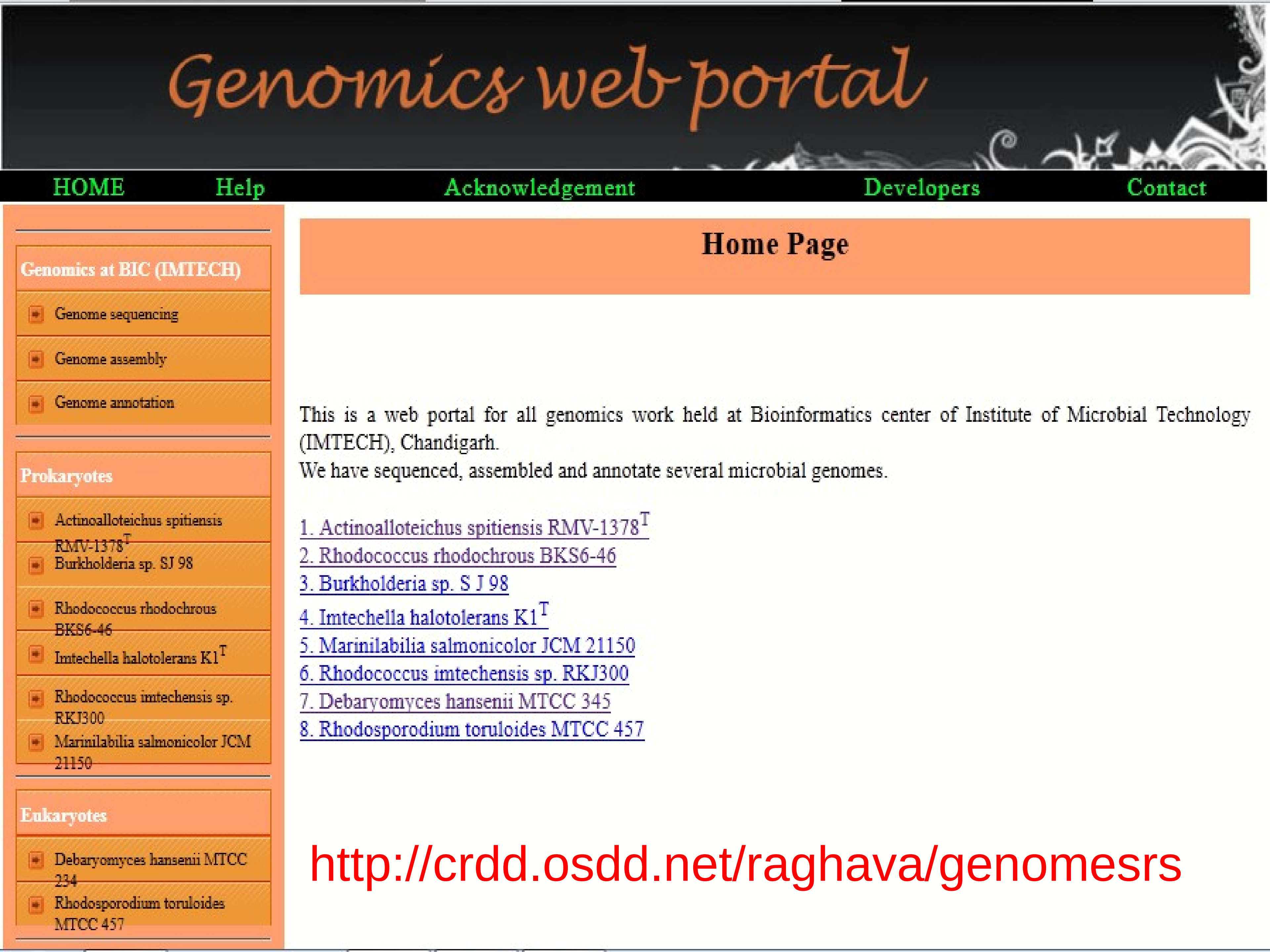
Genome size	4.98-Mb
Contigs produced	72
Protein coding genes	4,227
rRNAs	3
tRNAs	52
Closest neighbor	*Bacteroides* sp. 2_1_7

Debaryomyces hansenii var. hansenii MTCC234

- *D. hansenii is considered a* sodium includer, and the accumulation of a large amount of NaCl does not have any adverse effect on its physiology (Prista C. et. al. 2005).

- Besides xylitol, strains of *D. hansenii* are also known to produce arabitol and riboflavin (Breuer U et. al. 2006).

- Compared to *D. hansenii strain CBS767, whose genome was sequenced previously*,MTCC 234 is more halotolerant and it also produces riboflavin and arabitol.



Genomics at BIC (IMTECH)

- Genome sequencing
- Genome assembly
- Genome annotation

Prokaryotes

- Actinoalloteichus spitiensis RMV-1378^{T}
- Burkholderia sp. SJ 98
- Rhodococcus rhodochrous BKS6-46
- Imtechella halotolerans K1^{T}
- Rhodococcus imtechensis sp. RKJ300
- Marinilabilia salmonicolor JCM 21150

Eukaryotes

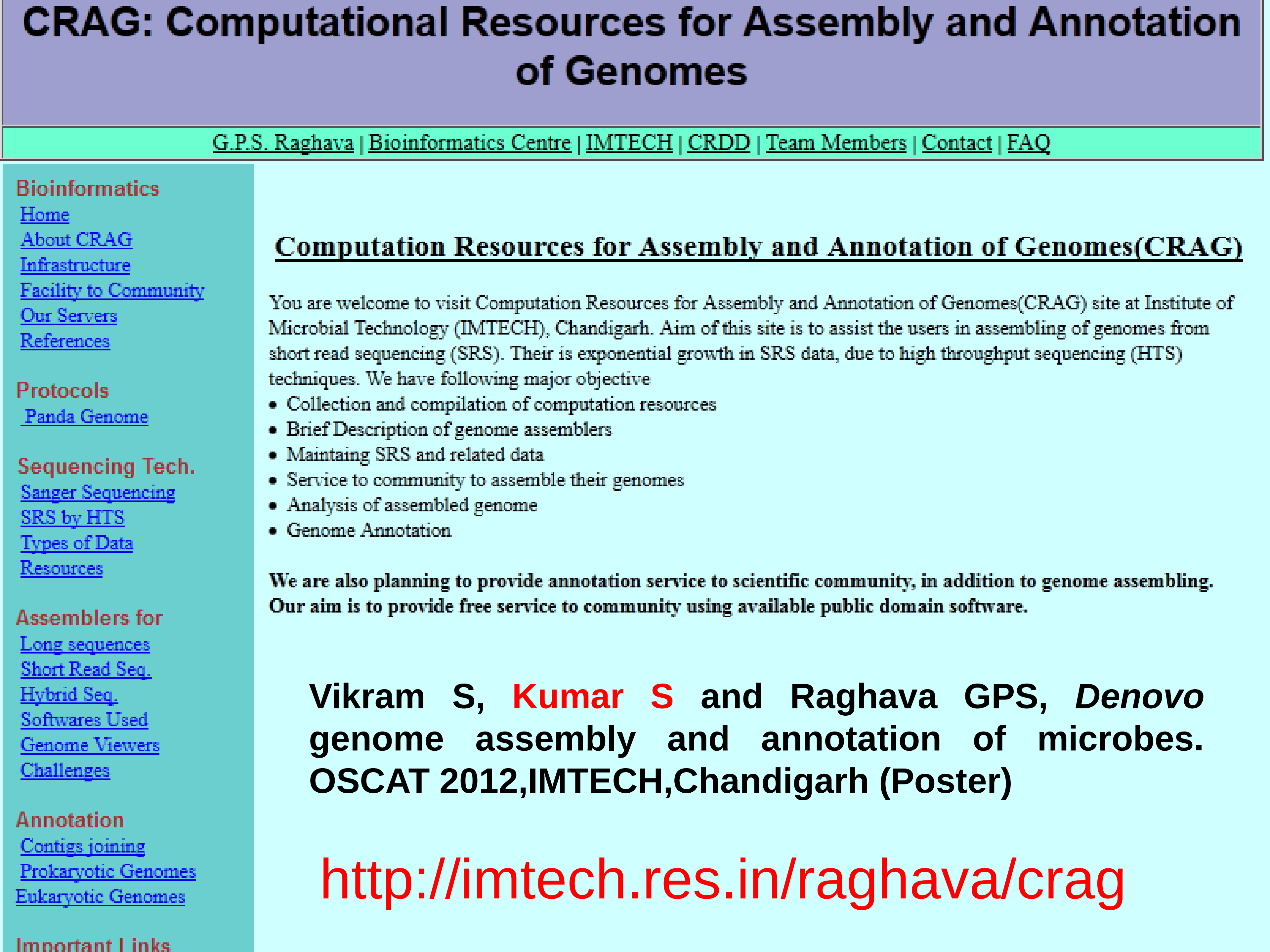
- Debaryomyces hansenii MTCC 234
- Rhodosporodium toruloides MTCC 457

Home Page

This is a web portal for all genomics work held at Bioinformatics center of Institute of Microbial Technology (IMTECH), Chandigarh.
We have sequenced, assembled and annotate several microbial genomes.

1. Actinoalloteichus spitiensis RMV-1378^{T}
2. Rhodococcus rhodochrous BKS6-46
3. Burkholderia sp. S J 98
4. Imtechella halotolerans K1^{T}
5. Marinilabilia salmonicolor JCM 21150
6. Rhodococcus imtechensis sp. RKJ300
7. Debaryomyces hansenii MTCC 345
8. Rhodosporodium toruloides MTCC 457

http://crdd.osdd.net/raghava/genomesrs



CRAG: Computational Resources for Assembly and Annotation of Genomes

G.P.S. Raghava | Bioinformatics Centre | IMTECH | CRDD | Team Members | Contact | FAQ

Bioinformatics
- Home
- About CRAG
- Infrastructure
- Facility to Community
- Our Servers
- References

Protocols
- Panda Genome

Sequencing Tech.
- Sanger Sequencing
- SRS by HTS
- Types of Data
- Resources

Assemblers for
- Long sequences
- Short Read Seq.
- Hybrid Seq.
- Softwares Used
- Genome Viewers
- Challenges

Annotation
- Contigs joining
- Prokaryotic Genomes
- Eukaryotic Genomes

Important Links

Computation Resources for Assembly and Annotation of Genomes(CRAG)

You are welcome to visit Computation Resources for Assembly and Annotation of Genomes(CRAG) site at Institute of Microbial Technology (IMTECH), Chandigarh. Aim of this site is to assist the users in assembling of genomes from short read sequencing (SRS). Their is exponential growth in SRS data, due to high throughput sequencing (HTS) techniques. We have following major objective

- Collection and compilation of computation resources
- Brief Description of genome assemblers
- Maintaing SRS and related data
- Service to community to assemble their genomes
- Analysis of assembled genome
- Genome Annotation

We are also planning to provide annotation service to scientific community, in addition to genome assembling. Our aim is to provide free service to community using available public domain software.

Vikram S, Kumar S and Raghava GPS, *Denovo* genome assembly and annotation of microbes. OSCAT 2012,IMTECH,Chandigarh (Poster)

http://imtech.res.in/raghava/crag

thank
you