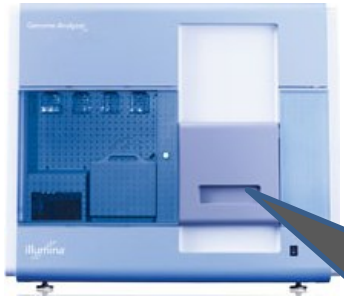
Whole Genome Assembly and annotation with Next Generation Sequencing datasets

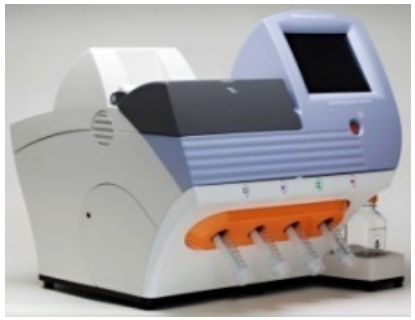
Shailesh Kumar, PhD
Staff Scientist,
National Institute of Plant Genome Research (NIPGR),
New Delhi

Next Generation Sequencing (NGS)



Illumina

454



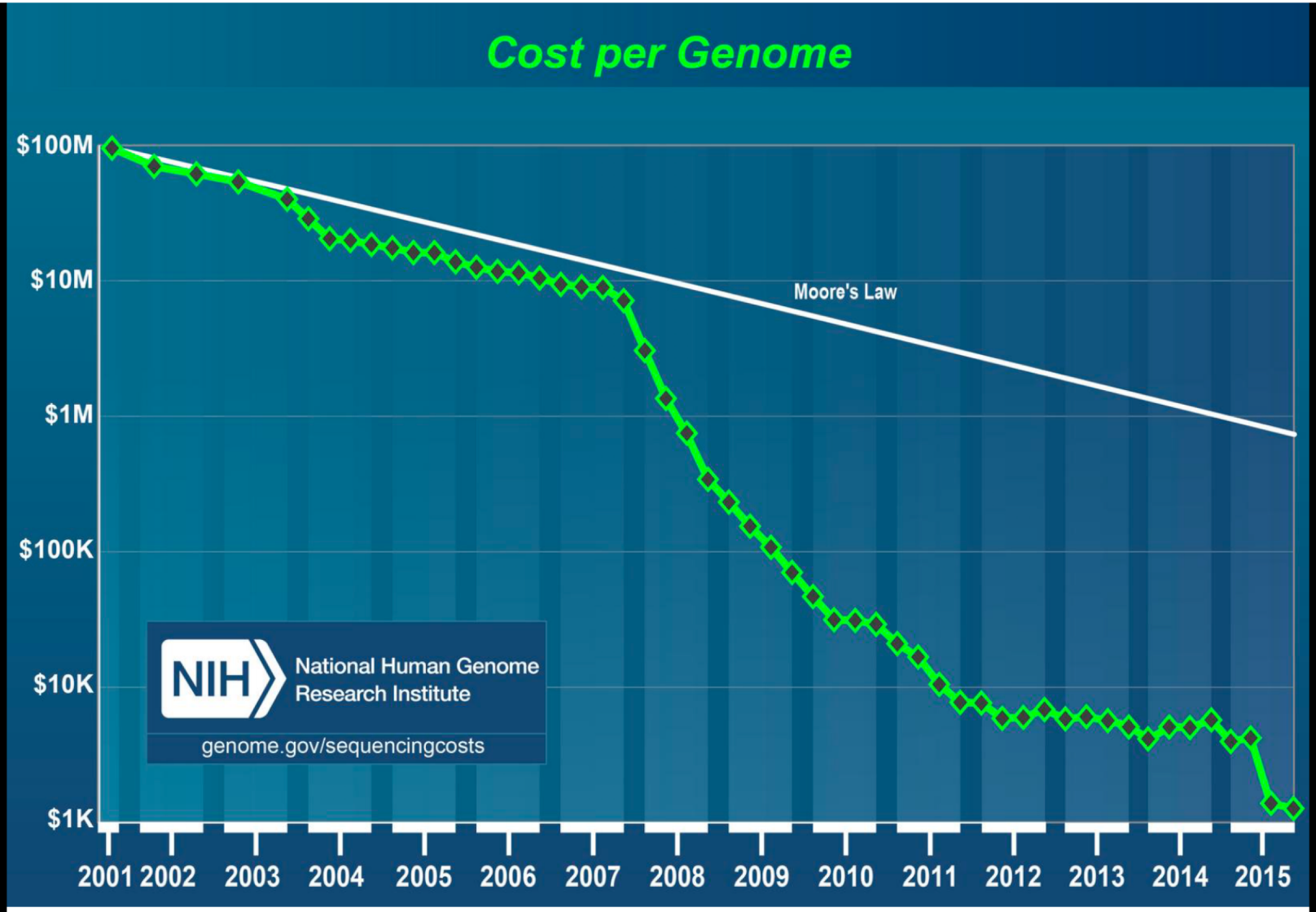
Ion torrent

Low cost &
Less time

Halicoscope

PacBio

Nanopore



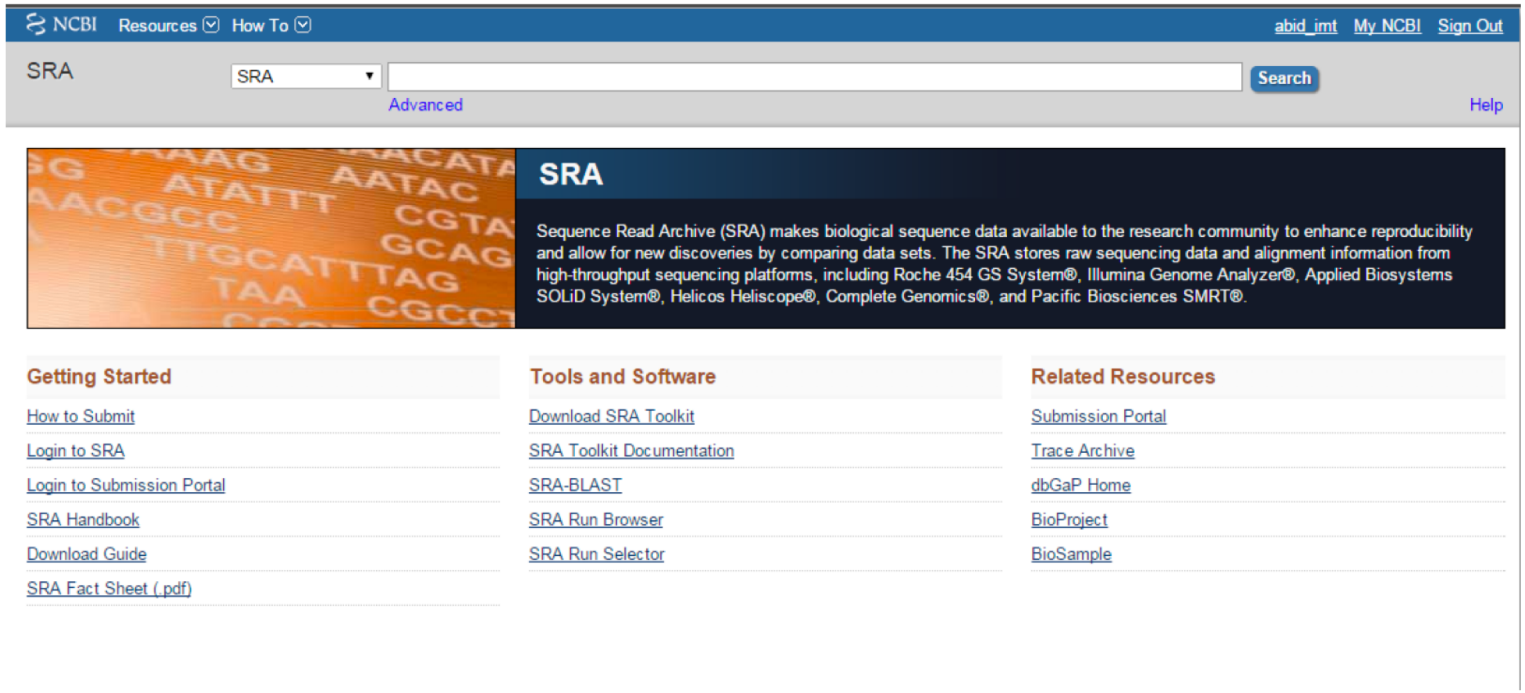
http://www.genome.gov/sequencingcosts/

RAW SEQUENCING DATA

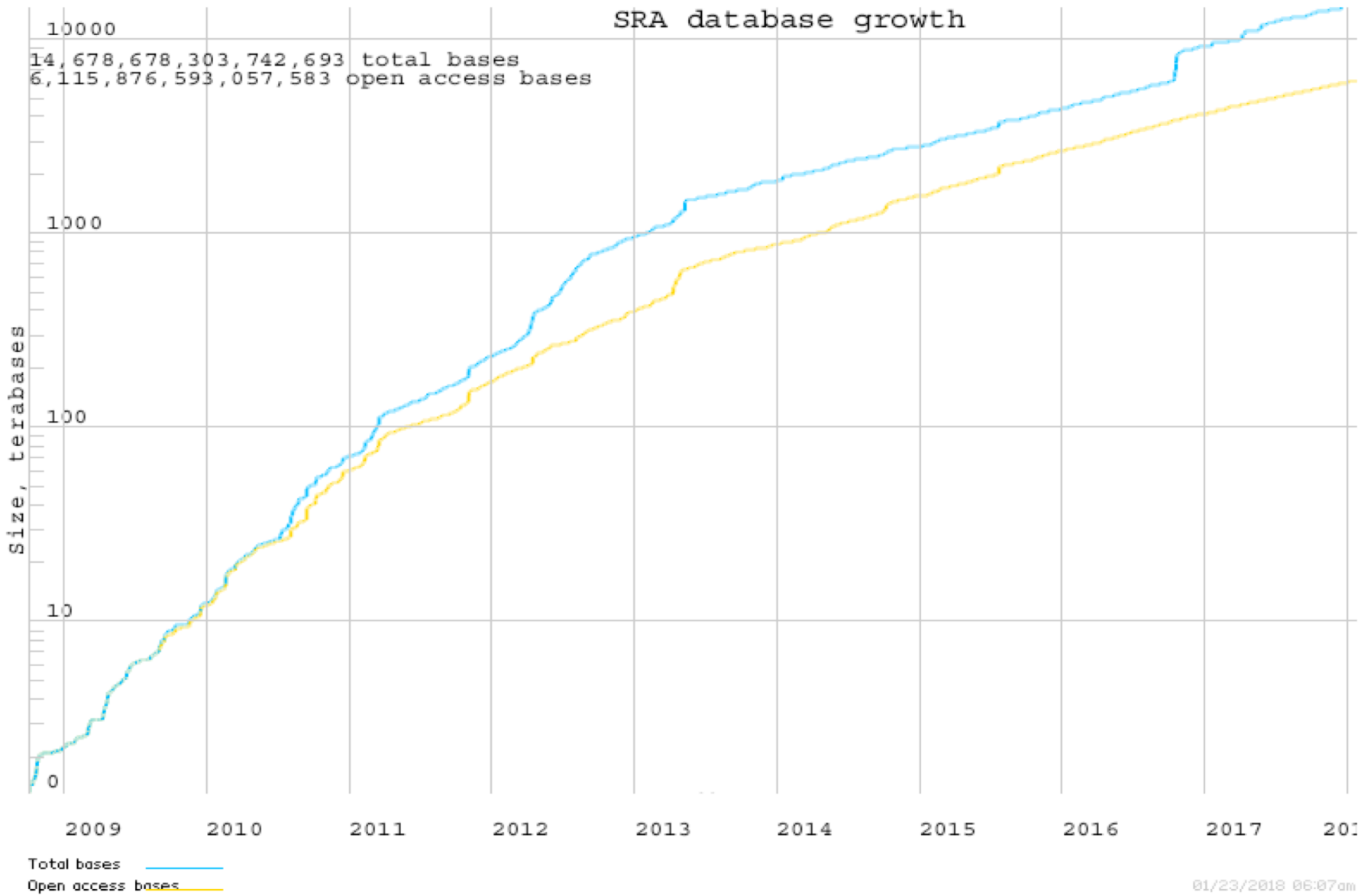
Sequence Read Archive (SRA)

http://www.ncbi.nlm.nih.gov/sra

SRA makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets.



Sequence Read Archive (SRA) of USA



Challenges

Drug-target relationship ??

Removal of artifacts in short reads ??

Genome assembly of short reads ??

Effect of detected variation ??

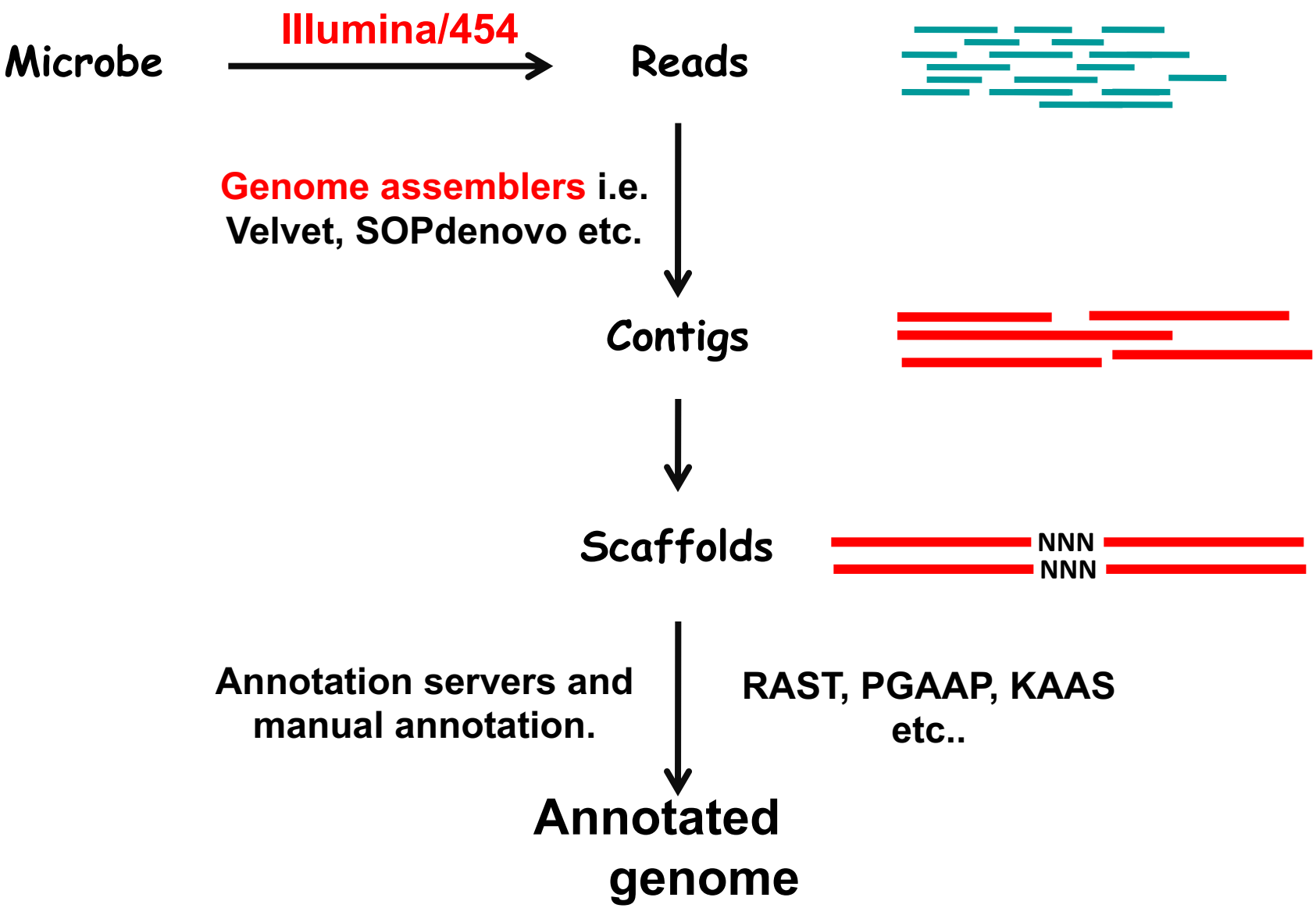
Annotation and validation of assembled genome ??

Variation detection, in human ??

Which assembler is best ??

Development of new algorithms ??

Microbial Genome assembly and annotation



Microbial Genome assembly and annotation

Quality Check (QC)

- Remove low quality bases
- Remove reads containing adaptor sequences
- Trim or remove reads containing Ns

Genome assembly

- Reference genome selection
- Reference assisted genome assembly
- De novo genome assembly, if reference genome not available

Genome finishing

- Aligning contigs/scaffolds to related species
- Assembled genome visualization
- Gap filling by Sanger's sequencing

Genome annotation

- Genome annotation by existing pipelines
- Genome annotation by different specific tools with existing data
- Manual annotation for a particular class of genes

Raw data (fastq format)

Mozilla Firefox

```
@read1
AGCTTATCCTCTGCTCACCCCCGGGTTAGCGCACTTGATGTATTCACAGC
+
BA1@CC7CBCCC9C8;B2@>C?B@B@B3=9?@B1:AB7B?B8B?B6B.7.
@read2
TTGGGCGGGATCTCCAGAAGCATATGGATGTGATCCACACAGCATTCTGC
+
?>?B@)<?@,AA7A@C<C?=@@B;+)?B5*@2=@+=BB,=B6C>AB@B24
@read3
TATGCTCAAGAAGGGGCTGATGAGTTGGTGTTTTACGATATCACTGCCTC
+
A3AB:B1:B;9/0BBBCBB<BB@AA0?BB9:BB<A@BB@7@6@<A@@@<3
```

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.......................................................
 ........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ...............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.......................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

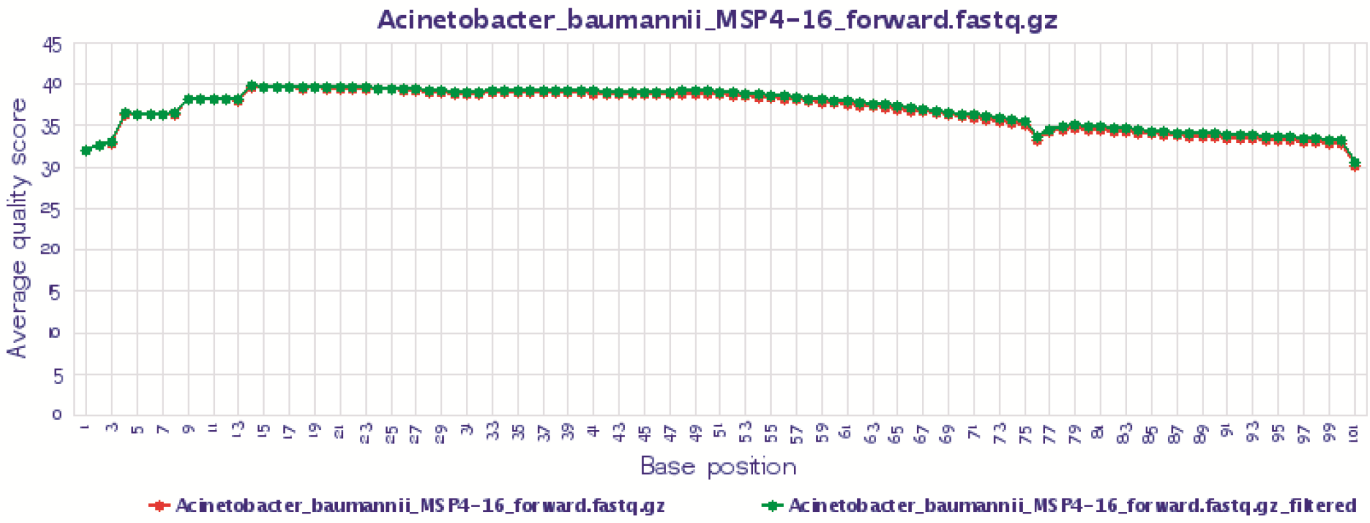
Filtering of Raw Reads

- Fastp
- NGSQC-Toolkit
- Fastx
- FastQC
- Cutadapt
- PRINSEQ
- Tagcleaner
- …… Many more

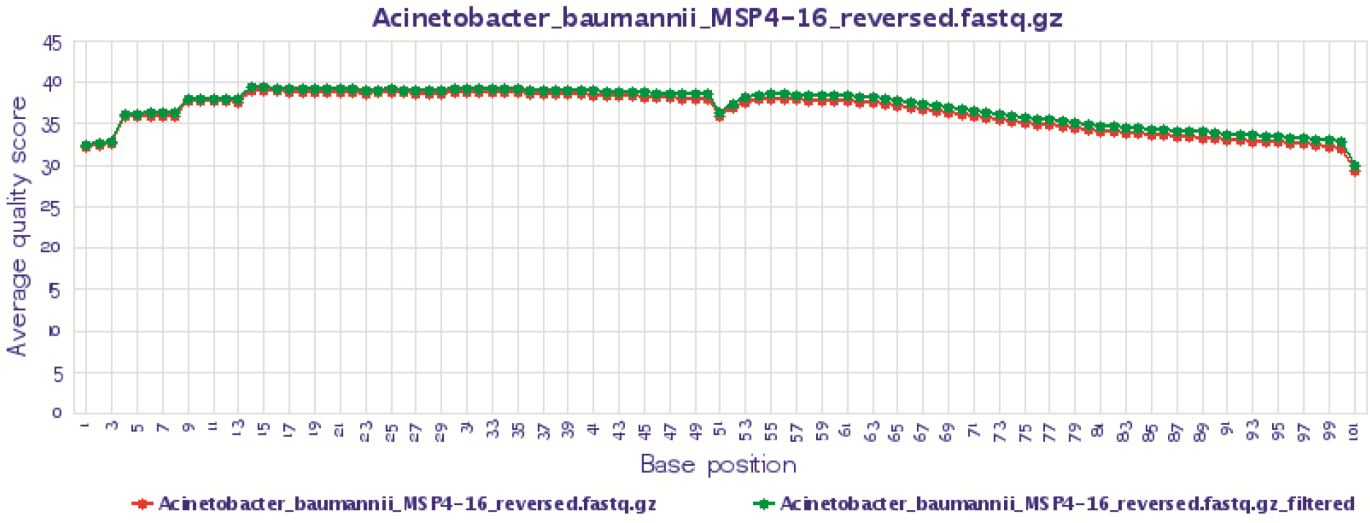
- Paired-end information
- Filter and Trim
- Automated for latest primer/adaptors
- Automated for different data type
- Keep unpaired (filtered) reads also
- Fast

Raw data filtering

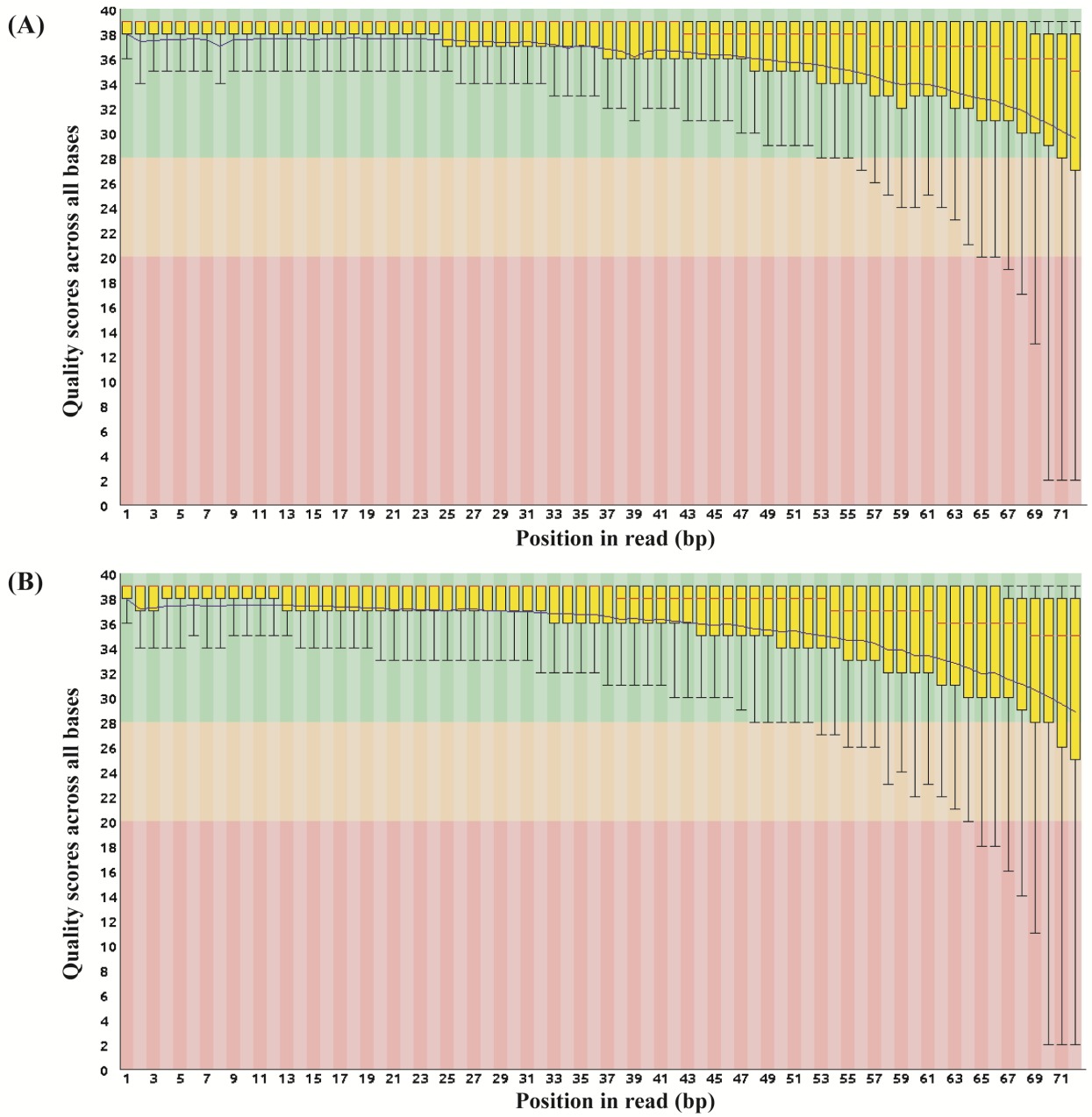
(A)



(B)



Raw data filtering



Fastq reads for genome assembly

```
@M01964:17:000000000-D03AH:1:1101:15930:1342 1:N:0:1
CTGCGCTATCGCGCGGCGGTTGACATCTCATCGTTTAGGTCGTGCTCTCGTTGGTGCTGTTCTTTCAGATCCTTGGGTTTGCGCTGTTCATCTTTTGGTTGCTGCTCATCGCTCGGGTGGTCGTCGAGTTCATCCGCTC
GTTCAGCCGGGACTGGCATCCCACCGGCCTCACTGTGG
+
1>11AAD@AF@AGECEEE?EE/0A1D1A12DA1B?//11B1////>1B1/??00?/1B1BB@FFF>222B11BF11/?FE/1/<>/?F>2BBFFGG1/FF/0?11?<1F<1/?<///>--><-A<---<C<0DD:.:A@
-AEC/009----:9..9/:99A09A.-99AAB9FBB/-
@M01964:17:000000000-D03AH:1:1101:15422:1344 1:N:0:1
TCCCTGGATCGTCGGGAATACCGATTCCGGTGCGTTGCGGGTGATCTGCGTCATCAGCTGGGCGAGCTTGGGTGAATCCGCCGCGACGTTGATTGCTCGGTGGAACTCATGATTGAGCCGCACCGTGCGCTCGTGGTCA
TCCCCGGCATAGGCCCGCTCCAGTTGGGCCTGAATGTCTTTGAGCTGGTGCAGCTGCTCGGGGGTGATGTTGACCGCGGCCCGCGCCGCCAGCTCGCCACCGACATGCGC
+
A>>>AC?1FFAFF?EC??F1B100AAA200A///A//////A//D@D11//>/BB211B10//////?F0//?//1B></<<////</>F//@111??/->...1<FD1>11=D0000<-<-::-;.:-:AB-@..CBB
FF099----:////99--9>9//BB//-9:E--/;/;BFFFB/-9B/-:-///BFBBFF?----9--;B9B//999=---99;-=-9->99--BF-999A9A-999//-9
@M01964:17:000000000-D03AH:1:1101:15722:1346 1:N:0:1
CTCTATGTGCGCAGGCTGTGTAGAGCCTTTAGCTAGCCCCGGAATGCCGCTATGCCTGCAGCCGATGCAGCATGACGTTGGTCACGGCGCCGCGCTCGACGGCGGCCATCATGTAGGTGCAGGTCGGTTGGCGCCGCCG
GTCCGTCGGCGATCCGGGATTGAGCAACCGCAGCCCCGTCTCGGTGGTGGTGTCCCACGGGATGTGGCTGTGCCCGAAGACCAGC
+
1>>A1DFFBF?AA1AAGCFBGB1311ABFF111B211AA0////A111/AA/A11AB10011>//>//1110@11@/BF/0?@21/////<///<<?//<//</--<-<<0=0;000;0000.;:-@;A--9-A@--9-
-;9@-A9--9--9:9----;:9//9//:A-9--9:9A;ABFB/-9--:--;-ABB9/9----9-://9A-://9BA----999-;
@M01964:17:000000000-D03AH:1:1101:15170:1355 1:N:0:1
GCCGACGCCTGCGACGGGTTCATCCTGGTGCCTCATCTGTCACCGCACGGTCTCGACGAGTTCGTCGACCGCGTGGTGCCACTAC
+
1>>A1>DDAFAFG?EECCEFC1FFAB10B010111DAD1222A1///A///>?@01////BB@/?>///////>//?/1<11B0<
@M01964:17:000000000-D03AH:1:1101:15043:1363 1:N:0:1
CTCGTTGATAGCGAATGCGACCGTTGTCCTTGTCGTTCACCTGATGGAACTTGATGTCTTGGTCCTCGGTTGCGCTGTACACCTTGACCGGCACGTTGACGAGCCCGAATGCGATCGAACCCTTCCAGATGGAACGCAT
GCAGCCAGTATGCCCATACCGTTGCCGGCAAAACAGTCCGACAGGCCCGCTGAACAGGGCAAGGGCGGTGCTAGCGCCGAGTTCGAGGCGAGACCGAGCCGAGAACGAGCCG
+
1AAAAAF1DD3DE11ECF00AA0BE0BA1DE11A/A/A1AAB00B1101ABF11D2DA2D11BAAFB//B//0//E/B1@2@0BF11@B//>/>/BF//</////?///?11//??//0?<<FG<<111>110<<.<-A
.<00/<..<0D0=0:0;CCC.;A.::--9--./9./;C9A.:-9.-9A-9@--/://--9-----:-----;/99-@9>--AABE---99--9999--99@----9---99-
@M01964:17:000000000-D03AH:1:1101:15365:1370 1:N:0:1
GCGCCAAGCACTATTACGCGACCTCAGACCGCGTCACCTTCCGCACGTTGCGTGGCAGCTTCGACCTGATCCTGAACACGGTGTCGGCGAACCTGCCGCTCGACGATTACGTGAACCTGCTCGACGTCGACGGCACGCT
CGTCGAATTGGGTATCCCGGAGCACCCCATGCAGCTCGCCGCGTTCCCGCTGGCGGTGATGCGGCGCAGCCTGTCCGGCTCGAACATCGGCAGCATCGCCGAGACCCAGGAG
+
1>1>AADAAFFFFGGFGCA00AAEF111AA/////AABFAA///A/BB///B///0//BB@//>>?00B>1B111>1>//?/?<//////<?A0<//<<//-<-->F.<.</..<<<..C;.-:-AA--9--:-:.9A
A?@B;--;A//-;/BF9B------999@-://///9B/A-@-9-;@A9B?=@---@-;-9://-------999-B/;--999-9-:-:/--9----BB--9-----99-----
@M01964:17:000000000-D03AH:1:1101:15553:1376 1:N:0:1
TTCCTGGCAAGGCTGGGAAGTGCTGTTCGACCTGGCGCGGTTCATCGGCGTCAGCCTGCCCGGCTACGACTTGCGCCGCGAGCACATCACCGACGCGCTAGGGGAGTTGGCCCAGGATGCGCTGACTCTGGTGGAAATC
CCGGCGCTGGCGATCTCGTCGACCGATTGCCGTAAGCGCGCCGCCCAGCGCCGGCCGCTGTGGTACCTGATGCCCGACGGTGTCGTGCAATACGTCTCCAAGCGCCGGCT
+
```

Genome Assemblers

S. No.	Genome assemblers	URL
1.	Velvet	http://www.ebi.ac.uk/~zerbino/velvet/
2.	SOAPdenovo	http://soap.genomics.org.cn/
3.	Euler-sr	http://euler-assembler.ucsd.edu/portal/
4.	ABySS	http://www.bcgsc.ca/platform/bioinfo/software/abyss
5.	Edena	http://www.genomic.ch/edena.php
6.	SSAKE	http://www.bcgsc.ca/platform/bioinfo/software/ssake
7.	ALLPATHS LG	http://software.broadinstitute.org/allpaths-lg/blog/
8.	FALCON	https://github.com/PacificBiosciences/FALCON
9.	SPAdes	http://cab.spbu.ru/software/spades/
10.	Hinge	https://github.com/HingeAssembler/HINGE
11.	CANU	https://canu.readthedocs.io/en/latest/quick-start.html

…… and Many More

Assembled Contig

File Edit View Search Terminal Help

```
>Contig_1 [organism=Mycobacterium microti OV254] [strain=OV254] [isolation-source=wild voles] [gcode=11]
GCGGTCCGGGGTCGCCATTGAGGCcTGGGACGTCCAAGCCCGCCGGGGTGGGAAGGACTT
CCCGCGGGGTAGgCGCCGACAGCGGCCCACCGCATACCGGCCAAGCGCCACGACCCTGCG
TCGCCAGCACGCGCTCGGCGACCGCGATCTGCTGTTCCCGGGTGGCCAGTTGGGCCGACG
GGGCGTACTCGCGACCACCGTGTGAAGCCcACGTGCTGGCATTGAACTGAACGCCACCGT
AGTAACCGTTGCCGGTGTTGATGCCCCAGTTGCCACCCGATTCGCACCGGGCTACCTGAT
CCCACTCGCTGTCCGTGGCCGCAGCCGCTTGACCCGCCAGGGCGATGCTGCCGCCACCAA
GAACCGCCCCGGTAAAGGCGATCTTGGCGACCTGAAGGTTGGATGTCGTGGGTTTGCGGT
GACGTCCGCTCATACGCGCCGAAATTCCTCTCTGCACGCGCCTGCGAGGTCAGCTGTCGG
GTTCGGGTTGGAGAGGCCACCCGGCCGACGTCGTCTTCAGTCATCTCCGAAGAGTCGACG
CCAGCTTCACCCCAAGGAGTCGCATGCGACTCCGGATCCGGCGGACCGGTGGGTCCCCCG
CCTCCATCCACATTCCGGGGGCCCTCACCCACTGGATGGAGCTCGGCGCTATGGGTGAGA
GAGGACCGCCATCGGTTCAGCTTGGCGAGCCTCCGGAGACGGTAACCGGTTCTGTCGACC
GCGTCACTTTCTGCGCGACTCGGCGTTTCCGGCGCGGCCGTCCGCGAAAGATGCAGGAAC
ATCAAGGATTTGCGCTGGTCACCACGGCTCGATATCGGGCCGTGTCCGCGCCGTTATCAT
TCCGTGACGTGAGCTATCTCACAGAATCAACCGCCCGCGAACGGGGGTAACACGTCGATC
GTGTTACCTGCCCGCAACGGTTTGGTTGCGTCCCGTACCGCGATGCCGTCGCAGAGATAG
GAGCATCGGCTCAACACGGTTGCAAGTCGGGTACCCGAAGCGGCGAGTCTTTGAACCAAC
TCGGCCACCGTGGTGCCCGGGCGAAGAACGACGGTTTCCGTTTCGACGCCCGCGGCGGCG
CGCGCGGCGGCGAAGTAGCGAACCGTCACCGCGATGCCCGCGGTGTCTTCGGTGTTGACG
CGAGACTCAGCCACCGATCGCACTCATCGGGCGGTCCGGCTGGACGAAGTTCGGGTCGTT
GATTCCGTGTCCGGCGGGCTTGCCCCACATCGCGGCGCGCCACGCTGTCTCGATCGCATC
GTCGCCGGCGCCGGCGCGCAGCAGCCCGCGCAGGTCGGTCTCGTCGGTGGCGAACAGGCA
ACTTCGGATCTGCCCGTCGGCCGTTAGCCGCGTGCGGTCACAGGCCGAGCAGAAGGCGTG
CGACACCGAGGCGATCACCCCGAACTTTCCGCTCGGCGTGCCCGGCCCCGTATCGACCAG
CCACAACTCGGCAGGCGCCGAACCACGTGGCGCCGGGTCTGGACGCAGCCGGAAGTGCGG
CCGCAACGCGGCCAGCACCTCGTCGGCGGTCAGCGCCGACTCGCGCCGCCATCGATGCCC
GGCGTCCAGCGGCATCTGCTCGATCACCCGCAACTGGTAGCCGTGCTCGAGACAGAACCC
CAACAGCTCGACGACGTCTGCACGGCCGGAGGCGGGATCCAGCACGGCATTCACCTTGAC
CGGCGTCAGACCGGCCTCCTTGGCGGCCACCAGGCCGGCTACCACGTCGgCGAGCCGGTC
GCGACGGGTGATCGCCTTGAAGCGATCCCGGTTCACGCTGTCCAGTGAGACATTCACGCG
GTTCAGCCCTGCCGCGGCGAGGCCGGCCGCCCGCCGGGCGAGGCCGACGCCGTTGGTGGT
CAGCGAAATCTGCGGCCGGGGCCGCAGTTCGGCCGCCGCCGCGACCACCTCTTCGAGGTG
GCGCGACAGGAGCGGCTCACCGCCGGTGAAGCGCACGCTGGTGATGCCCAGCCGGGTGAC
GGCAATGCGCATCAGCCTGGACAACTCGTCTGAGCGCAGCAGCTGGTCACCGGGCAGCCA
GTTCAGGCCCTCGGCCGGCATGCAGTAATTGCACCGAAGATTGCAACGATCGGTCAGCGA
TACCCGCAGATCGGTTGCGGCCCGGCCGTAGGTGTCGCGCAGCGGGCCCTCGGTCGGCAT
CTCGTCGGCGACCGTGCCAGGGGGCCGGGCCGTCGCGACGCTGGGCATCCTCGGCAGCCC
```

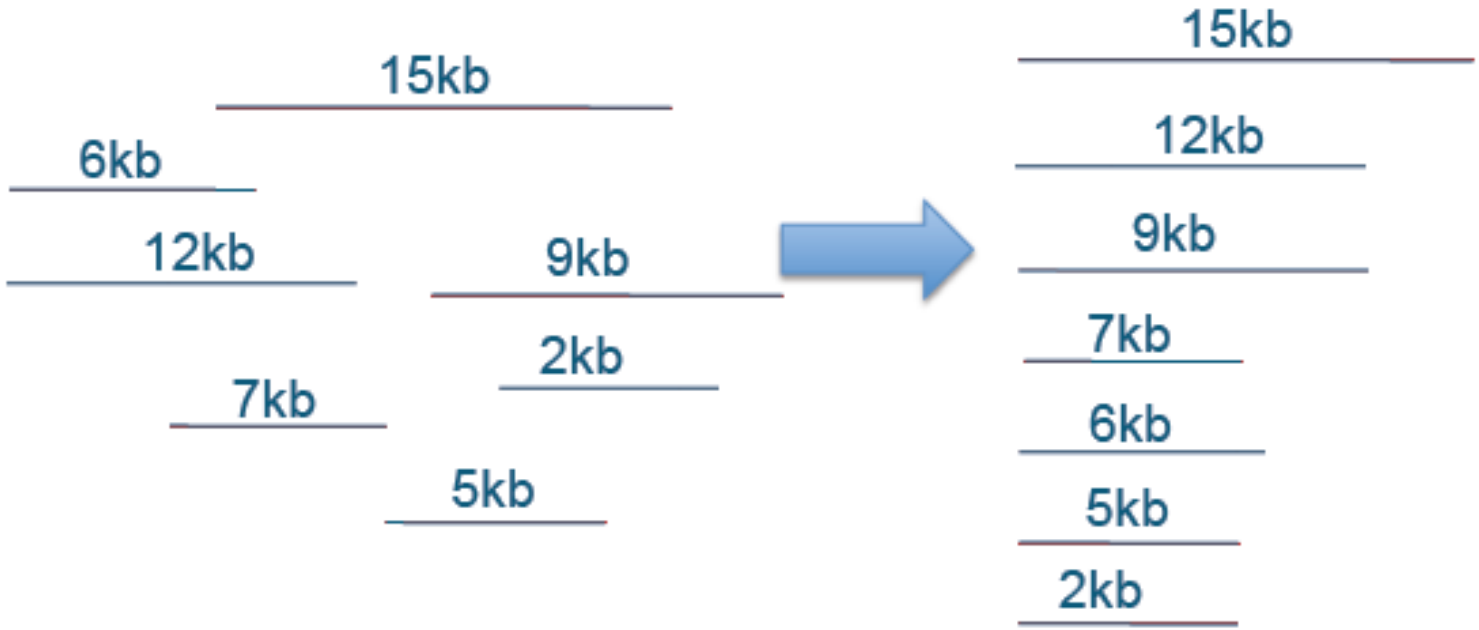
How to measure Genome assembly, N 50 ??

N50 has traditionally been used to compare assemblies

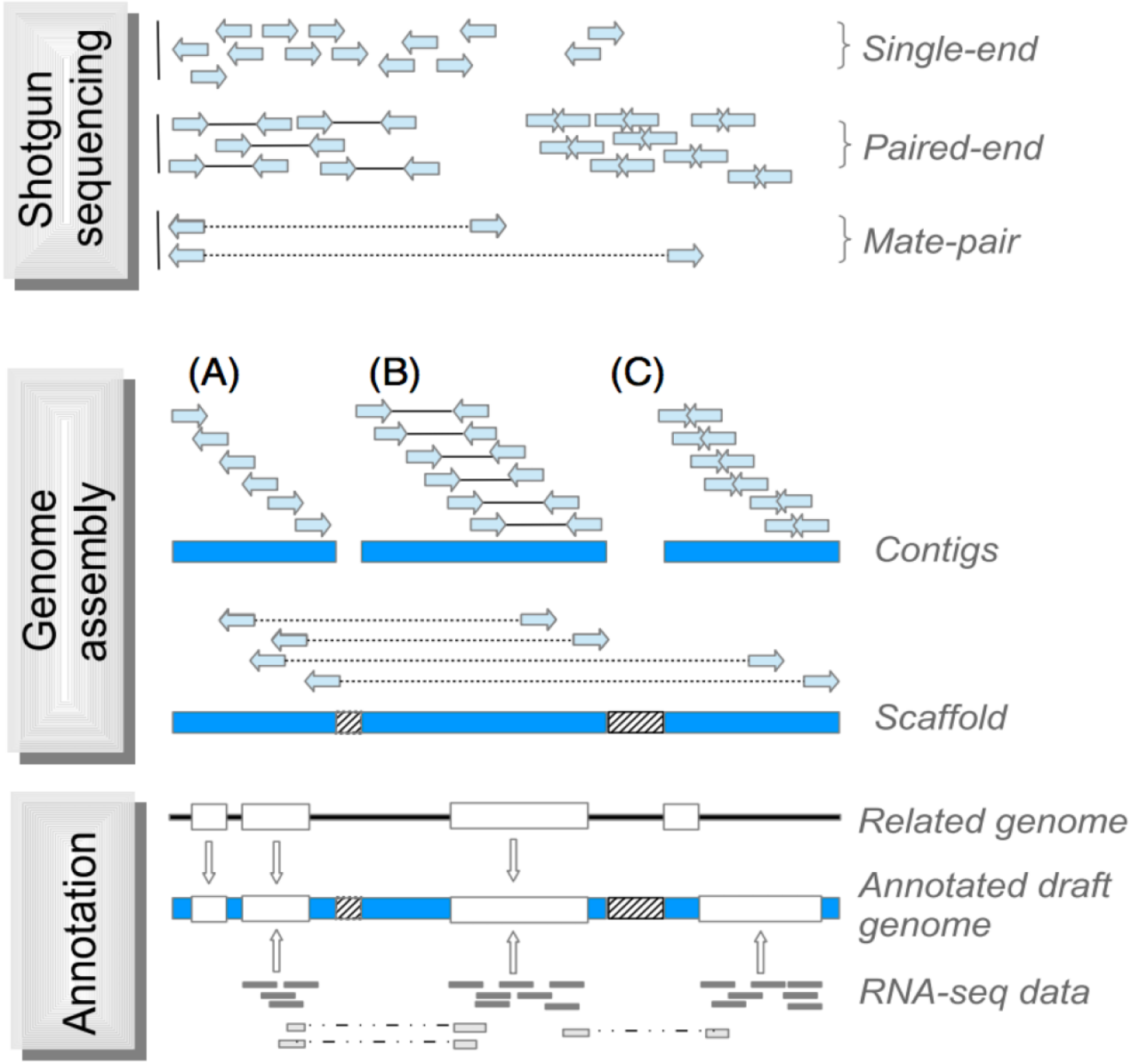
If you order the set of contigs produced by the assembler by size

- N50 is the size of the contig such that 50% of the total bases are in contigs of equal or greater size

E.g.



= 56/2 = 28 => 9Kb N50



Robert Ekblom et al., Evol Appl. 2014 Nov; 7(9): 1026–

Which assembler is best ?

THE ASSEMBLATHON

About

An offshoot of the Genome 10K project, and primarily organized by the UC Davis Genome Center, Assemblathons are contests to assess state-of-the-art methods in the field of genome assembly.

Assemblathon 1 occurred at the end of 2010 and the results were published in late 2011. A second

Background

What is the Assemblathon?

The Assemblathon is a set of periodic collaborative efforts that all help improve methods of genome assembly. It will hopefully become an annual event that will spur improvements in this computationally intensive field. The overall goal of each Assemblathon event is to have participating groups try to use their own software to each assemble one or more genomes that

Which assembler is best ?

CSH PRESS GENOME RESEARCH

Assemblathon 1: A competitive assessment of de novo short read assembly methods

Dent Earl[1,2], Keith Bradnam[3], John St. John[1,2], Aaron Darling[3], Dawei Lin[3,4], Joseph Fass[3,4], Hung On Ken Yu[3], Vince Buffalo[3,4], Daniel R. Zerbino[2], Mark Diekhans[1,2], Ngan Nguyen[1,2], Pramila Nuwantha Ariyaratne[5], Wing-Kin Sung[5,6], Zemin Ning[7], Matthias Haimel[8], Jared T. Simpson[7],

OPEN ACCESS ARTICLE

This Article

Published in Advance September 16, 2011, doi: 10.1101/gr.126599.111

Genome Res. 2011. 21: 2224-2241

Which assembler is best ?

OXFORD ACADEMIC

National Institute of Plant Genome Research ▾ Sign In ▾ Registe

(GIGA)n SCIENCE

BGI华大

Articles Submit ▾ Alerts About ▾

Volume 2, Issue 1
December 2013

Article Contents

Abstract

Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species

Keith R Bradnam ✉, Joseph N Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, Jarrod A Chapman, Guillaume Chapuis, Rayan Chikhi ... Show more

Author Notes

GigaScience, Volume 2, Issue 1, December 2013, 2047-217X-2-10, https://doi.org/10.1186/2047-217X-2-10

Published: 22 July 2013 **Article history** ▾

PDF Split View Cite Permissions Share ▾

PDF Help

Which assembler is best ?

Methodology article | Open Access | Published: 11 September 2019

dnAQET: a framework to compute a consolidated metric for benchmarking quality of de novo assemblies

Gokhan Yavas, Huixiao Hong & Wenming Xiao

BMC Genomics **20**, Article number: 706 (2019) | Cite this article

> F1000Res. 2019 Dec 23;8:2138. doi: 10.12688/f1000research.21782.2. eCollection 2019.

Benchmarking of long-read assemblers for prokaryote whole genome sequencing

Ryan R Wick [1], Kathryn E Holt [1 2]

Affiliations + expand

PMID: 31984131 PMCID: PMC6966772 DOI: 10.12688/f1000research.21782.2

Comparison of long read methods for sequencing and assembly of a plant genome

Valentine Murigneux, Subash Kumar Rai, Agnelo Furtado, Timothy J.C. Bruxner, Wei Tian, Qianyu Ye, Hanmin Wei, Bicheng Yang, Ivon Harliwong, Ellis Anderson, Qing Mao, Radoje Drmanac, Ou Wang, Brock A. Peters, Mengyang Xu, Pei Wu, Bruce Topp, Lachlan J.M. Coin, Robert J. Henry

doi: https://doi.org/10.1101/2020.03.16.992933

A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies

Wenyu Zhang [1], Jiajia Chen, Yang Yang, Yifei Tang, Jing Shang, Bairong Shen

Affiliations + expand

PMID: 21423806 PMCID: PMC3056720 DOI: 10.1371/journal.pone.0017915

MICROBIAL GENOMICS

RESEARCH ARTICLE

De Maio *et al.*, *Microbial Genomics* 2019;5

DOI 10.1099/mgen.0.000294

MICROBIOLOGY SOCIETY

OPEN DATA OPEN MICROBIOLOGY

Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes

Genome annotation pipelines/servers for Prokaryotes

The SEED Viewer

SEED Viewer version 2.0

Welcome to the SEED Viewer - a read-only browser of the curated SEED data.
For more information about The SEED please visit theSEED.org.
For daily updates on SEED activity visit the Daily SEED

»Navigate »Help find

Welcome to the PubSEED.

Type a search string:

Search

MENU SCIENTIFIC REPORTS

Open Access | Published: 10 February 2015

RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes

Thomas Brettin, James J. Davis, Terry Disz, Robert A. Edwards, Svetlana Gerdes, Gary J. Olsen, Robert Olson, Ross Overbeek, Bruce Parrello, Gordon D. Pusch, Maulik Shukla, James A. Thomason III, Rick Stevens, Veronika Vonstein, Alice R. Wattam & Fangfang Xia

Database | Open Access | Published: 08 February 2008

The RAST Server: Rapid Annotations using Subsystems Technology

Ramy K Aziz, Daniela Bartels, [...] Olga Zagnitko

BMC Genomics **9**, Article number: 75 (2008) | Cite this article

102k Accesses | **5766** Citations | **10** Altmetric | Metrics

MG-RAST

metagenomics analysis server

cite us

version 4.0.3

430,846 metagenomes containing 1,703 billion sequences and 244.48 Tbp processed for 32,098 registered users.

for programmatic access visit our API site

search string e.g. mgp128 or mgm4447970.3 search

upload download analyze

Victorian Bioinformatics Consortium

Prokka

Description

Prokka is a software tool for the rapid annotation of prokaryotic genomes. A typical 4 Mbp genome can be fully annotated in less than 10 minutes on a quad-core computer, and scales well to 32 core SMP systems. It produces GFF3, GBK and SQN files that are ready for editing in Sequin and ultimately submitted to Genbank/DDJB/ENA.

Download

Prokka v1.12 — 14 March 2017 — Download (360MB) — MD5 — Changes — Docs — Paper — GitHub

> Bioinformatics. 2014 Jul 15;30(14):2068-9. doi: 10.1093/bioinformatics/btu153. Epub 2014 Mar 18.

Prokka: rapid prokaryotic genome annotation

Torsten Seemann [1]

Affiliations + expand

PMID: 24642063 DOI: 10.1093/bioinformatics/btu153

PGAP pipeline of NCBI, available as tool

NCBI Prokaryotic Genome Annotation Pipeline

The NCBI Prokaryotic Genome Annotation Pipeline (PGAP) is designed to annotate bacterial and archaeal genomes (chromosomes and plasmids).

Genome annotation is a multi-level process that includes prediction of protein-coding genes, as well as other functional genome units such as structural RNAs, tRNAs, small RNAs, pseudogenes, control regions, direct and inverted repeats, insertion sequences, transposons and other mobile elements.

NCBI has developed an automatic prokaryotic genome annotation pipeline that combines *ab initio* gene prediction algorithms with homology based methods. The first version of NCBI Prokaryotic Genome Pipeline was developed in 2001 and is regularly upgraded to improve structural and functional annotation quality (Haft DH et al 2018, Tatusova T et al 2016). Recent improvements utilize curated protein profile hidden Markov models (HMMs), including TIGRFAMS and new HMMs for antimicrobial resistance proteins, and curated complex domain architectures for functional annotation of proteins.

> Nucleic Acids Res. 2016 Aug 19;44(14):6614-24. doi: 10.1093/nar/gkw569. Epub 2016 Jun 24.

NCBI prokaryotic genome annotation pipeline

Tatiana Tatusova [1], Michael DiCuccio [1], Azat Badretdin [1], Vyacheslav Chetvernin [1], Eric P Nawrocki [1], Leonid Zaslavsky [1], Alexandre Lomsadze [2], Kim D Pruitt [1], Mark Borodovsky [3], James Ostell [1]

Affiliations + expand

PMID: 27342282 PMCID: PMC5001611 DOI: 10.1093/nar/gkw569

JGI-IMG

JGI IMG/M INTEGRATED MICROBIAL GENOMES & MICROBIOMES

JGI HOME CONTACT US LOGIN / SIGN-ON

Home IMG/M Find Genomes Find Genes Find Functions Compare Genomes OMICS My IMG Help

Due to the emerging COVID-19 pandemic, JGI will not be accepting or processing any physical samples because of reduced onsite staffing until further notice.
IMG is still accepting bioinformatic data sets via IMG Submission Site

We need your feedback IMG survey IMG Webinar YouTube Playlist

Integrated Microbial Genomes and Microbiomes

The **mission** of the Integrated Microbial Genomes & Microbiomes(IMG/M) system is to support the annotation, analysis and distribution of microbial genome and microbiome datasets sequenced at DOE's Joint Genome Institute (JGI).

IMG/M is also open to scientists worldwide for the annotation, analysis, and distribution of their own genome and microbiome datasets, as long as they agree with the IMG/M data release policy and follow the metadata requirements for integrating data into IMG/M (see IMG/M submission site).

If you use IMG web resources or data to assist in research publications or proposals. Please cite:
IMG - Chen et al., 2019 (PMID: 30289528).
GOLD v7 - Mukherjee et al., 2018 (NAR doi: 10.1093/nar/gky977).

Review Data Usage Policy **Submit Your Data**

Announcements

NEW Metagenome Bin Search IMG has developed a new Metagenome Bin Search functionality. This new feature is available in IMG/M and IMG/M ER.

Read more...

IMG Content

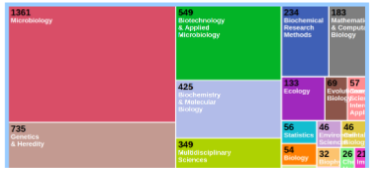
Genes
60,857,847,421

Bases
18,596,713,944,242

Scaffolds
55,296,944,330

IMG Statistics

IMG & GOLD Citations



Genome annotation pipelines/servers for Eukaryotes

MAKER-2

Yandell Lab

Department of Human Genetics - University of Utah

Home | People | Research | Software | Publications | About | Links | Utah | Contact | Internal

MAKER

Last Software Update

v3.01.03 (April 7, 2020)

Overview

MAKER is a portable and easily configurable genome annotation pipeline. Its purpose is to allow smaller eukaryotic and prokaryotic genome projects to independently annotate their genomes and to create genome databases. MAKER identifies repeats, aligns ESTs and proteins to a genome, produces ab-initio gene predictions and automatically synthesizes these data into gene annotations having evidence-based quality values. MAKER is also easily trainable: outputs of preliminary runs can be used to automatically retrain its gene prediction algorithm, producing higher quality gene-models on seusequent runs. MAKER's inputs are minimal and its ouputs can be directly loaded into a GMOD database. They can also be viewed in the Apollo genome browser; this feature of MAKER provides an easy means to annotate, view and edit individual contigs and BACs without the overhead of a database. MAKER should prove especially useful for emerging model organism projects with minimal bioinformatics expertise and computer resources.

MAKER-P

Yandell Lab

Department of Human Genetics - University of Utah

Home | People | Research | Software | Publications | About | Links | Utah | Contact | Internal

MAKER-P

Overview

Sequencing diverse plant species of evolutionary, agricultural, and medicinal interest is becoming routine for even small groups - genome annotation and analysis is much less so. The MAKER-P pipeline is designed to make the annotation of novel plant genomes tractable for small groups with limited bioinformatics experience and resources, and faster and more transparent for large groups with more experience and resources. The MAKER-P pipeline generates species-specific repeat libraries, as well as structural annotations of protein coding genes, non-coding RNAs, and pseudogenes.

MAKER-P consists of the main engine (standard MAKER) together with a number of accessory script and protocols that are downloaded separately.

- MAKER (MAKER versions 2.29+ incorporate a scalable parallelization scheme for large genomes and for deployment within the iPlant Cyberinfrastructure at TACC. NSF IOS-1126998, Developing an effective, portable annotation engine for plant genomes funds its development).

BRAKER1

BRAKER1

BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS

Download from:

- BRAKER1
- AUGUSTUS
- GeneMark-ES/ET

Katharina J. Hoff, Simone Lange, Alexandre Lomsadze, Mark Borodovsky and Mario Stanke
"BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS"
Bioinformatics, 2015, Nov 11
PubMed | Article

NOTE: We have to correct one important reference in the BRAKER1 publication.
In computations of the gene prediction accuracy for the D. melanogaster genome we used the r6.07 version of the fly genome and annotation. However, the Supplementary materials to the paper (available at the "Bioinformatics" journal website) incorrectly cite the earlier r5.55 version of the D. melanogaster genome.

BLAST2GO

- For all type of genomes.
- InterPro, KEGG pathways and GO etc. integrated.

KASS server for Pathways

KAAS - KEGG Automatic Annotation Server

for ortholog assignment and pathway mapping

Request **Help**

About KAAS

KAAS (KEGG Automatic Annotation Server) provides functional annotation of genes by BLAST or GHOST comparisons against the manually curated KEGG GENES database. The result contains KO (KEGG Orthology) assignments and automatically generated KEGG pathways.

- KAAS Help

Complete or Draft Genome

KAAS works best when a complete set of genes in a genome is known. Prepare query amino acid sequences and use the BBH (bi-directional best hit) method to assign orthologs.

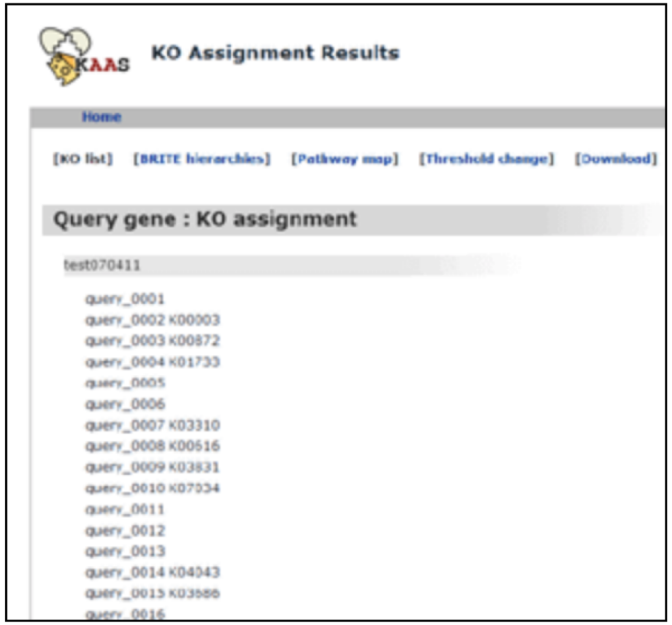
- KAAS job request (BBH method)

Partial Genome

KAAS can also be used for a limited number of genes. Prepare query amino acid sequences and use the SBH (single-directional best hit) method to assign orthologs.

Example of Results

KO assignment



KEGG pathway mapping

Tools

Research Article | Open Access
Volume 2019 | Article ID 4767354 | 12 pages |
https://doi.org/10.1155/2019/4767354

GAAP: A Genome Assembly + Annotation Pipeline

Jinhwa Kong,[1] Sun Huh,[2] **Jung-Im Won**,[3] **Jeehee Yoon**,[4] Baeksop Kim,[4] and Kiyong Kim[5]

Gene Prediction pp 29-51 | Cite as

Structural and Functional Annotation of Eukaryotic Genomes with GenSAS

Authors | Authors and affiliations

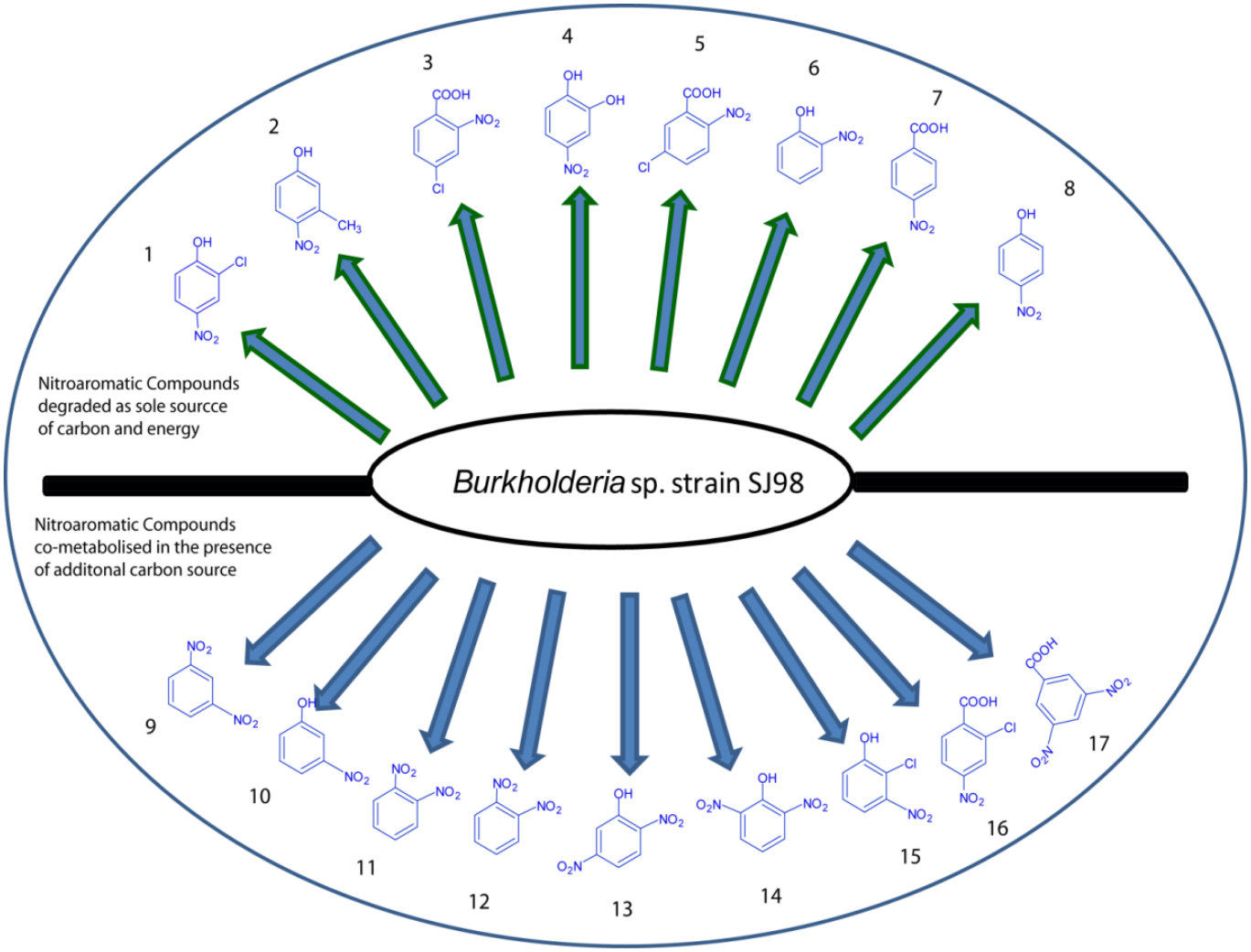
Jodi L. Humann, Taein Lee, Stephen Ficklin, Dorrie Main

FILE FORMATS

Common file formats

- FASTA — Nucleotide sequence (file extension .fas or .fa)
- FASTQ — Nucleotide sequence including quality scores
- SAM — Sequence alignment
- BAM — Binary version of SAM
- GFF3 — Annotation
- GTF — Annotation
- BED — Annotation
- VCF — Variant calling

Burkholderia sp. SJ98



(Bhushan, et al., 2000, Samanta, et al., 2000; Pandey et al., 2012)

Microbial Genome assembly and annotation

Burkholderia sp. SJ98 Genome

Roche's 454

Reads

Genome assembler (e.g. Newbler)

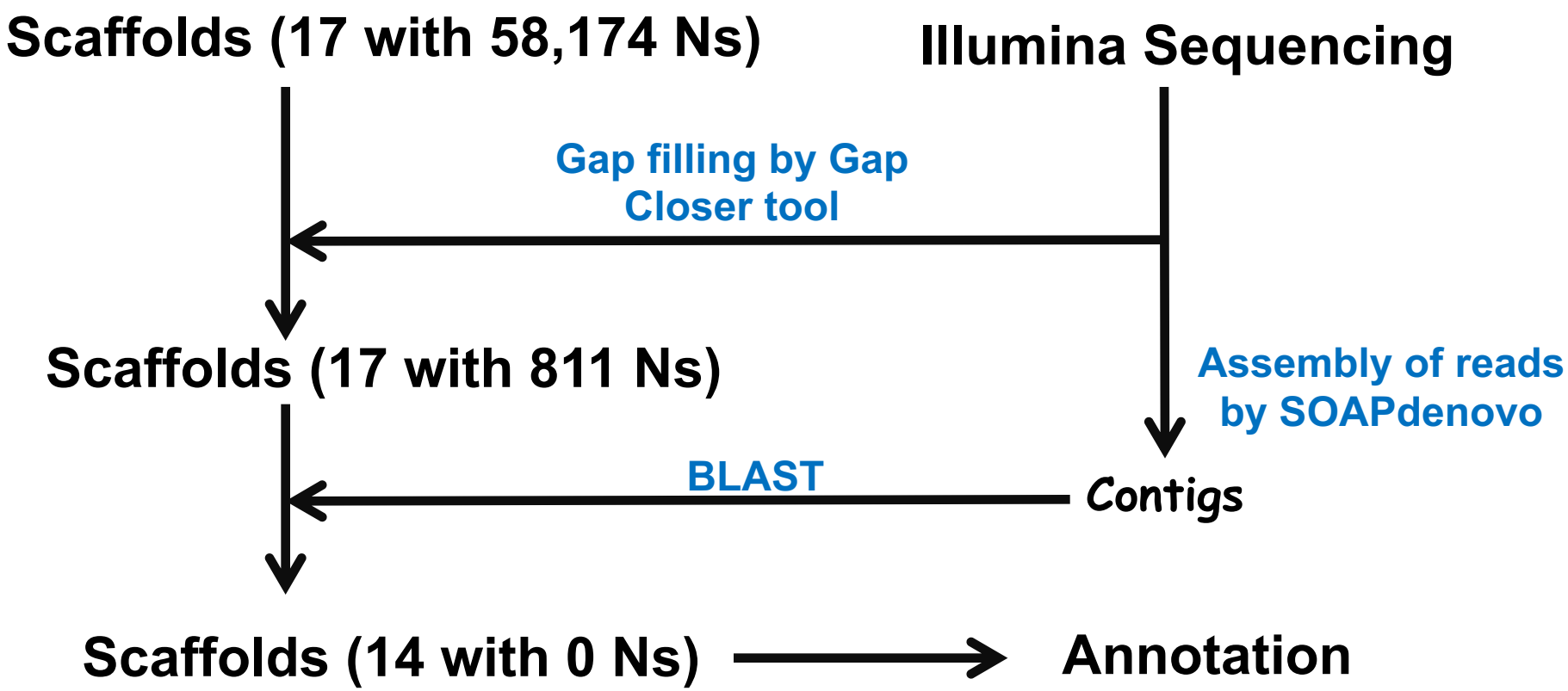
Contigs (79)

Scaffolds (17)

Annotation servers and manual annotation.

RAST, PGAAP, KAAS etc..

Annotated genome



Genome assembly	Sequences	Size (bp)	N 50	Ns	GC (%)
Assembly-1*	17	7,894,128	1,315,287	58,174	62.23
Assembly-2**	17	7,884,563	1,314,594	811	62.68
Assembly-3***	14	7,878,727	1,314,594	0	62.68

*Scaffolds produced by assembly of Roche's 454 FLX data.
**Sequences (16 contigs and 1 scaffold) produced after gap filling of Assembly-1 by Illumina GA IIX data.
***Contigs produced after the finishing of Assembly-3 (Sanger's sequencing and manually by BLAST), final assembly.
doi:10.1371/journal.pone.0070624.t001

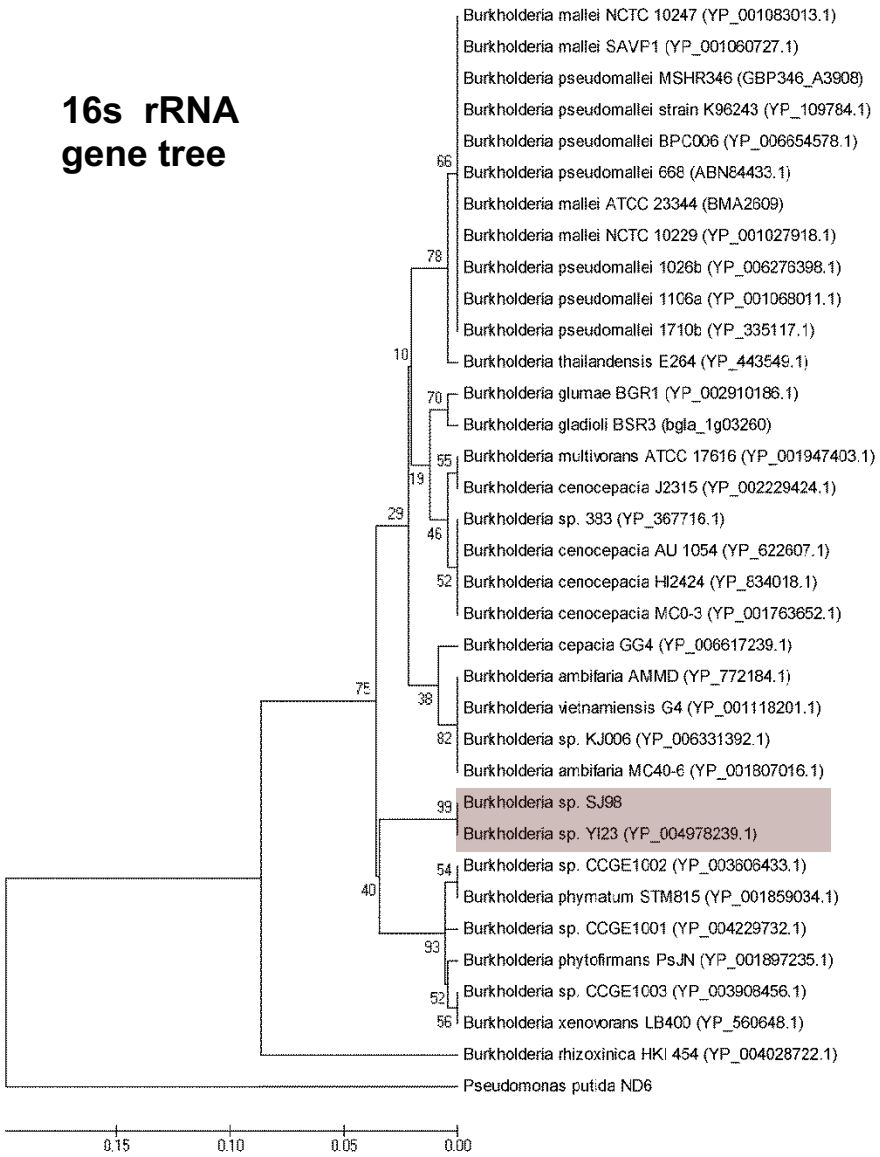
Burkholderia sp. SJ98

- Sequenced by Illumina and Roche's 454 technologies.
- Assembly by Hybrid approach with SOAPdenovo, Gapcloser and Newbler software packages.
- Some gaps were filled by Sanger's sequencing.
- Genome annotation by PGAAP pipeline and RAST.
- Phylogenomics on the basis of rpoB gene.
- Identification of chemotaxis genes.
- Comparison of chemotaxis gene clusters with related strains.

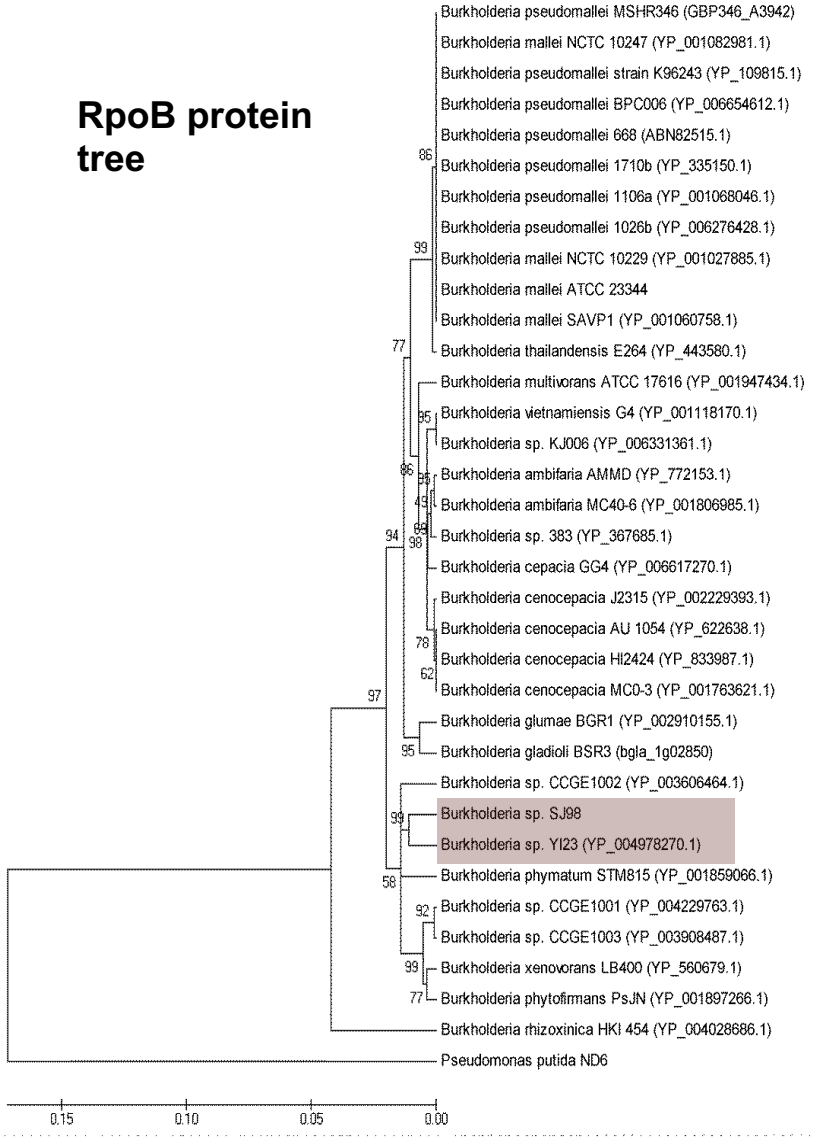
(Kumar et al., J. Bacteriol. June 2012; Kumar et al., PLoS ONE 2013)

Phylogenomics of *Burkholderia* sp. SJ98

16s rRNA gene tree



RpoB protein tree



Genome characterization of *Burkholderia* sp. SJ98 and compared strains.

Characteristics	*Burkholderia sp.* SJ98	*Burkholderia sp.* YI23	*Burkholderia sp.* CCGE 1001	*Burkholderia sp.* CCGE 1002	*Burkholderia sp.* CCGE 1003
Length (bp)	7,878,727	8,896,411	6,833,751	7,884,858	7,043,595
GC content	62.68%	63.26%	63.63%	63.27%	63.25%
No. of protein coding genes	7,268	7,804	5,965	6,889	5,998
No. of tRNA genes	52	64	62	73	63

(Kumar, et al., PLoS ONE 2013)

Order and orientation of contigs – more errors in one assembly than in another

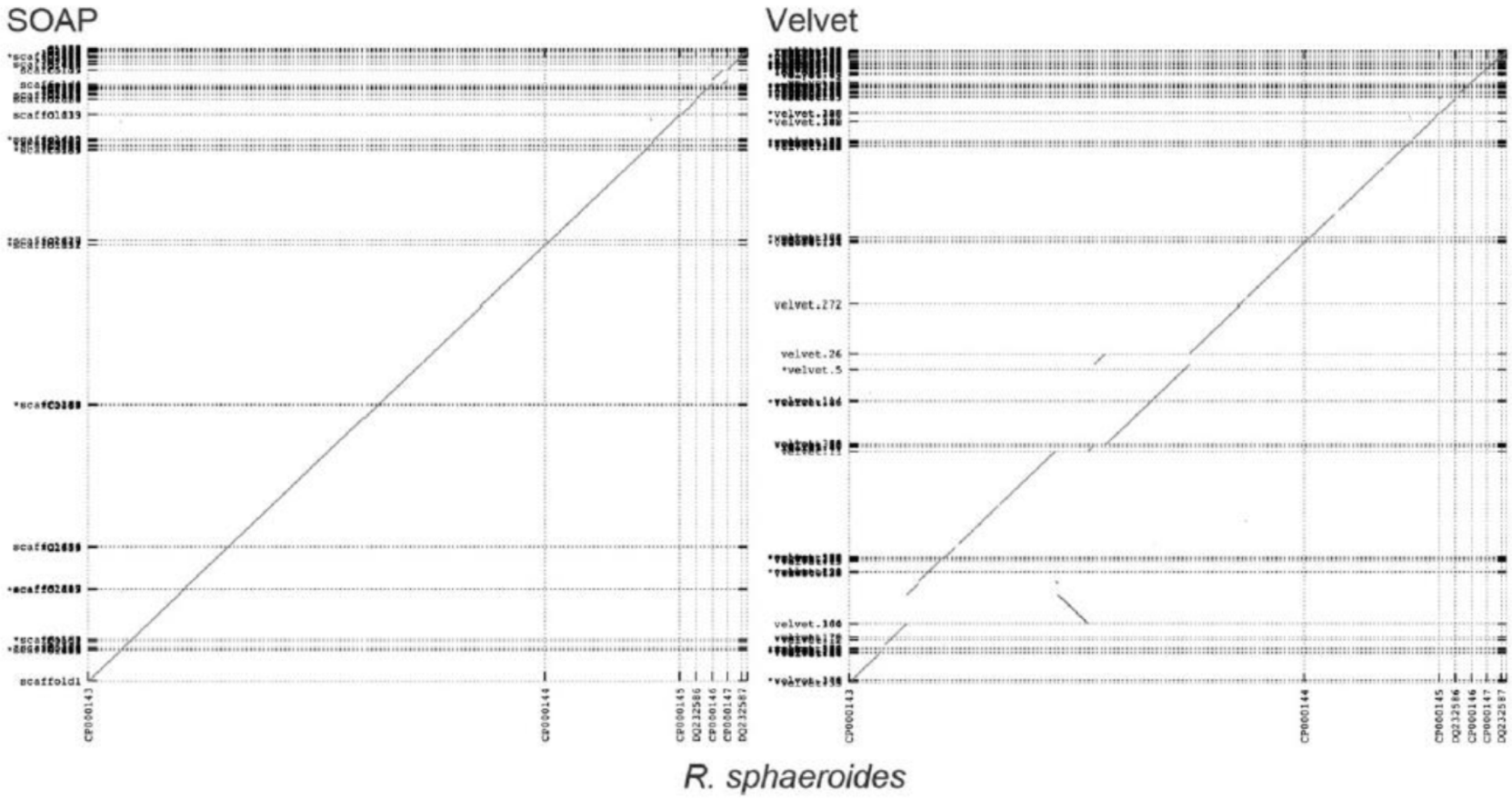


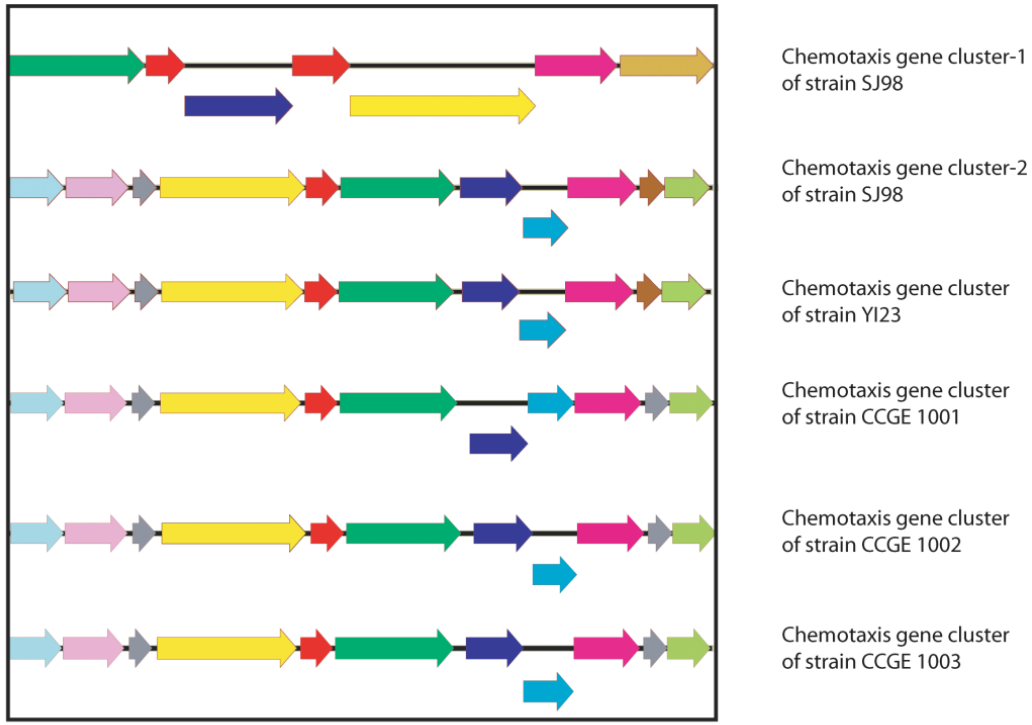
Figure 2. A dot-plot comparison of the SOAPdenovo and Velvet scaffolds of *R. sphaeroides*. The finished reference chromosomes are plotted on the *x*-axis and the assembly scaffolds on the *y*-axis. Dotted lines indicate scaffold or chromosome boundaries. The apparent rearrangement at the *top right* of the SOAPdenovo plot is an artifact of the circular reference plasmid.

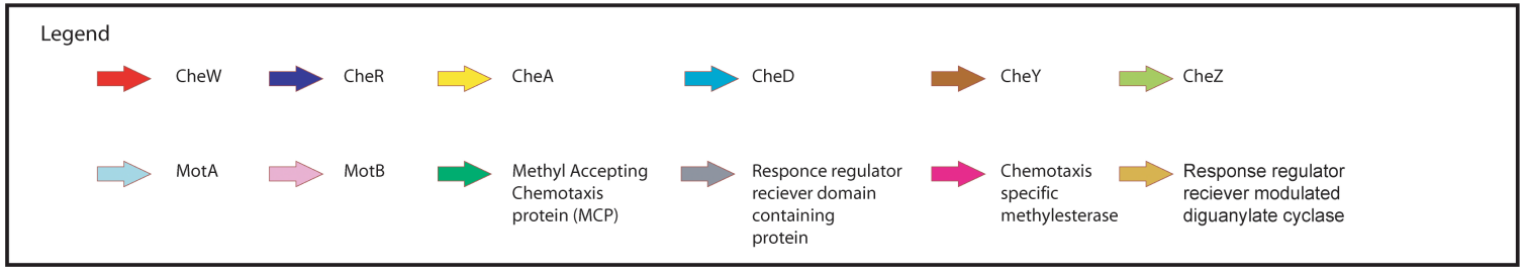
Number of chemotaxis genes in different species

Gene	Species					
	E.coli	*Burkholderia sp.* SJ98	*Burkholderia sp.* YI23	*Burkholderia sp.* CCGE 1001	*Burkholderia sp.* CCGE 1002	*Burkholderia sp.* CCGE 1003
CheA	1	2	2	2	2	3
CheB	1	4	2	4	3	3
CheC	0	1	0	2	1	2
CheR	1	3	3	2	2	3
CheW	1	5	4	2	2	3
CheY	1	2	1	0	0	0
CheZ	1	1	1	1	2	1
MCPs	4	19	12	22	21	32
Total	**10**	**37**	**25**	**35**	**33**	**47**

(Kumar, et al., PLoS ONE 2013)

Comparison of chemotaxis gene clusters in *Burkholderia* strains





(*Kumar, et al.*, PLoS ONE 2013)

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Bacteriology

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

Shailesh Kumar, Surendra Vikram and Gajendra Pal Singh Raghava

OPEN ACCESS Freely available online

PLOS | ONE

Gen
Spe

12 other bacterial genome drafts (Published).....

Shailesh Kumar, Surendra Vikram, Gajendra Pal Singh Raghava*

Bioinformatics Centre, Council of Scientific and Industrial Research - Institute of Microbial Technology, Sector 39-A, Chandigarh, India

OPEN ACCESS Freely available online

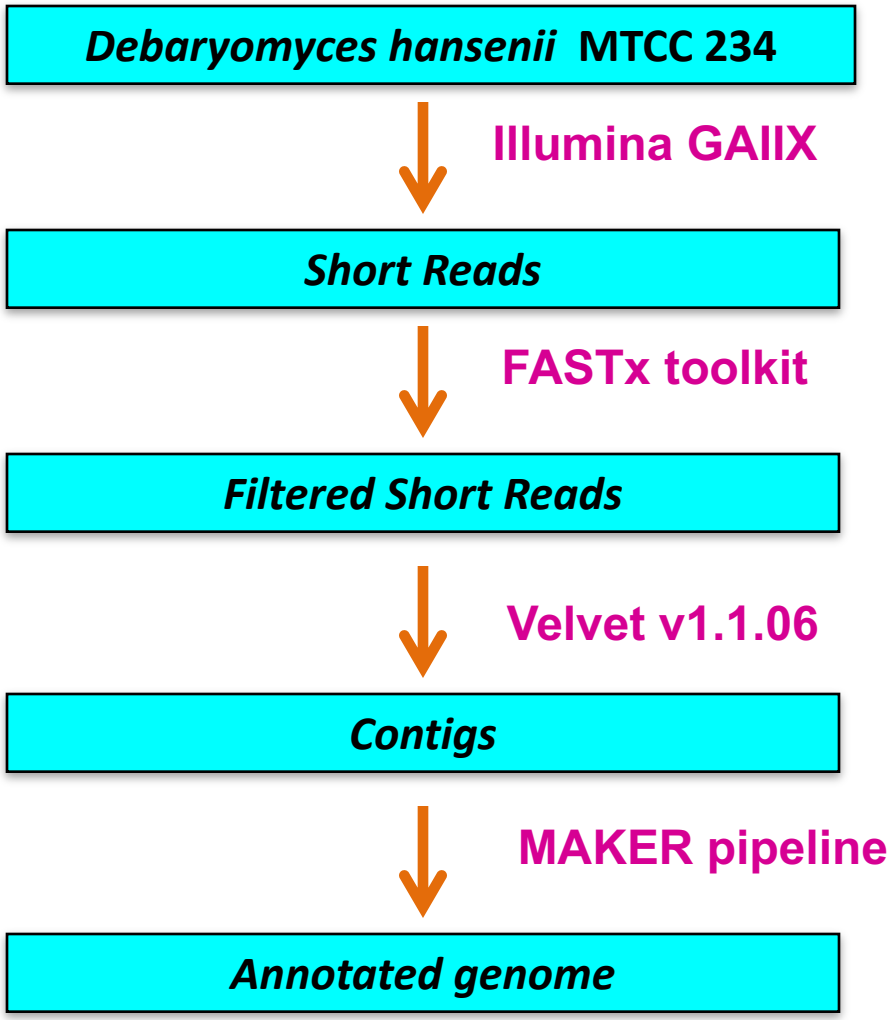
PLOS | ONE

Genes Involved in Degradation of *para*-Nitrophenol Are Differentially Arranged in Form of Non-Contiguous Gene Clusters in *Burkholderia* sp. strain SJ98

Surendra Vikram[1], Janmejay Pandey[2¤], Shailesh Kumar[1], Gajendra Pal Singh Raghava[1*]

1 Bioinformatics Center, CSIR-Institute of Microbial Technology, Chandigarh, India, **2** Microbial Type Culture Collection Center, CSIR-Institute of Microbial Technology, Chandigarh, India

Debaryomyces hansenii MTCC234

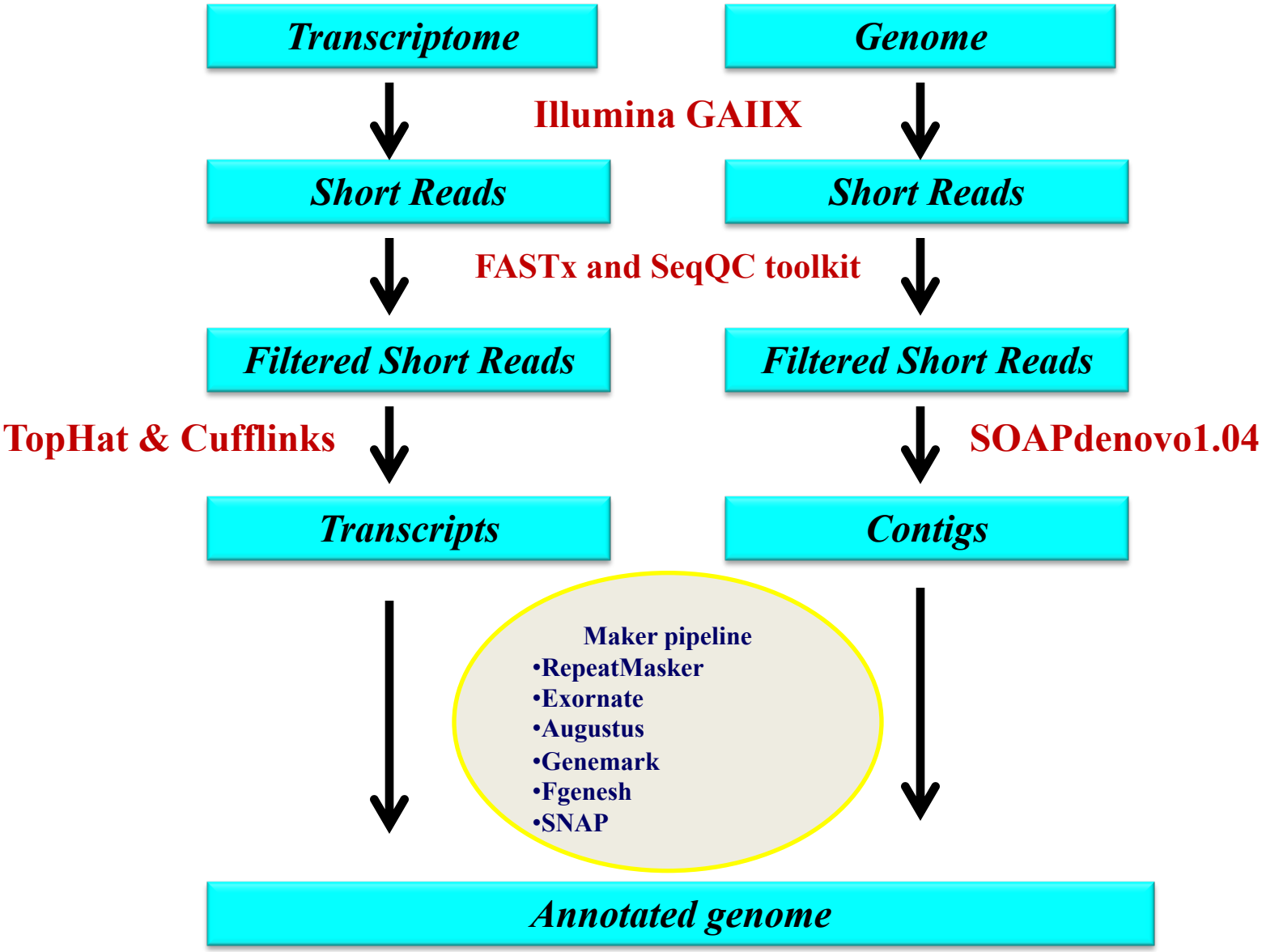


Genome size	11.46-Mb
Contigs produced	542
Protein coding genes	5,294
rRNAs	3
tRNAs	69

- **Of these, 5,069 proteins could be mapped to the UniProt database.**
- **Genes for riboflavin metabolism and, pentose and glucuronate inter conversion pathway have been found.**

Kumar et al., Eukaryot Cell. 2012

Rhodosporidium toruloides MTCC 457



Kumar et al., Eukaryot Cell. 2012

AMERICAN SOCIETY FOR MICROBIOLOGY | Eukaryotic Cell Published 2002-2015

Genome Sequence of the Oleaginous Red Yeast *Rhodosporidium toruloides* MTCC 457

Shailesh Kumar[a], Hariom Kushwaha[a], Anand Kumar Bachhawat[a,b], Gajendra Pal Singh Raghava[a] and Kaliannan Ganesan[a]

Draft Genome Sequence of Salt-Tolerant Yeast *Debaryomyces hansenii* var. *hansenii* MTCC 234

Shailesh Kumar, Anmoldeep Randhawa, Kaliannan Ganesan, Gajendra Pal Singh Raghava and Alok K. Mondal

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

Burkholderia sp. SJ98 database

Home | Team | Contact us | IMTECH | CRDD

0

About

- Introduction
- Sequencing
- Assembly
- Annotation method

Annotation

- Pathways
- Phylogenomics
- Comparison
- Chemotaxis
- MCPs
- Gene clusters
- Annotation data
- Genome Browser

BLAST

- Contigs
- Genes
- Proteins

Links

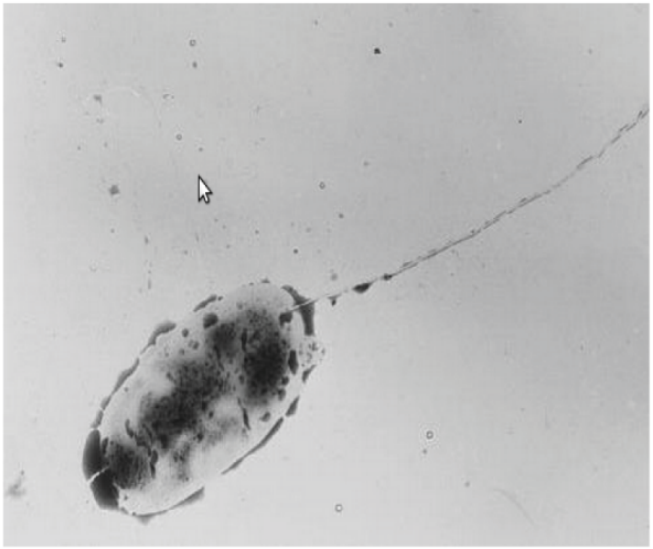
Introduction

Burkholderia strain sp. SJ98 is a gram -ve bacterium, isolated from a pesticide contaminated soil sample from Assam agricultural fields, India by using an enrichment technique developed by Samanta et al. (2000). This strain is known to degrade a variety of nitroaromatic compounds including p-nitrophenol, 2-chloro-4-nitrophenol (2C4NP), 4-chloro-2-nitrobenzoate (4C2NB), 5-chloro-2-nitrobenzoate (5C2NB) and transform 2-chloro-3-nitrophenol (2C3NP) and 2-chloro-4-nitrobenzoate (2C4NB).

We have carried out whole genome sequencing, assembly and annotation of this strain and studied the genes involved in biogeradation of nitroaromatic compounds.

Chemotaxis (*Che*) genes and Methyl accepting chemotaxis proteins (i.e. MCPs) have been studied in the genome of *Burkholderia* strain sp. SJ98.

All genome assembly and annotation data of *Burkholderia* sp. SJ98 is available at this platform.



A Transmission electron microscopy (TEM) image of *Burkholderia* sp. SJ98.

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

Genomics web portal

HOME CSIR IMTECH Developers Contact

Genomics at BIC (IMTECH)

- Genome sequencing
- Genome assembly
- Genome annotation

Prokaryotes

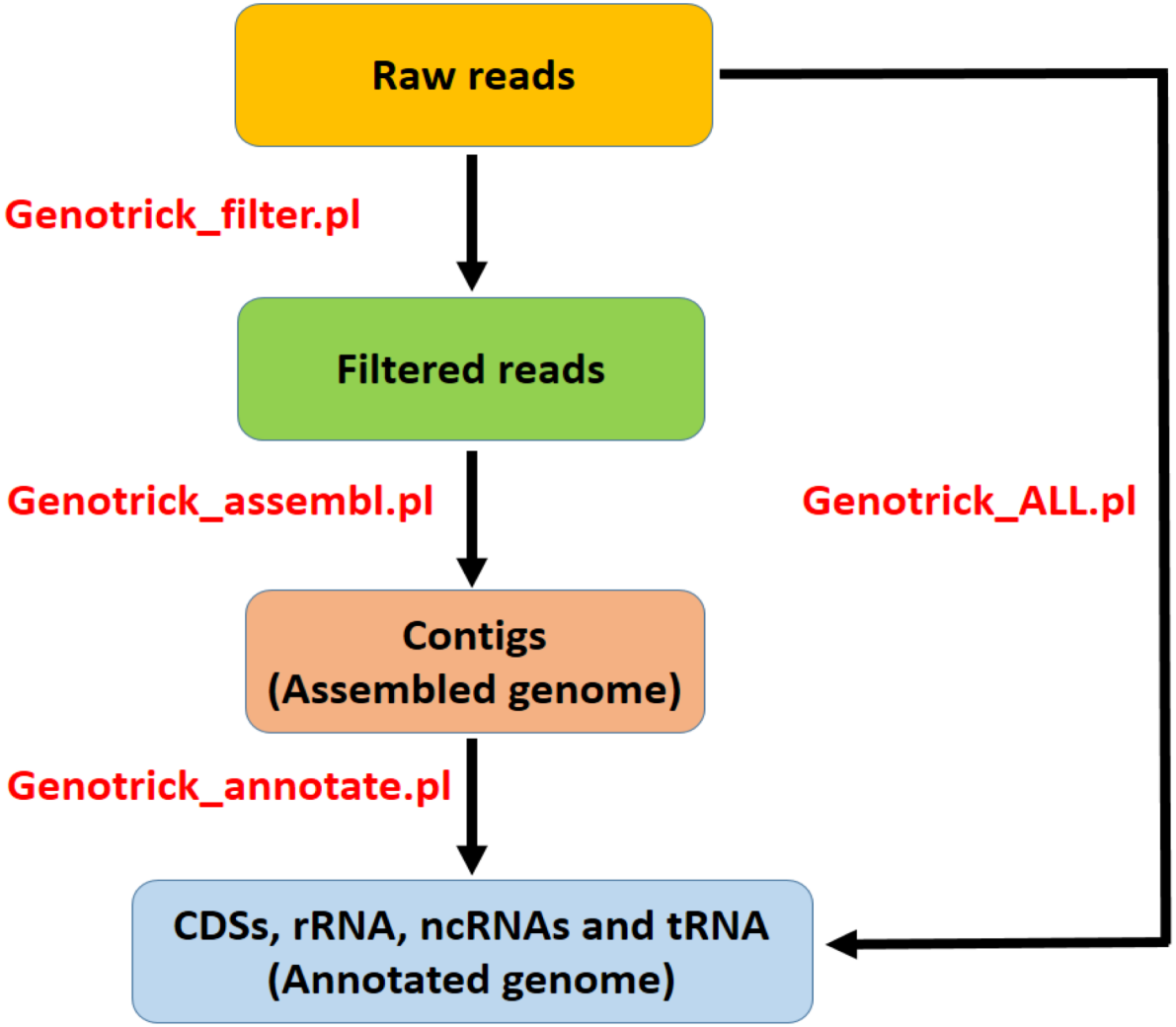
- Actinoalloteichus spitiensis RMV-1378T
- Burkholderia sp. SJ 98
- Rhodococcus rhodochrous BKS6-46
- Imtechella halotolerans K1T
- Rhodococcus imtechensis sp. RKJ300
- Marinilabilia salmonicolor JCM 21150
- Citrobacter freundii MTCC 1658T
- Arthrobacter sp. SJCon
- Rhodococcus qingshengii strain BKS 20-40
- Rhodococcus triatomae BKS 15-14
- Acinetobacter baumannii MSP4-16
- Amycolatopsis decaplanina DSM 44594T
- Rhodococcus ruber strain BKS 20-38
- Streptomyces gancidicus strain BKS 13-15
- Citrobacter freundii strain GTC 09479

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-1378T
2. Rhodococcus rhodochrous BKS6-46
3. Burkholderia sp. S J 98
4. Imtechella halotolerans K1T
5. Marinilabilia salmonicolor JCM 21150
6. Rhodococcus imtechensis sp. RKJ300
7. Debaryomyces hansenii MTCC 345
8. Rhodosporodium toruloides MTCC 457
9. Citrobacter freundii MTCC 1658T
10. Arthrobacter sp. SJCon
11. Rhodococcus qingshengii BKS 20-40
12. Rhodococcus triatomae BKS 15-14
13. Acinetobacter baumannii MSP4-16
14. Amycolatopsis decaplanina DSM 44594 T
15. Rhodococcus ruber BKS 20-38
16. Streptomyces gancidicus strain BKS 13-15
17. Citrobacter freundii strain GTC 09479 (GTC14897)

Genotrick- A Pipeline for whole genome assembly and annotation



- **All in one pipeline for :-**
  - Filtering of short reads
  - Genome assembly from filtered reads
  - Genome annotation from contigs.
- **Software packages:-**
  - NGSQC toolkit (Patel & Jain, 2012) to filter NGS raw data.
  - Velvet (Zerbino & Birney, 2008) for assembly.
  - PROKKA (Prokka: Prokaryotic Genome Annotation System - http://vicbioinformatics.com/) for annotation.

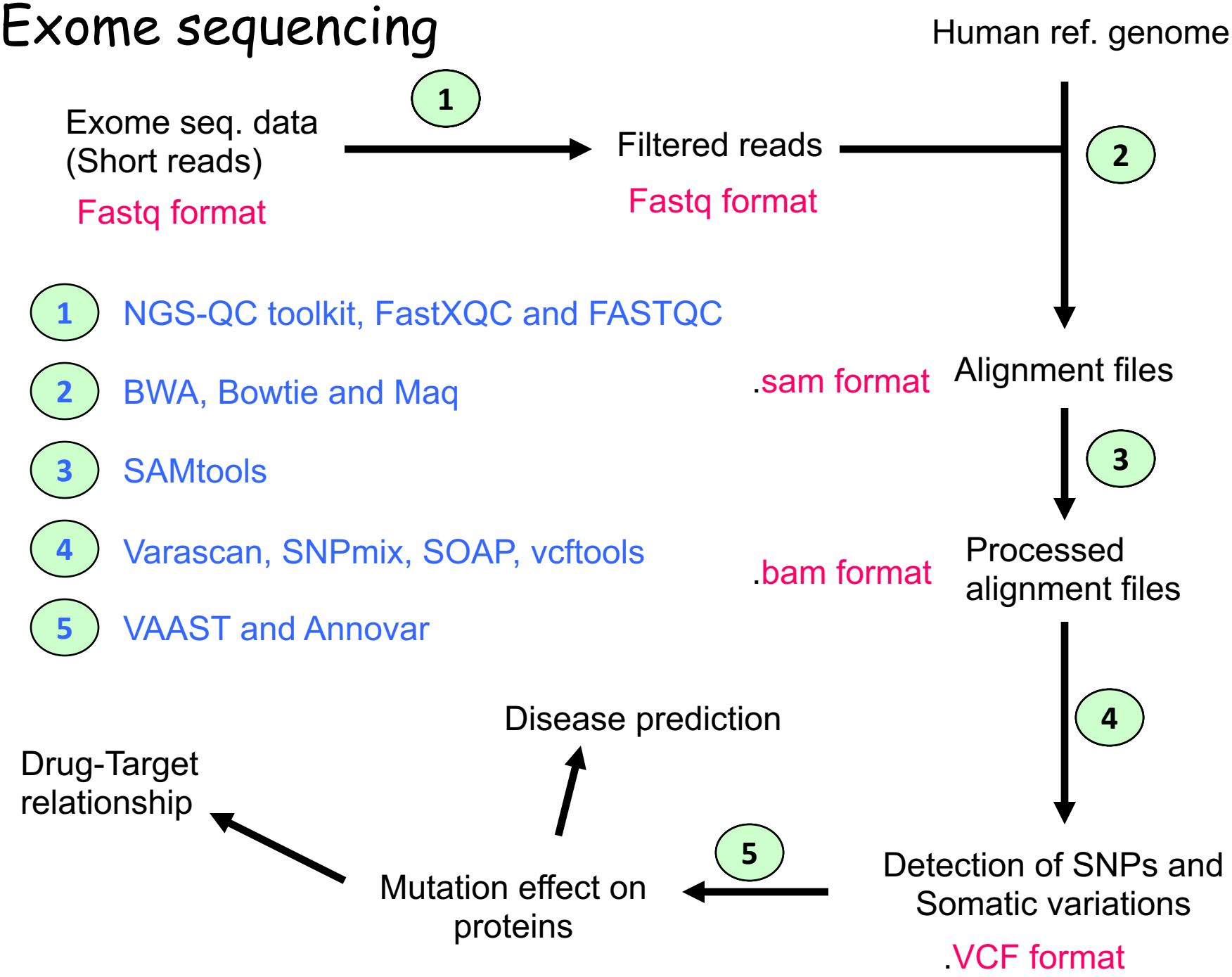
http://crdd.osdd.net/raghava/genomesrs/genotrick/

Genomics data for Human

Applications of NGS for human

1. Whole genome sequencing
2. RNA-seq
3. Exome capture
4. Small RNA sequencing for Non coding RNA study
5. Degradome sequencing
6. CHIPseq-for specific protein binding site; genomewide.
7. Hi-C data analysis for 3D architecture of genome
8. Many more …….

Exome sequencing



1. NGS-QC toolkit, FastXQC and FASTQC
2. BWA, Bowtie and Maq
3. SAMtools
4. Varascan, SNPmix, SOAP, vcftools
5. VAAST and Annovar

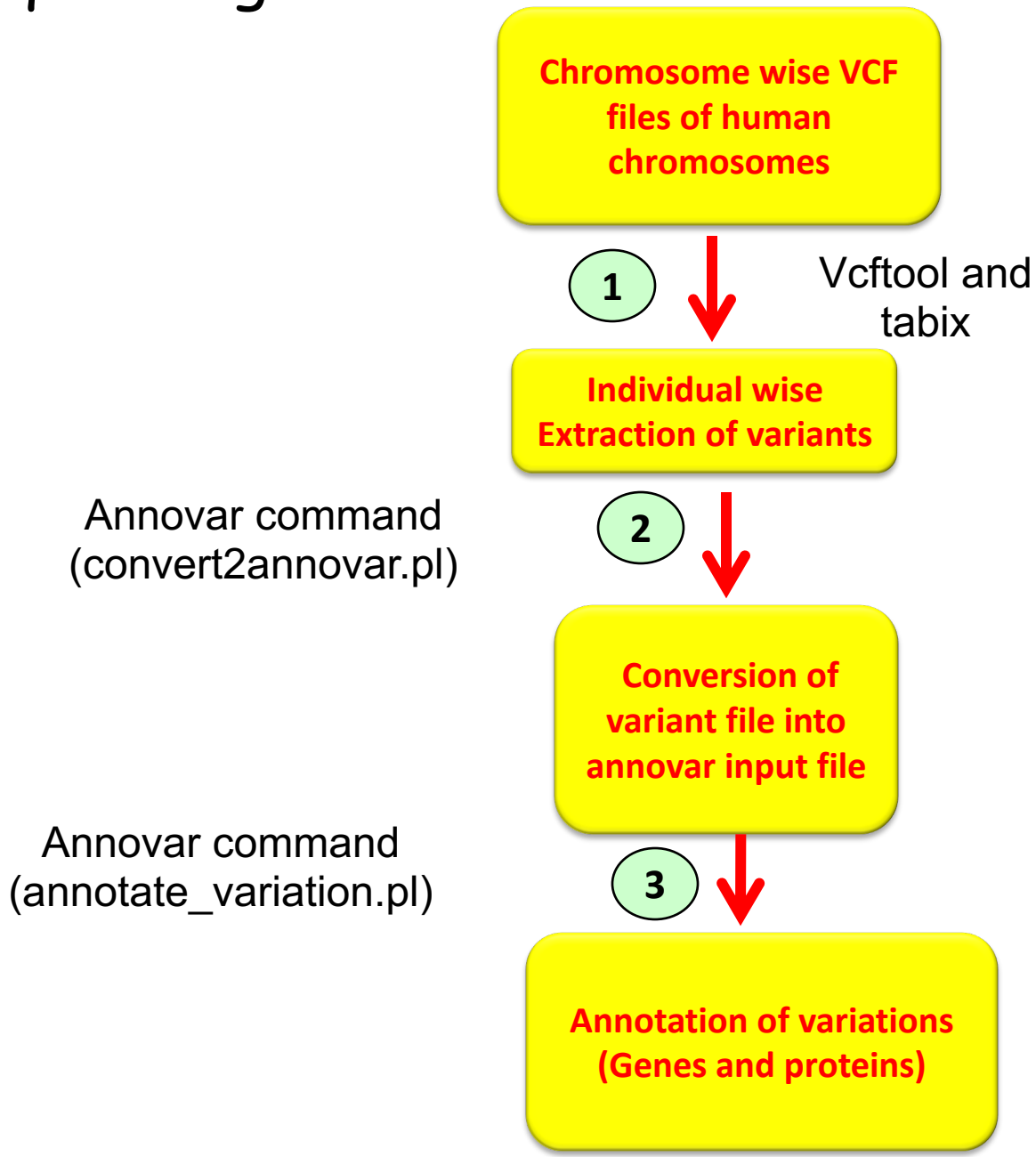
Whole genome sequencing data

1000 Genome Project

Total Samples – 1,182

Processed samples – 1,092

Coding genes having variations – 13,144



PLOS | ONE

RESEARCH ARTICLE

A Web-Based Platform for Designing Vaccines against Existing and Emerging Strains of *Mycobacterium tuberculosis*

Sandeep Kumar Dhanda, Pooja Vir, Deepak Singla, Sudheer Gupta, Shailesh Kumar, Gajendra P. S. Raghava*

PLOS | ONE

RESEARCH ARTICLE

A Platform for Designing Genome-Based Personalized Immunotherapy or Vaccine against Cancer

Sudheer Gupta, Kumardeep Chaudhary, Sandeep Kumar Dhanda, Rahul Kumar, Shailesh Kumar, Manika Sehgal, Gandharva Nagpal, Gajendra P. S. Raghava*

Identifying fusion transcripts using Next Generation Sequencing

What are Fusion transcripts ?

a) Fusion of two transcripts, may be coding of non coding.

a) Traditionally, detected in various tumors and stabilized as biomarkers
   - BCR-ABL: Chronic myelogenous leukemia
   - TMPRSS2-ERG : Prostate cancer
   - EML4-ALK : Lung cancer

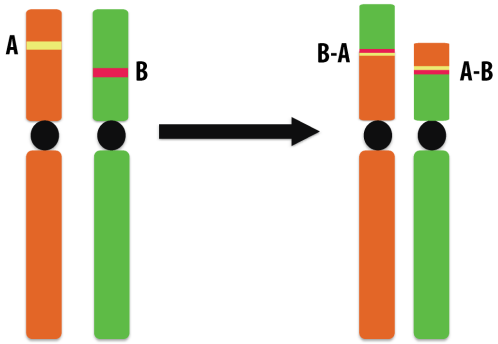
b) Fusion transcripts have also been found in non-neoplastic tissues too (Qin, F. et al. 2015).

a) In other organisms also i.e. Mouse and Fruit fly (Frenkel-Morgenstern M et al. 2013).

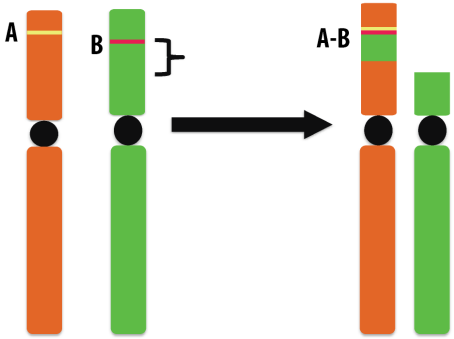
Fusion transcripts formation at DNA level

Balanced rearrangements

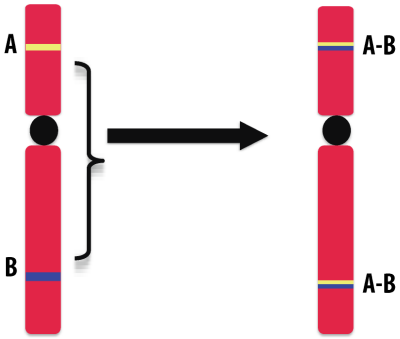
Translocation



Insertion

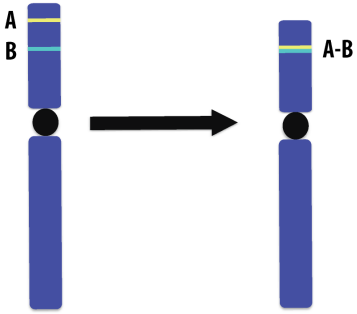


Inversion



Unbalanced rearrangement

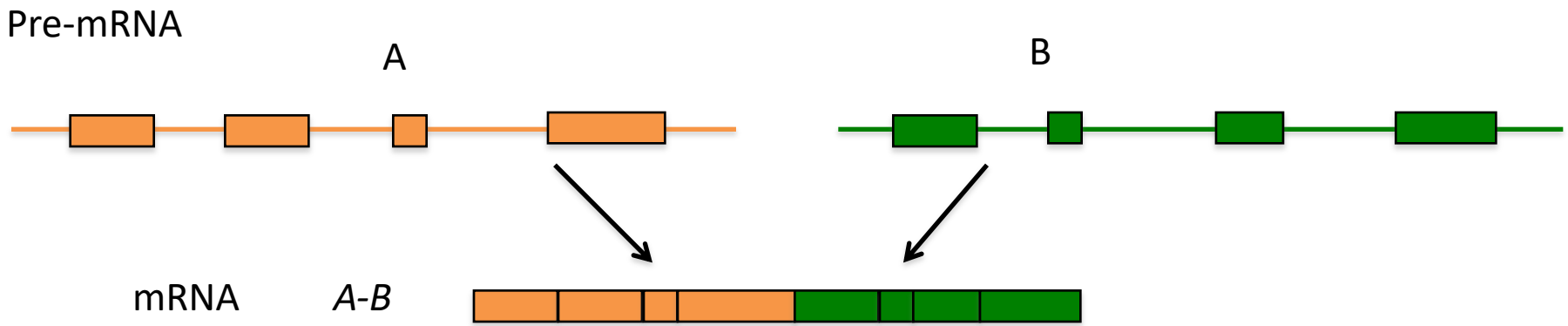
Deletion



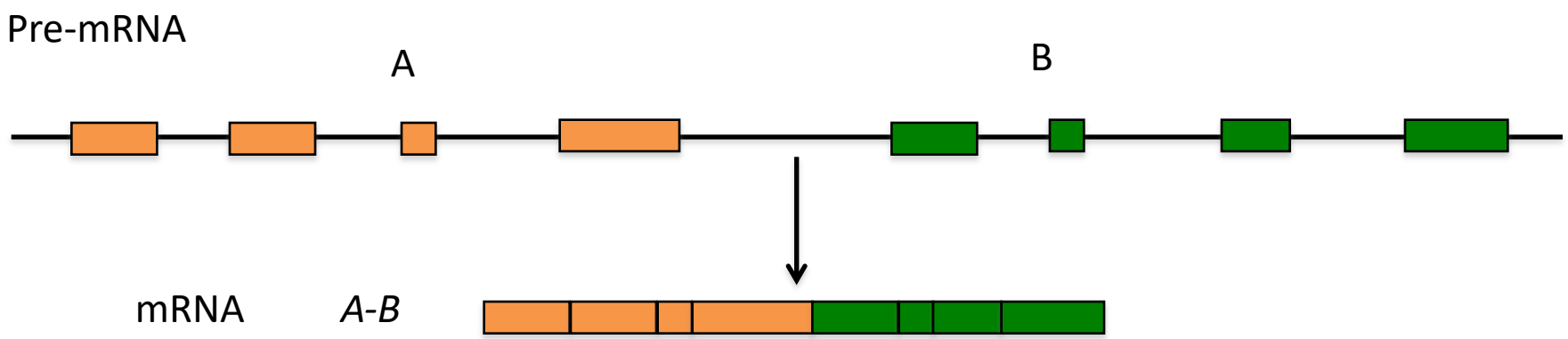
Wiley Interdisciplinary Reviews: RNA **7** (6), 811-823

RNA level: Trans-Splicing vs. Cis-Splicing

a) Trans-Splicing of different transcripts



b) Cis-Splicing of neighboring genes



Modification of Genome Research **22** (7): 1231–1242.

Fusion transcripts detection tools

a) Several tools available for RNA-Seq reads

b) Some tools requires both RNA-Seq and Whole Genome Sequencing reads

c) Some consider single-end reads and other require paired-end reads

d) Latest tools are also accepting reads produced by Third generation sequencers i.e. PacBio.

Our questions …

a) Which tool produce maximum true fusion ?
b) Overlap between the results of different tools ?
c) Which is taking less
    1. Computational time ?
    2. Memory (RAM) ?
d) Is there any detection in the data that does not have fusion ?
    1. If yes, then which tool produce minimum false fusions ?
    2. Factors alter the false fusion detection ?

Any previous attempt for this ?

a) In 2013, Carrara et. al. compared only six tools with positive and negative datasets.
b) No time and RAM comparison
c) Latest tools had not compared

Datasets for comparison study

a) Positive dataset (by Fusionmap developers)
   1. Simulated paired-end RNA-Seq reads (~60,000 pairs of reads, 75nt length
   2. Representing 50 fusions
b) Negative dataset (by Carrara et. al. 2013)
   1. Simulated reads of length 100nt, 75nt and 50nt prepared
   2. For each length, we have two sets of different quality scores
c) Mix dataset (Positive + Negative)
   1. Positive dataset mixed with 75nt negative data (70,000,000 paired-end reads)
   2. Represents 50 fusions, embedded in reads, does not have fusion
d) Test data (Our real data)
   1. Data from our previous study (Qin, F. et al. 2015)
   2. 6 RNA-Seq runs of prostate cancer cell line
   3. Two large and four small RNA-Seq data

Publication

SCIENTIFIC REPORTS

OPEN

Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data

Received: 08 October 2015

Shailesh Kumar[1], Angie Duy Vo[1], Fujun Qin[1] & Hui Li[1,2]

TITLE	CITED BY	YEAR
Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data S Kumar, AD Vo, F Qin, H Li Scientific reports 6, 21597	89	2016

Another research

Published online 17 November 2015

Nucleic Acids Research, 2016, Vol. 44, No. 5 ***e47***
doi: 10.1093/nar/gkv1234

Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data

Silvia Liu[1,2,†], Wei-Hsiang Tsai[3,†], Ying Ding[1,2,†], Rui Chen[1], Zhou Fang[1], Zhiguang Huo[1], SungHwan Kim[1], Tianzhou Ma[1], Ting-Yu Chang[4], Nolan Michael Priedigkeit[5], Adrian V. Lee[6], Jianhua Luo[7], Hsei-Wei Wang[3,4,8,*], I-Fang Chung[3,8,*] and George C. Tseng[1,2,*]

[1]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, 130 De Soto Street, Pittsburgh, PA 15261, USA, [2]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Biomedical Science Tower 3, 3501 Fifth Avenue, Pittsburgh, PA 15213, USA, [3]Institute of Biomedical Informatics, National Yang-Ming University, No. 155, Sec. 2, Linong Street, Beitou District, Taipei 112, Taiwan, [4]Institute of Microbiology and Immunology, National Yang-Ming University, No. 155, Sec. 2, Linong Street, Beitou District, Taipei 112, Taiwan, [5]Molecular Pharmacology, School of Medicine, University of Pittsburgh, 3550 Terrace Street, Pittsburgh, PA 15261, USA, [6]Magee-Women's Research Institute, 204 Craft Avenue, Pittsburgh, PA 15213, USA, [7]Department of Pathology, School of Medicine, University of Pittsburgh, 3550 Terrace Street, Pittsburgh, PA 15261, USA and [8]Center for Systems and Synthetic Biology, National Yang-Ming University, No. 155, Sec. 2, Linong Street, Beitou District, Taipei 112, Taiwan

Our Review

- ✓ Fusion genes
  - Identification
  - Importance
- ✓ Fusion finders
  - Mechanism
  - Comparison
  - Future
- ✓ Benchmarking studies
  - Pros and cons
  - Importance
- ✓ Future directions
  - Tool Development
  - Data types
  - Further comparisons

WIREs RNA

WILEY

Explore this journal >

Advanced Review

Identifying fusion transcripts using next generation sequencing

Shailesh Kumar, Sundus Khalid Razzaq, Angie Duy Vo, Mamta Gautam, Hui Li

Kumar, S., Razzaq, S. K., Vo, A. D., Gautam, M. and Li, H. (2016), Identifying fusion transcripts using next generation sequencing. WIREs RNA. doi:10.1002/wrna.1382. PMID: 27485475

Volume 7, Number 6, November/December 2016

WIREs RNA

wires.wiley.com/rna

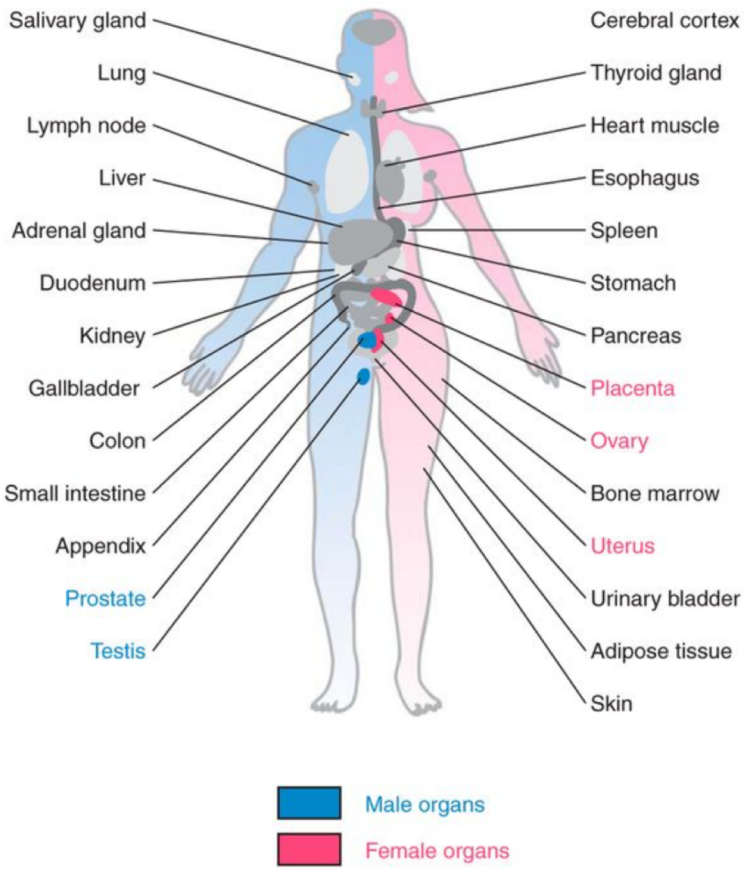
WILEY

Cover Image…

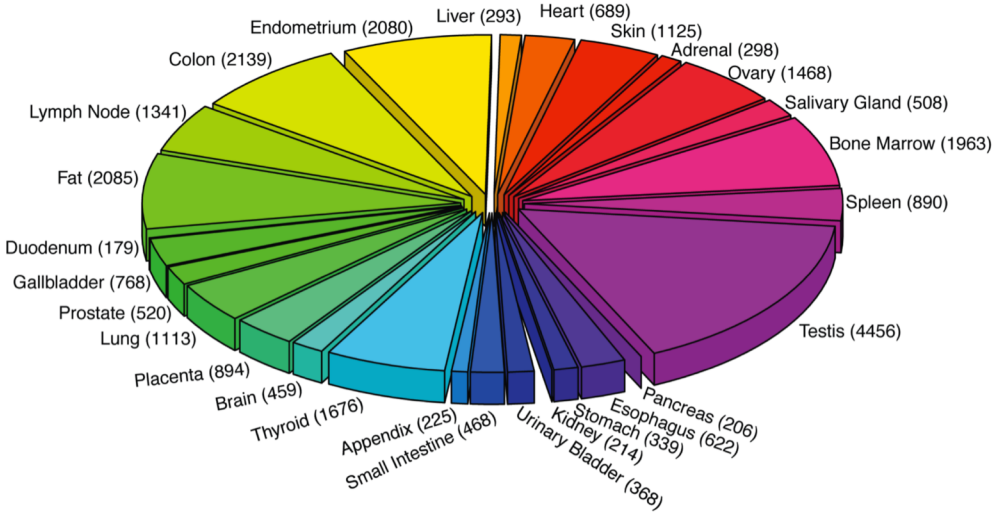
Fusion transcript in normal tissue samples

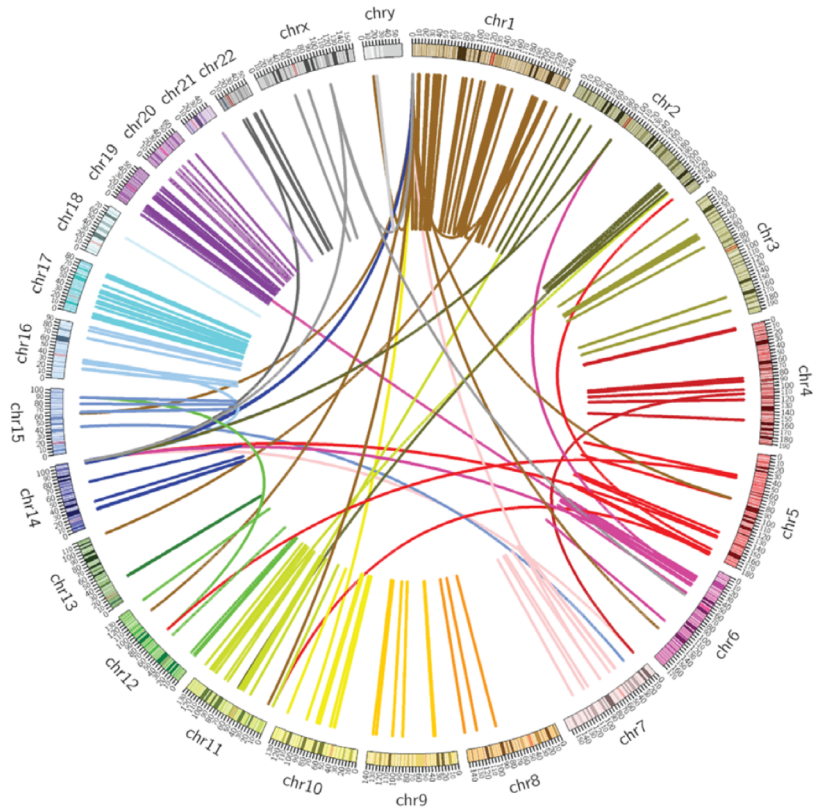
Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics.
Mol Cell Proteomics 2014 13: 397-406.

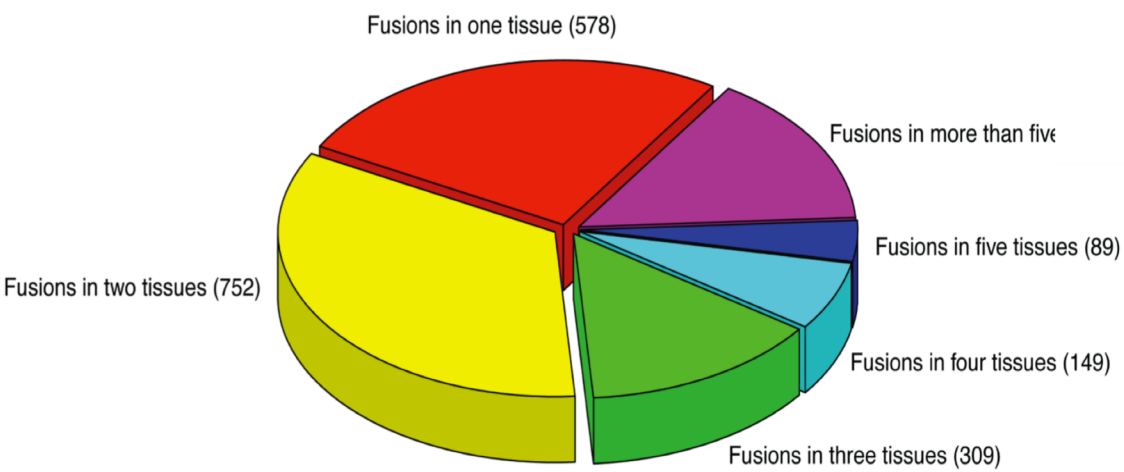
- The human tissues and organs analyzed by the transcriptomics analysis.
- ✓ 27 Human Tissues RNA-Seq samples.
- ✓ 201 RNA-Seq Runs.
- ✓ 22,107 unique fusions identified by Ericscript.



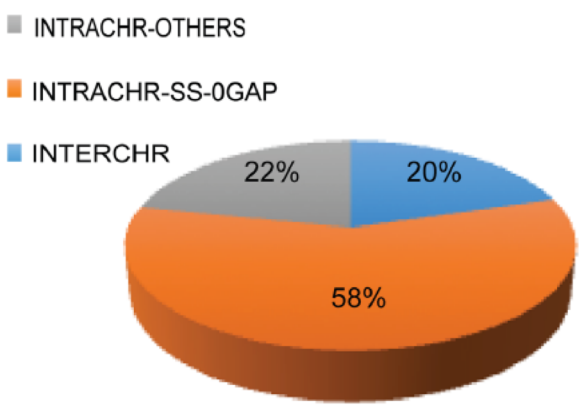
Distribution of fusion RNAs



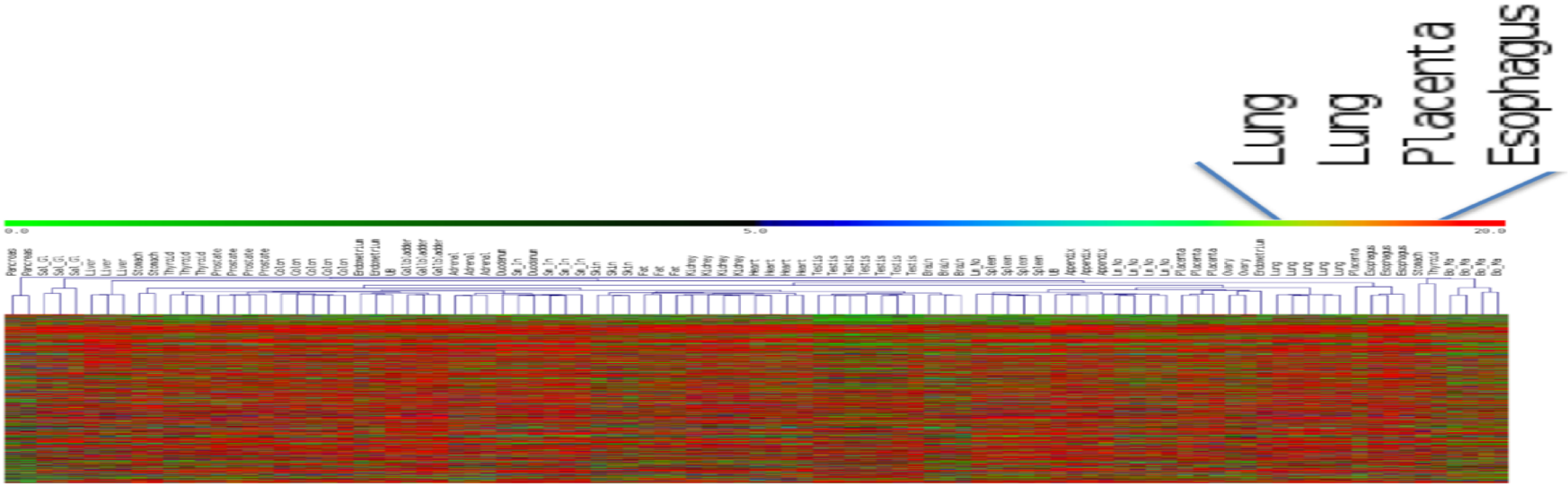




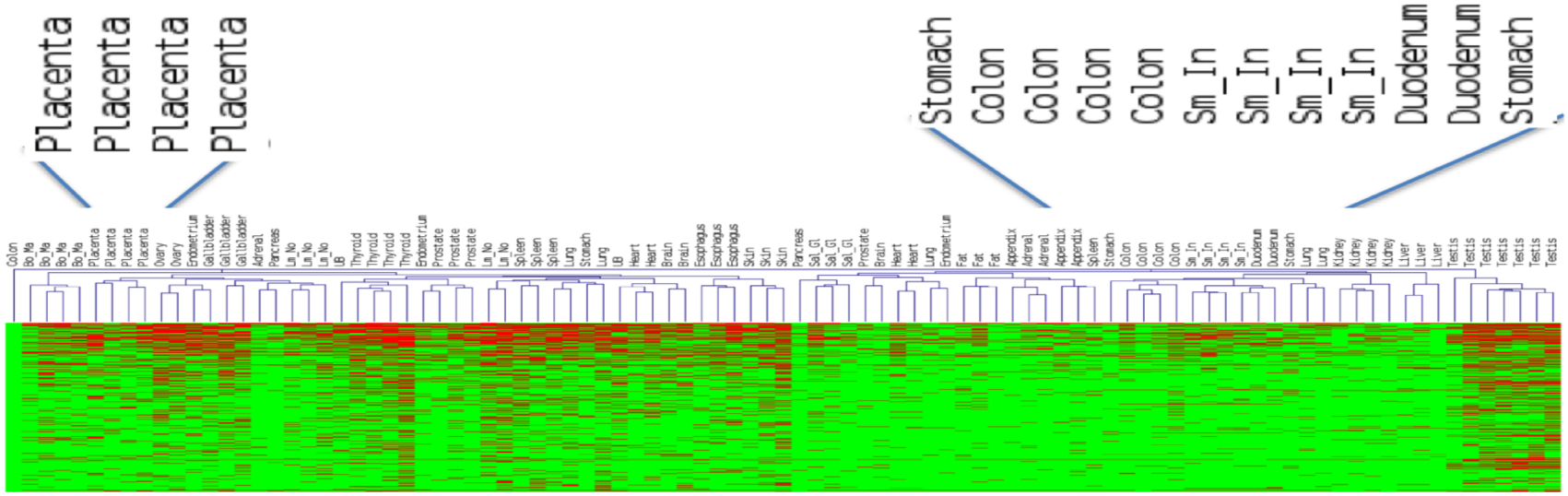
Recurrent (in >1 samples) gene fusions



Expression based clustering



Fusion based clustering



Normal tissues

Nucleic Acids Research

Recurrent chimeric fusion RNAs in non-cancer tissues and cells

Mihaela Babiceanu[1], Fujun Qin[1], Zhongqiu Xie[1], Yuemeng Jia[1], Kevin Lopez[1], Nick Janus[2], Loryn Facemire[1], Shailesh Kumar[1], Yuwei Pang[1], Yanjun Qi[2], Iulia M. Lazar[3] and Hui Li[1,4,*]

[1]Department of Pathology, School of Medicine, University of Virginia, Charlottesville, VA 22908, USA, [2]Department of Computer Science, University of Virginia, Charlottesville, VA 22908, USA, [3]Department of Biological Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA and [4]Department of Biochemistry and Molecular Genetics, School of Medicine, University of Virginia, Charlottesville, VA 22908, USA

genes

MDPI

Article

Absence of Correlation between Chimeric RNA and Aging

Reyna Huang [1], Shailesh Kumar [1,2] and Hui Li [1,3,*]

http://gtexportal.org/home/

GTEx Portal

GTEx | Datasets | Gene Association | eQTL Browser | Sample Data | Documentation | Publications | Contact | Search Gene or SNP ID... | Login | Register

2016-9-9

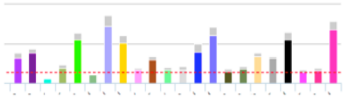
V6p Data Released

Read More >>

Current Release

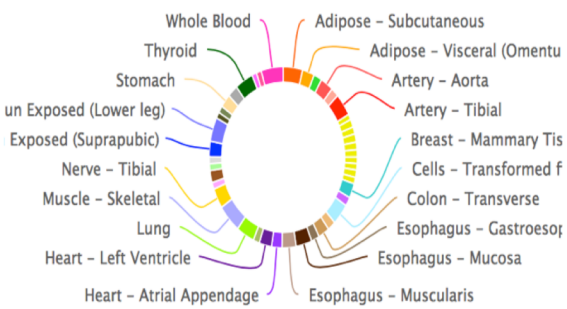
Latest Version: V6p

Dataset Summary Statistics Report



Browse eQTL Tissues

Total samples in all eQTL tissues: 7051

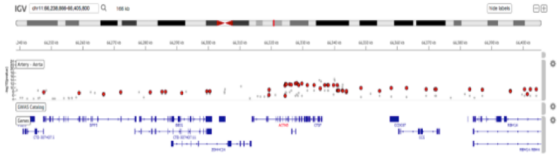


Genetic Association

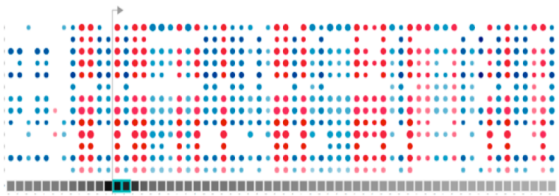
Single Tissue eQTLs

Search eQTL by gene or SNP ID

eQTL IGV Browser



Gene eQTL Visualizer



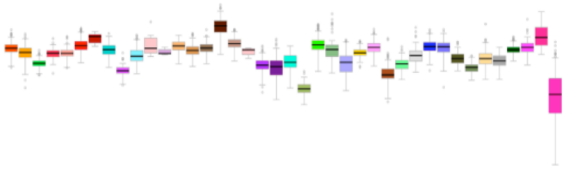
View eQTL data of a gene...

Transcriptome

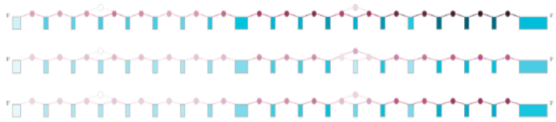
Search expression by gene ID...

Top 100 Expressed Genes in a Tissue (e.g. Blood)

Gene Expression in Tissues



Exon and Isoform Expression



News

A total of ~10,000 RNA-Seq samples analyzed

Publication

Nucleic Acids Research

The Landscape of Chimeric RNAs in Non-Diseased Tissues and Cells

Sandeep Singh[1#], Fujun Qin[1#], Shailesh Kumar[2], Justin Elfman[1,3], Emily Lin[1], Lam-Phong Pham[1], Amy Yang[1], Hui Li[1,2,*]

[1]Department of Pathology, School of Medicine, University of Virginia, Charlottesville, VA 22908

[2]National Institute of Plant Genome Research (NIPGR), New Delhi, India 110067

[3]Department of Biochemistry and Molecular Genetics, School of Medicine, University of Virginia, Charlottesville, VA 22908

# These authors contributed equally to this work.

* Corresponding author: Hui Li, 345 Crispell Dr., MR6-B524, School of Medicine, University of Virginia, Charlottesville, VA 22908, 434-9826624, hl9r@virginia.edu

Fusions in Cancer tissues

PNAS

Fusion transcriptome profiling provides insights into alveolar rhabdomyosarcoma

Zhongqiu Xie[a], Mihaela Babiceanu[a,1], Shailesh Kumar[a], Yuemeng Jia[a], Fujun Qin[a], Frederic G. Barr[b], and Hui Li[a,c,2]

[a]Department of Pathology, University of Virginia, Charlottesville, VA 22908; [b]Laboratory of Pathology, National Cancer Institute, Bethesda, MD 20892; and [c]University of Virginia Cancer Center, Charlottesville, VA 22908

Edited by Peter K. Vogt, The Scripps Research Institute, La Jolla, CA, and approved September 27, 2016 (received for review August 2, 2016)

ELSEVIER

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

EBioMedicine

Published by THE LANCET

The Landscape and Implications of Chimeric RNAs in Cervical Cancer

Peng Wu [a,b,1], Shuo Yang [a,1], Sandeep Singh [b], Fujun Qin [b], Shailesh Kumar [b,c], Ling Wang [a], Ding Ma [a,*], Hui Li [b,d,*]

shailesh@nipgr.ac.in