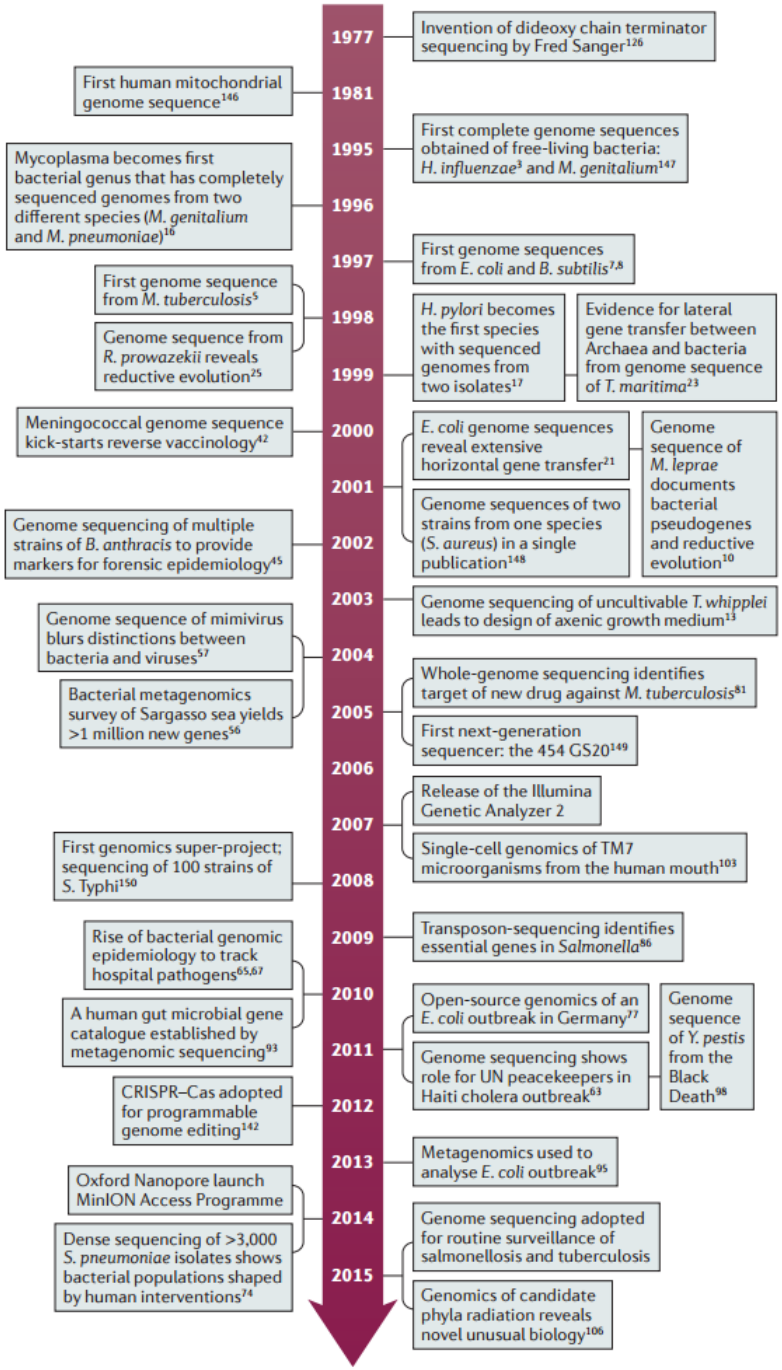
NGS를 이용한 미생물 유전체의 해독과 분석

Haeyoung Jeong
Super-Bacteria Research Center
KRIBB



Twenty years of bacterial genome sequencing

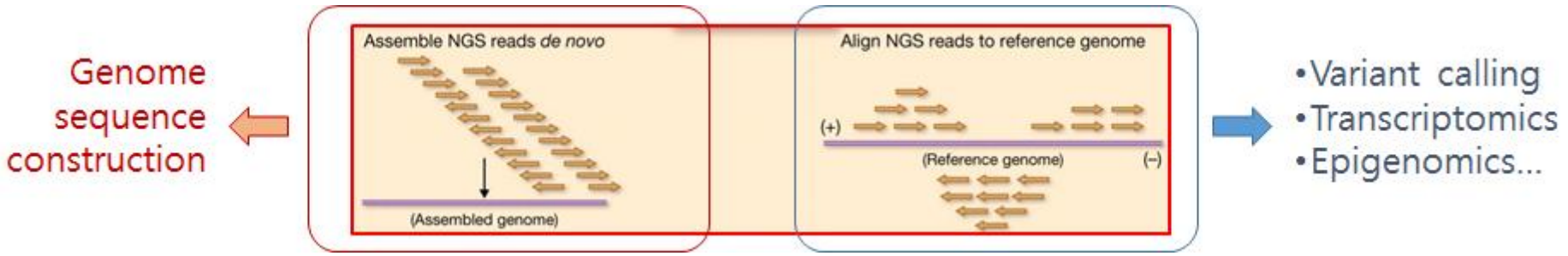
- The first revolution
  - ➢ Whole-genome shotgun
- The second revolution
  - ➢ High-throughput sequencing (the advent of next-generation sequencing)
- The third revolution
  - ➢ Single-molecule sequencing (long-read sequencing)

- *Comparative genomics*
- *Metagenomics*
- *Translational clinical bacterial genomics*

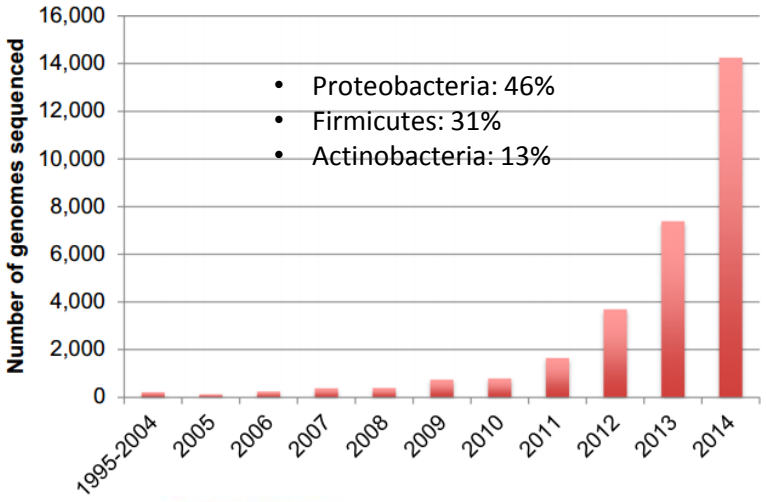
Nature Rev. Microbiol. 13:787 (2015)

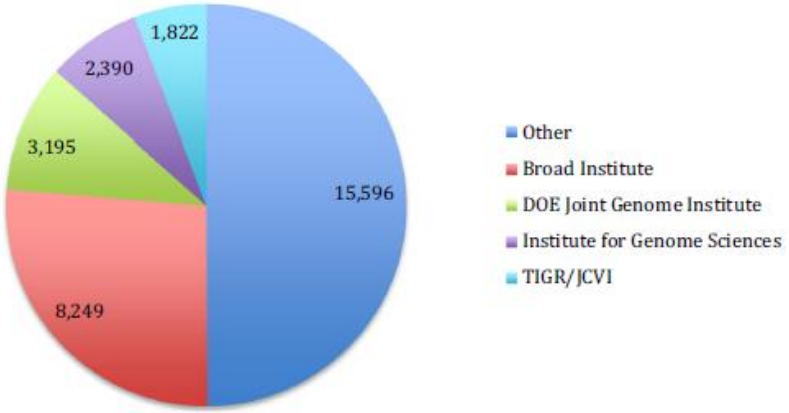
Two types of genome sequencing

- ***De novo* sequencing** [software: assembler]
  - The initial sequencing that results in the primary genetic sequence of organisms (sequencing a novel genome for the first time)
  - Depends on **de Bruijn graph-based assembly** (for NGS-based short reads) or **overlap-layout-consensus** procedure (for Sanger sequencing)
- **(Targeted) resequencing** [software: mapper or aligner]
  - The sequencing of (part of) an individual's genome in order to detect differences
  - Depends on mapping reads on the reference genome sequence
- Bioinformatic challenges posed by Next-Gen Sequencing (NGS)
  - Short read size
  - Huge data size
  - New technologies (different error models, read length, data size, etc.)



Insights from 20 years of bacterial genome sequencing





Genome sequence sources

- Genome size variation, protein-coding contents
- Genetic diversity is much greater than we thought
- Diversity in what all bacteria needs: tRNAs, codons, and codon usage
- Important roles for DNA sequence repeats in bacterial genomes
- Defense systems in archaea and bacteria
- Bacterial microcompartment organelles
- Genome comparisons and phylogeny
- Taxonomic enigmas can be resolved by comparative genomics

Funct. Integr. Genomics 15(2): 141 (2015)

극복해야 할 문제점

- The distribution of sequenced genome is **quite skewed towards a few phyla** that contain model organisms (50 difference bacteria phyla and 11 different archaeal phyla)
- Second-generation sequencing has produced **a large number of draft genomes** (close to 90% of bacterial genomes in GenBank are currently not complete)

Table 1 Number of sequenced genomes for 6 selected phyla and the percent of all genomes found in the phyla

Phyla	Number genomes	% of total
Actinobacteria	4059	13
Bacteroidetes/Chlorobi group	932	3
Cyanobacteria	340	1
Firmicutes	9628	31
Proteobacteria	14,268	46
Spirochaetes	525	2
Other	1500	5

Source: GenBank prokaryotes.txt file downloaded 4 February 2015

Table 4 Number of complete and permanent draft genomes and the percent of those genomes with each project status

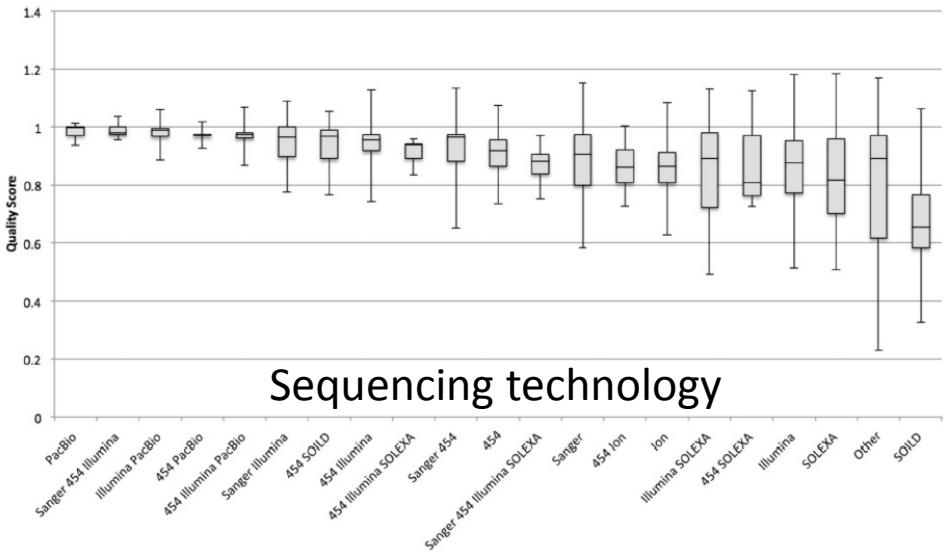
Project status	Bacteria	Archaea	Plasmids	Total
Finished	3060	173	1186	4419
Permanent draft	19,696	312	9	20,017
Draft	672	4	1	677
Total	23,428	489	1196	25,113

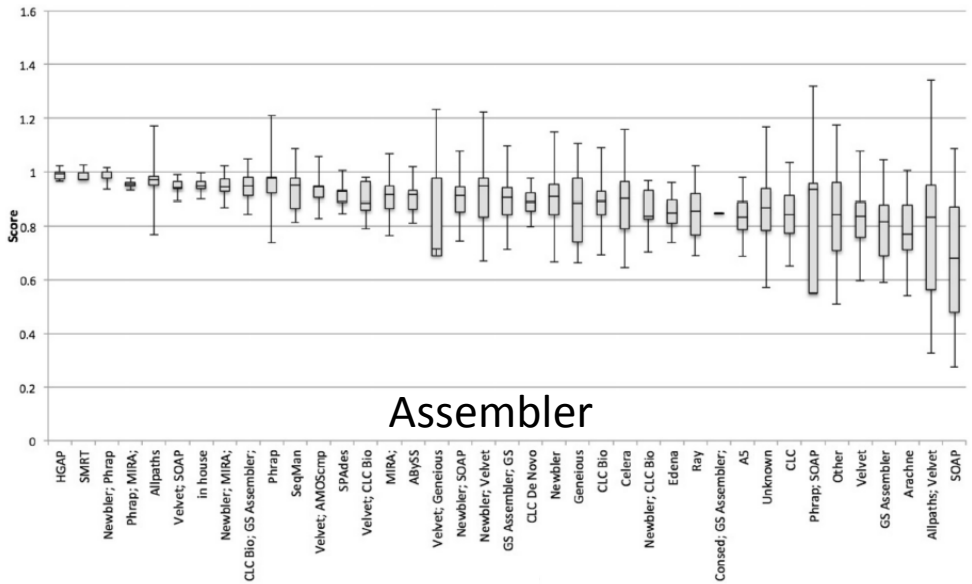
Source: IMG Statistics, accessed 4 February 2015

Draft genomes are fragmented representations which lack clarity on structural elements such as the orientation of certain sequence regions, or whether a plasmid is integrated into a genome [Science 326: 236, 2009]

Funct. Integr. Genomics 15(2): 141 (2015)

Quality scores for 32,000 genomes





- Categories for scoring
  - The completeness assembly
  - The presence of full-length RNA genes
  - tRNA composition
  - The presence of a set of 102 conserved ("essential") genes based on Pfam-A domain in prokaryotes – Refer to additional files 1 from the journal article website
- tRNAscan-SE, RNAmmer, Prodical, HMM3 were used for the analysis

Stand. Genomc Sci. 9:20 (2014)

3세대 시퀀싱 기술의 시대

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly

Sergey Koren and Adam M Phillippy

Curr. Opin. Microbiol. 23:110 (2015)

Long read sequencing 기법의 리뷰

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
GENOME ASSEMBLY ALGORITHMS
BACTERIOLOGY

Completing bacterial genome assemblies: strategy and performance comparisons

Yu-Chieh Liao, Shu-Hung Lin & Hsin-Hung Lin

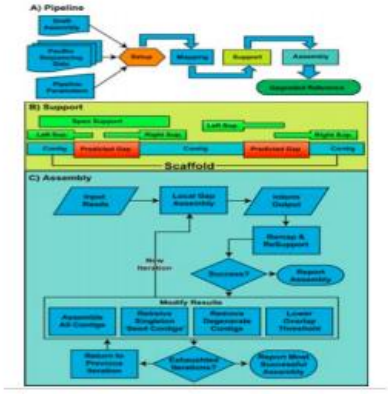
Sci. Rep. 5:8747 (2015)

Single PacBio SMRT cell run(~100x)으로 세균 유전체의 완성 수준 해독 가능

PacBio data 활용 기법의 발전

<5X Coverage >50x

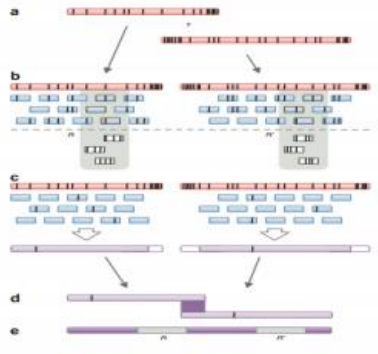
PBJelly



Gap Filling and Assembly Upgrade

English *et al* (2012)
PLOS One. 7(11): e47768

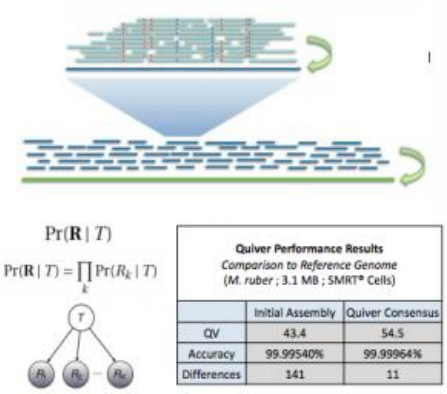
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

HGAP & Quiver



$$\Pr(\mathbf{R} \mid T)$$

$$\Pr(\mathbf{R} \mid T) = \prod_k \Pr(R_k \mid T)$$

Quiver Performance Results
Comparison to Reference Genome
(M. ruber ; 3.1 MB ; SMRT® Cells)

	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

PB-only Correction & Polishing

Chin *et al* (2013)
Nature Methods. 10:563–569

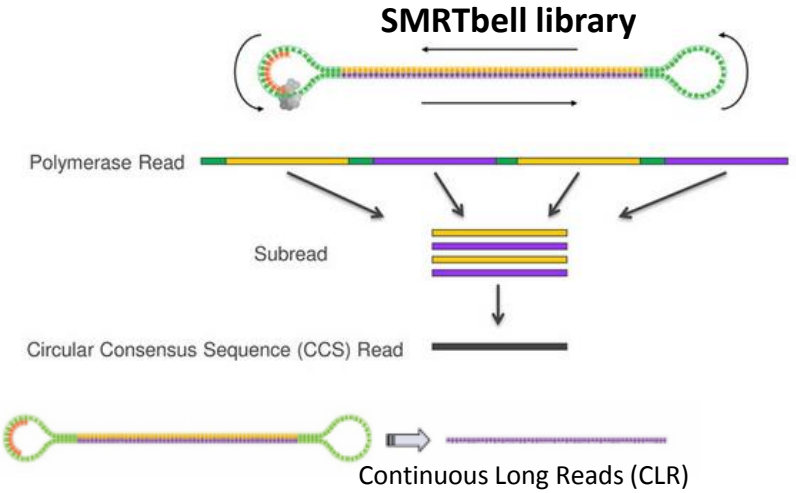
```
Query  16       TTATCA-CGCGGGAAGAAGATGACAGCCCGTTTGCTG-TAATACCGCTGTGCT-ACGG-C  71
                ||||||  |||||  ||||   ||||||||||||||||  |||  ||||   |||||  |||| |
Sbjct  2528414  TTATCAACGCGG-AAGAGATTGACAGCCCGTTTGCTGGTAAAACCGTGGTGCTTACGGGC  2528356
```

- Identities: 87%
- Gaps: 8%

PAG 2013: Michael Schatz, “De novo assembly of complex genomes using single molecule sequencing”
http://schatzlab.cshl.edu/presentations/2014-01-14.PAG.Single%20Molecule%20Assembly.pdf

RS_HGAP_Assembly protocol

SMRT analysis server
웹 포탈 이용



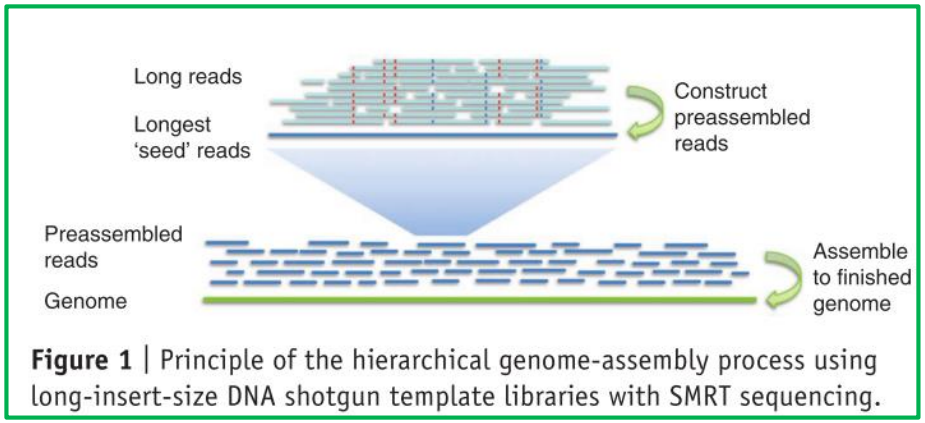


Figure 1 | Principle of the hierarchical genome-assembly process using long-insert-size DNA shotgun template libraries with SMRT sequencing.

* Library size가 너무 크면 plasmid 염기서열 재구성이 어려워질 수 있음

<Typical results>

Polished Contigs	**2**
Adapter Dimers (0-10bp)	**0.01%**
Short Inserts (11-100bp)	**0.01%**
Number of Bases	**854,061,799**
Number of Reads	**57,597**
N50 Read Length	**20,440**
Mean Read Length	**14,828**
Mean Read Score	**0.84**
Mapped Reads	**53,810**
Mapped Read Length of Insert	**8,334**
Average Reference Length	**2,948,050**
Average Reference Bases Called	**100.0%**
Average Reference Consensus Concordance	**100.0%**
Average Reference Coverage	**115.52**

SMRT cells (.H5 file)

↓

Filtered subreads

↓

Preassembled reads

↓ *Celera assembler (draft assembly)*

Draft assembly

↓ *QUIVER (resequencing)*

Polished assembly

공개된 genome sequence 파악하기

- **NCBI**
  - http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html
  - ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/assembly_summary.txt ←2016년 5월 2일 현재 61145건(archaea: 527건)
  - [data download] ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/
- **JGI Genomes OnLine Database**
  - https://gold.jgi.doe.gov/downloadexceldata ←2016년 5월 2일 현재 70687건 (모든 organism)
- **Ensemble Bacteria**
  - http://bacteria.ensembl.org/species.html ←2016년 5월 2일 현재 29777건(종 단위 집계이므로 수치가 다름)
  - [data download] http://bacteria.ensembl.org/info/website/ftp/index.html

RefSeq: NCBI Reference Sequence Database

- A comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins
- Prokaryotic RefSeq genomes
  http://www.ncbi.nlm.nih.gov/genome/browse/reference/
  - **Reference genomes** (120 entries – 2016년 5월 2일 현재)
    - NCBI의 스탭진이나 공동연구그룹의 참여를 통해 만들어진 가장 고품질의 데이터
    - 임학적 중요성(중요한 감염병 원인균), 어셈블리와 주석화의 품질, 실험적 증거의 유무 등이 자격 요건
    - 단백질 ID는 YP_ 또는 NP_로 시작
  - **Representative genomes** (4150 entries – 2016년 5월 2일 현재)
    - 클러스터링을 거쳐 대표성(예: type strain)이나 어셈블리의 품질 측면에서 각 종마다 representative genome을 선정
    - 단백질 ID는 WP_로 시작(non-redundant accessions across multiple species)
  - **Variant genomes (**나머지 유전체들로서 단백질 ID는 WP_로 시작)
- Prokaryotic RefSeq genome은 일부 고품질의 reference genome을 제외하고 NCBI의 annotation pipeline을 통하여 재주석화를 거침
  - ftp://ftp.ncbi.nlm.nih.gov/refseq/release/announcements/WP-proteins-06.10.2013.pdf (New RefSeq protein product and data model)
  - 재주석화를 거쳐 새롭게 예측된 유전자의 locus tag은 Prefix_*RS*00001의 형식을 갖추며, 예전 기록과는 old_locus_tag으로 연결

RefSeq Release 75 (58,776 organisms) 2016-03-14 : ftp://ftp.ncbi.nlm.nih.gov/refseq/release/

NCBI에서 유전체 정보 검색하기[1]

1. 계층적 탐색
   - "Genome" → "Microbes" or "Prokaryotic reference genomes"
   - http://www.ncbi.nlm.nih.gov/genome/browse/
2. FTP site에서 직접 다운로드
   - ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/assembly_summary.txt의 맨 마지막 필드에서 ftp path 확인
     (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF~ RefSeq, GCA~: GenBank)
   - Accession number를 아는 경우: 소수점 이하(버전)은 생략 가능
     http://www.ncbi.nlm.nih.gov/nuccore/CP000154.2 (complete)
     http://www.ncbi.nlm.nih.gov/nuccore/ALEG00000000.1 (WGS)
3. Entrez search

"Paenibacillus polymyxa"[organism]"

"Paenibacillus polymyxa"[organism] AND "E681"[strain] → Next slide

"Paenibacillus polymyxa"[organism] AND KRIBB[submitter]

"Paenibacillus polymyxa"[organism] AND "complete genome"[assembly level]

NCBI에서 유전체 정보 검색하기[2]

Genomes

Assembly	2	genome assembly information
BioProject	1	biological projects providing data to NCBI
BioSample	1	descriptions of biological source materials
Clone	0	genomic and cDNA clones
dbVar	0	genome structural variation studies
Epigenomics	0	epigenomic studies and display tools
Genome	1	genome sequencing projects by organism
GSS	0	genome survey sequences
Nucleotide	17	DNA and RNA sequences
Probe	0	sequence-based probes and primers
SNP	0	short genetic variations
SRA	0	high-throughput DNA and RNA sequence read archive
Taxonomy	0	taxonomic classification and nomenclature catalog

Next slide

ASM14687v2 Assembly 신버전(v2)

1. Organism: **Paenibacillus polymyxa E681** (firmicutes)
Infraspecific name: Strain: **E681**
Submitter: Korea Research Institute of Bioscience and Biotechnology (KRIBB)
Date: 2015/03/18
Assembly level: Complete Genome
Genome representation: full
GenBank assembly accession: GCA_000146875.2 (latest)
RefSeq assembly accession: GCF_000146875.3 (latest)
Release type: Minor
IDs: 360171 [UID] 1712878 [GenBank] 1855058 [RefSeq]

ASM14687v1 Assembly 구버전(v1)

2. Organism: **Paenibacillus polymyxa E681** (firmicutes)
Infraspecific name: Strain: **E681**
Submitter: Korea Research Institute of Bioscience and Biotechnology (KRIBB)
Date: 2010/09/03
Assembly level: Complete Genome
Genome representation: full
GenBank assembly accession: GCA_000146875.1 (replaced)
RefSeq assembly accession: GCF_000146875.1 (replaced)
IDs: 274838 [UID] 169438 [GenBank] 274838 [RefSeq]

Display Settings: Full Report

Send to:

ASM14687v2

Organism name: Paenibacillus polymyxa E681 (firmicutes)
Infraspecific name: Strain: E681
BioSample: SAMN02603484
Submitter: Korea Research Institute of Bioscience and Biotechnology (KRIBB)
Date: 2015/03/18
Release type: Minor
Assembly level: Complete Genome
Genome representation: full
GenBank assembly accession: GCA_000146875.2 (latest)
RefSeq assembly accession: GCF_000146875.3 (latest)
RefSeq assembly and GenBank assembly identical: yes

See Genome Information for **Paenibacillus polymyxa**

There are 20 assemblies for this organism
See more

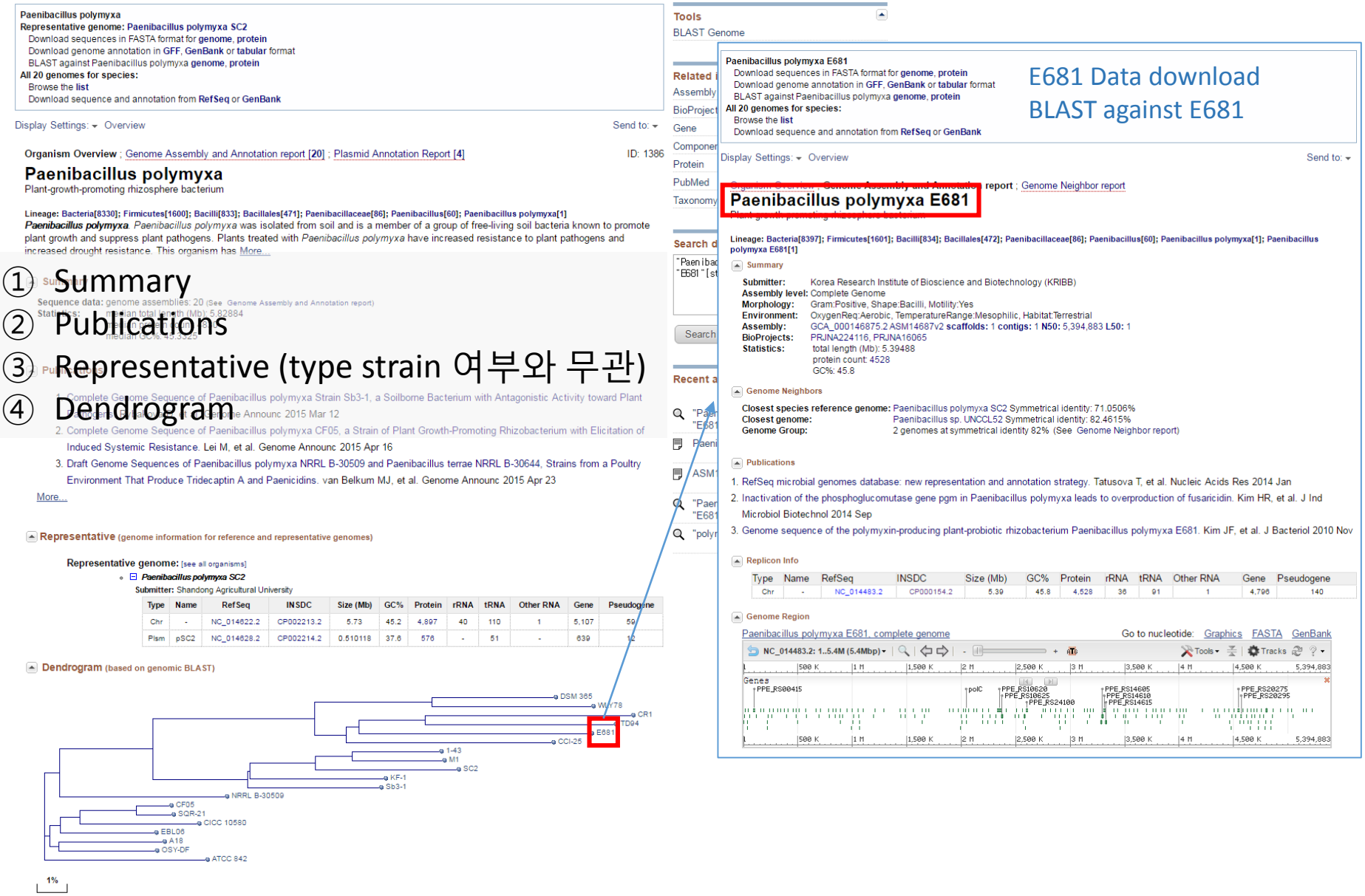
Access the data
Download the RefSeq assembly
Download the GenBank assembly
Download the full sequence report
Download the statistics report

Assembly Information
Assembly Help
Assembly Basics
NCBI Assembly Data Model

NCBI에서 유전체 정보 검색하기[3]

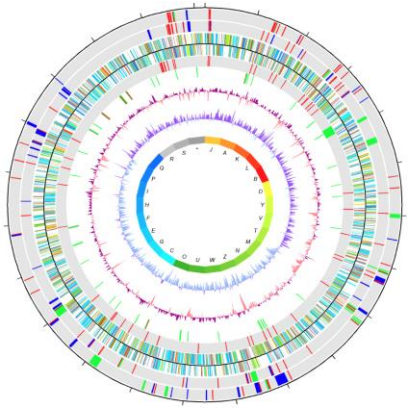
① Summary
② Publications
③ Representative (type strain 여부와 무관)
④ Dendrogram

E681 Data download
BLAST against E681

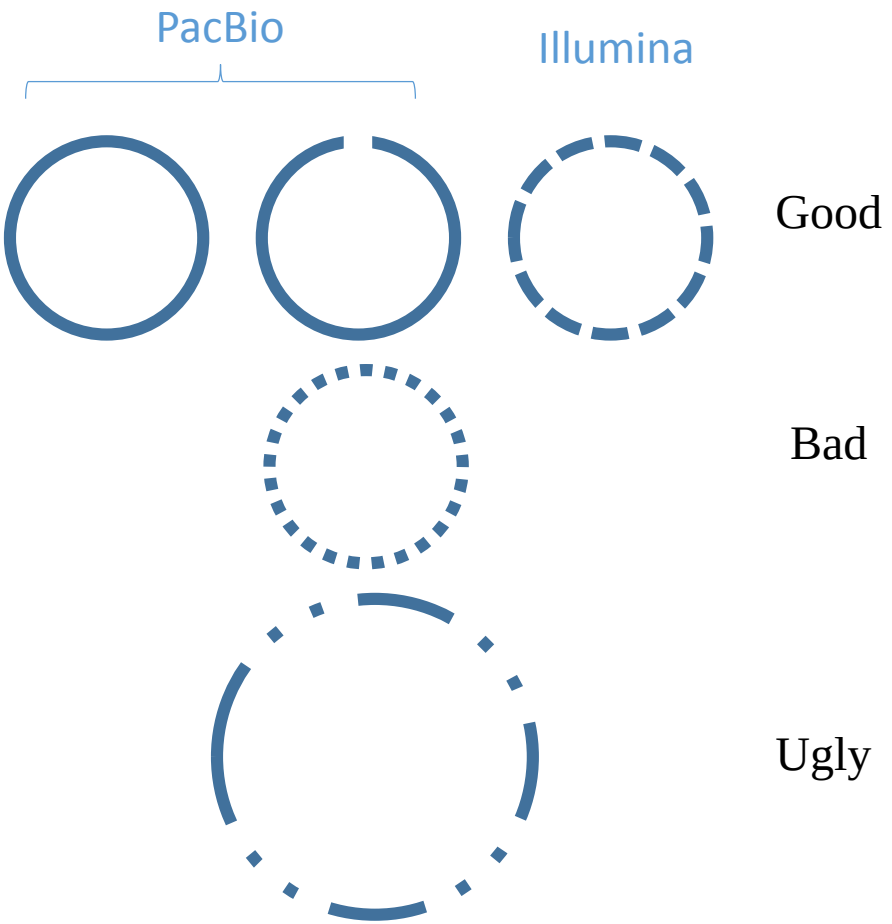


The Good, the Bad, and the Ugly

- Contig number (small)
- Total contig length (close to target genome size)
- N50 (long) and others
- Can be compared using QUAST



De novo assembly



N50 (Nx) definition

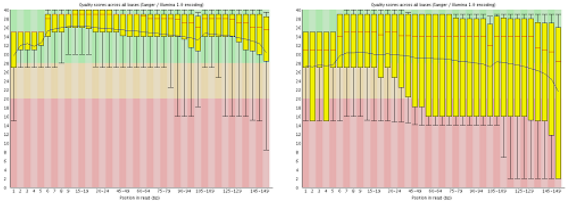
Assemblies: the good, the bad, the ugly

Ewan Birney NATURE METHODS | VOL.8 NO.1 | JANUARY 2011 | 59

The low cost of short-read sequencing has motivated the development of *de novo* assemblies from only short-read data; impressively, assemblies for large mammalian genomes are now available. However, this is still a developing field, and these *de novo* assemblies have many artifacts, as do all *de novo* assemblies.

일루미나 데이터의 전처리

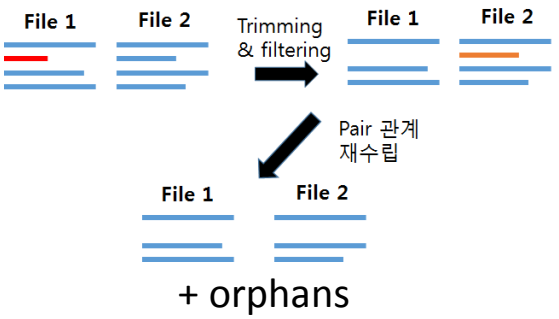
*FastQC는 단순 QC report만 생성



Tool	PE reads handling	Parallel processing	NGS artifacts handling	Quality score-based trimming
ngsShoRT (2.1)	Yes	Yes	Yes	Yes: 3'-end, quality window and filter out low quality reads
NGS QC toolkit (v.2.3.2) [15]	Yes	Yes	Yes	Yes: filter out low quality reads
FASTX toolkit (v. 0.0.13.2) [26]	No	No	No	Yes: filter out low quality reads
SeqTrim [25]	No	No	No	Yes: filter out low quality reads
CutAdapt (v.1.3) [14]	No	No	No	Yes: filter out low quality reads
Btrim [27]	No	No	No	Yes: quality window
SolexaQA (v.2.2) [8]	Yes	No	No[7]	Yes: quality window and filter out low quality reads
Sickle [28]	Yes	No	No	Yes: quality window
Scythe [24]	No	No	Yes, but only 3'	No
Trimmomatic (v.0.32) [16]	Yes	Yes	Yes	Yes: quality window and filter out low quality reads

[Source Code for Biology and Medicine 2014, 9:8]

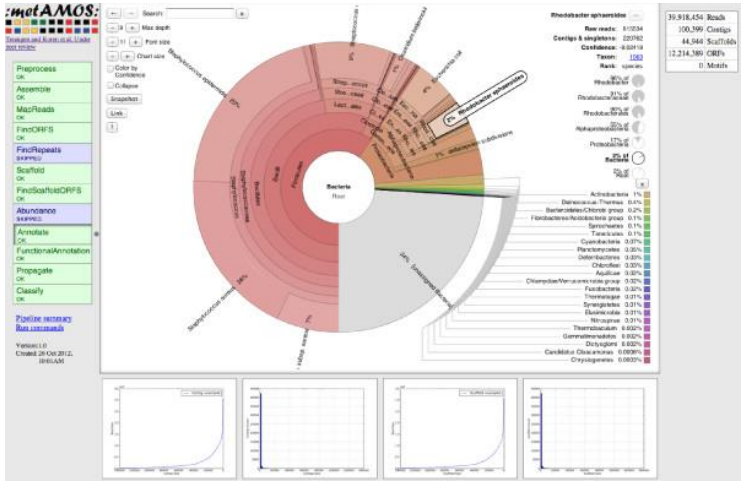
- Sequencing artifacts removal (including adaptor sequences)
- Dealing with "N" bases
- Quality score-based trimming
- 5'/3'-end bases trimming
- Paired end reads handling
- [Error correction: refer to Brief Bioinform 2013 14(1):56-66]



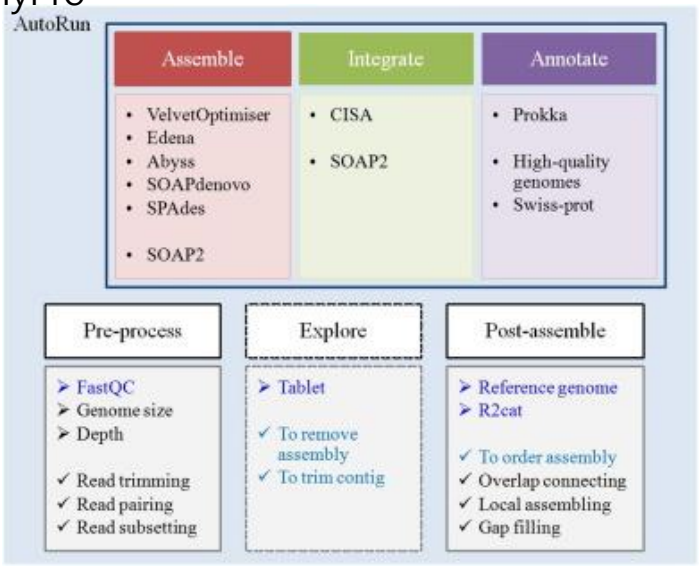
일루미나 데이터를 이용한 assembly

1. Command-line interface에서 실행되는 단위 프로그램
2. Read data의 전처리와 변환 등의 작업을 위한 보조 프로그램(또는 기능) 포함
   - SGA, IDBA_UD, SPAdes
3. K-mer 크기를 달리해 나가면서 최적의 assembly 결과 선택
   - VelvetOptimiser, IDBA_UD
4. 전처리, error correction, de novo assembly, scaffolding 등 일련의 작업을 수행하는 반자동화 파이프라인
   - A5-miseq
5. GUI 형태의 통합 환경에 de novo assembly·reference mapping이 포함
   - Unipro UGENE, CLC Genomics Workbench, Geneious Pro
6. **MetAMOS:** a modular framework for (meta)genomic assembly, analysis and validation (Genome Biol. 2013, 14:R2)
7. **MyPro:** a seamless pipeline for automated prokaryotic genome assembly annotation (J. Microbiol. Methods. 2015, 113:72)

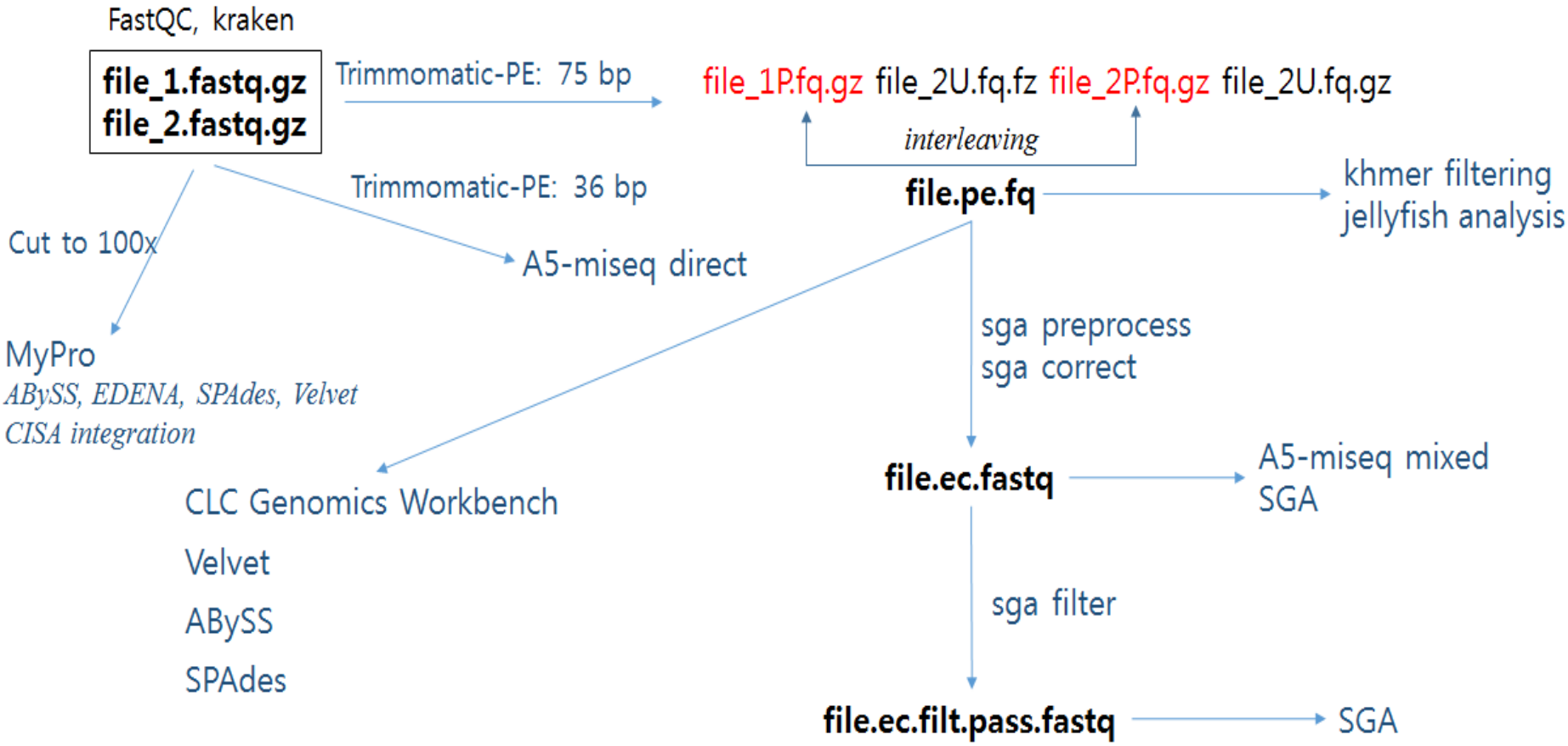
MetAMOS



MyPro



Illumina data processing pipeline의 사례



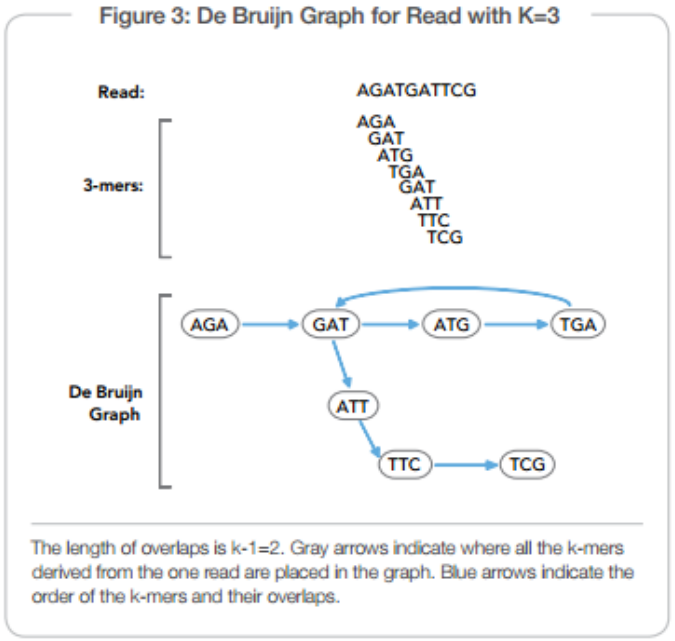
- FastQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- Trimmomatic: http://www.usadellab.org/cms/?page=trimmomatic
- KRAKEN: https://ccb.jhu.edu/software/kraken/

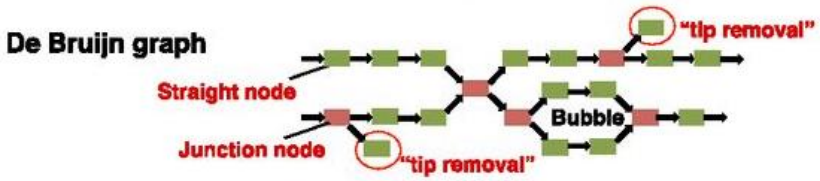
K-mer based sequence analysis

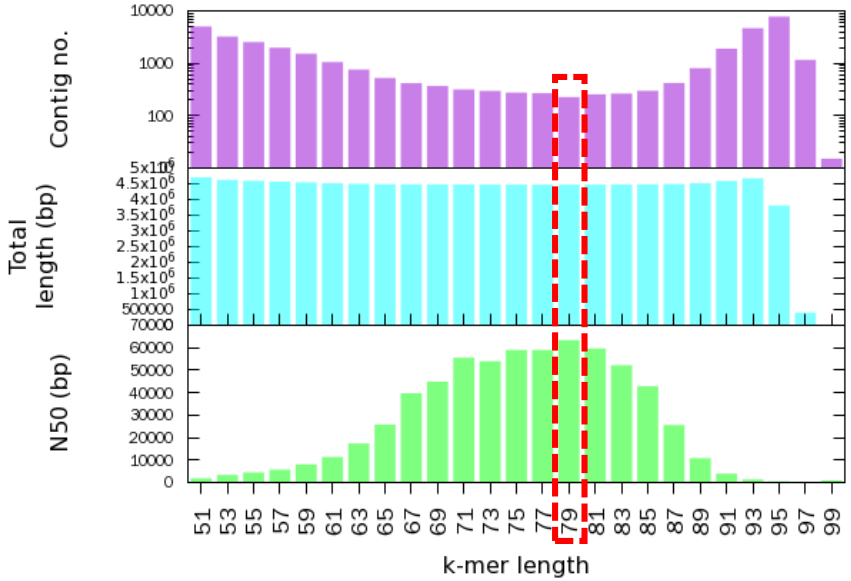
- A k-mer is a DNA word of a fixed length, or **a substring of length k**. (there are 17,179,869,184 unique 17-mer)
- Counting the occurrence of all such substring is a central step in many DNA sequence analysis steps
  - Estimating sequencing depth and genome size
  - Exploring repetitive genomes
  - Exploring the heterozygous genome
  - Dealing with sequencing error and coverage bias
  - The basic foundation of de Bruijn graphs
- K-mer based *digital normalization* discards redundant data and both sampling variation and the number of errors present in deep sequencing data set
- Fast k-mer counting and implementation of data structures for k-mer storage are challenging tasks

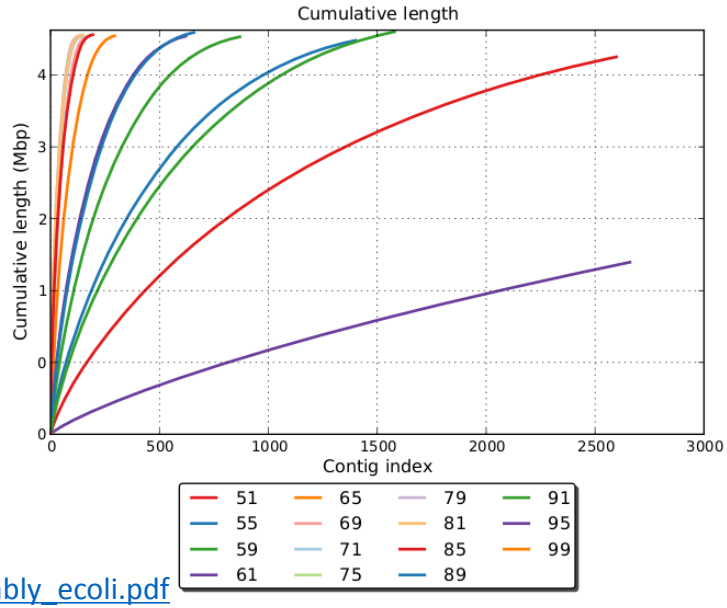
http://arxiv.org/ftp/arxiv/papers/1308/1308.2012.pdf

K-mers dramatically affect de novo assembly









E. coli BL21 (TaKaRa) real data, 100x subsample (475.9 Mb, 101 nt x 2), no pretreatment

http://www.illumina.com/Documents/products/technotes/technote_denovo_assembly_ecoli.pdf

Genome Res. 2014 Aug;24(8):1384-95.

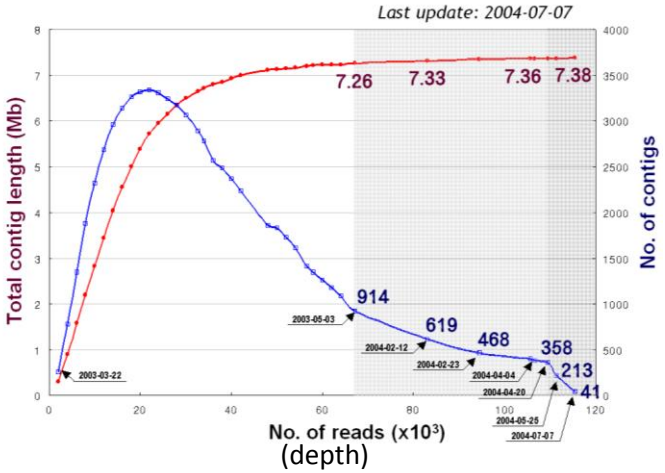
Velvet v1.2.10

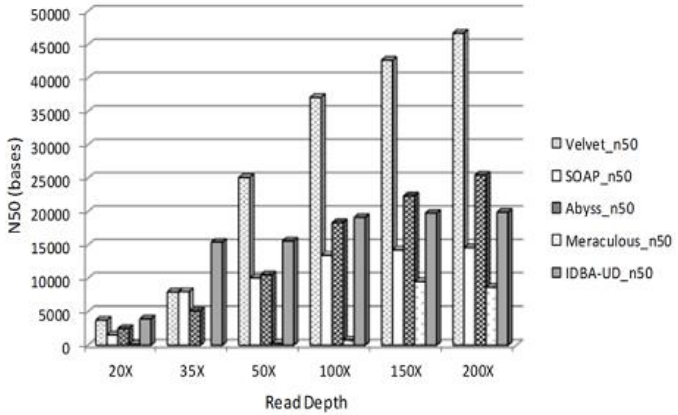
Is there optimal read depth (Illumina)?

Sequencing depth D =

$$\frac{\text{(Read length)x(read number)}}{\text{(Estimated genome size)}}$$

Sanger sequencing example





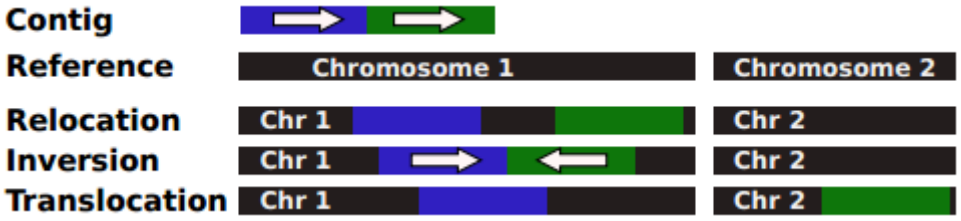
- For large mammalian genomes, the depth of coverage achieved is typically low (20-60X)
- For small genome assemblies using de Bruijn graph-based assemblers
  - 50X is enough to get a "good" genome assembly
  - Sequencing at a depth greater than 100X does not provide any additional benefits
  - High depth requires larger computational resources, even cases "bad" assemblies (why?)

PLoS One 8:4 (2013)

Genome assembly evaluation

- Without a reference sequence
  - QUAST(http://bioinf.spbau.ru/quast): assembly metrics의 비교
  - REAPR(http://www.sanger.ac.uk/science/tools/reapr): read pair alignment를 이용
  - BUSCO(http://busco.ezlab.org/) : assessing assembly in terms of "universal" gene content
- With a reference sequence
  - QUAST: assembly metrics + # misassemblies, # misassembled contigs, misassembled contig length, # local misassemblies, # unaligned contigs, unaligned length, unaligned fraction (%), # N's per 100 kbp…

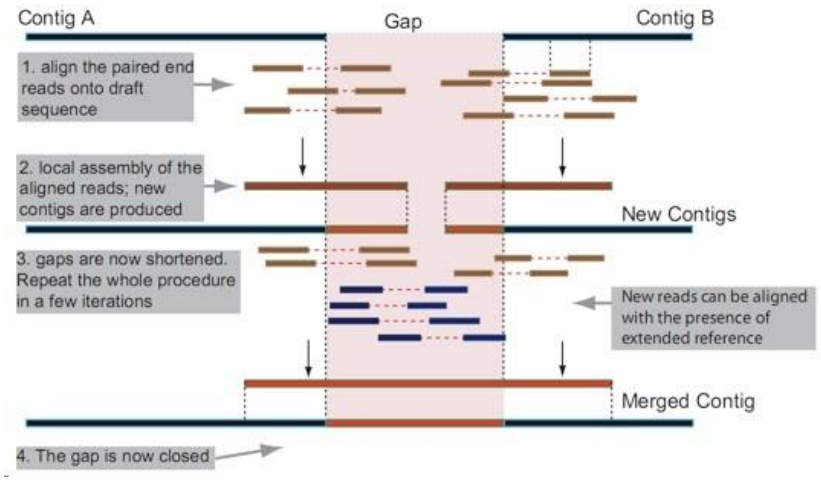
http://busco.ezlab.org/



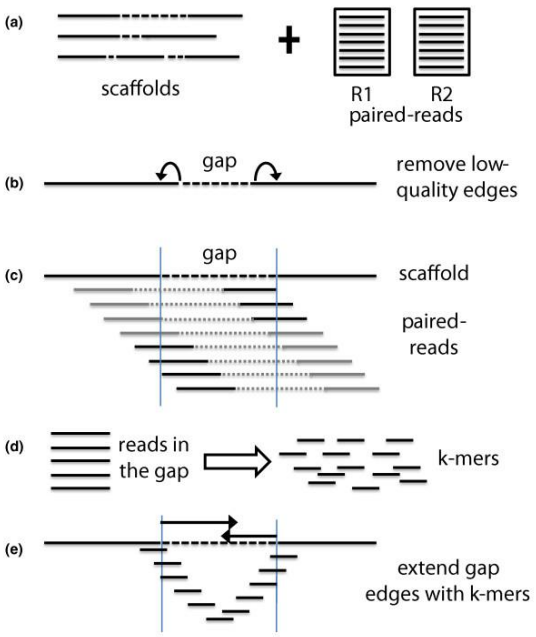
Plantagora defines a *misassembly breakpoint* as a position in the assembled contigs where the left flanking sequence aligns over 1 kb away from the right flanking sequence on the reference, or they overlap by more than 1 kb, or the flanking sequences align on opposite strands or different chromosomes

Scaffolding and automatic gap filling

- 최신 de novo assembler는 scaffold 작성 기능을 대부분 갖추고 있음
  - Gap 영역을 N으로 표현한 pseudomolecule 형태로 나타내거나 AGP format으로 출력
  - AGP specification: https://www.ncbi.nlm.nih.gov/assembly/agp/AGP_Specification/
  - "A comprehensive evaluation of assembly scaffolding tools" Genome Biol. 2014, 15:R42 PMID: 24581555
- Automatic gap filling software
  - GapCloser for SOAPdenovo (http://soap.genomics.org.cn/soapdenovo.html)
  - IMAGE (Genome Biol. 2010, 11:R41 PMID: 20388197)
  - GapFiller (Genome Biol. 2012, 13:R56 PMID:22731987)



IMAGE

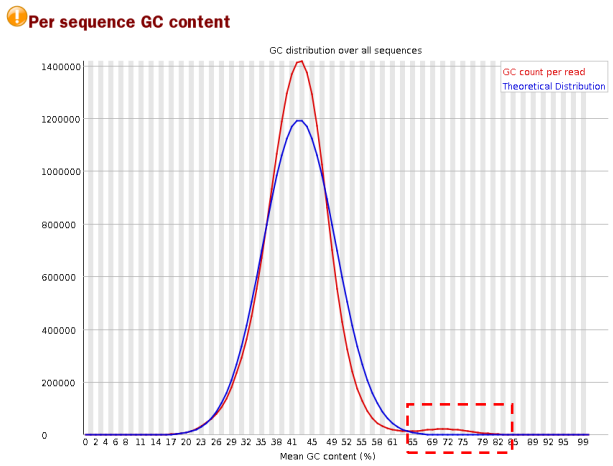


GapFiller

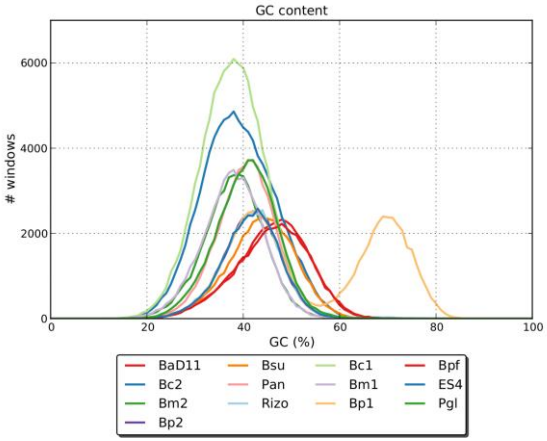
Bad assembly의 주요 발생 원인

- Short-read를 이용한 de novo assembly는 근본적으로 정확성에 한계가 있음
  - “Genome assembly is not a solved problem” – Ewan Birney
- Errors and/or inadequate sequencing depth
- Contamination
  - ➢ 검출 방법:
    1. %G+C의 차이를 이용(read 혹은 assembly level)
    2. K-mer abundance profile analysis
    3. Read classification using database search (e.g., KRAKEN)
- Heterozygosity or polymorphism for eukaryotic genomes

Read level

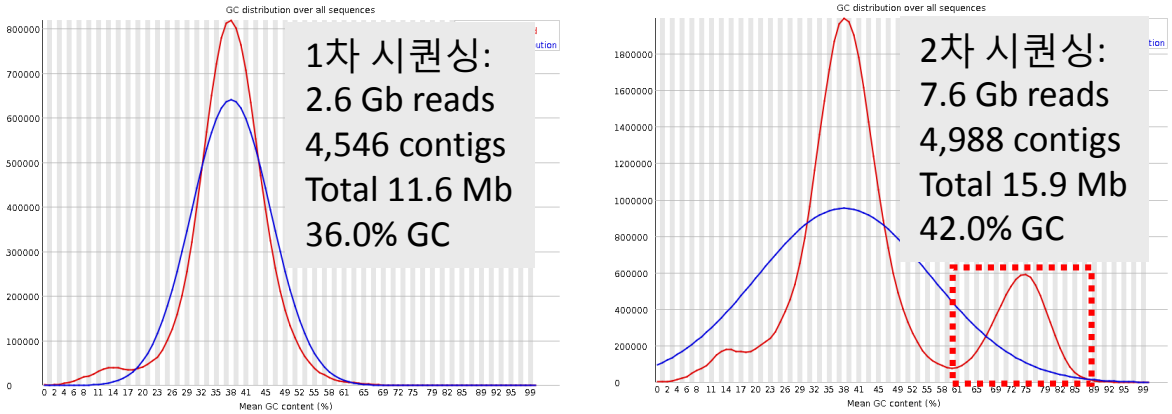


Assembly level



Contamination is not uncommon

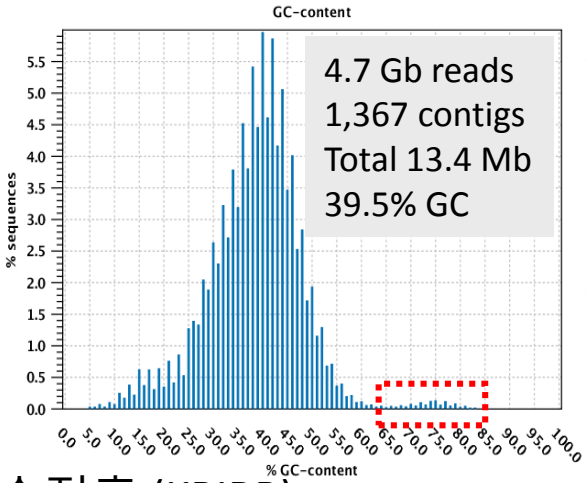
Acid-tolerance yeast [1]: 극명한 %GC 차이와 가까운 reference sequence를 이용



- 알려진 reference genome에 매핑하여 남은 read(13.7%)를 조립하니 888 contig(4.38 Mb, 72.9% GC)의 세균 유전체 서열 생성
- *Cellulomonas fimi* ATCC 484와 유사

다른 효모 assembly에서는 high GC contig로부터 최소 2종의 bacterial rRNA 서열 검출

Acid-tolerance yeast [2]: coverage 활용



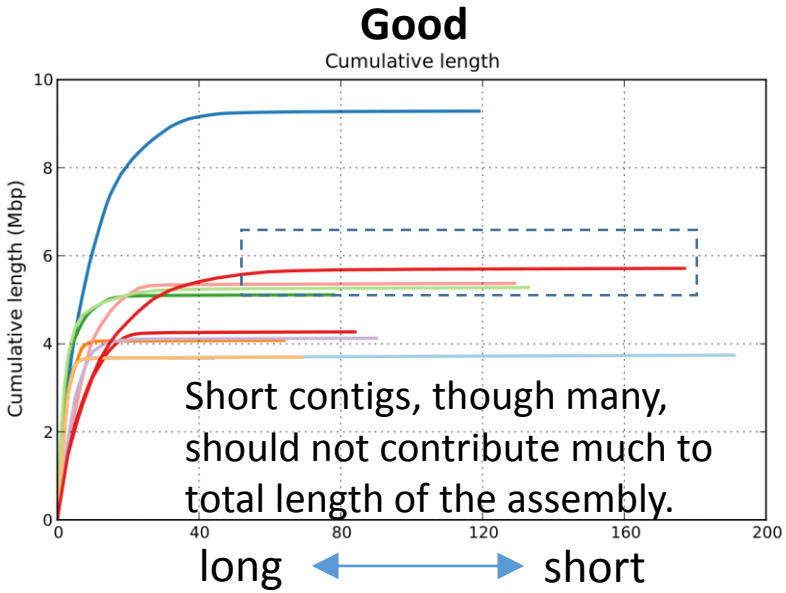
- Cutoff 기준: coverage 100
  - ➢ High coverage: 100 contigs, total 9.4 Mb
  - ➢ Low coverage: 1267 contigs, total 4.0 Mb
- Blast 검색 결과 low read coverage contig는 bacterial sequence로 확인

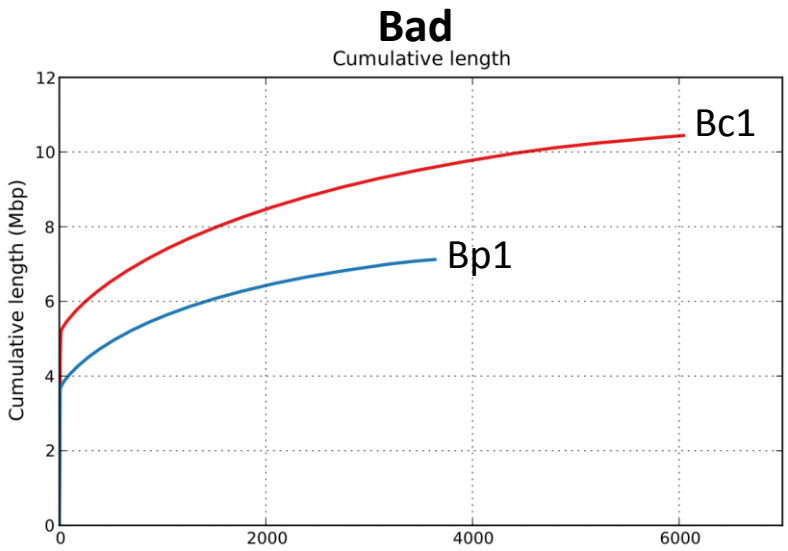
How to eliminate contamination?

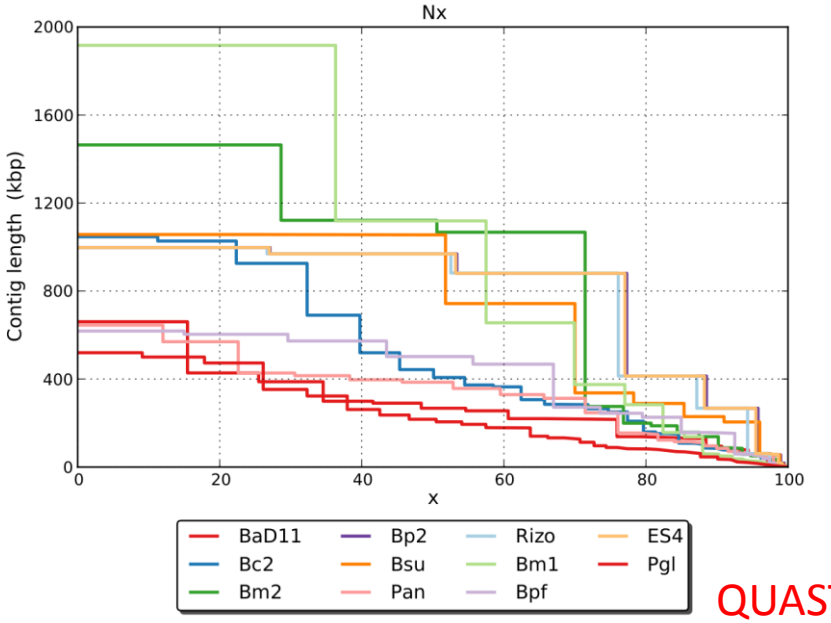
- At read or assembly level
- By reference mapping or differences in %GC and/or read coverage
- Detection or elimination of contamination is not always feasible!

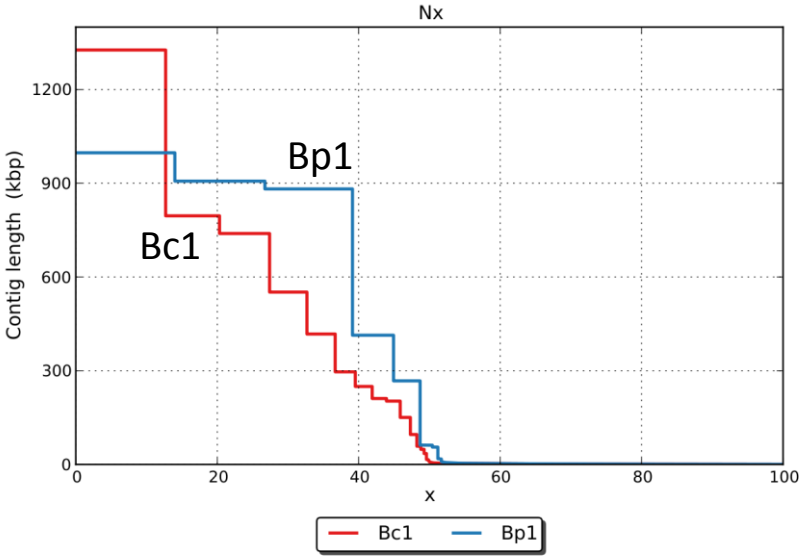
Dr. 손정훈 (KRIBB)

Bad assemblies have too many short contigs







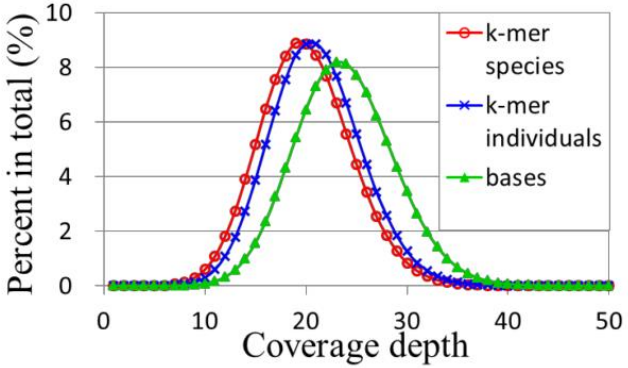


QUAST diagnostic plots

Drs. 박승환, 반재구 (KRIBB)

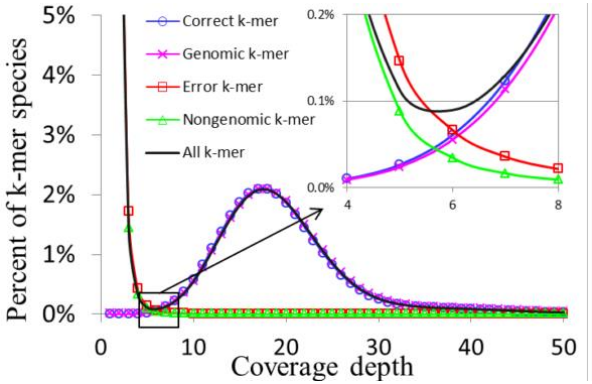
k-mer spectrum examples for sequencing reads

"Ideal" data (without error)

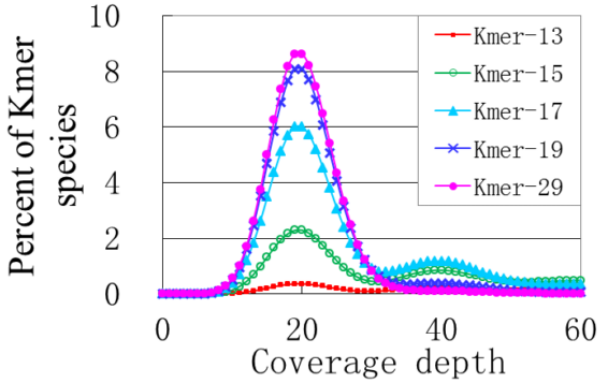


23.8x simulated sequencing data from the 10-Mb ideal genome (L=100, k=17)

With 1% error

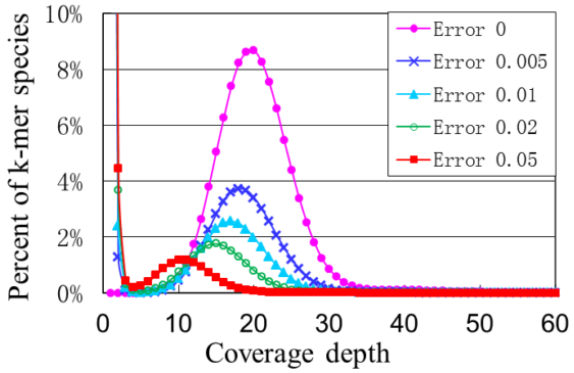


Simulated reads from non-heterozygous *Arabidopsis* genome (k=17)



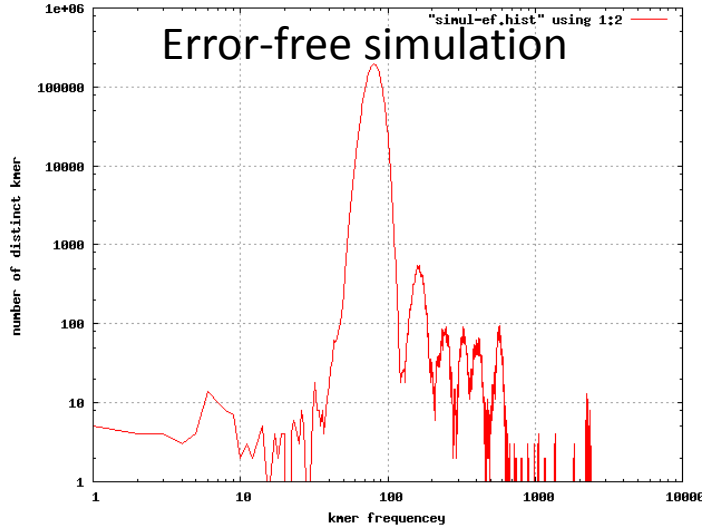
Simulated sequencing data from the human reference genome

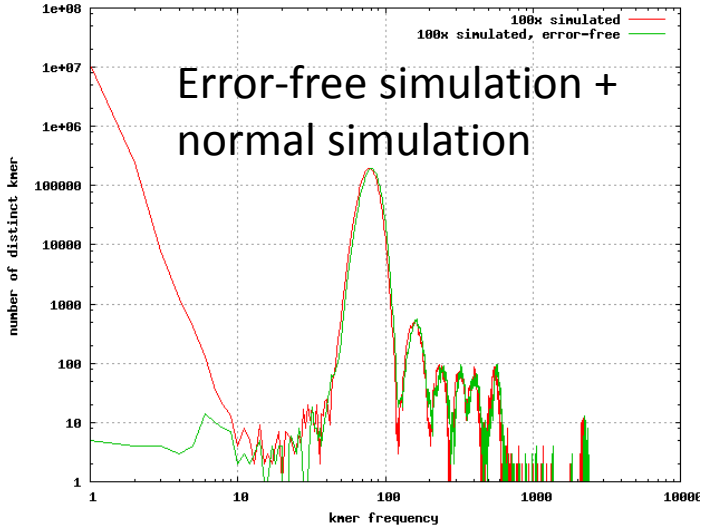
With various sequencing error rate

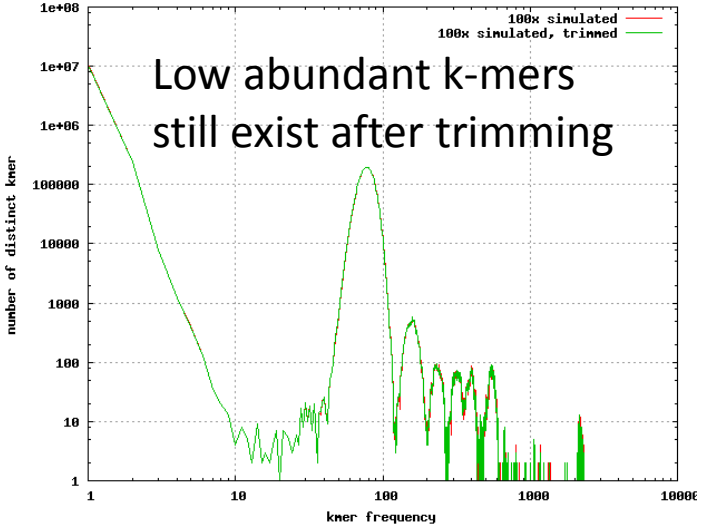


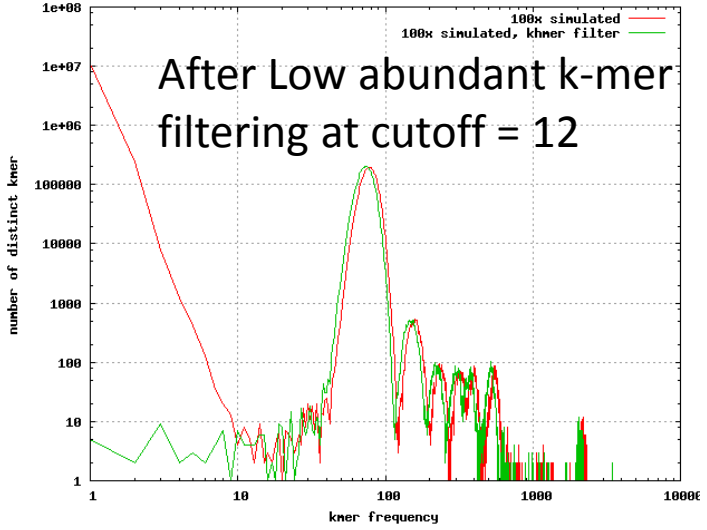
http://arxiv.org/ftp/arxiv/papers/1308/1308.2012.pdf

Quality trimming cannot remove errors





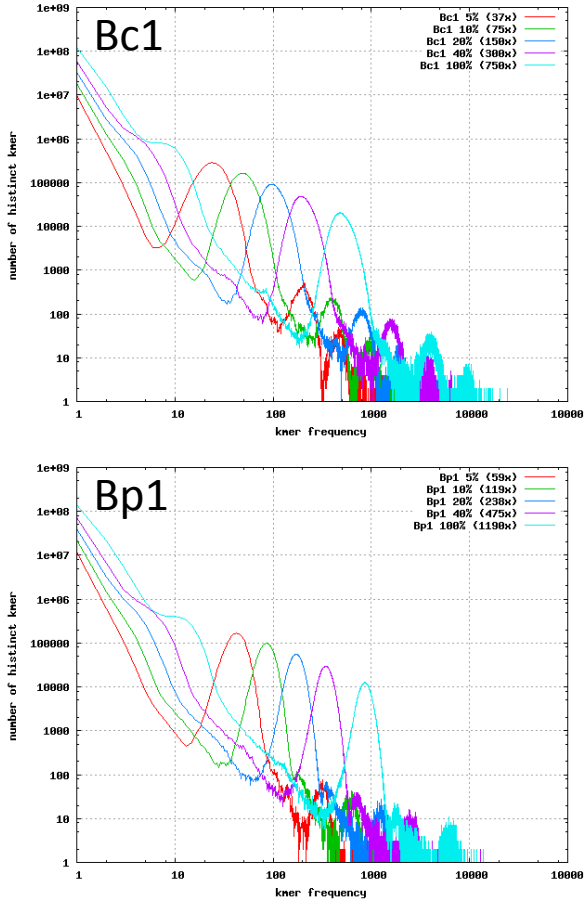




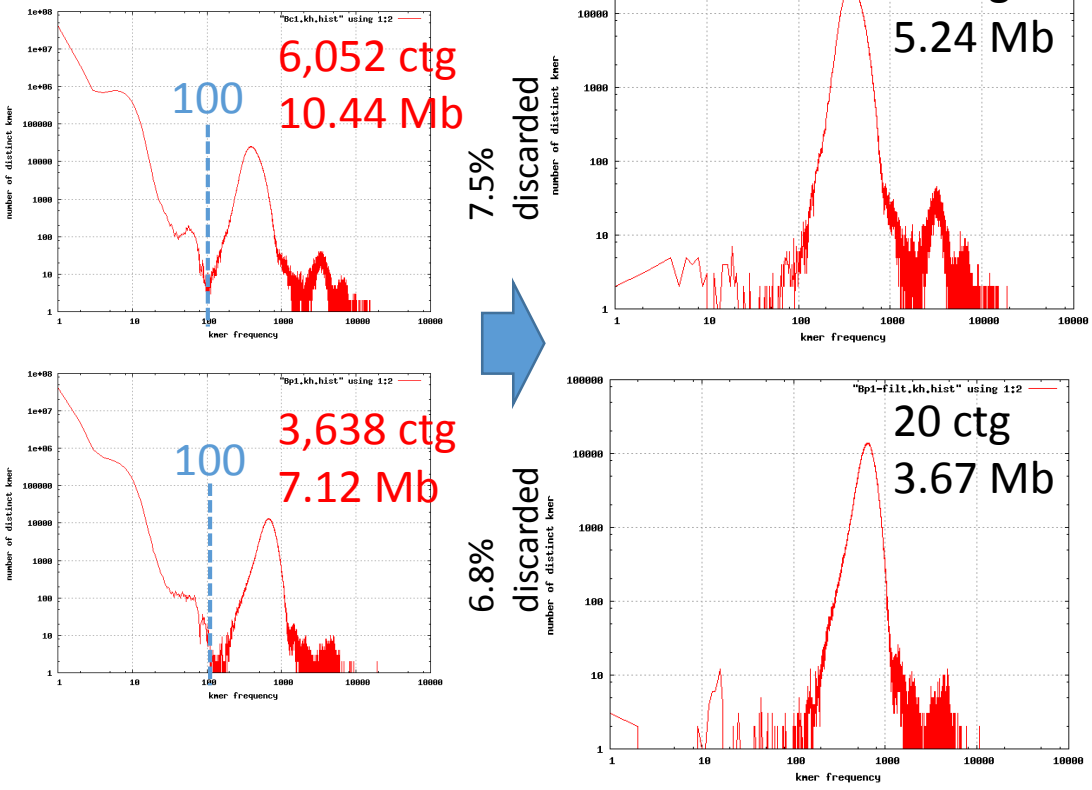
100x read simulation using art_illumina v2.3.7

Effect of pretreatments on k-mer profiles

1. Subsampling



2. Filtering low-abundance k-mers

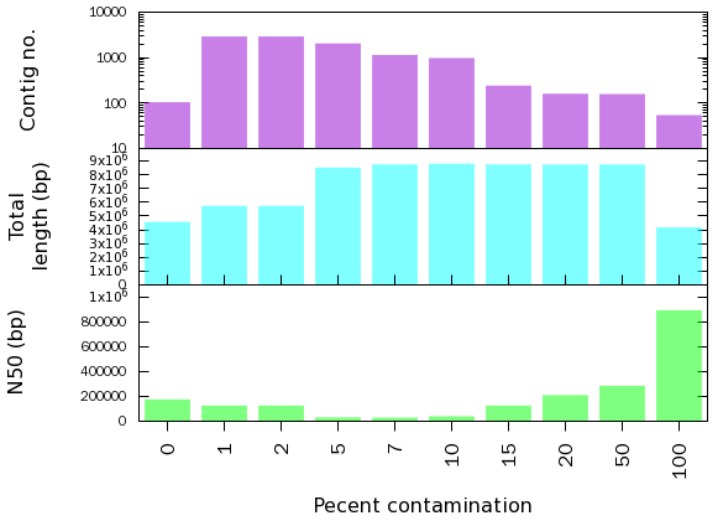


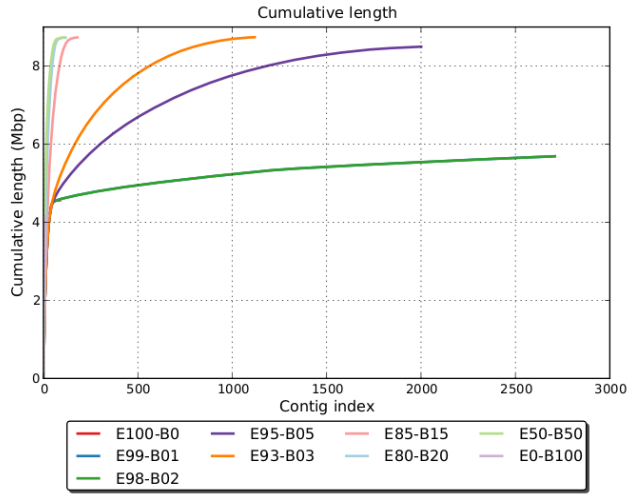
- 5/10/20/40% subsampling by CLC GW
- Analysis by khmer (k-mer = 20)

- [khmer] filter-abund.py (k-mer =21)
- CLC Genomics Workbench (word size 64, bubbles size auto, fast mode

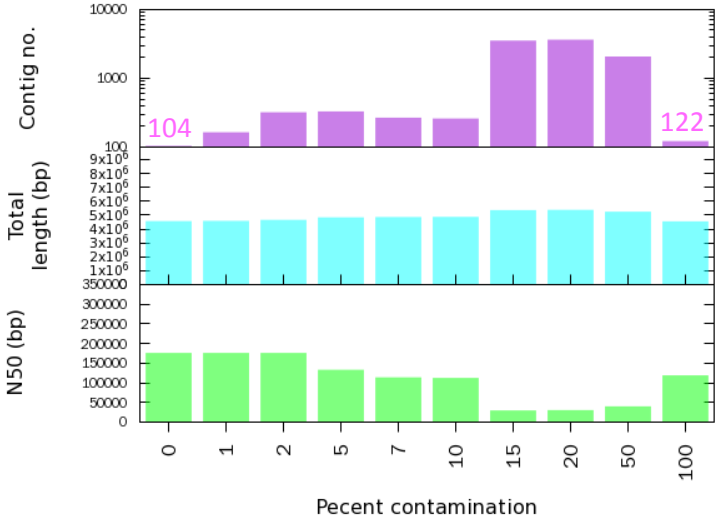
Assembly comparison using simulated data

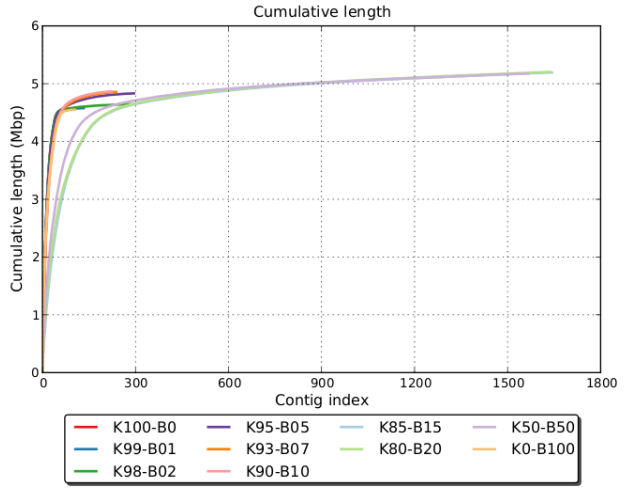
1. Species level: *Escherichia coli* K12 MG1655 vs. *Bacillus subtilis* 168





2. Strain level: *Escherichia coli* K12 MG1655 vs. *Escherichia coli* B REL606

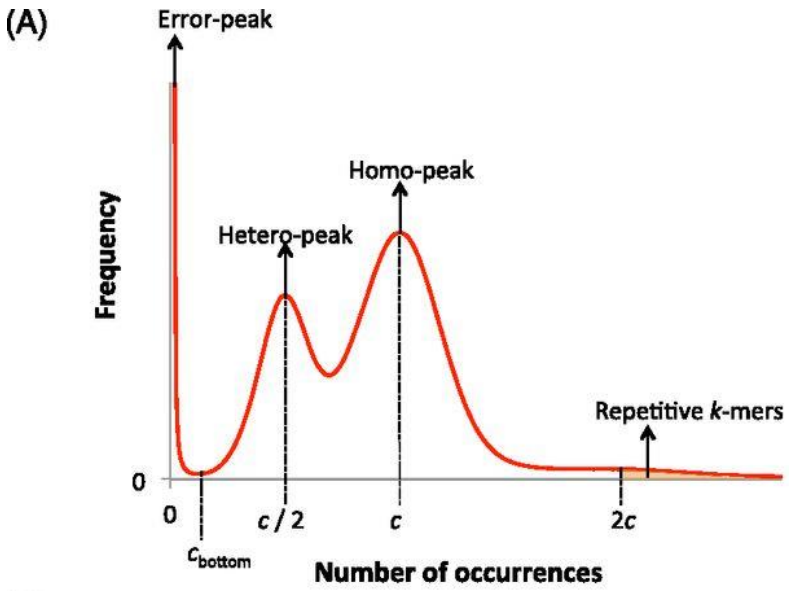




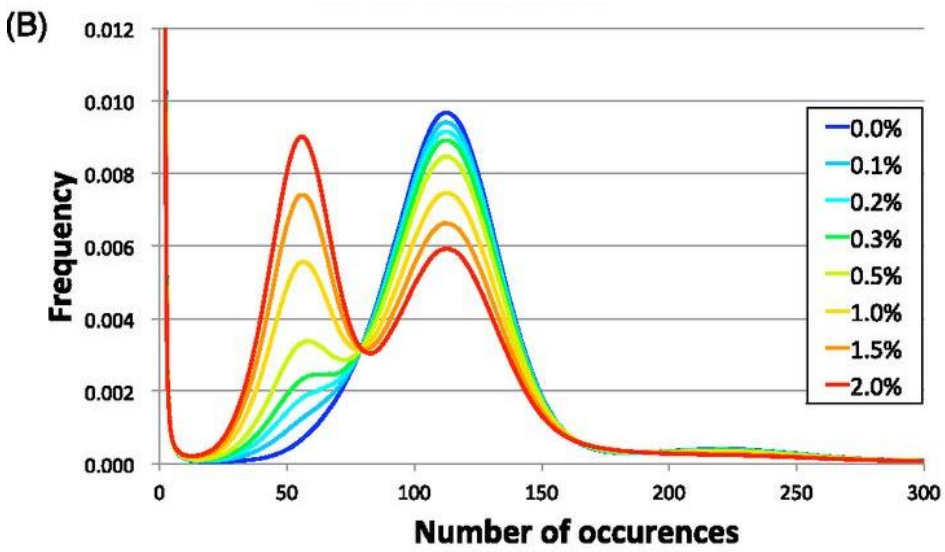
SPAdes v3.5.0

17-mer distribution in heterozygous (diploid) genomes

Schematic model



Simulated heterozygous data from *C. elegans*



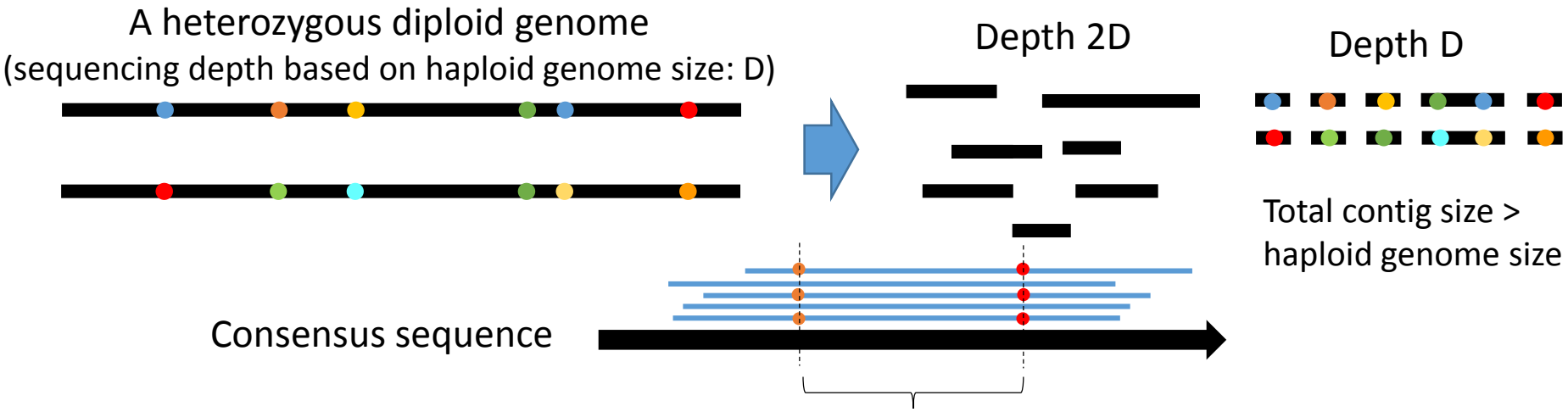
Genome Res. 2014. 24: 1384-1395

A diploid genome, GENOME = (GENOME$_1$ ∪ GENOME$_2$), can be viewed as two similar double-stranded haplomes, GENOME$_1$ and GENOME$_2$. Typically, differences between haplomes are represented as a collection of SNPs and short indels. Given a pairwise alignment, we use *percent identity* (percent of matches among all columns of the alignment) to measure similarity between sequences.

Heterozygosity or polymorphic rate in percent

J. Comput. Biol. 2015. 6:5280545

Checking for diploidity using variant calling



Calculation of percent heterozygosity

Samples	contig length	Basic variant calling				Fixed ploidy variant calling		Heterozygosity
		ploidy=1		ploidy=2		ploidy=2		
		all	hetero	all	hetero	all	hetero	
Species 5%	N.A.	58	34	58	34	85	**61**	N.A.
Strain 50%	5,742,551	11328	11162	11328	11162	11911	**11798**	0.21%.
E. coli K-12	4,616,069	55	31/55	55	31	71	**47**	1.01818E-05
BY25573	10,813,448	99	73	99	73	168	**153**	1.4149E-05
Cz (no filter)	9,372,581	238	167	236	167	379	**331**	3.53158E-05
Cz (filter)	9,372,581	208	148	208	148	352	**306**	3.26484E-05
KCTC7118	11,734,634	850	717	850	717	1285	**1201**	0.01%
KCTC7524	11,162,302	9751	9581	9751	9581	31717	**31614**	0.28%
E	11,689,042	7044	6896	7044	6896	8139	**8067**	0.07%
J	11,670,968	7004	6802	7004	6802	7885	**7727**	0.07%

CLC Genomics Workbench 8.5.1

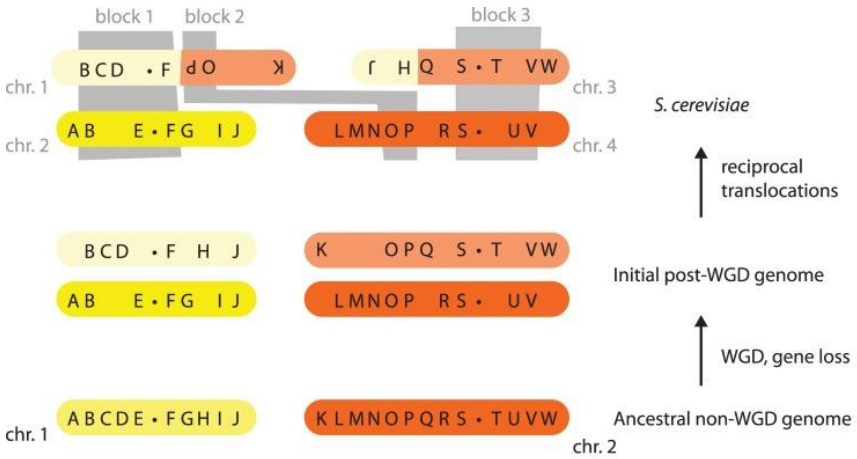
Simulation results from yeast with known ploidy

Species	*De novo* assembly	Basic variant calling (heterozygous/all)		Fixed ploidy variant calling (heterozygous/all)	
		Ploidy = 1	Ploidy = 2	Ploidy = 1	Ploidy = 2
***Saccharomyces cerevisae* S288C haploid**	362 contigs 11,662,944	403/628	403/628	4/316	590/798
***Candida albicans* SC5314 heterozygous diploid**	6,170 contigs 15,724,733 (> haploid size)	42,024/42,568	42,024/42,568	48/2,766	43,874/44,264 Heterozygosity estimate: 0.28%

- *Saccharomyces cerevisiae* S288C: 17 chromosomes + mitochondrion (12,157,105 bp)
- *Candida albicans* SC5314 assembly 22: (8 chromosomes x 2) + mitochondrion (28,605,418 bp)

Candida albicans

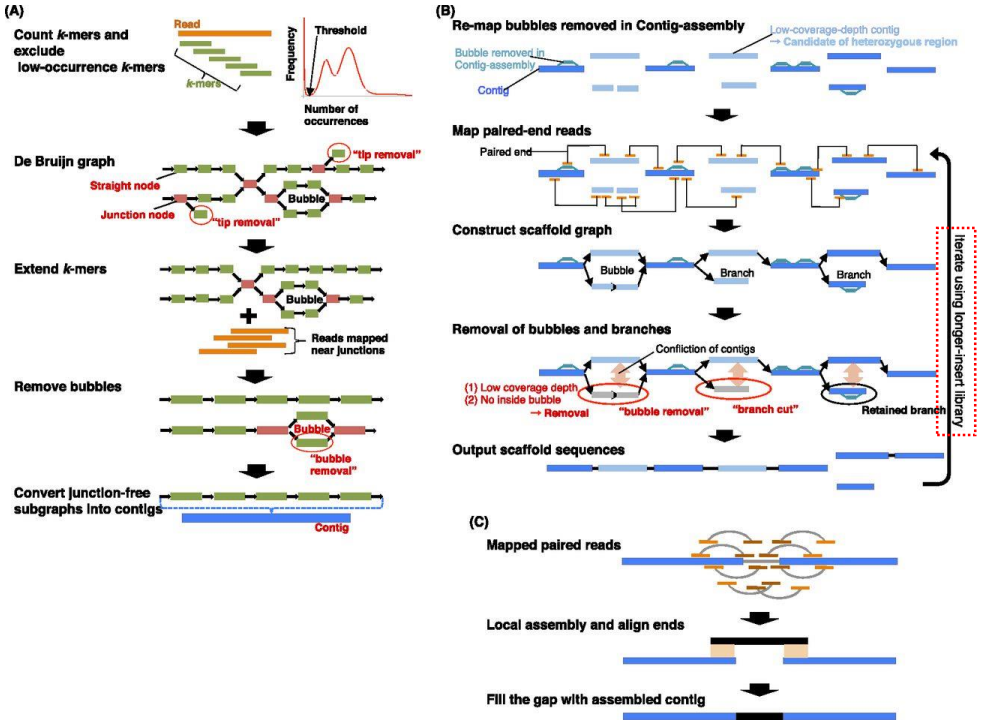
- A pre-whole genome duplication (WGD) yeast (including *Kluyveromyces lactis*)
- The heterozygosity far exceeds that found in other polymorphic genomes such as human and *Anopheles* and is widespread among the clinical isolates
- 4.21 polymorphisms per Kb (Proc Natl Acad Sci USA 2004, 101: 7329)



PLoS Biol. 2015 13(8): e1002221

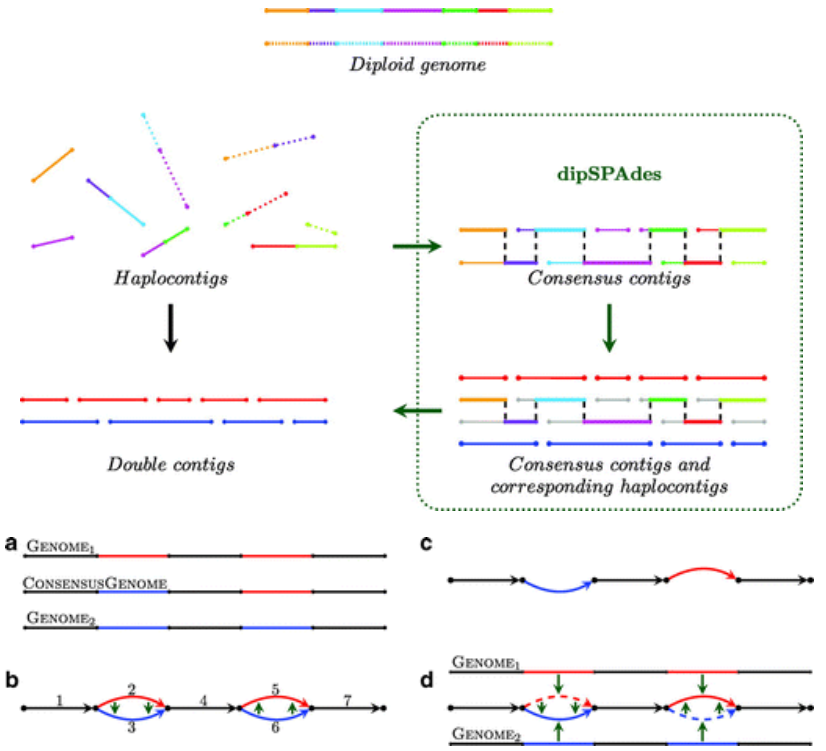
How to assemble diploid genomes?

Platanus
Genome Res. (2014)



Separate contigs are constructed from each haplotype (highly heterozygous regions rich in SNVs and structural variations); dependent on long-insert library.

dipSPAdes
J. Comput. Biol. (2015)



Consensus contigs + haplocontigs

Heterozygosity(%)와 polymorphic rate가 염색체 및 집단 수준에서 혼용되고 있음

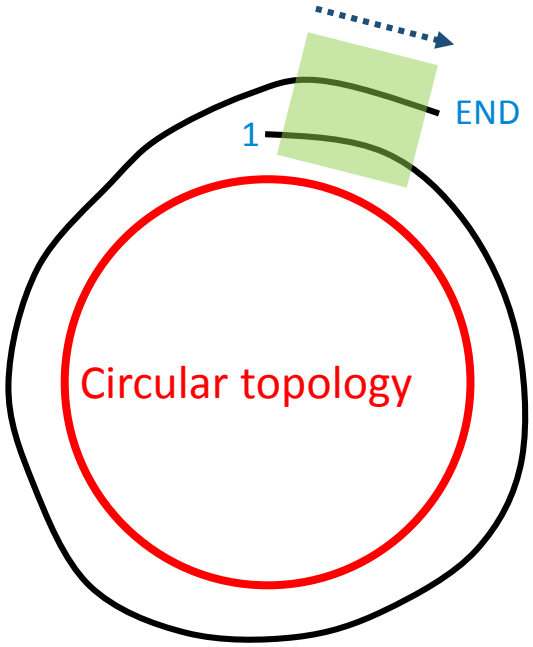
SMRT analysis 활용을 위한 자료

- https://github.com/PacificBiosciences/SMRT-Analysis
- https://github.com/PacificBiosciences/SMRT-Analysis/wiki/Official-Documentation
- http://www.pacificbiosciences.com/Tutorials/Bacterial_Assembly_Epigenetic_Analysis_HGAP/story.html (Video tutorial)
- https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/Evaluating-Assemblies
- Jon Badalamenti "Approaches for analyzing and assembling PacBio single-molecule real-time (SMRT) sequencing data", https://www.msi.umn.edu/sites/default/files/Badalamenti_PacBio_tutorial_12-10-2014.pdf
  - Always run with 100x coverage of longest reads
  - Tune key parameters (minimum subread length, minimum polymerase read quality, and anticipated genome size)
  - Checkpoints for final assembly:
    - Check coverage plots
    - Check for plasmids
    - BLAST any small contigs
    - Try HGAP.2
    - Quiver is not perfect

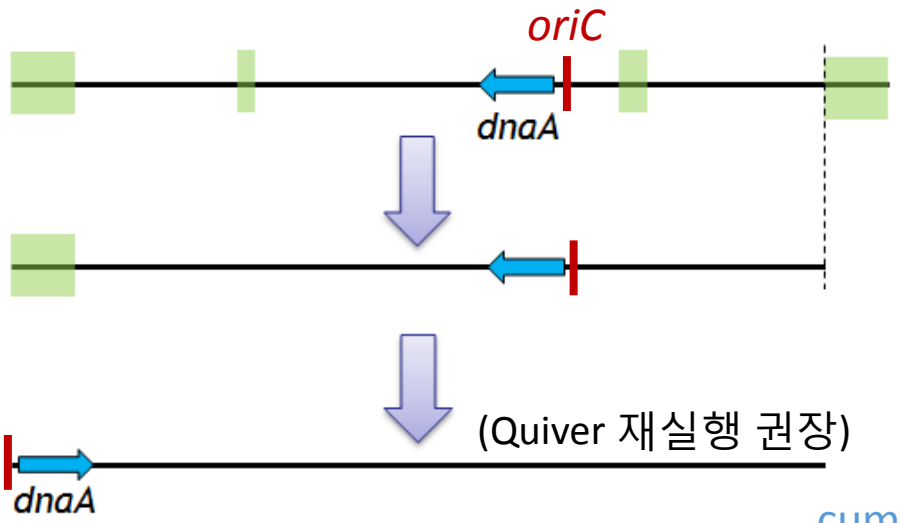
HGAP.3	HGAP.2	HGAP.1
RS_HGAP_Assembly.3	RS_HGAP_Assembly.2	RS_HGAP_Assembly.1
New in SMRT Analysis 2.2		Deprecated in SMRT Analysis 2.2
Performance improvements by replacing the consensus step with pbutgcns	Performance improvements by replacing the correction step with pbdagcon	Initial HGAP production implementation

Single contig의 후처리

1. Terminal overlap 제거
2. *oriC*를 기준으로 재정렬
3. 남아있는 low quality base 개선



말단은 염기서열 quality가 일반적으로 나쁘므로, 뒷부분은 정확히 +1 position부터 맞지 않을 수도 있고, 100% 동일하지 않을 수도 있음

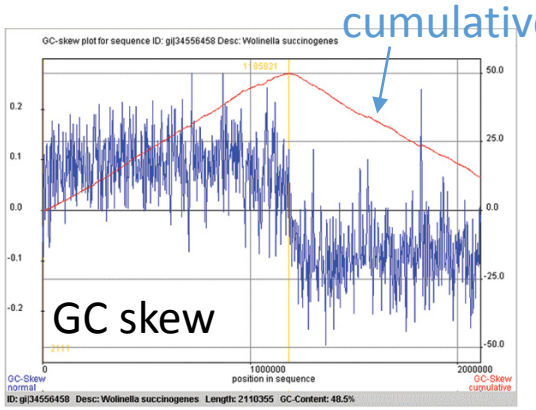


기타 처리할 내용:

- **신뢰하기 곤란한 짧은 contig의 제거**
- **일루미나 read가 존재할 경우 추가적인 염기 단위의 교정**

oriC 영역 예측의 근거:

- dnaA 유전자의 위치
- GC skew: (G-C)/(G+C)
- 타 genome과의 비교



Species identification using genome data

Genome sequence

Prokka or
RAST annotation

1. 16S rRNA gene
2. Genes & proteins
3. Closest neighbors (RAST)

EzTaxon (16S rRNA gene similarity >97%*)

specI (http://vm-lux.embl.de/~mende/specI/)
- based on 40 universal single-copy marker genes

A list of closest neighbors with available genome sequences

ANI (95~96%)

- JSpecies
- ANI.pl (https://github.com/chjp/ANI/blob/master/ANI.pl) or other ANI calculators

or

DDH estimate using the Genome-to Genome Distance Calculator (http://ggdc.dsmz.de/distcalc2.php)

*Pairwise 16S rRNA gene sequence similarity >97% resulted in DDH values ≥70%

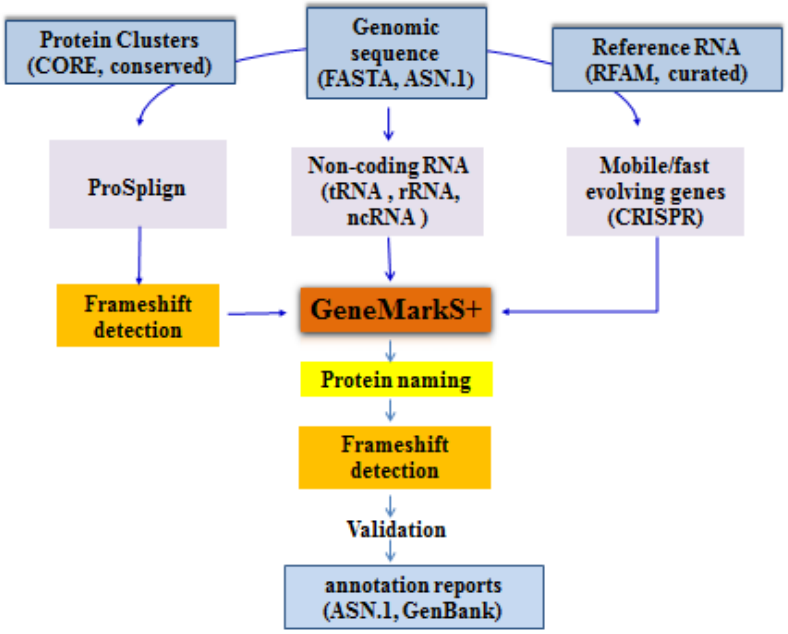
유전체 주석화 서비스

	NCBI	RAST	JGI
Database	GenBank (INSDC)	Seed	IMG/M
Annotation	PGAP	RAST server	IMG-ER
프로젝트 사전 등록	필요(BioProject)	불필요	필요(GOLD)
제출 정보	Complete Genome WGS Genome	FASTA file 등록(gene prediction 포함 가능)	FASTA file 등록 (gene prediction 포함 가능)
비공개 상태의 정보 활용	불가능	가능(공유 포함)	가능

- **PGAP** http://www.ncbi.nlm.nih.gov/genome/annotation_prok/
- **RAST server** http://rast.nmpdr.org/
- **IMG-ER** https://img.jgi.doe.gov/cgi-bin/mer/main.cgi
- **Prokka** (standalone tool) https://github.com/tseemann/prokka [PMID: 24642063]

Prokaryotic Genome Annotation Pipeline (PGAP)

- http://www.ncbi.nlm.nih.gov/genome/annotation_prok/
- http://www.ncbi.nlm.nih.gov/books/NBK174280/ (detailed description)
- Minimum standards for complete genomes
  - Structural RNA (5S, 16S, 12S): at least one copy of each with appropriate length
  - tRNA: at least one copy for each amino acids
  - # protein coding genes/genome length ratio is closet to 1
  - No genes completely contained in another gene on the same or opposite strand
  - No partial features



Release note (2015/09/17)

May 2013 Version 2.0

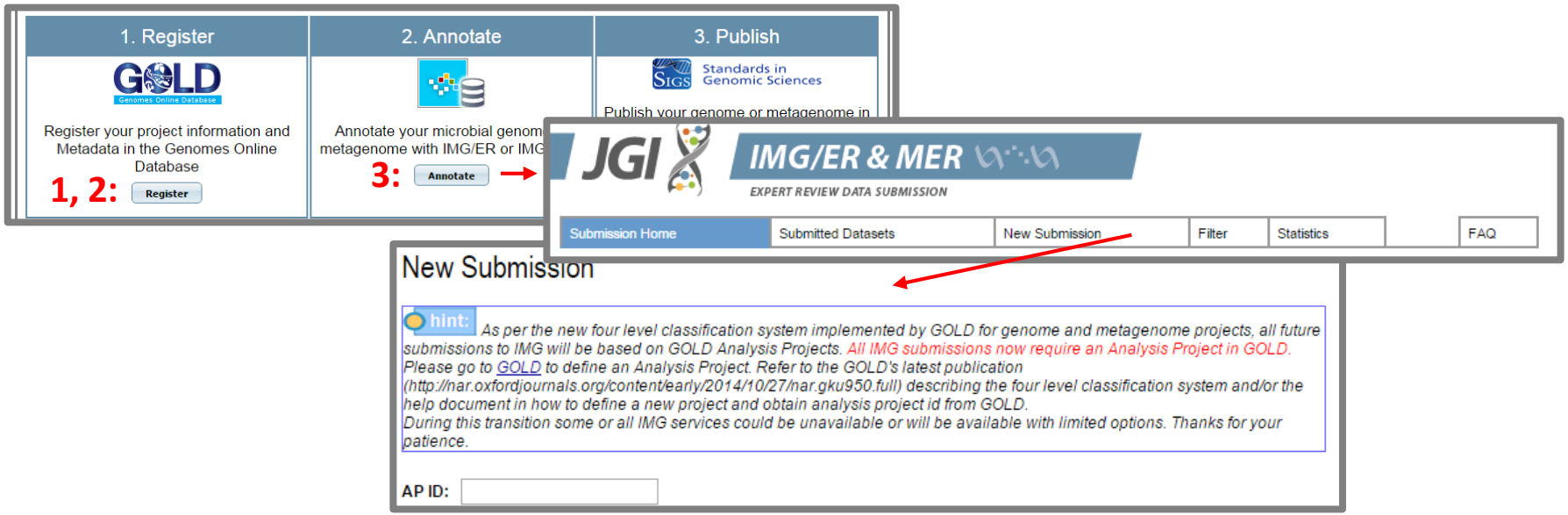
Version 2.0 uses protein homology and GeneMarkS+ prediction program.

Features annotated: Gene; CDS; rRNA; tRNA; repeats in CRISPR region

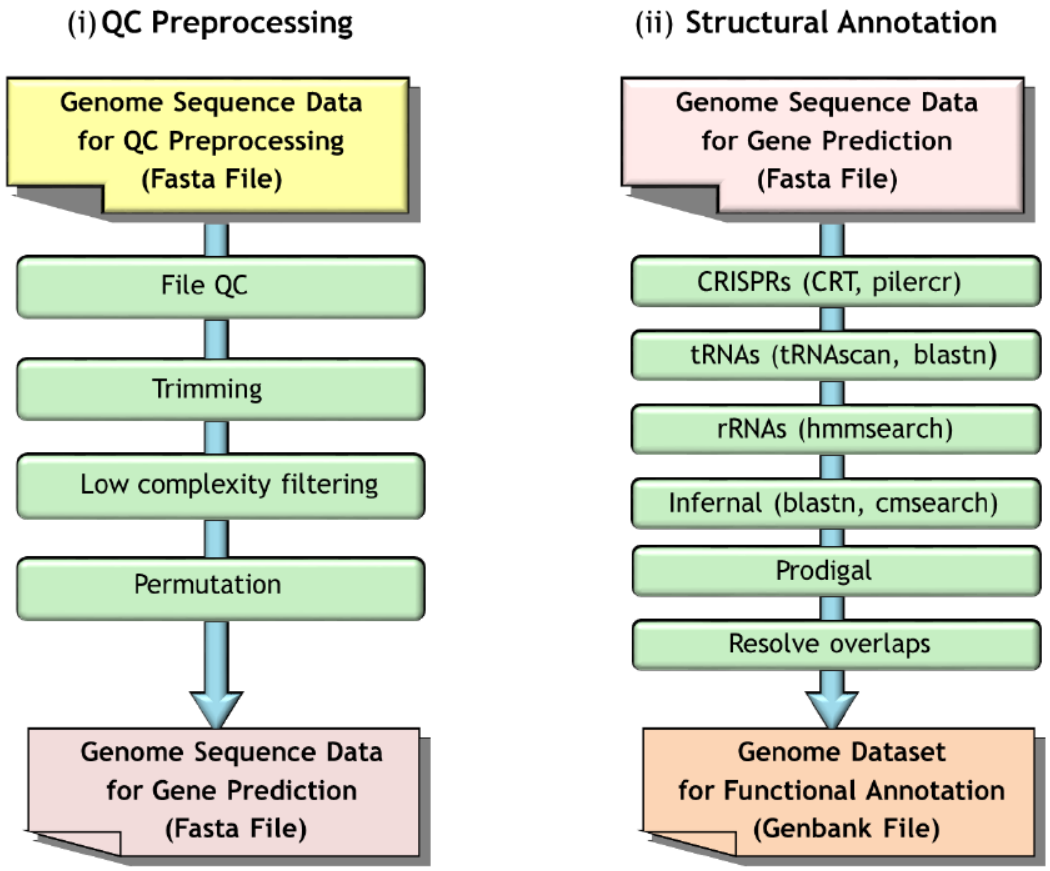
This version does not include: small non-coding RNA (ncRNA)

IMG/ER

- IMG (Integrated Microbial Genomes) system is a community resource for analysis and annotation of genomes and metagenomes
- IMG/ER (IMG Expert Review) provides users with tools for analyzing their private genome/metagenome dataset [Nucleic Acids Res. 2014, 42:D560]
- Genome annotation procedure:
  1. Project registration is required at GOLD https://gold.jgi.doe.gov/
  2. Create a new Analysis Project (AP) for submission to IMG
  3. Submit dataset to IMG (select an AP ID)



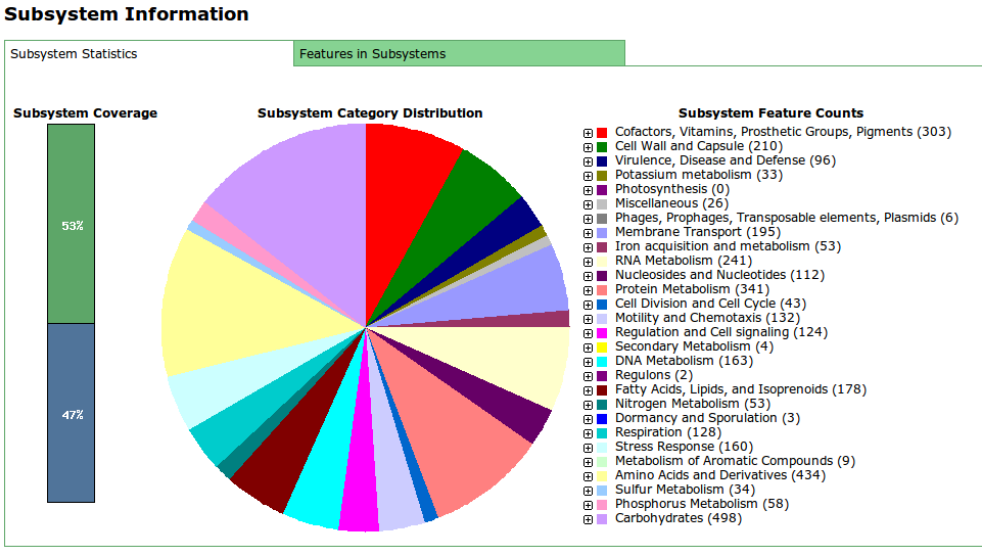
DOE-JGI SOP for microbial genome annotation

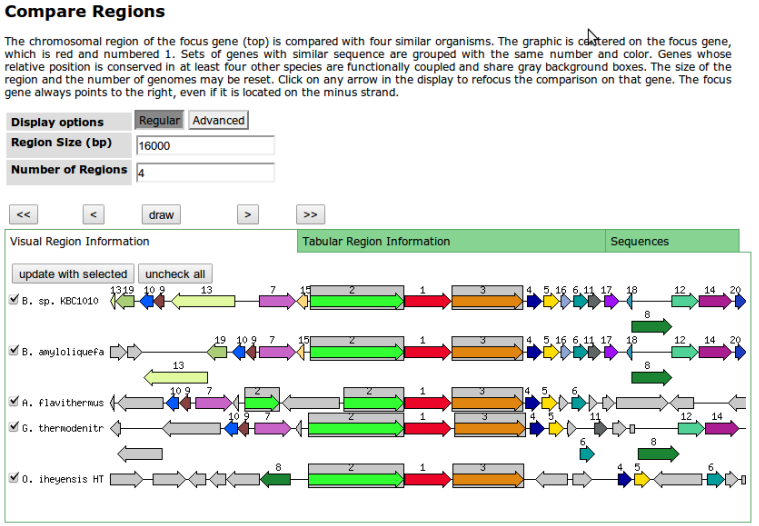


http://img.jgi.doe.gov/w/doc/MGAandDI_SOP.pdf

RAST annotation server

- http://rast.nmpdr.org/
- It leverages the data and procedures established within the SEED framework to provide automated high quality gene calling and functional annotation
- The service normally makes the annotated genome available within 12-24 hours of submission
- Annotation results can be shared by specifying registered users





유전체 정보 등록 절차(NCBI)

1. BioProject 등록(필수)
   - 균주 분리 장소 및 일시, 연구의 목적 등
2. BioSamples 및 raw data의 SRA 등록(권장)
3. **WGS** https://submit.ncbi.nlm.nih.gov/subs/wgs/
4. **Complete genome sequence**
   - Guide: http://www.ncbi.nlm.nih.gov/genbank/genomesubmit/
   - Submitter template file 생성: sequin 혹은 https://submit.ncbi.nlm.nih.gov/genbank/template/submission/
   - Genome sequence file의 확장자는 .fsa로 지정
   - ASN.1 파일 생성: tbl2asn -p path_to_files -t template -M n -Z discrep -j "[organism=Clostridium difficile] [strain=ABDC] [gcode=11]"
   - 오류가 없다면 GenomesMacroSend에서 파일 전송: http://www.ncbi.nlm.nih.gov/projects/GenomeSubmit/genome_submit.cgi
5. 공개 일자는 등록일 기준 최장 5년 후까지 지정 가능하며, 공개일 이전이라도 accession no.를 언급한 논문이 출판되면 즉시 공개됨

*논문 투고 시점에 유전체 정보가 public DB에서 먼저 공개되기를 요구하는 학술지도 있음

Genome sequencing has come of age, and genomics will become central to microbiology's future. It may appear at the moment that the human genome is the main focus and primary goal of genome sequencing, but do not be deceived. The real justification in the long run, is **microbial genomics**.

Carl Woese, 1998-2012