Next-generation Sequencing Data Analysis

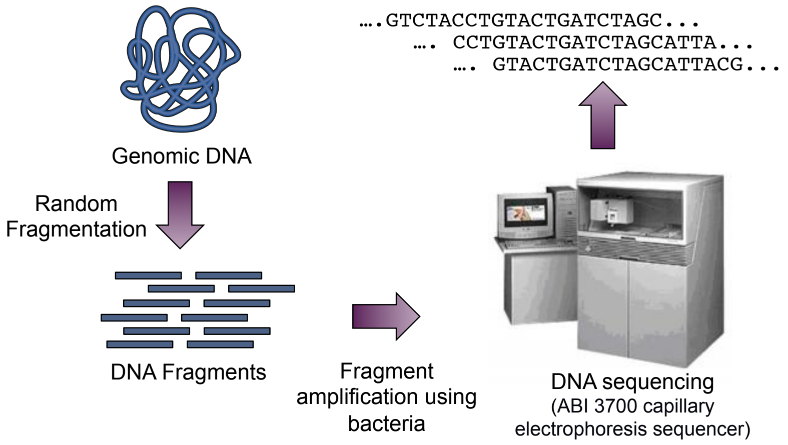
CSBio 2015 Pre-conference workshop: 22 November 2015

CSBio 2015 Bangkok, Thailand

Weerayuth Kittichotirat
weerayuth.kit@kmutt.ac.th

Systems biology and Bioinformatics Research Group
Pilot Plant Development and Training Institute
King Mongkut's University of Technology Thonburi

Overview of the previous generation of DNA sequencing



….GTCTACCTGTACTGATCTAGC...
…. CCTGTACTGATCTAGCATTA...
…. GTACTGATCTAGCATTACG...

Genomic DNA

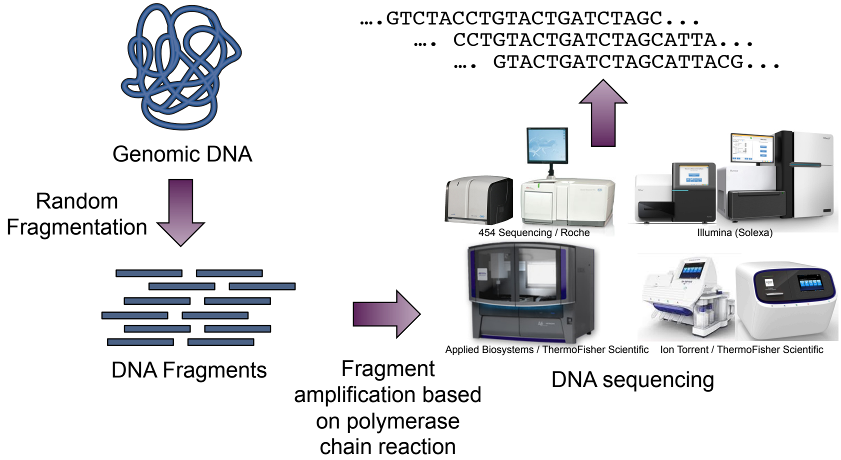
Random Fragmentation

DNA Fragments

Fragment amplification using bacteria

DNA sequencing (ABI 3700 capillary electrophoresis sequencer)

Overview of the next generation DNA sequencing



….GTCTACCTGTACTGATCTAGC...
…. CCTGTACTGATCTAGCATTA...
…. GTACTGATCTAGCATTACG...

Genomic DNA

Random Fragmentation

DNA Fragments

Fragment amplification based on polymerase chain reaction

454 Sequencing / Roche

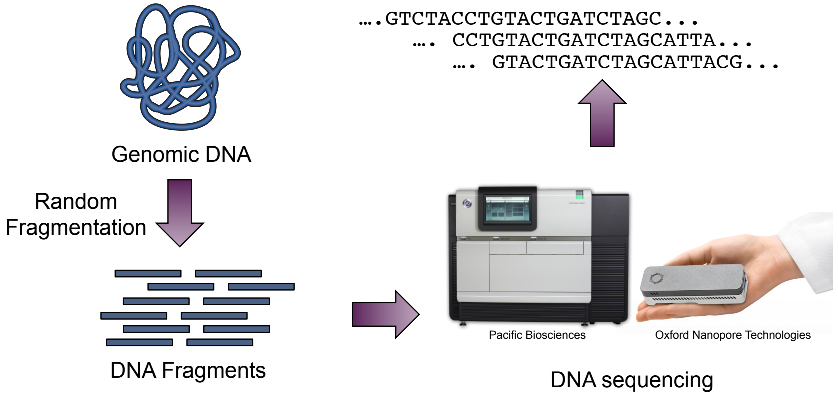
Illumina (Solexa)

Applied Biosystems / ThermoFisher Scientific

Ion Torrent / ThermoFisher Scientific

DNA sequencing

Overview of the third generation DNA sequencing



….GTCTACCTGTACTGATCTAGC...
…. CCTGTACTGATCTAGCATTA...
…. GTACTGATCTAGCATTACG...

Genomic DNA

Random Fragmentation

DNA Fragments

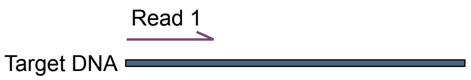
Pacific Biosciences

Oxford Nanopore Technologies

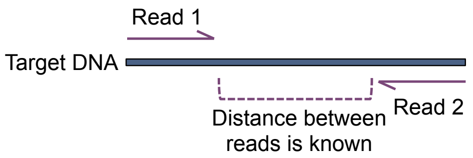
DNA sequencing

Sequencing run types

- Single-end sequencing

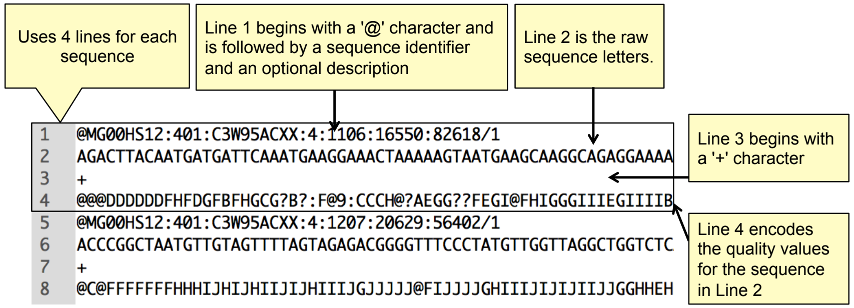


- Paired-end sequencing



FASTQ file format

- FASTQ format is a text-based format for storing both
  - a biological sequence (usually nucleotide sequence)
  - and its corresponding quality scores[1].



Uses 4 lines for each sequence

Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description

Line 2 is the raw sequence letters.

Line 3 begins with a '+' character

Line 4 encodes the quality values for the sequence in Line 2

```
1  @MG00HS12:401:C3W95ACXX:4:1106:16550:82618/1
2  AGACTTACAATGATGATTCAAATGAAGGAAACTAAAAAGTAATGAAGCAAGGCAGAGGAAAA
3  +
4  @@@DDDDDDFHFDGFBFHGCG?B?:F@9:CCCH@?AEGG??FEGI@FHIGGGIIIEGIIIIB
5  @MG00HS12:401:C3W95ACXX:4:1207:20629:56402/1
6  ACCCGGCTAATGTTGTAGTTTTAGTAGAGACGGGGTTTCCCTATGTTGGTTAGGCTGGTCTC
7  +
8  @C@FFFFFFFHHHIJHIJHIIJIJHIIIJGJJJJJ@FIJJJJGHIIIJIJIJIIJJGGHHEH
```

- Quality values are encoded using ASCII scheme

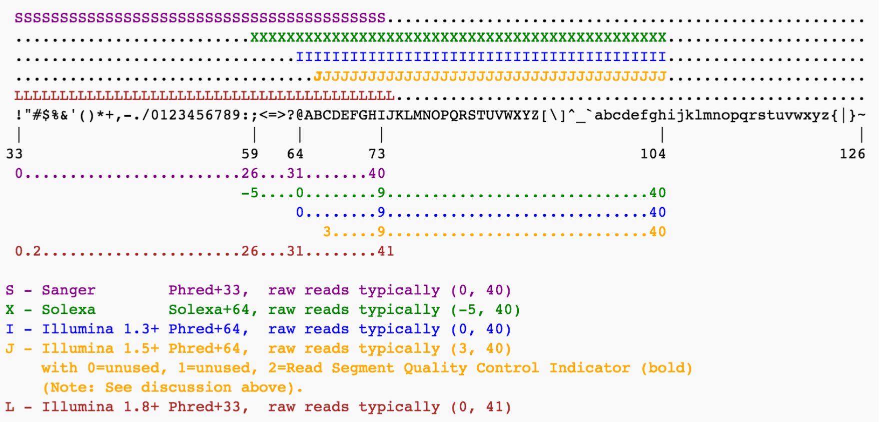
```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

Lowest quality value

Highest quality value

[1] https://en.wikipedia.org/wiki/FASTQ_format

Quality values are encoded differently for different platform



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./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)
```

https://en.wikipedia.org/wiki/FASTQ_format

Quality value

- A quality value Q is an integer mapping of p (i.e., the probability that the corresponding base call is incorrect)[1]

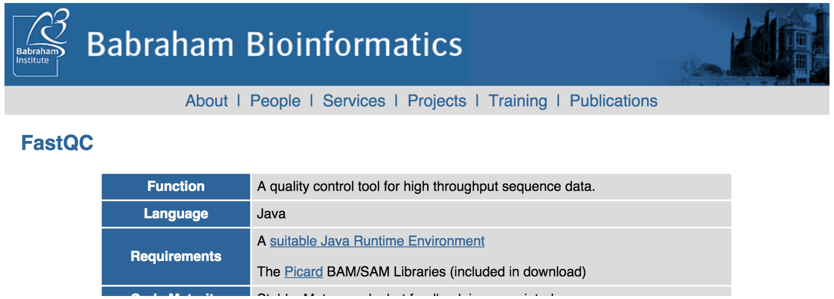
$$Q_{\text{sanger}} = -10 \log_{10} p$$

Quality score	Probability of incorrect bases	Base call accuracy
10	1 in 10	90%
17	1 in 50	98%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%

[1] https://en.wikipedia.org/wiki/FASTQ_format

Quality control check using FastQC

- A quality control tool for high throughput sequence data.
- Developed by the Bioinformatics Group at the Babraham Institute, United Kingdom
- Available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/



Babraham Bioinformatics

About | People | Services | Projects | Training | Publications

FastQC

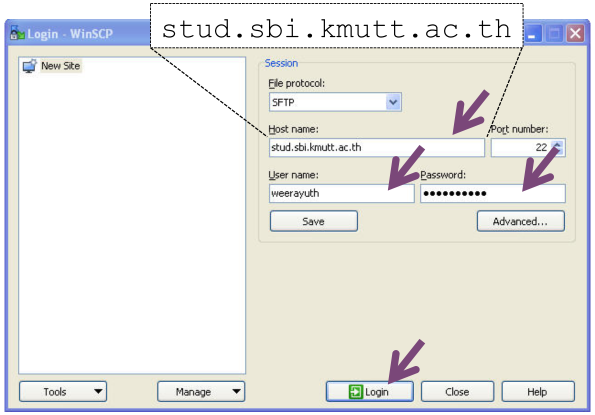
Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)

Quality control check using FastQC

- Transfer raw sequencing data on to the Linux server

1) Open WinSCP

2) Fill in Host Name, User name, Password, and click Login



stud.sbi.kmutt.ac.th

Login - WinSCP

New Site

Session

File protocol: SFTP

Host name: stud.sbi.kmutt.ac.th

Port number: 22

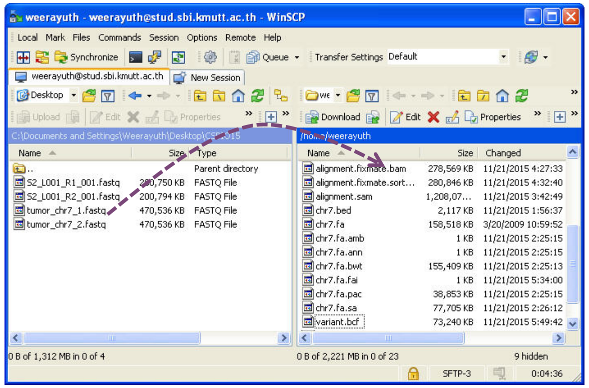
User name: weerayuth

Password: ••••••••••

Quality control check using FastQC

- Transfer raw sequencing data on to the Linux server

3) Click and Drop raw sequence read files from the left to right panel



weerayuth - weerayuth@stud.sbi.kmutt.ac.th - WinSCP

Local Mark Files Commands Session Options Remote Help
Synchronize
Queue
Transfer Settings Default
weerayuth@stud.sbi.kmutt.ac.th
New Session
Desktop
Upload
Edit
Properties
Download
Edit
Properties

C:\Documents and Settings\Weerayuth\Desktop\CSB2015

Name	Size	Type
..		Parent directory
S2_L001_R1_001.fastq	200,750 KB	FASTQ File
S2_L001_R2_001.fastq	200,794 KB	FASTQ File
tumor_chr7_1.fastq	470,536 KB	FASTQ File
tumor_chr7_2.fastq	470,536 KB	FASTQ File

/home/weerayuth

Name	Size	Changed
alignment.fixmate.bam	278,569 KB	11/21/2015 4:27:33
alignment.fixmate.sort...	280,846 KB	11/21/2015 4:32:40
alignment.sam	1,208,07...	11/21/2015 3:42:49
chr7.bed	2,117 KB	11/21/2015 1:56:37
chr7.fa	158,518 KB	3/20/2009 10:59:52
chr7.fa.amb	1 KB	11/21/2015 2:25:15
chr7.fa.ann	1 KB	11/21/2015 2:25:15
chr7.fa.bwt	155,409 KB	11/21/2015 2:25:13
chr7.fa.fai	1 KB	11/21/2015 5:34:00
chr7.fa.pac	38,853 KB	11/21/2015 2:25:15
chr7.fa.sa	77,705 KB	11/21/2015 2:26:12
variant.bcf	73,240 KB	11/21/2015 5:49:42

0 B of 1,312 MB in 0 of 4

0 B of 2,221 MB in 0 of 23 9 hidden

SFTP-3 0:04:36

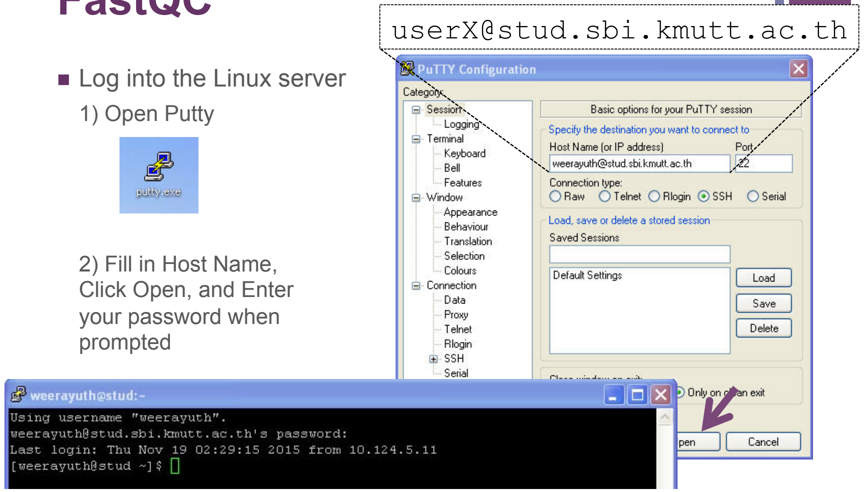
Quality control check using FastQC

- Log into the Linux server

1) Open Putty

putty.exe

2) Fill in Host Name, Click Open, and Enter your password when prompted



userX@stud.sbi.kmutt.ac.th

PuTTY Configuration

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address): weerayuth@stud.sbi.kmutt.ac.th

Port: 22

Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

Load, save or delete a stored session

Saved Sessions

Default Settings

weerayuth@stud:~

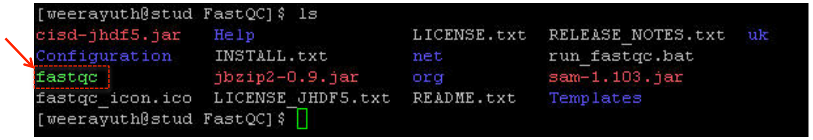
```
Using username "weerayuth".
weerayuth@stud.sbi.kmutt.ac.th's password:
Last login: Thu Nov 19 02:29:15 2015 from 10.124.5.11
[weerayuth@stud ~]$
```

Quality control check using FastQC

- A few useful commands
  - `ls` : Show all files and directories (folders)
  - `mkdir` : Make a new directory (folder)
  - `cd` : Change directory (folder)

Quality control check using FastQC

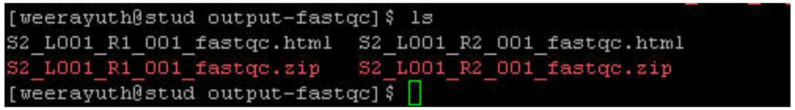
- Download and install **FastQC**
  - `mkdir fastqc`
  - `cd fastqc`
  - `wget http://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.4.zip`
  - `unzip fastqc_v0.11.4.zip`
  - `cd FastQC`
  - `chmod 755 fastqc`
  - `ls`



```
[weerayuth@stud FastQC]$ ls
cisd-jhdf5.jar    Help              LICENSE.txt  RELEASE_NOTES.txt  uk
Configuration     INSTALL.txt       net          run_fastqc.bat
fastqc            jbzip2-0.9.jar    org          sam-1.103.jar
fastqc_icon.ico   LICENSE_JHDF5.txt README.txt   Templates
[weerayuth@stud FastQC]$
```

Quality control check using FastQC

- Run **FastQC** to check quality of sequencing data
  - `cd`
  - `mkdir output-fastqc`
  - `fastqc/FastQC/fastqc S2_L001_R1_001.fastq S2_L001_R2_001.fastq -o output-fastqc`
  - `cd output-fastqc`
  - `ls`



```
[weerayuth@stud output-fastqc]$ ls
S2_L001_R1_001_fastqc.html  S2_L001_R2_001_fastqc.html
S2_L001_R1_001_fastqc.zip   S2_L001_R2_001_fastqc.zip
[weerayuth@stud output-fastqc]$
```

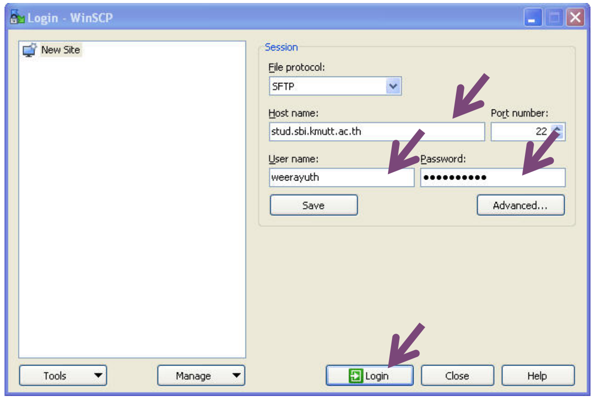
Quality control check using FastQC

- Copy FastQC result back

1) Open WinSCP

WinSCP.exe
WinSCP: SFTP, FTP
Martin Prikryl

2) Fill in Host Name, User name, Password, and click Login



Login - WinSCP

Session

File protocol: SFTP

Host name: stud.sbi.kmutt.ac.th

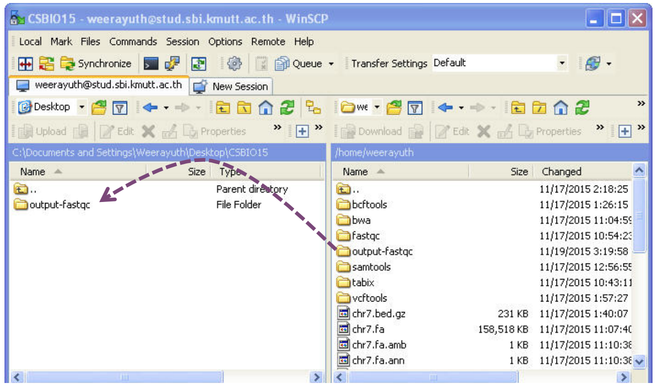
Port number: 22

User name: weerayuth

Password: ••••••••••

+ Quality control check using FastQC

- Copy FastQC result back
  3) Click and Drop folder output-fastqc from the right to left panel



CSBIO15 - weerayuth@stud.sbi.kmutt.ac.th - WinSCP

Local Mark Files Commands Session Options Remote Help
Synchronize
Queue
Transfer Settings Default
weerayuth@stud.sbi.kmutt.ac.th
New Session
Desktop
we
Upload Edit Properties
Download Edit Properties

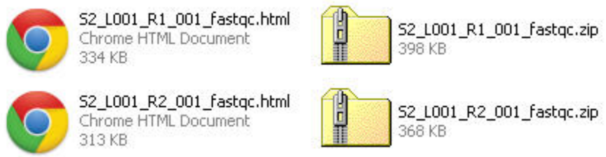
C:\Documents and Settings\Weerayuth\Desktop\CSBIO15

Name	Size	Type
..		Parent directory
output-fastqc		File Folder

/home/weerayuth

Name	Size	Changed
..		11/17/2015 2:18:25
bcftools		11/17/2015 1:26:15
bwa		11/17/2015 11:04:59
fastqc		11/17/2015 10:54:23
output-fastqc		11/19/2015 3:19:58
samtools		11/17/2015 12:56:55
tabix		11/17/2015 10:43:11
vcftools		11/17/2015 1:57:27
chr7.bed.gz	231 KB	11/17/2015 1:40:07
chr7.fa	158,518 KB	11/17/2015 11:07:40
chr7.fa.amb	1 KB	11/17/2015 11:10:38
chr7.fa.ann	1 KB	11/17/2015 11:10:38

+ FastQC output files



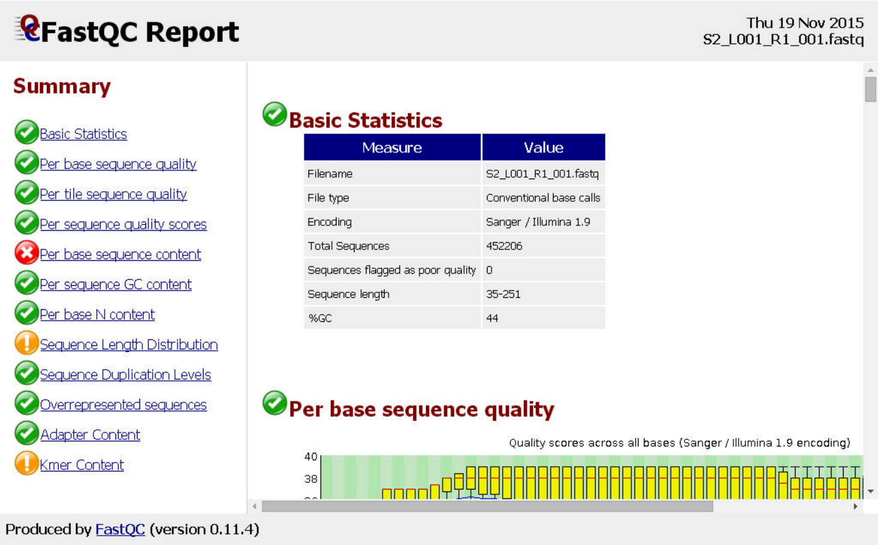
S2_L001_R1_001_fastqc.html
Chrome HTML Document
334 KB

S2_L001_R1_001_fastqc.zip
398 KB

S2_L001_R2_001_fastqc.html
Chrome HTML Document
313 KB

S2_L001_R2_001_fastqc.zip
368 KB

- FastQC generates a HTML report and a zip file containing individual graphs for each input file



FastQC Report

Thu 19 Nov 2015
S2_L001_R1_001.fastq

Summary

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

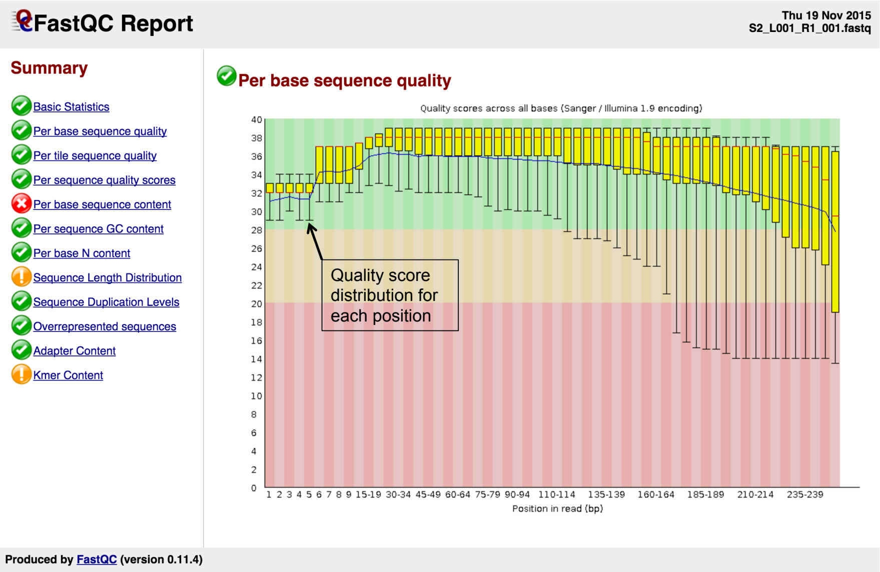
✓ Basic Statistics

Measure	Value
Filename	S2_L001_R1_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	452206
Sequences flagged as poor quality	0
Sequence length	35-251
%GC	44

✓ Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

Produced by FastQC (version 0.11.4)



FastQC Report

Thu 19 Nov 2015
S2_L001_R1_001.fastq

Summary

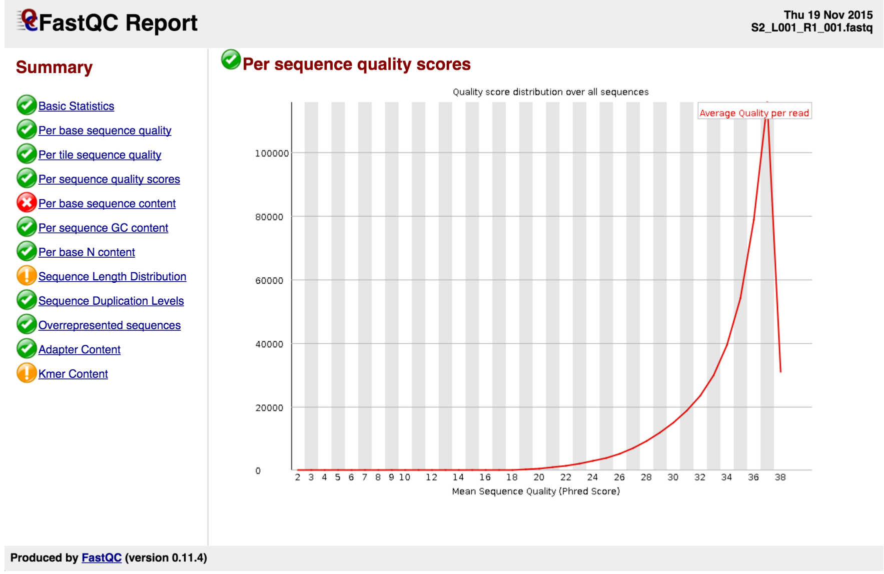
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter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

✓ Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

Quality score distribution for each position

Produced by FastQC (version 0.11.4)



FastQC Report

Thu 19 Nov 2015
S2_L001_R1_001.fastq

Summary

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

✓ **Per sequence quality scores**

Produced by FastQC (version 0.11.4)

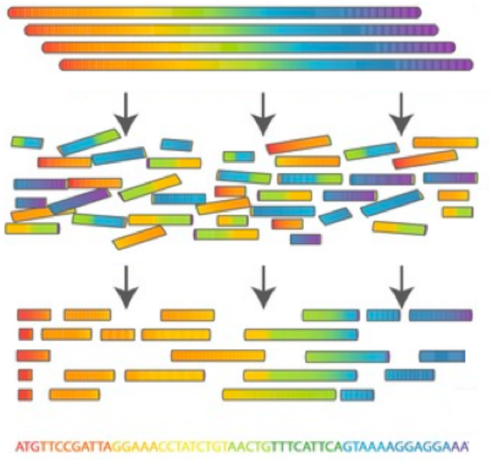
- Explanation of each quality check can be found at
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/

+

Coffee Break

+

De novo genome assembly



Genomic DNA

Randomly fragmented and sequenced

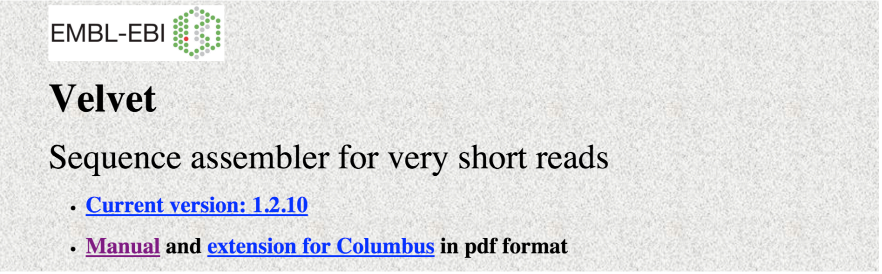
Sequence reads are stitched back based on overlapping sequences into longer contig sequences

http://people.mpi-inf.mpg.de/~sven/images/assembly.png

+

De novo assembly using Velvet

- Velvet is a *de novo* genomic assembler specially designed for short read sequencing technologies
- Developed by Daniel Zerbino and Ewan Birney at the European Bioinformatics Institute (EMBL-EBI), United Kingdom
- Available at https://www.ebi.ac.uk/~zerbino/velvet/



EMBL-EBI

Velvet

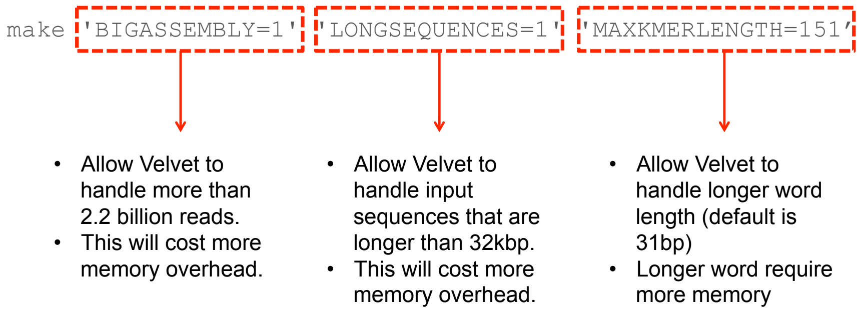
Sequence assembler for very short reads

- **Current version: 1.2.10**
- **Manual and extension for Columbus in pdf format**

De novo assembly using Velvet

- Download and install **Velvet**
  - `cd`
  - `mkdir velvet`
  - `cd velvet`
  - `wget https://www.ebi.ac.uk/~zerbino/velvet/velvet_1.2.10.tgz`
  - `tar -xzf velvet_1.2.10.tgz`
  - `cd velvet_1.2.10`
  - `make 'BIGASSEMBLY=1' 'LONGSEQUENCES=1' 'MAXKMERLENGTH=151’`

De novo assembly using Velvet



`make 'BIGASSEMBLY=1' 'LONGSEQUENCES=1' 'MAXKMERLENGTH=151’`

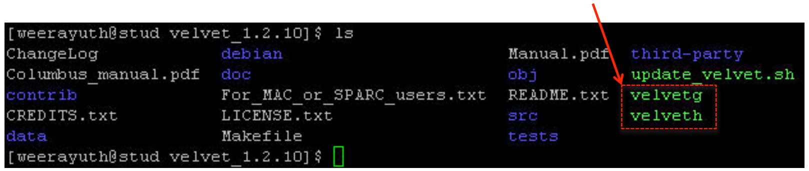
- Allow Velvet to handle more than 2.2 billion reads.
- This will cost more memory overhead.

- Allow Velvet to handle input sequences that are longer than 32kbp.
- This will cost more memory overhead.

- Allow Velvet to handle longer word length (default is 31bp)
- Longer word require more memory

De novo assembly using Velvet

- Download and install **Velvet**
  - `cd`
  - `mkdir velvet`
  - `cd velvet`
  - `wget https://www.ebi.ac.uk/~zerbino/velvet/velvet_1.2.10.tgz`
  - `tar -xzf velvet_1.2.10.tgz`
  - `cd velvet_1.2.10`
  - `make 'BIGASSEMBLY=1' 'LONGSEQUENCES=1' 'MAXKMERLENGTH=151'`
  - `ls`



```
[weerayuth@stud velvet_1.2.10]$ ls
ChangeLog            debian                       Manual.pdf   third-party
Columbus_manual.pdf  doc                          obj          update_velvet.sh
contrib              For_MAC_or_SPARC_users.txt   README.txt   velvetg
CREDITS.txt          LICENSE.txt                  src          velveth
data                 Makefile                     tests
[weerayuth@stud velvet_1.2.10]$
```

De novo assembly using Velvet

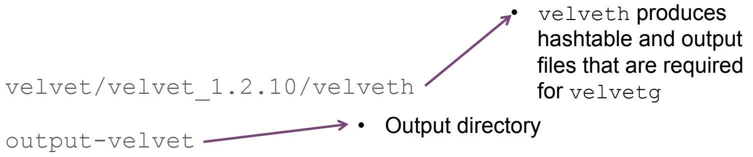
- Running *de novo* assembly
  - `cd`
  - `mkdir output-velvet`
  - `velvet/velvet_1.2.10/velveth output-velvet 83 -fastq -shortPaired -separate S2_L001_R1_001.fastq S2_L001_R2_001.fastq`

+ *De novo* assembly using Velvet

`velvet/velvet_1.2.10/velveth`

- `velveth` produces hashtable and output files that are required for `velvetg`

+ *De novo* assembly using Velvet



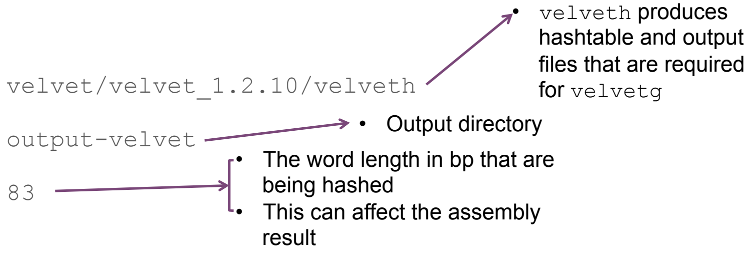
`velvet/velvet_1.2.10/velveth`

- `velveth` produces hashtable and output files that are required for `velvetg`

`output-velvet`

- Output directory

+ *De novo* assembly using Velvet



`velvet/velvet_1.2.10/velveth`

- `velveth` produces hashtable and output files that are required for `velvetg`

`output-velvet`

- Output directory

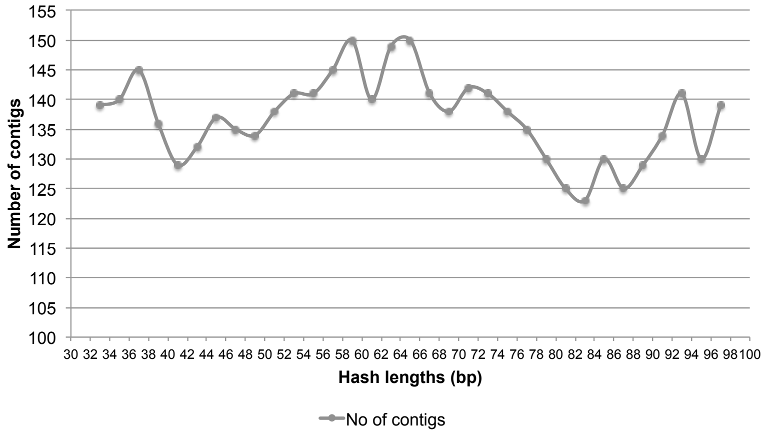
`83`

- The word length in bp that are being hashed
- This can affect the assembly result

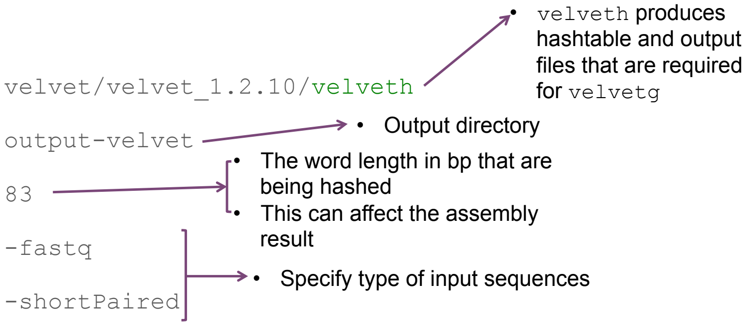
+ Choice of hash length

- It must be an odd number. If an even number is entered, Velvet will just decrement it and proceed
- It must be below or equal to MAXKMERHASH length
- It must be shorter than the read length, otherwise you simply will not observe any overlaps between reads
- Longer hash length allows more specific overlap but fewer reads will be used in the assembly resulting in a decrease coverage
- Shorter hash length allows more reads to be used in the assembly and result in an increase in sensitivity and coverage but will also introduce more errors and higher computation overhead
- Choice of hash length can affect the assembly output and therefore tests with different lengths are usually carried out to find the length that works best

+ Choice of hash length



+ *De novo* assembly using Velvet



```
velvet/velvet_1.2.10/velveth
output-velvet
83
-fastq
-shortPaired
```

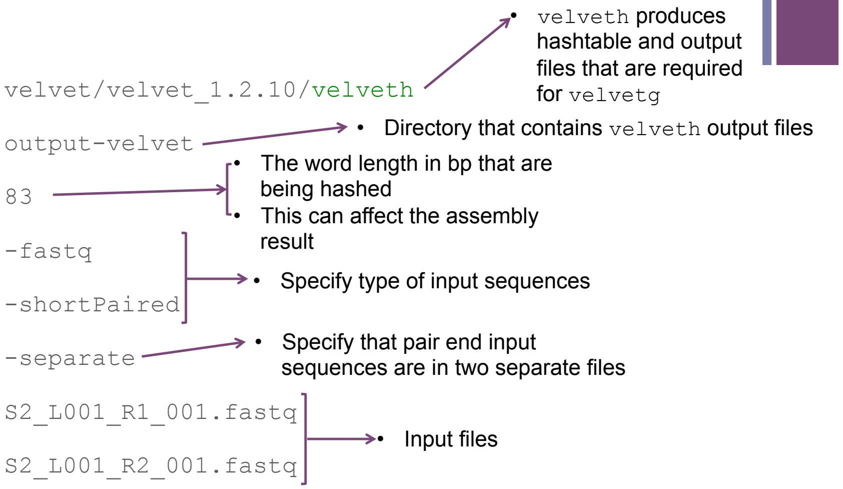
- `velveth` produces hashtable and output files that are required for `velvetg`
- Output directory
- The word length in bp that are being hashed
- This can affect the assembly result
- Specify type of input sequences

+ Types of input sequences

Supported file formats are:	Read categories are:
fasta (default)	**short** (default)
fastq	**shortPaired**
fasta.gz	**short2** (same as short, but for a separate insert-size library)
fastq.gz	**shortPaired2** (see above)
sam	**long** (for Sanger, 454 or even reference sequences)
bam	
eland	
gerald	

https://www.ebi.ac.uk/~zerbino/velvet/Manual.pdf

+ *De novo* assembly using Velvet



```
velvet/velvet_1.2.10/velveth
output-velvet
83
-fastq
-shortPaired
-separate
S2_L001_R1_001.fastq
S2_L001_R2_001.fastq
```

- `velveth` produces hashtable and output files that are required for `velvetg`
- Directory that contains `velveth` output files
- The word length in bp that are being hashed
- This can affect the assembly result
- Specify type of input sequences
- Specify that pair end input sequences are in two separate files
- Input files

De novo assembly using Velvet

- Running *de novo* assembly
  - `cd`
  - `mkdir output-velvet`
  - `velvet/velvet_1.2.10/velveth output-velvet 83 -fastq -shortPaired -separate S2_L001_R1_001.fastq S2_L001_R2_001.fastq`
  - `velvet/velvet_1.2.10/velvetg output-velvet -exp_cov auto -min_contig_lgth 300`

De novo assembly using Velvet

`velvet/velvet_1.2.10/velvetg` → `velvetg` makes use of `velveth` output files to create a sequence assembly

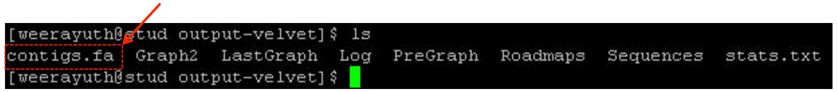
`output-velvet` → Output directory

`-exp_cov auto` →
- Let Velvet automatically determined the expected coverage
- This option is intended mainly for standard genomic sequencing

`-min_contig_lgth 300` → Specify the mininum contig length in the output contigs.fa file

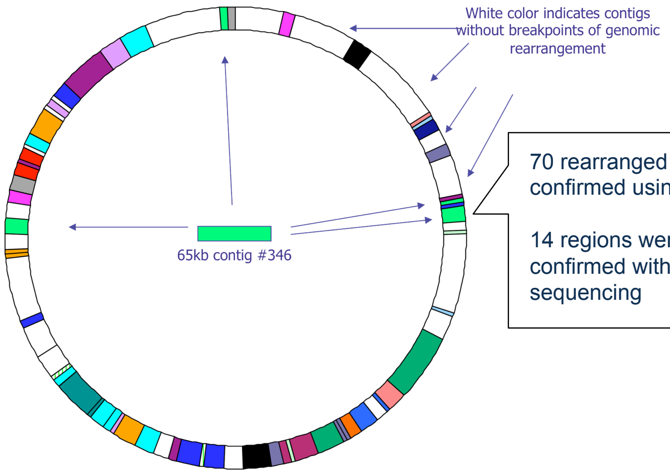
De novo assembly using Velvet

- Running *de novo* assembly
  - `cd`
  - `mkdir output-velvet`
  - `velvet/velvet_1.2.10/velveth output-velvet 83 -fastq -shortPaired -separate S2_L001_R1_001.fastq S2_L001_R2_001.fastq`
  - `velvet/velvet_1.2.10/velvetg output-velvet -exp_cov auto -min_contig_lgth 300`
  - `cd output-velvet`
  - `ls`



```
[weerayuth@stud output-velvet]$ ls
contigs.fa  Graph2  LastGraph  Log  PreGraph  Roadmaps  Sequences  stats.txt
[weerayuth@stud output-velvet]$
```

Initial results of genome sequencing of *Aggregatibacter actinomycetemcomitans* strain D7S-1

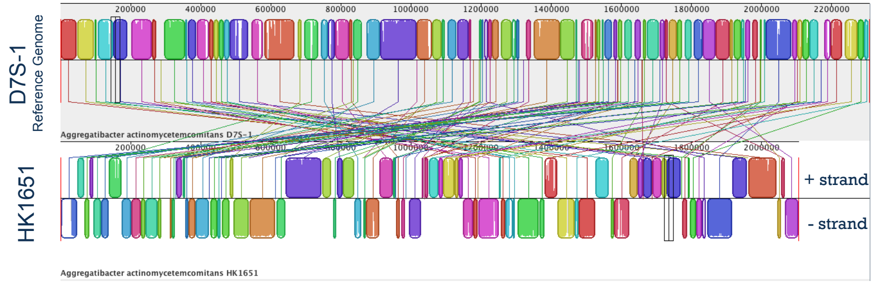

White color indicates contigs without breakpoints of genomic rearrangement
65kb contig #346

70 rearranged regions were confirmed using PCR

14 regions were further confirmed with Sanger sequencing

Alignment of D7S contigs based on the genomic map of HK1651.

Genomic rearrangement between strain D7S-1 and HK1651 of *Aggregatibacter actinomycetemcomitans*



- Whole genome sequence alignment created using the Mauve progressive alignment software

Special challenges with next generation sequencing data

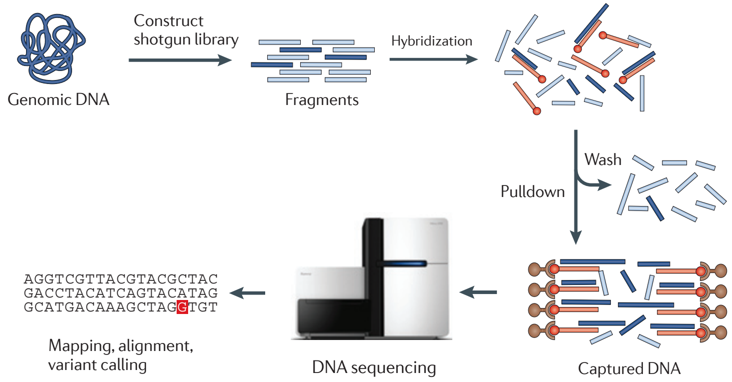
- Typically, only an incomplete genome is generated
- The cost of closing all gaps to produce a complete genome is still high
  - More incomplete genome sequences will be in public databases in the future
- Each technology is prone to making certain type of errors
  - Roach/454 and Ion Torrent tends to produce insertion/deletion in homopolymer regions
- Mapping to a reference genome may not be possible or even misleading
- Incomplete genome and sequence error produce “even greater” challenges in downstream analysis such as gene prediction and annotation

+ Whole-exome sequencing

- The targeted sequencing of the subset of the human genome that code for RNA or amino-acid.
- About 1% (30Mb) of the human genome.
- It is estimated that 85% of the disease-causing mutations are located in coding and functional regions of the genome

- Rabbani, B., Tekin, M., & Mahdieh, N. (2013). The promise of whole-exome sequencing in medical genetics. Journal of human genetics, 59(1), 5-15.
- van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. Trends in genetics, 30(9), 418-426.

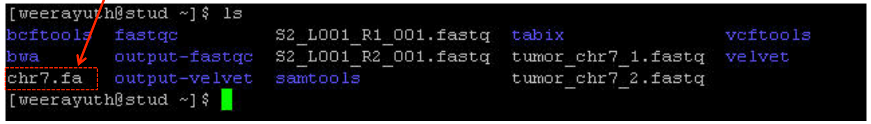
Whole-exome sequencing



1. Bamshad, Michael J., et al. "Exome sequencing as a tool for Mendelian disease gene discovery." *Nature Reviews Genetics* 12.11 (2011): 745-755.
2. Human All Exon. (n.d.). http://www.genomics.agilent.com/en/SureSelect-DNA-Target-Enrichment-Baits-/Human-All-Exon/?cid=AG-PT-124&tabId=AG-PR-1308

+
Whole-exome sequencing data analysis

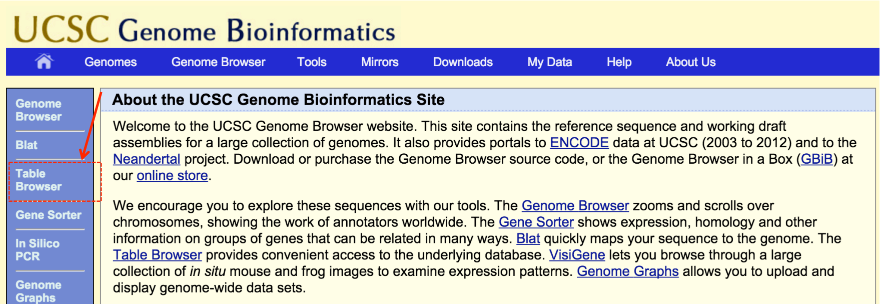
- Preparation of a reference human genome sequence
  - `cd`
  - `wget http://hgdownload.cse.ucsc.edu/goldenpath/hg19/chromosomes/chr7.fa.gz`
    - More information can be found at http://hgdownload.cse.ucsc.edu/goldenpath/hg19/chromosomes/
  - `gunzip chr7.fa.gz`
  - `ls`



```
[weerayuth@stud ~]$ ls
bcftools  fastqc          S2_L001_R1_001.fastq  tabix                 vcftools
bwa       output-fastqc   S2_L001_R2_001.fastq  tumor_chr7_1.fastq    velvet
chr7.fa   output-velvet   samtools              tumor_chr7_2.fastq
[weerayuth@stud ~]$
```

+
Whole-exome sequencing data analysis

- Preparation of the human genome annotation information
  - Go to https://genome.ucsc.edu/
  - Click Table Browser



About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to ENCODE data at UCSC (2003 to 2012) and to the Neandertal project. Download or purchase the Genome Browser source code, or the Genome Browser in a Box (GBiB) at our online store.

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VisiGene lets you browse through a large collection of *in situ* mouse and frog images to examine expression patterns. Genome Graphs allows you to upload and display genome-wide data sets.

- Retrieve annotation data for Chromosome 7 as shown below

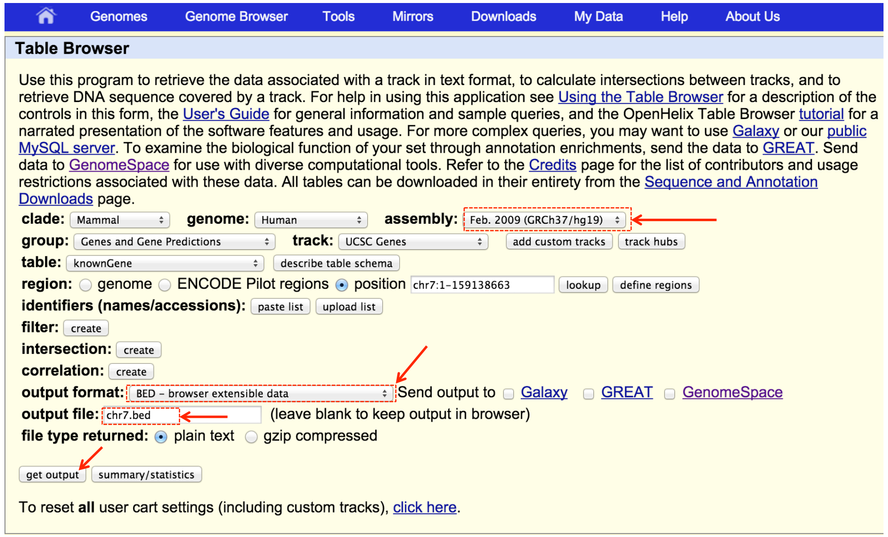


Table Browser

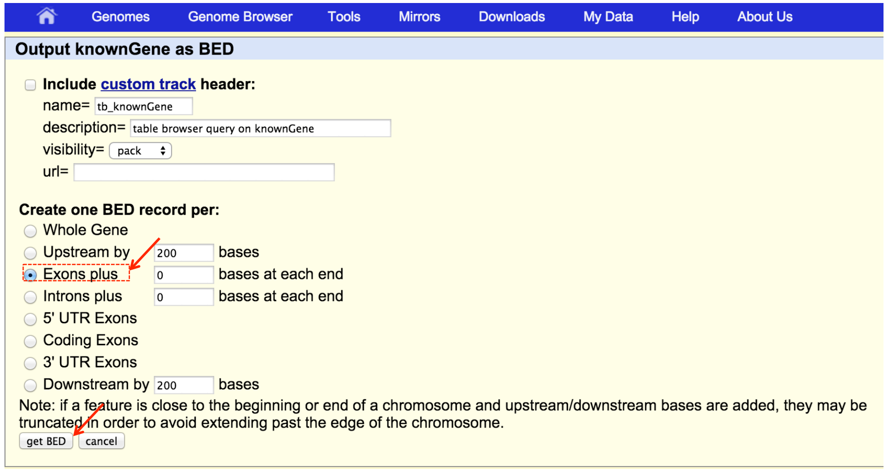
Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)
group: Genes and Gene Predictions **track:** UCSC Genes [add custom tracks] [track hubs]
table: knownGene [describe table schema]
region: ○ genome ○ ENCODE Pilot regions ● position chr7:1-159138663 [lookup] [define regions]
identifiers (names/accessions): [paste list] [upload list]
filter: [create]
intersection: [create]
correlation: [create]
output format: BED - browser extensible data Send output to □ Galaxy □ GREAT □ GenomeSpace
output file: chr7.bed (leave blank to keep output in browser)
file type returned: ● plain text ○ gzip compressed

[get output] [summary/statistics]

To reset **all** user cart settings (including custom tracks), click here.

- Select Exons and click 'get BED' as shown below



Output knownGene as BED

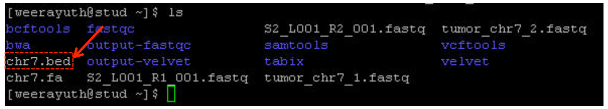
□ **Include custom track header:**
name= tb_knownGene
description= table browser query on knownGene
visibility= pack
url=

Create one BED record per:
○ Whole Gene
○ Upstream by 200 bases
● Exons plus 0 bases at each end
○ Introns plus 0 bases at each end
○ 5' UTR Exons
○ Coding Exons
○ 3' UTR Exons
○ Downstream by 200 bases

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

[get BED] [cancel]

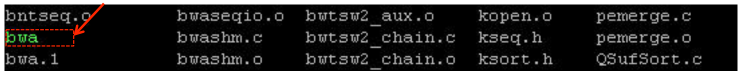
- Transfer file 'chr7.bed' to the Linux server using WinSCP



```
[weerayuth@stud ~]$ ls
bcftools  fastqc                S2_L001_R2_001.fastq  tumor_chr7_2.fastq
bwa       output-fastqc         samtools              vcftools
chr7.bed  output-velvet         tabix                 velvet
chr7.fa   S2_L001_R1_001.fastq  tumor_chr7_1.fastq
[weerayuth@stud ~]$
```

+ Whole-exome sequencing data analysis

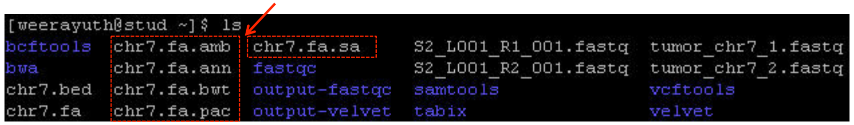
- Download and install Burrows-Wheeler Aligner http://bio-bwa.sourceforge.net/
  - `cd`
  - `mkdir bwa`
  - `cd bwa`
  - `wget 'http://downloads.sourceforge.net/project/bio-bwa/bwa-0.7.12.tar.bz2?r=http%3A%2F%2Fsourceforge.net%2Fprojects%2Fbio-bwa%2Ffiles%2F&ts=1447657569&use_mirror=jaist' -O bwa-0.7.12.tar.bz2`
  - `bzip2 -d bwa-0.7.12.tar.bz2`
  - `tar -xf bwa-0.7.12.tar`
  - `cd bwa-0.7.12/`
  - `make`
  - `ls`



```
bntseq.o     bwaseqio.o  bwtsw2_aux.o    kopen.o   pemerge.c
bwa          bwashm.c    bwtsw2_chain.c  kseq.h    pemerge.o
bwa.1        bwashm.o    bwtsw2_chain.o  ksort.h   QSufSort.c
```

+ Whole-exome sequencing data analysis

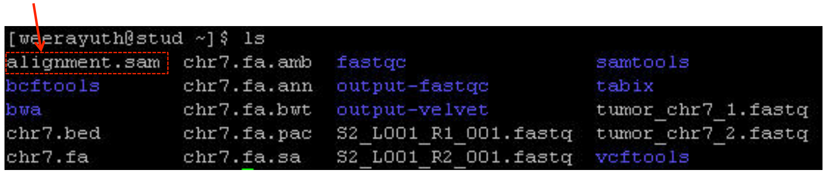
- Index reference genome
  - `cd`
  - `bwa/bwa-0.7.12/bwa index chr7.fa`
    - More information http://bio-bwa.sourceforge.net/bwa.shtml
  - `ls`



```
[weerayuth@stud ~]$ ls
bcftools  chr7.fa.amb  chr7.fa.sa      S2_L001_R1_001.fastq  tumor_chr7_1.fastq
bwa       chr7.fa.ann  fastqc          S2_L001_R2_001.fastq  tumor_chr7_2.fastq
chr7.bed  chr7.fa.bwt  output-fastqc   samtools              vcftools
chr7.fa   chr7.fa.pac  output-velvet   tabix                 velvet
```

+ Whole-exome sequencing data analysis

- Align sequence reads to reference genome
  - `cd`
  - `bwa/bwa-0.7.12/bwa mem chr7.fa tumor_chr7_1.fastq tumor_chr7_2.fastq > alignment.sam`
    - More information http://bio-bwa.sourceforge.net/bwa.shtml
    - SAM format specification https://samtools.github.io/hts-specs/SAMv1.pdf
  - `ls`

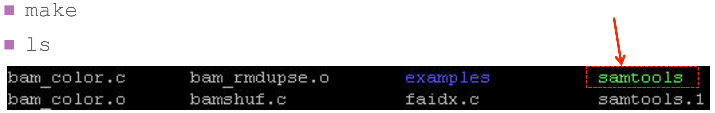


```
[weerayuth@stud ~]$ ls
alignment.sam  chr7.fa.amb  fastqc                samtools
bcftools       chr7.fa.ann  output-fastqc         tabix
bwa            chr7.fa.bwt  output-velvet         tumor_chr7_1.fastq
chr7.bed       chr7.fa.pac  S2_L001_R1_001.fastq  tumor_chr7_2.fastq
chr7.fa        chr7.fa.sa   S2_L001_R2_001.fastq  vcftools
```

+ Whole-exome sequencing data analysis

- Download and install Samtools (http://www.htslib.org/)

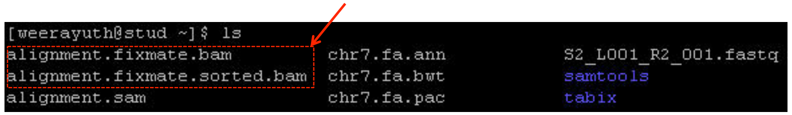
- Samtools is a suite of programs for interacting with high-throughput sequencing data
  - `cd`
  - `mkdir samtools`
  - `cd samtools`
  - `wget https://github.com/samtools/samtools/releases/download/1.2/samtools-1.2.tar.bz2`
  - `bzip2 -d samtools-1.2.tar.bz2`
  - `tar -xf samtools-1.2.tar`
  - `cd samtools-1.2/`
  - `make`
  - `ls`



```
bam_color.c    bam_rmdupse.o    examples    samtools
bam_color.o    bamshuf.c        faidx.c     samtools.1
```

Whole-exome sequencing data analysis

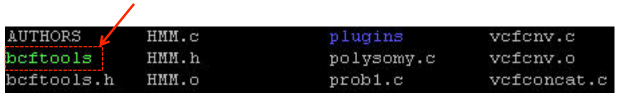
- “Clean” alignment result using Samtools
  - `cd`
  - `samtools/samtools-1.2/samtools fixmate -O bam alignment.sam alignment.fixmate.bam`
    - BWA can sometimes leave unusual FLAG information on SAM records, it is helpful when working with many tools to first clean up read pairing information and flags
  - `samtools/samtools-1.2/samtools sort -O bam -o alignment.fixmate.sorted.bam -T alignment.fixmate.temp alignment.fixmate.bam`
    - Sort records from name order into coordinate order
  - `ls`



```
[weerayuth@stud ~]$ ls
alignment.fixmate.bam          chr7.fa.ann              S2_L001_R2_001.fastq
alignment.fixmate.sorted.bam   chr7.fa.bwt              samtools
alignment.sam                  chr7.fa.pac              tabix
```

Whole-exome sequencing data analysis

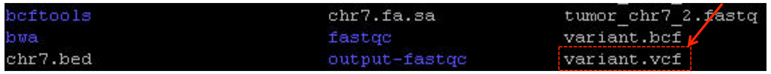
- Download and install BCFtools - utilities for variant calling and manipulating VCFs and BCFs (http://www.htslib.org/doc/bcftools.html)
  - `cd`
  - `mkdir bcftools`
  - `cd bcftools`
  - `wget https://github.com/samtools/bcftools/releases/download/1.2/bcftools-1.2.tar.bz2`
  - `bzip2 -d bcftools-1.2.tar.bz2`
  - `tar -xf bcftools-1.2.tar`
  - `cd bcftools-1.2/`
  - `make`
  - `ls`



```
AUTHORS      HMM.c          plugins       vcfcnv.c
bcftools     HMM.h          polysomy.c    vcfcnv.o
bcftools.h   HMM.o          prob1.c       vcfconcat.c
```

Whole-exome sequencing data analysis

- Call sequence variants from alignment data
  - `cd`
  - `samtools/samtools-1.2/samtools mpileup -go variant.bcf -f chr7.fa -Q 30 -l chr7.bed alignment.fixmate.sorted.bam`
    - Use mpileup to produce a BCF file that contains all of the locations in the genome.
    - http://www.htslib.org/doc/samtools.html
  - `bcftools/bcftools-1.2/bcftools call -vmO v -o variant.vcf variant.bcf`
    - Call genotypes and reduce our list of sites to those found to be variant by passing this file into bcftools call
    - http://www.htslib.org/doc/bcftools.html
  - `ls`



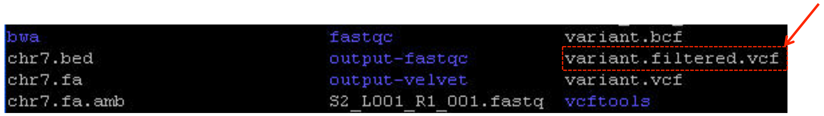
```
bcftools                   chr7.fa.sa          tumor_chr7_2.fastq
bwa                        fastqc              variant.bcf
chr7.bed                   output-fastqc       variant.vcf
```

Whole-exome sequencing data analysis

- Download and install VCFtools (https://vcftools.github.io/index.html)
- VCFtools provides easily accessible methods for working with complex genetic variation data in the form of VCF files
  - `cd`
  - `mkdir vcftools`
  - `cd vcftools`
  - `wget 'http://downloads.sourceforge.net/project/vcftools/vcftools_0.1.13.tar.gz?r=http%3A%2F%2Fsourceforge.net%2Fprojects%2Fvcftools%2Ffiles%2F&ts=1448009724&use_mirror=jaist' -O vcftools_0.1.13.tar.gz`
  - `tar -xzf vcftools_0.1.13.tar.gz`
  - `cd vcftools_0.1.13`
  - `make`
  - `export PERL5LIB=/home/weerayuth/vcftools/vcftools_0.1.13/perl`

+ Whole-exome sequencing data analysis

- Filter variant result using vcftools
  - `cd`
  - `cat variant.vcf | vcftools/vcftools_0.1.13/bin/vcf-annotate --filter MinDP=20/RefN -H > variant.filtered.vcf`
    - Filter out variants that are supported by less than 20 reads
    - Filter out variants where reference sequence is N
    - https://vcftools.github.io/perl_module.html#vcf-annotate
  - `ls`



```
bwa                                fastqc               variant.bcf
chr7.bed                           output-fastqc        variant.filtered.vcf
chr7.fa                            output-velvet        variant.vcf
chr7.fa.amb                        S2_L001_R1_001.fastq vcftools
```

+ Whole-exome sequencing data analysis

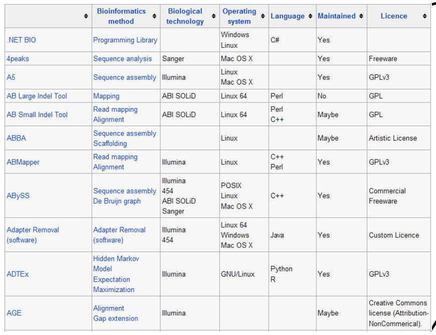
- More information on vcf format can be found at http://www.1000genomes.org/wiki/analysis/variant%20call%20format/vcf-variant-call-format-version-41
- The resulting vcf file can be further annotated to add more functional information using variant annotation tools
- This can be done by using command-line or web-based variant annotation tools
- An example of a web-based variant annotation tool is wANNOVAR by Wang Genomics Lab at University of Southern California http://wannovar.usc.edu/

+ Final comments

- This workshop only introduce open-source software for doing NGS data analysis
- Advantages of open-source software
  - Free
  - Clear methods
  - Most run on Linux platforms (stable, can easily make your own pipelines)
- Disadvantage of open-source software
  - Most run on Linux platforms (Requires knowledge in Linux system and command line)
  - Lots of small software that only do one specific job
  - Can become obsolete very quickly

+ Final comments

- A good collection of software packages for next generation sequence data analysis can be found at http://seqanswers.com/wiki/Software



	Bioinformatics method	Biological technology	Operating system	Language	Maintained	Licence
.NET BIO	Programming Library		Windows Linux	C#	Yes	
4peaks	Sequence analysis	Sanger	Mac OS X		Yes	Freeware
A5	Sequence assembly	Illumina	Linux Mac OS X		Yes	GPLv3
AB Large Indel Tool	Mapping	ABI SOLiD	Linux 64	Perl	No	GPL
AB Small Indel Tool	Read mapping Alignment	ABI SOLiD	Linux 64	Perl C++	Maybe	GPL
ABBA	Sequence assembly Scaffolding		Linux		Maybe	Artistic License
ABMapper	Read mapping Alignment	Illumina	Linux	C++ Perl	Yes	GPLv3
ABySS	Sequence assembly De Bruijn graph	Illumina 454 ABI SOLiD Sanger	POSIX Linux Mac OS X	C++	Yes	Commercial Freeware
Adapter Removal (software)	Adapter Removal (software)	Illumina 454	Linux 64 Windows Mac OS X	Java	Yes	Custom Licence
ADTEx	Hidden Markov Model Expectation Maximization	Illumina	GNU/Linux	Python R	Yes	GPLv3
AGE	Alignment Gap extension	Illumina			Maybe	Creative Commons license (Attribution-NonCommerical).

- This lecture is merely a small introduction to the big world of next generation sequence data analysis
- The endless possibility and new discovery is waiting for you

Acknowledgements

KMUTT King Mongkut's University of Technology Thonburi

W UNIVERSITY of WASHINGTON

My students from the Bioinformatics and Systems biology program at KMUTT

Thank you for your attention...