高雄醫學大學
Kaohsiung Medical University

高階人才研習營

次世代定序及生物資訊分析研習營/課程編號D5

高教深耕計劃－構面二＆核心議題三
發展學校特色＆培育特色研發人才系列活動

Date	Time	Topic	Speaker	Place
July 10 2019	09:00~09:30	Opening	Vice-president Hung Director Hsu	IR334
	09:30~10:30	Sequencing technologies (lecture)	Dr. H. Sunny Sun	
	10:30~12:30	**Exome sequencing data analysis:** Data cleaning, mapping, variant calling and annotation (practical course)	Dr. Alan Tang	
	12:30~14:00	Lunch		
	14:00~16:00	**Exome sequencing data analysis:** structural variant identification and analysis (practical course)	Dr. Shuen-Lin Jeng	IR334
	16:00~16:30	Questions and Answers	Drs. H. Sunny Sun Dr. Alan Tang Dr. Shuen-Lin Jeng	

Date	Time	Topic	Speaker	Place
July 11 2019	09:00~09:30	Opening	Vice-president Hung Director Hsu	Audiovisual center
	09:30~10:00	**High-throughput genomic data analysis:** challenges and opportunities	Dr. I-Shou Chang	
	10:00~10:30	Study design and technical issues on RNA sequencing	Dr. Shih Sheng Jiang	
	10:30~11:20	Preprocessing for high throughput genomic data analysis	Dr. Chung-Hsing Chen	
	11:20~12:00	What after RNA-seq?	Dr. Shih Sheng Jiang	
	12:00~14:00	Lunch		
	14:00~14:30	RNA sequencing data analysis (demonstration)	Ms. Fang-Yu Tsai	IR334
	14:30~16:00	Pathway analysis for transcriptome profiling studies(array and RNA-seq) using GSEA (hands-on practice)	Ms. Fang-Yu Tsai Dr. Chung-Hsing Chen Dr. Shih Sheng Jiang	
	16:00~16:30	Questions and Answers	Drs. I-Shou Chang Dr. Shih Sheng Jiang Dr. Chung-Hsing Chen Ms. Fang-Yu Tsai	

主辦單位：高雄醫學大學研究資源整合發展中心
協辦單位：科技部生技醫藥生物資訊核心設施、衛福部第三期癌症研究跨機構合作平台及其整合應用計畫

Curriculum Vitae

H. Sunny Sun, Ph.D.



Current position and professional experiences

2016/8 ~	Director/Professor, Institute of Molecular Medicine, National Cheng Kung University
2011/8 ~	Director, Center for Genomic Medicine, National Cheng Kung University
2015/2 ~2016/7	Associated Vice President of Office of Research and Development, National Cheng Kung University

Research interest

1. Genomic medicine
2. Bioinformatics
3. Molecular genetics
4. Comparative genomics

Short research summary

Dr. Sunny Sun is a Professor and Director of Molecular Medicine at the National Cheng Kung University (NCKU), Taiwan. Dr. Sun received her Ph.D degree from the University of Wisconsin-Madison in the United State and has been involved in genomic researches for the past ~20 years. Dr. Sun's research interest focuses on the application of computational strategy and advanced molecular biology techniques to identify genes responsible for various human diseases and study the regulation of gene expression in physiological and pathological conditions. In addition to her primary role in the University, Dr. Sun is also the Director of Center for Genomic Medicine in NCKU, in which she leads a team to provide services for clinical diagnosis and supports for clinical research.

Publications (Selective publications in the past 5 years, * Corresponding author)

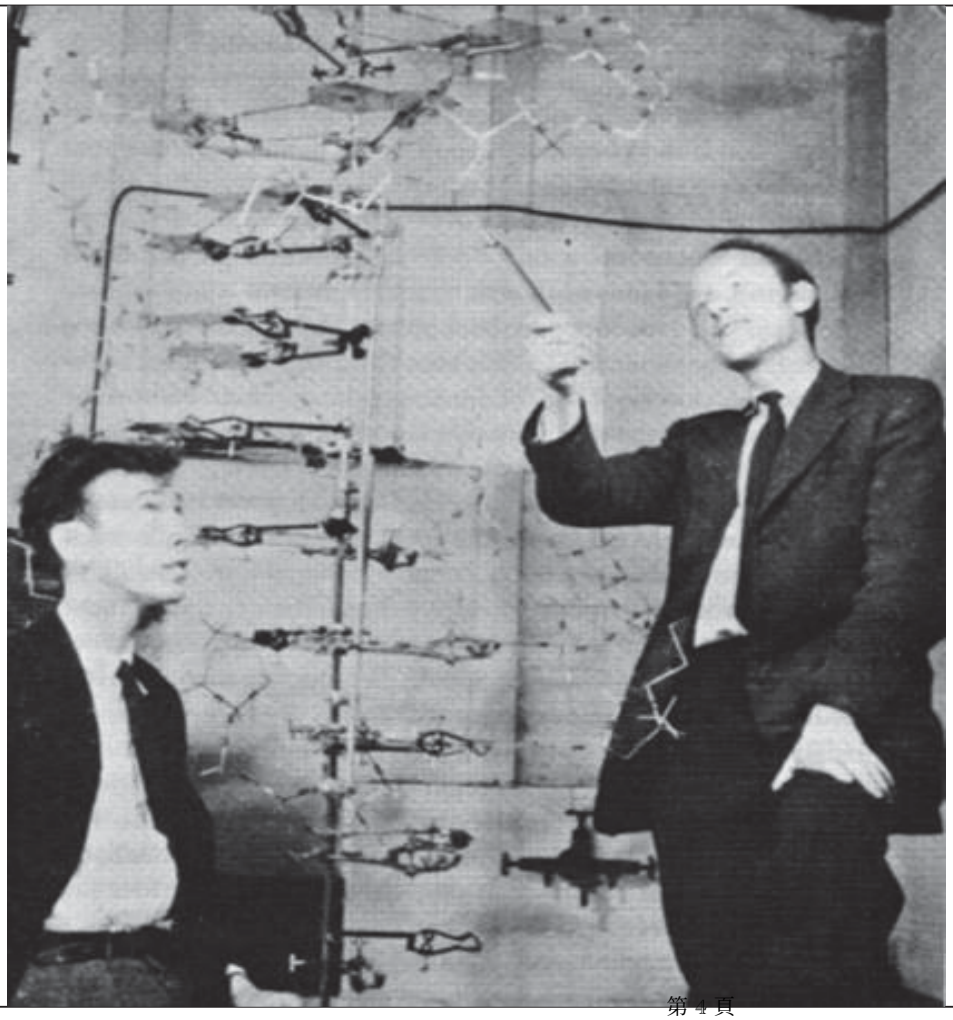
1. Jia-Shing Chen, Wei-Shiang Hung, Hsiang-Han Chan, Shaw-Jenq Tsai, **H. Sunny Sun***. 2013. In silico identification of oncogenic potential of fyn-related kinase in hepatocellular carcinoma. Bioinformatics 29(4):420-7.
2. Ming-Chih Lai, Shainn-Wei Wang, Lie Cheng, Woan-Yuh Tarn, Shaw-Jenq Tsai* and **H. Sunny Sun***. 2013. Human DDX3 participates in translation of HIV-1 mRNAs. PLoS ONE 8(7):e68665.
3. Chi-Jung Wu, Hsuan-Chen Wang, Po-Lin Chen, Ming-Chung Chang, H. Sunny Sun, Pei-Hsin Chou, Wen-Chien. 2013. AQU-1, a chromosomal class C-lactamase, among clinical Aeromonas dhakensis isolates: Distribution and clinical significance. International Journal of Antimicrobial Agents. Int J Antimicrob Agents. 42(5):456-461.
4. Chung CL, Lu CW, Cheng YS, Lin CY, **Sun HS***, Lin YM*. 2013. Association of aberrant expression of sex-determining gene fibroblast growth factor 9 with Sertoli cell-only syndrome. Fertility and Sterility 100(6):1547-1554.
5. Tsung-Ming Chen, Yu-Heng Shih, Joseph T. Tseng, Chih-Hao Wu, Yi-Han Lee, Shaw-Jenq Tsai*, and **H. Sunny Sun***. 2014. Overexpression of FGF9 in colon cancer cells is mediated by hypoxia-induced translational activation. NAR 42(5):2932-44.
6. Wan-Lin Liao, Shao-Chieh Lin, **H. Sunny Sun***, Shaw-Jenq Tsai*. 2014. Hypoxia-induced tumor malignancy and drug resistance: Role of microRNAs. Biomarkers and Genomic Medicine 6:1-11.
7. Ya-Chi Lin, Joseph T. Tseng, Shuen-Lin Jeng and **H. Sunny Sun***. 2014. Comprehensive analysis of common coding sequence variants in Taiwanese Han population. Biomarkers and Genomic Medicine 6:133-143.
8. Yi-Mei Joy Lin, Cheng-Da Hsu, Hsiao-Yen Hsieh, Chia-Chih Alex Tseng, **H. Sunny Sun***. 2014.

Sequence Variants of the HTR3A Gene Contribute to Genetic Prediction of Postoperative Nausea in Taiwan. Journal of Human Genetics, 59(12):655-60.
9. Kai-Chiang Liang, Joseph T. Tseng1, Shaw-Jenq Tsai and **H. Sunny Sun***. 2015. Characterization and distribution of repetitive elements in association with genes in the human genome. Computational Biology and Chemistry 57:29-38.
10. Kuei-Yang Hsiao, Meng-Hsing Wu, Ning Chang1, Shang-Hsun Yang, Chun-Wei Wu , **H. Sunny Sun*** and Shaw-Jenq Tsai*. 2015. Coordination of AUF1 and miR-148a to destabilize DNA methyltransferase 1 mRNA under hypoxia in endometriosis. Mol Hum Reprod. 21(12):894-904.
11. Ming-Chih Lai, Chiao-May Chang, and **H. Sunny Sun***. 2016. Hypoxia Induces Autophagy through Translational Up-Regulation of Lysosomal Proteins in Human Colon Cancer Cells. PLoS ONE 11(4):e0153627.
12. Wei-Hsuan Yen, Wu-Sian Ke, Jan-Jong Hung, Tsung-Ming Chen, Jia-Shing Chen, **H. Sunny Sun***. 2016. Sp1-mediated ectopic expression of T-cell lymphoma invasion and metastasis 2 in hepatocellular carcinoma. Cancer Medicine 5(3):465-77.
13. Meng-Shao Lai, Chia-Yih Wang, Shang-Hsun Yang, Chia-Ching Wu, **H. Sunny Sun**, Shaw-Jenq Tsai, Jih-Ing Chuang, Yung-Chia Chen, Bu-Miin Huang. 2016. The expression profiles of fibroblast growth factor 9 and its receptors in developing mice testes. Organogenesis 12(2):61-77.
14. Ming-Chih Lai, Chiao-May Chang, and **H. Sunny Sun*.** 2016. Hypoxia Induces Autophagy through Translational Up-Regulation of Lysosomal Proteins in Human Colon Cancer Cells. PLoS ONE 11(4):e0153627.
15. Chun-Wei Chien, Pei-Chi Hou, Han-Chung Wu, Yu-Ling Chang, Shih-Chieh Lin, Shau-Chieh Lin, Bo-Wen Lin, Jenq-Chang Lee, Yu-Jia Chang, and **H. Sunny Sun*** and Shaw-Jenq Tsai*. 2016. Targeting TYRO3 inhibits epithelial-mesenchymal transition and increases drug sensitivity in colon cancer. Oncogene 10;35(45):5872-5881.
16. Jing-Yi Chen, You-Syuan Lai, Hui-Jen Tsai, Cheng-Chin Kuo, B. Linju Yen, Su-Peng Yeh, **H. Sunny Sun**, Wen-Chun Hung. 2016. The oncometabolite R-2-hydroxyglutarate activates NF-κB-dependent tumor-promoting stromal niche for acute myeloid leukemia cells. Sci Rep. 31;6:32428.
17. Yi-Hsuan Lee, Ya-Wen Hsueh, Yao-Hung Peng, Kung-Chao Chang, Kuen-Jer Tsai, **H. Sunny Sun**, Ih-Jen Su and Po-Min Chiang. 2017. Low-cell-number, single-tube amplification (STA) of total RNA revealed miRNA changes from pluripotency to endothelium. BMC Biology 15(1):22.
18. Kuei-Yang Hsiao, Ya-Chi Lin, Sachin Kumar Gupta, Ning Chang, Laising Yen, **H. Sunny Sun***, and Shaw-Jenq Tsai*. 2017. Non-coding effects of circular RNA CCDC66 promotes colon cancer growth and metastasis. Cancer Research 77(9):2339-2350.
19. Nithila A. Joseph, Shiow-Her Chiou*, Zoe Lung, Cheng-Lin Yang, Tze-Yi Lin, Hui-Wen Chang, **H. Sunny Sun*** Sachin Kumar Gupta, Laising Yen, Shulhn-Der Wang, and Kuan-Chih Chow*. 2018. The role of HGF-MET pathway and CCDC66 cirRNA expression in EGFR resistance and epithelial-to-mesenchymal transition of lung adenocarcinoma cells. Journal of Hematology & Oncology 11(1):74.
20. Yi-Mei J. Lin, I-Lun Hsin, **H. Sunny Sun**, Shankung Lin, Yen-Ling Lai, Hsuan-Ying Chen, Ting-Yu Chen, Ya-Ping Chen, Yi-Ting Shen, and Hung-Ming Wu. 2018. NTF3 Is a Novel Target Gene of the Transcription Factor POU3F2 and Is Required for Neuronal Differentiation. Molecular Neurobiology, 55(11):8403-8413.
21. Tsung-Ming Chen, Ming-Chih Lai, Yi-Han Li, Ya-Ling Chan, Chih-Hao Wu, Yu-Ming Wang, Chun-Wei Chien, San-Yuan Huang, **H. Sunny Sun***, and Shaw-Jenq Tsai*. 2019. hnRNPM induces translation switch under hypoxia to promote colon cancer development. EBioMedicine. 41: 299–309.
22. Chen TM, Chen YH, Sun HS, Tsai SJ. 2019. Fibroblast growth factors: Potential novel targets for regenerative therapy of osteoarthritis. Chin J Physiol. 62(1):2-10.
23. Chun-Hsien Chu, Jia-Shing Chen, Pei-Chin Chuang, Chia-Hao Su, Ya-Ling Chan, Ying-Ju Yang, Yu-Ting Chiang, Yu-Ya Su, Po-Wu Gean, **H. Sunny Sun***. 2019. Oncoprotein TIAM2S as a novel regulator of serotonin level, brain plasticity, and behavior. Molecular Neurobiology (In revision).

The Journey of DNA Sequencing

H. Sunny Sun

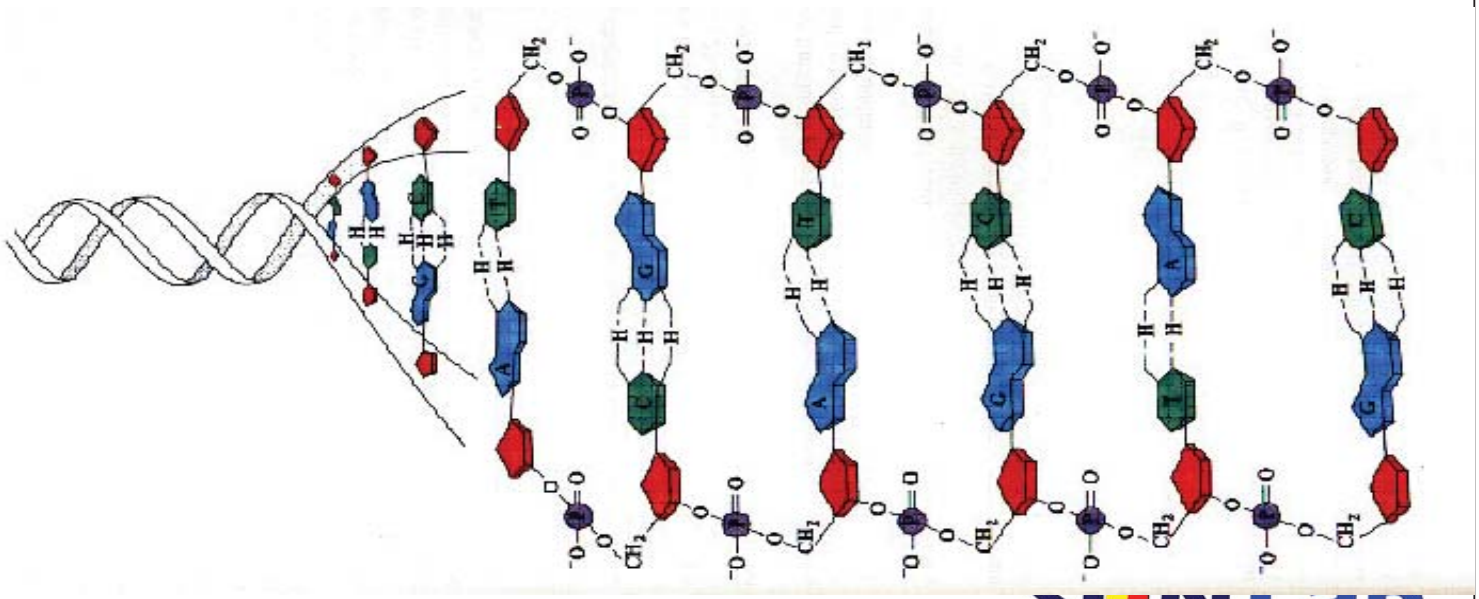
SUNLab
Molecular Genetics & Bioinformatics



"*Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid*"
- Francis Crick and James D. Watson. Nature 17 (1): 737–738 (dated 25 April 1953).
- It was the first publication which described the discovery of the double helix structure of DNA.

SUNLab
Molecular Genetics & Bioinformatics

Double Helix

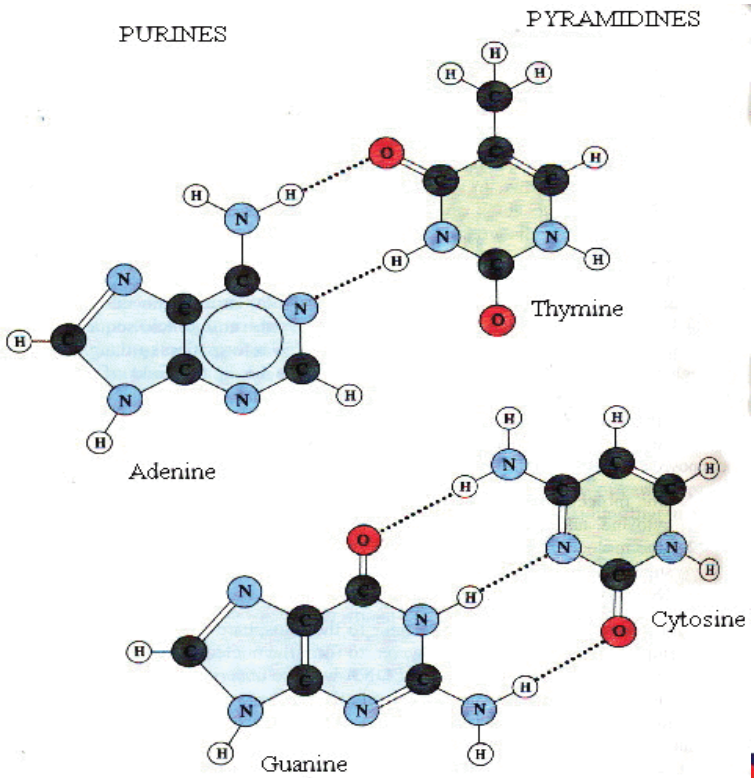


Molecular Genetics & Bioinformatics

DNA Base Pairing

A-T pairing

2 H-Bonds

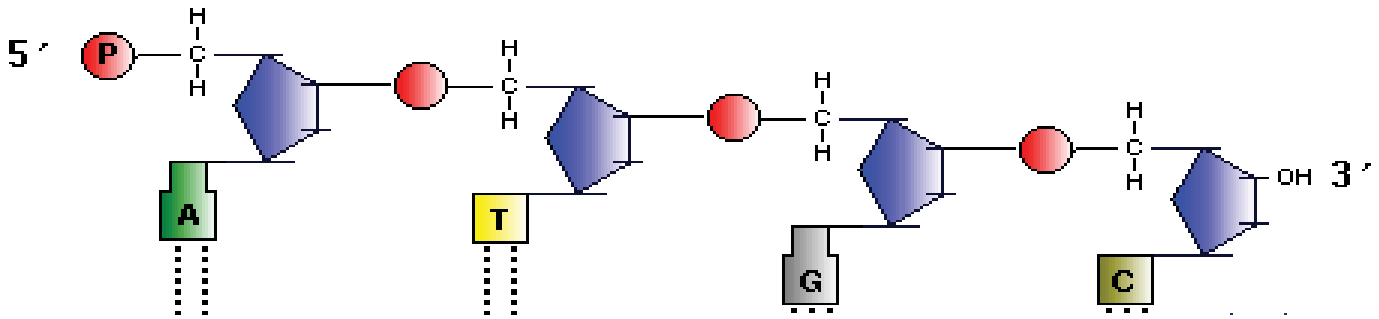


G-C pairing

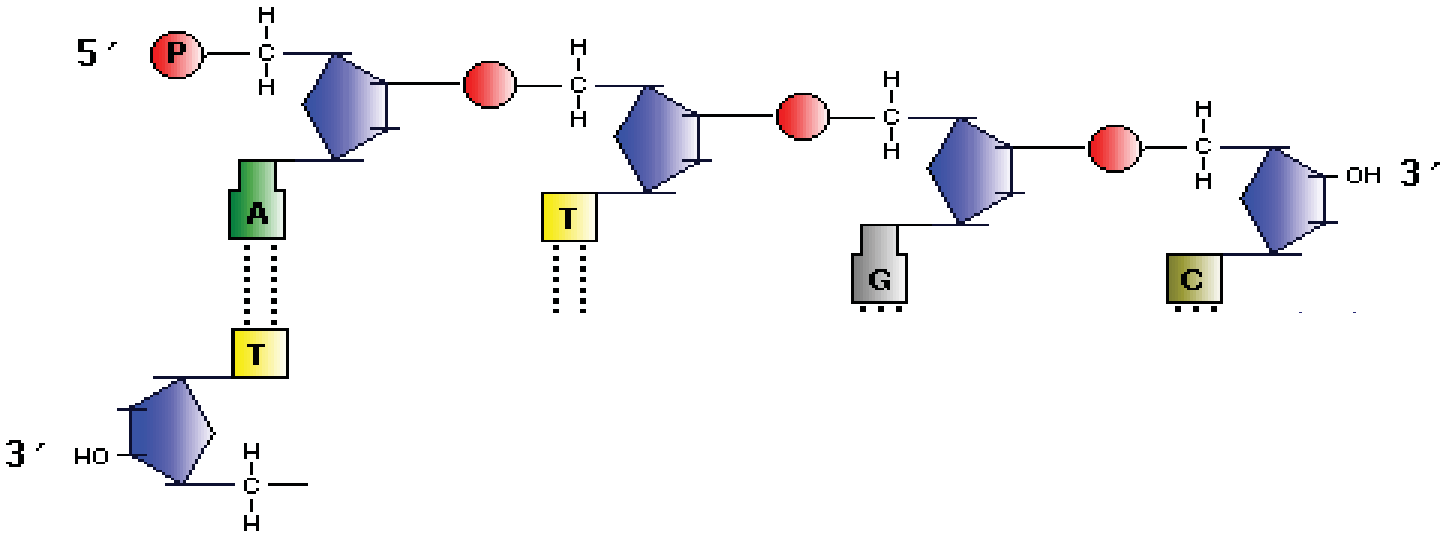
3 H-bonds

Molecular Genetics & Bioinformatics

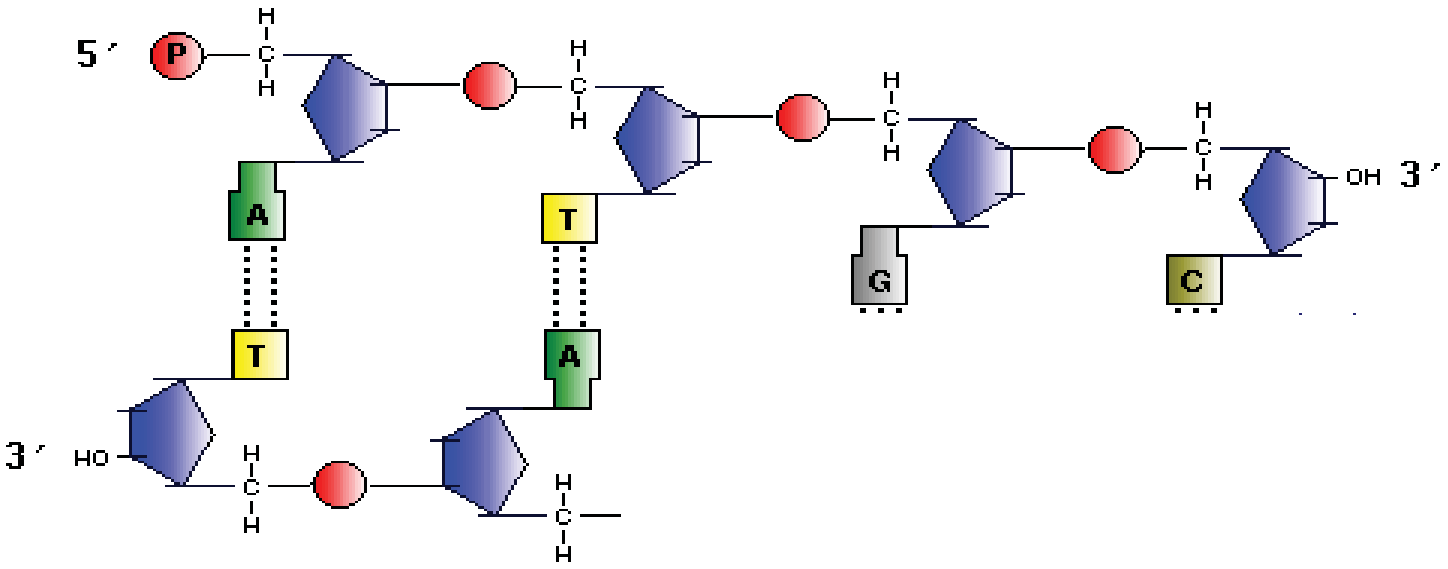
A-T/G-C Base Pairs Hold the DNA helices together



A-T/G-C Base Pairs Hold the DNA helices together

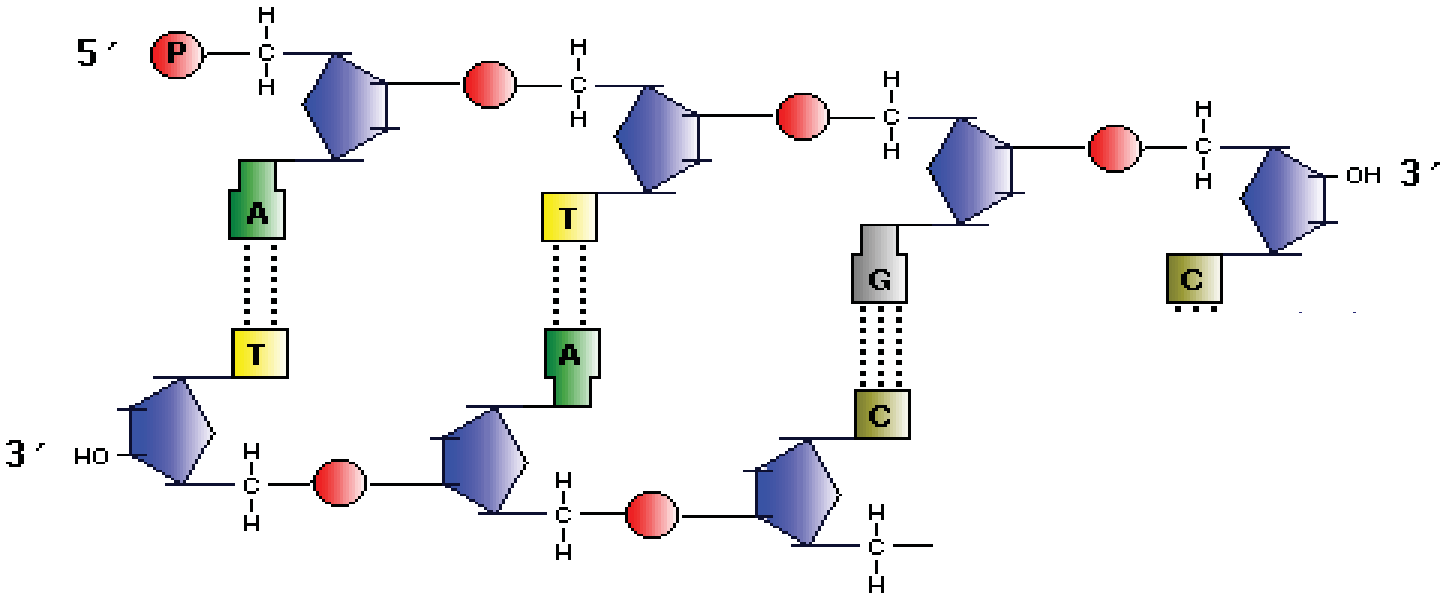


A-T/G-C Base Pairs Hold the DNA helices together



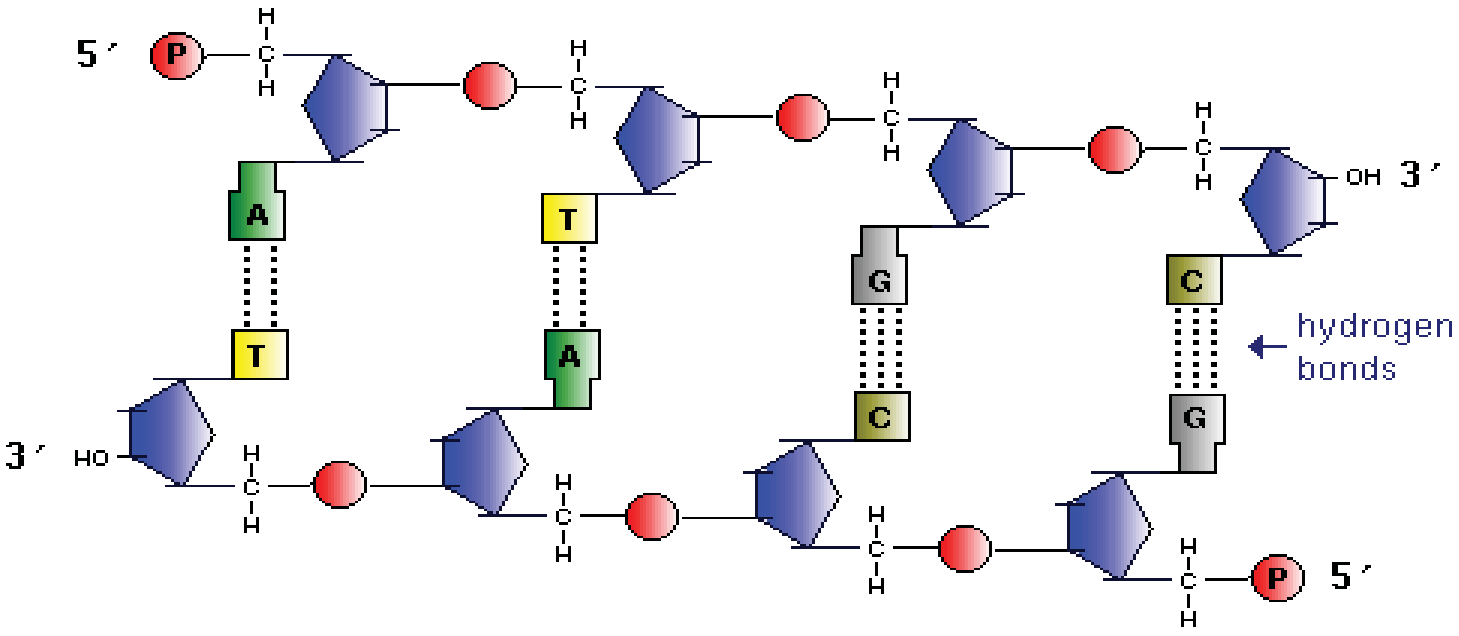
Molecular Genetics & Bioinformatics

A-T/G-C Base Pairs Hold the DNA helices together



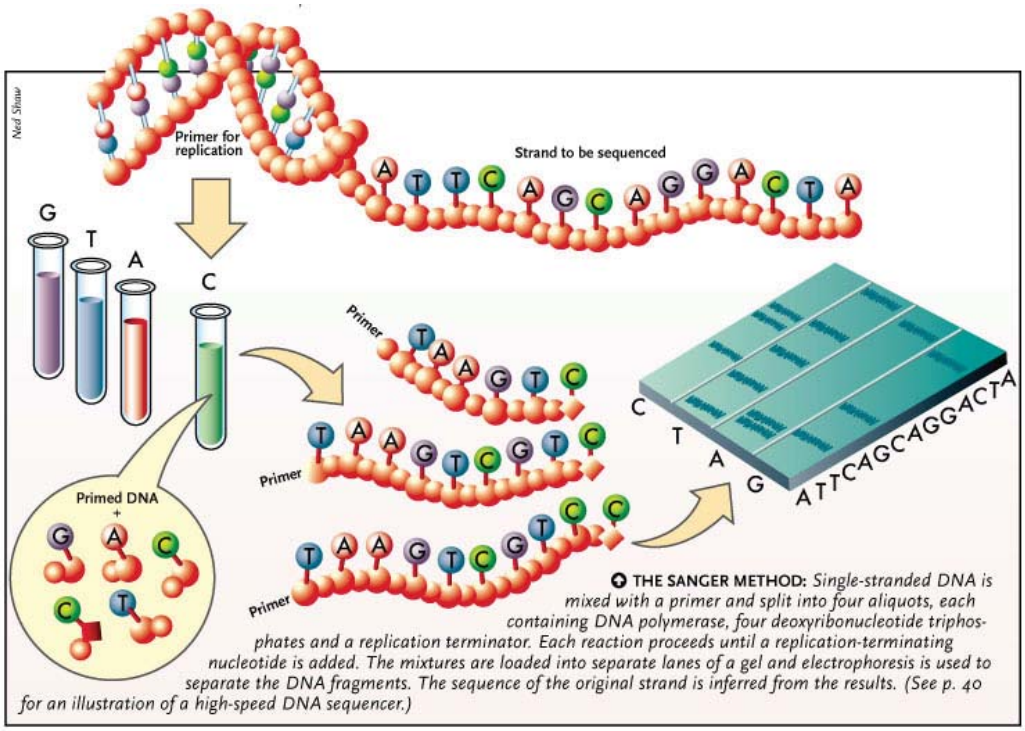
Molecular Genetics & Bioinformatics

A-T/G-C Base Pairs Hold the DNA helices together



Molecular Genetics & Bioinformatics

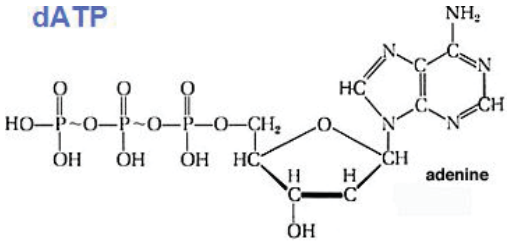
First generation sequencing: The Sanger Sequencing



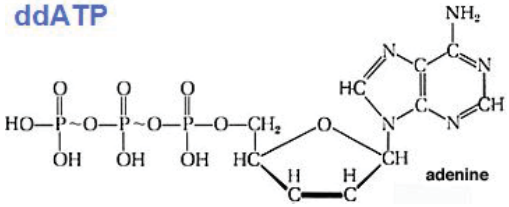
Molecular Genetics & Bioinformatics

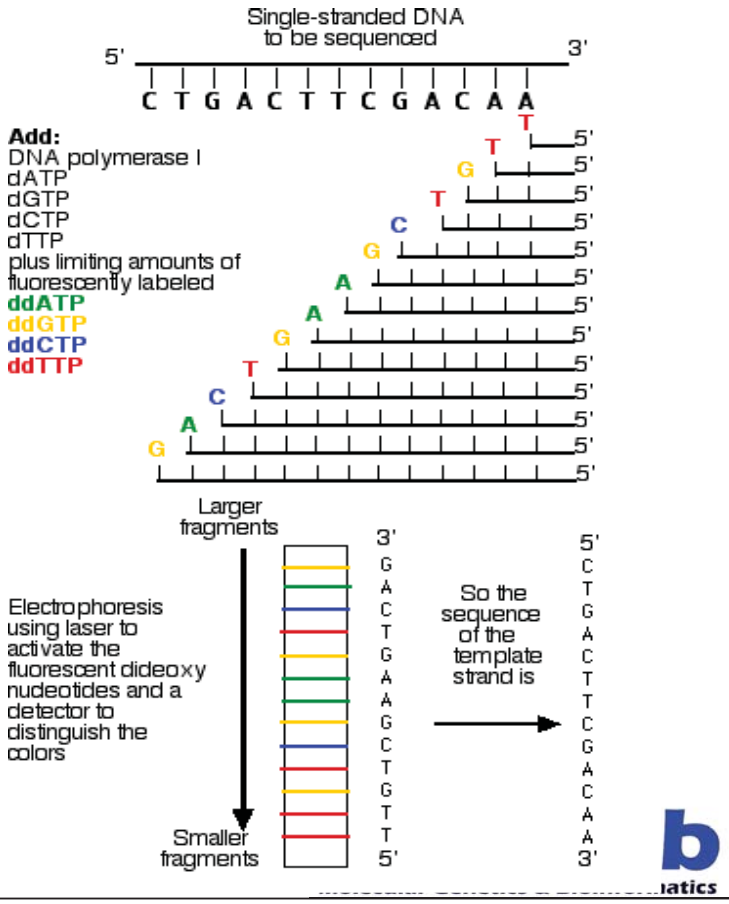
The Sequencing terminator

dATP

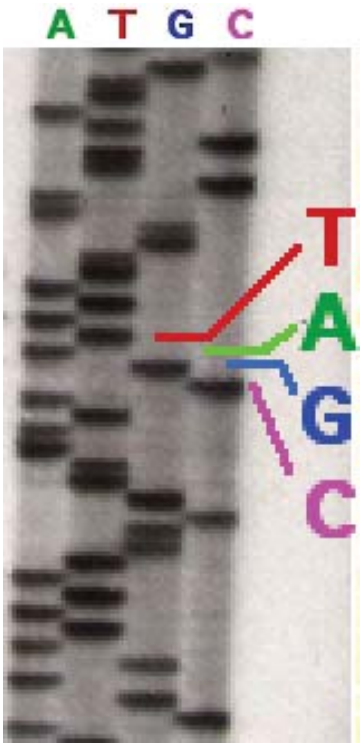


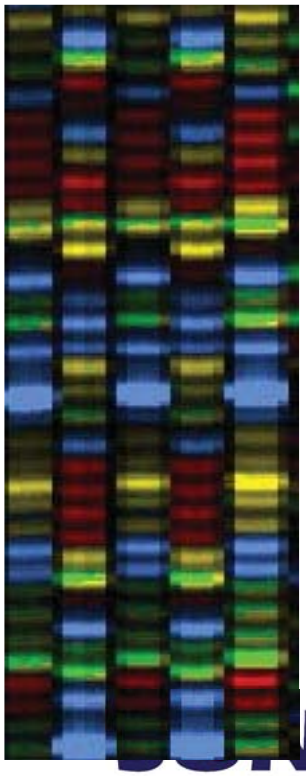
ddATP

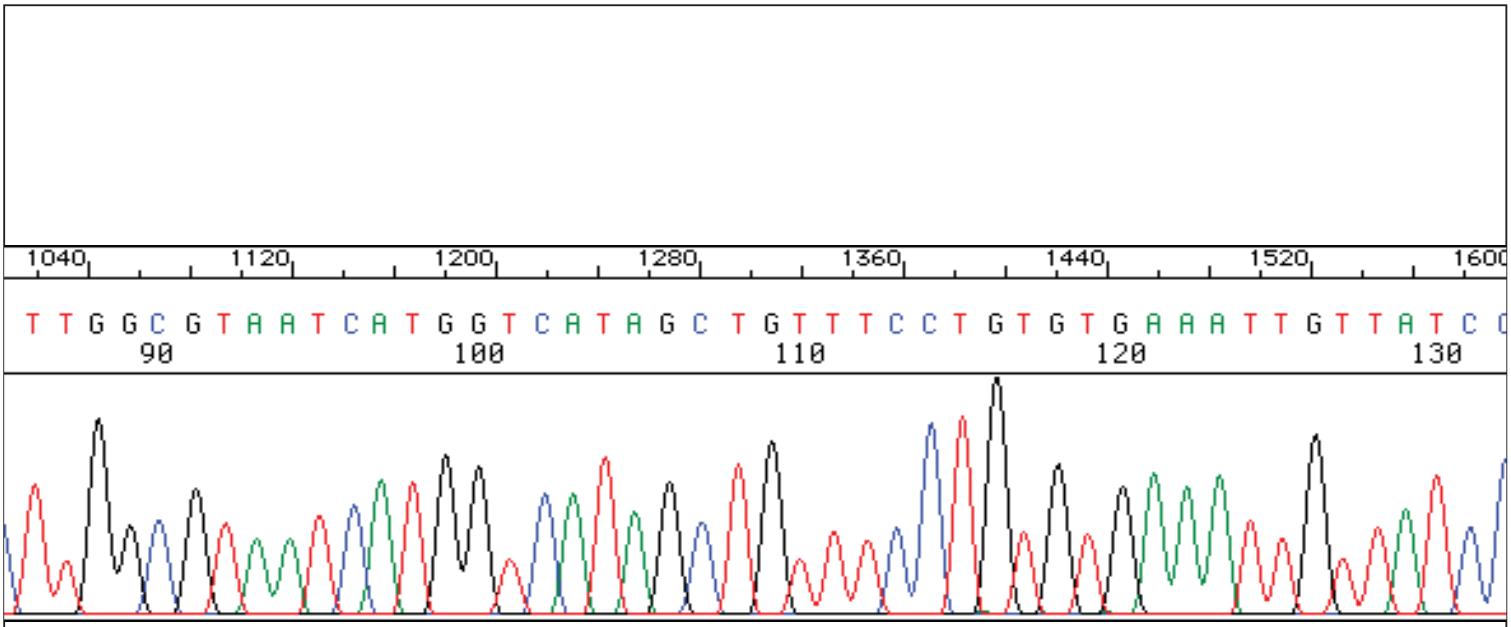




Sequencing gels



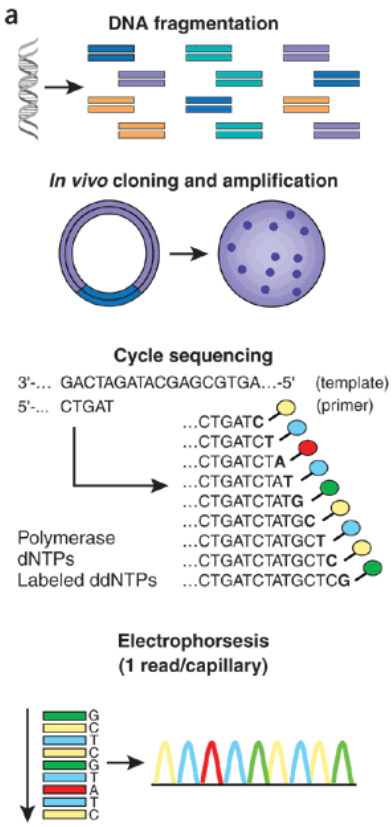




http://www.shsu.edu/~chm_tgc/sounds/flashfiles/GE.swf

SUNLab
Molecular Genetics & Bioinformatics

Work flow of Sanger sequencing



The Sanger sequencing
- Long read length
- Labor-intensive
- Time-consuming

SUNLab
Molecular Genetics & Bioinformatics

Nat. Biotech. 26, 1135 - 1145 (2008)

Second generation sequencing Technologies

Roche 454 SEQUENCING

Cost-effective
Fast
Ultra throughput
Cloning-free
Short reads

illumina®

life™

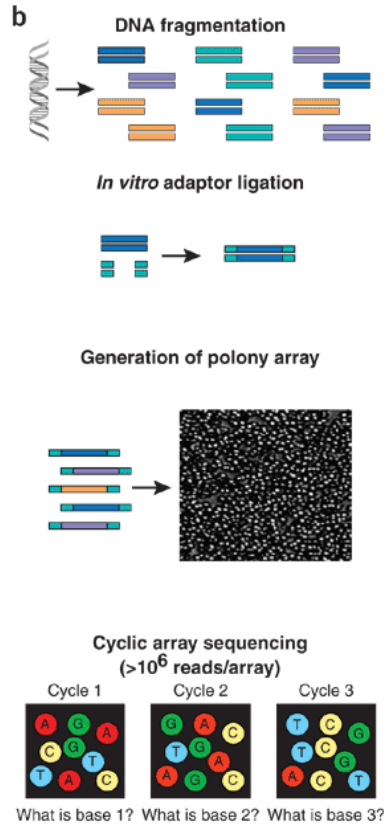
PACIFIC BIOSCIENCES™

ion torrent
by life technologies

Helicos BioSciences Corporation

SUNLab
Molecular Genetics & Bioinformatics

Work flow of NGS



Next Generation sequencing (NGS)
- Deep sequencing
- High-throughput sequencing

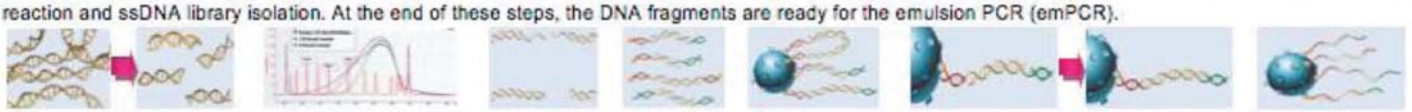
Nat. Biotech. 26, 1135 - 1145 (2008)

SUNLab
Molecular Genetics & Bioinformatics

Roche 454 Pyrosequencing

Principel

Preparation of the DNA includes : DNA fragmentation (nebulization), DNA size selection, Fragment end polishing, Adaptor ligation, Library immobilization, fill in reaction and ssDNA library isolation. At the end of these steps, the DNA fragments are ready for the emulsion PCR (emPCR).



fragmentation | size selection | Polishing | Adaptor ligation | Immobilization | Fill in reaction | ssDNA library isolation

emPCR include the immobilisation of the DNA fragments on capture beads (1 molecule / bead), emulsification (1bead / aqueous microreactor), amplification and indirect enrichment resulting in an immobilized and amplified library.



Immobilisation on capture beads | emulsification | amplification | Indirect enrichment : melt, enrichment with biotinilated primer, streptavidin magnetic beads, enrich and melt

Sequencing includes a prewash, the loading DNA library beads, enzyme beads (PPiase) and packing beads on the picotiter plate (PTP). Run over night. At the end of these steps you get your data.



DNA library loading | PPiase + enzyme + packing beads loading | assembly of the PTP | sequencing

Roche 454/GS FLX/Junior



Throughput	400-600 million high-quality, filter-passed bases per run* 1 billion bases per day
Run Time	10 hours
Read Length	Average length = 400 bases
Accuracy	Q20 read length of 400 bases (9 higher for prior bases)
Reads per run	>1 million high-quality reads
Data	Trace data accepted by NCBI sir
Computing Requirements	Cluster recommended (Roche G available)
Robustness	No complex optics or lasers; rea



System Performance	
Throughput	35 million high-quality, filtered bases per run*
Run Time	10 hours sequencing 2 hours data processing
Avg. Read Length	400 bases*
Accuracy	Q20 read length of 400 bases (99% accuracy at 400 bases)
Reads per Run	100,000 shotgun, 70,000 amplicon
Sample Input	gDNA, amplicons, cDNA, or BACs depending on the application
Physical Dimensions	40 cm wide x 60 cm deep x 40 cm high (the size of a laser printer) Weight = 55 lbs.
Computing	Linux-based OS on HP desktop computer included. All software is point-and-click.

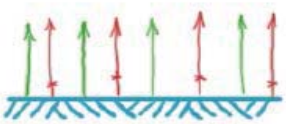
***Typical results.** Average read length and number of reads depend on specific sample and genomic characteristics*

Illumina/Solexa-GA

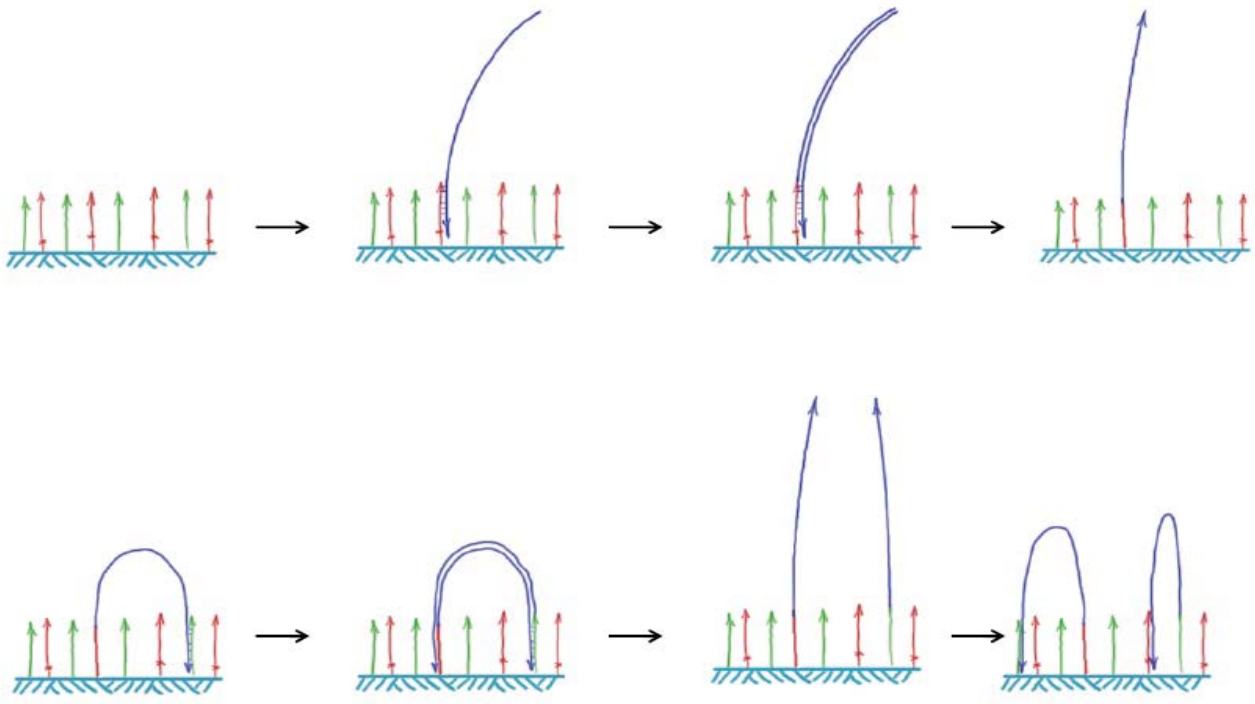
Clonal amplification by bridge PCR isothermally



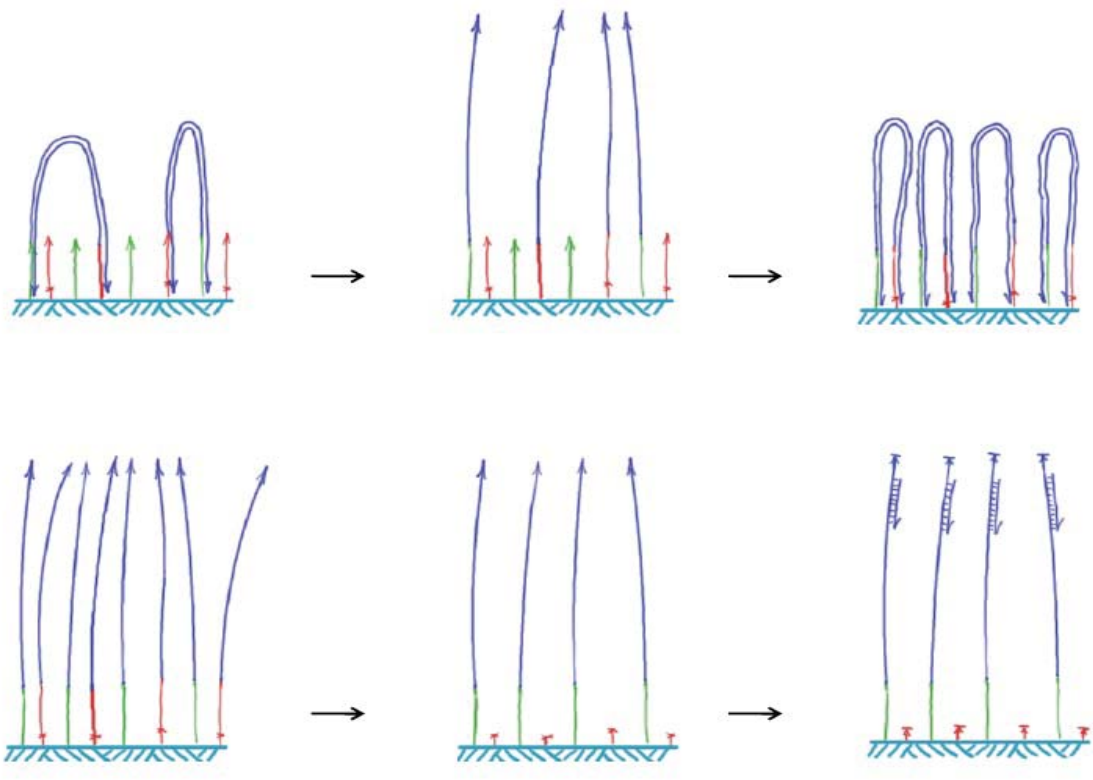




Molecular Genetics & Bioinformatics

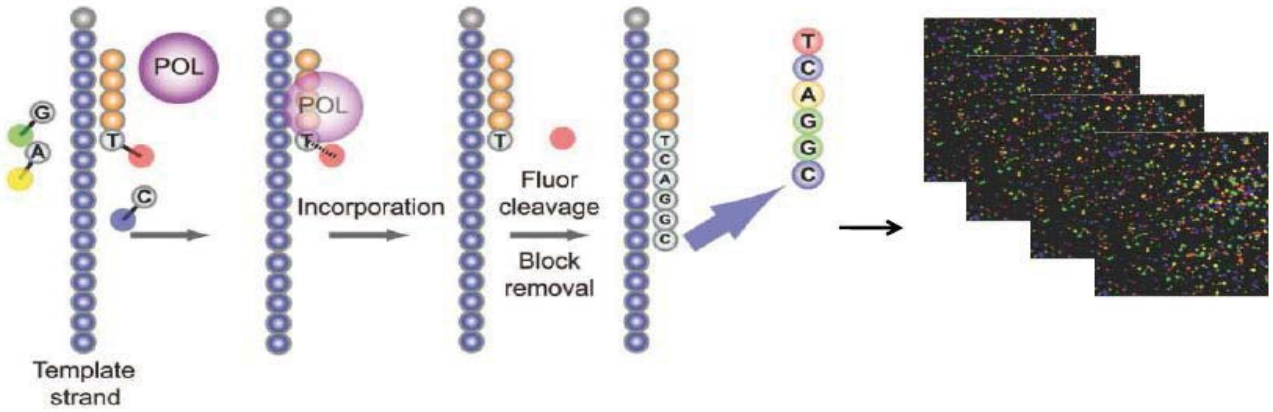


SUNLab
Molecular Genetics & Bioinformatics



SUNLab
Molecular Genetics & Bioinformatics

Sequence by reversible dye-terminators



SUNLab
Molecular Genetics & Bioinformatics

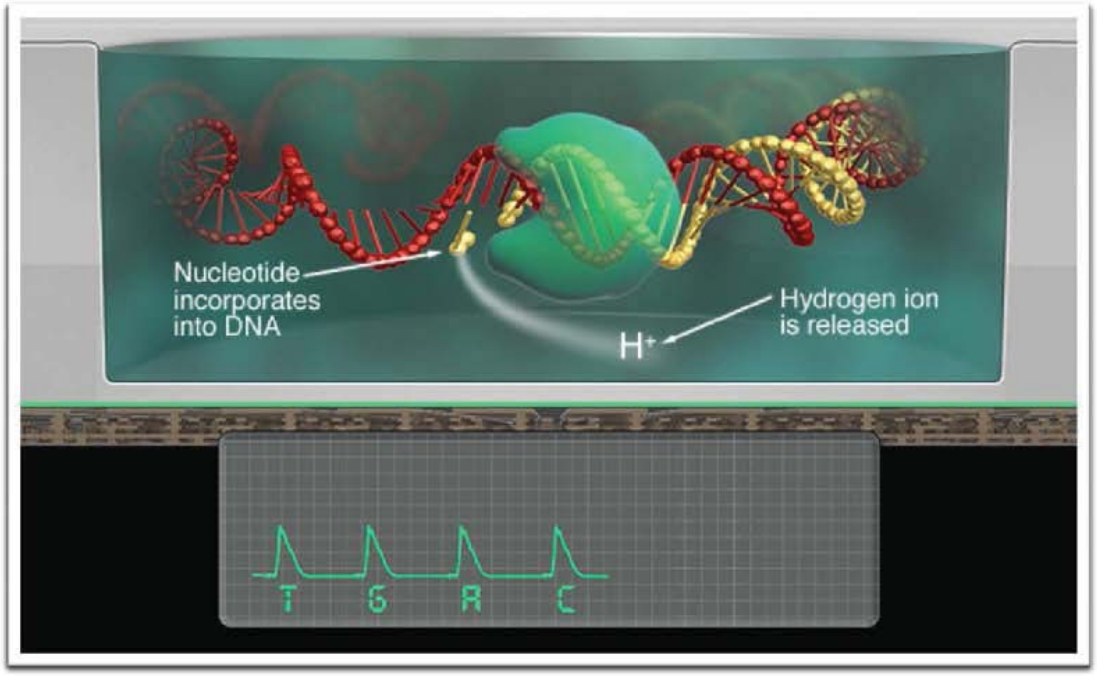
Illumina sequencers



SUNLab
Molecular Genetics & Bioinformatics

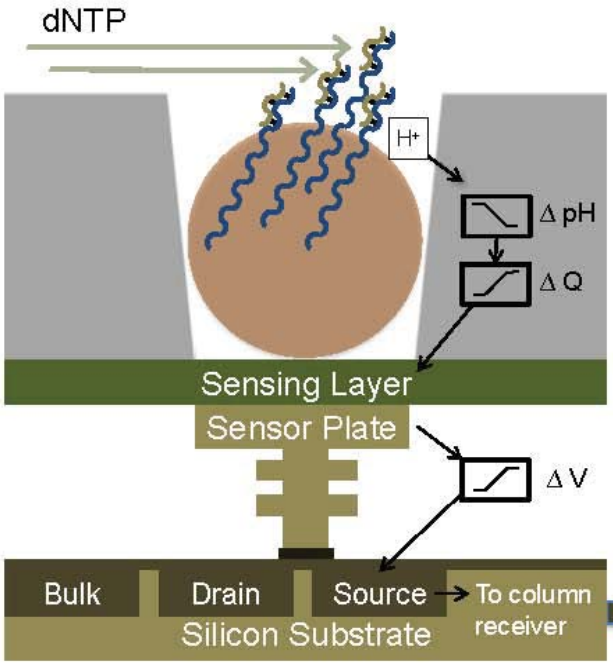
Ion Chemistry

H+ ion is released during base incorporation. Individual polymerases attached to beads are positioned in tiny wells that rest on a tiny pH meter.



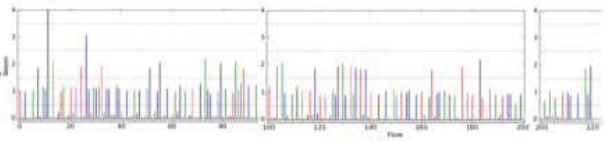
SUNLab
Molecular Genetics & Bioinformatics

Fast Direct Detection



DNA → Ions → Sequence

- Nucleotides flow sequentially over Ion semiconductor chip
- One sensor per well per sequencing reaction
- Direct detection of natural DNA extension
- Millions of sequencing reactions per chip
- Fast cycle time, real time detection



ion torrent

Confidential and Proprietary—DO NOT DUPLICATE

The Ion Series…..



Inquire
Get started with targeted NGS



Discover
Choose flexibility to support small and large projects



Decide
Save time with greater throughput

Ion chip compatibility		
Ion 510 Chip Ion 520 Chip Ion 530 Chip Ion 540 Chip	Ion 510 Chip Ion 520 Chip Ion 530 Chip Ion 540 Chip Ion 550 Chip	Ion 510 Chip Ion 520 Chip Ion 530 Chip Ion 540 Chip Ion 550 Chip
Maximum throughput/day (chip type)		
15 Gb (1 Ion 540 Chip)	30 Gb (2 Ion 540 chips)	50 Gb (2 Ion 550 chips)

Molecular Genetics & Bioinformatics

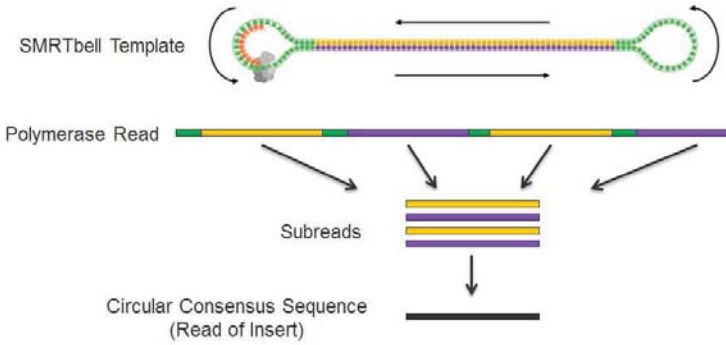
Third generation sequencing

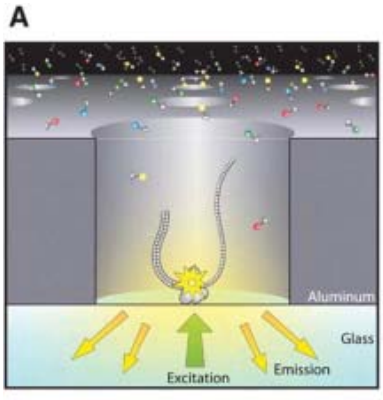


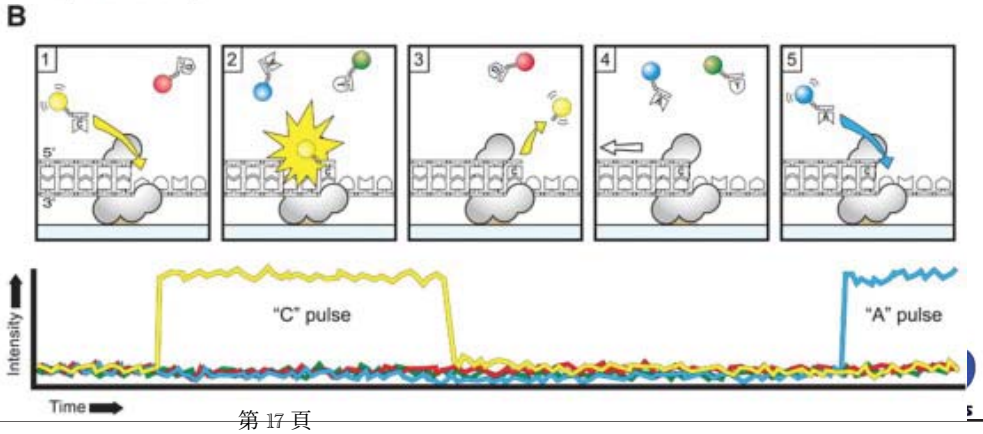
PacBio RS

SUNLab
Molecular Genetics & Bioinformatics

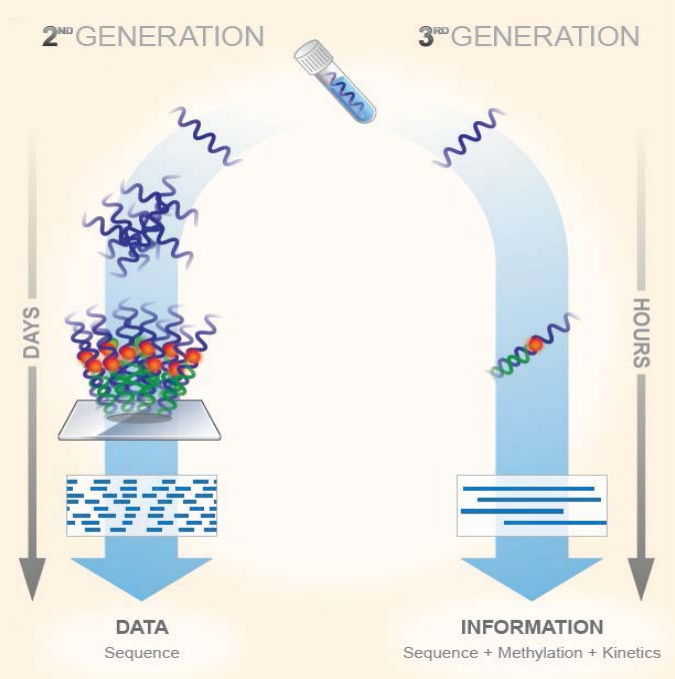
Single Molecule, Real-Time (SMRT) Sequencing







PacBio *RS*



- Pacific Biosciences has developed **Single Molecule Real Time** (SMRT™) DNA sequencing technology: PacBio *RS*.
- This technology enables, for the first time, the observation of natural DNA synthesis by a DNA polymerase as it occurs.
- This technology delivers **long reads** at single molecule level and fast time to result, enabling a new paradigm in genomic analysis.

http://www.youtube.com/watch?v=v8p4ph2MAvI

SUN Lab
Molecular Genetics & Bioinformatics

Emerging Technologies

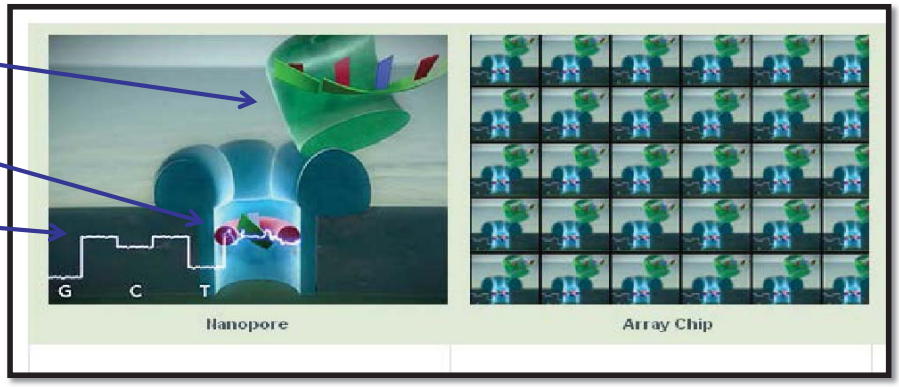
Oxford Nanopore Technology

DNA, RNA and Protein analysis

Exonuclease

Cyclodextrin

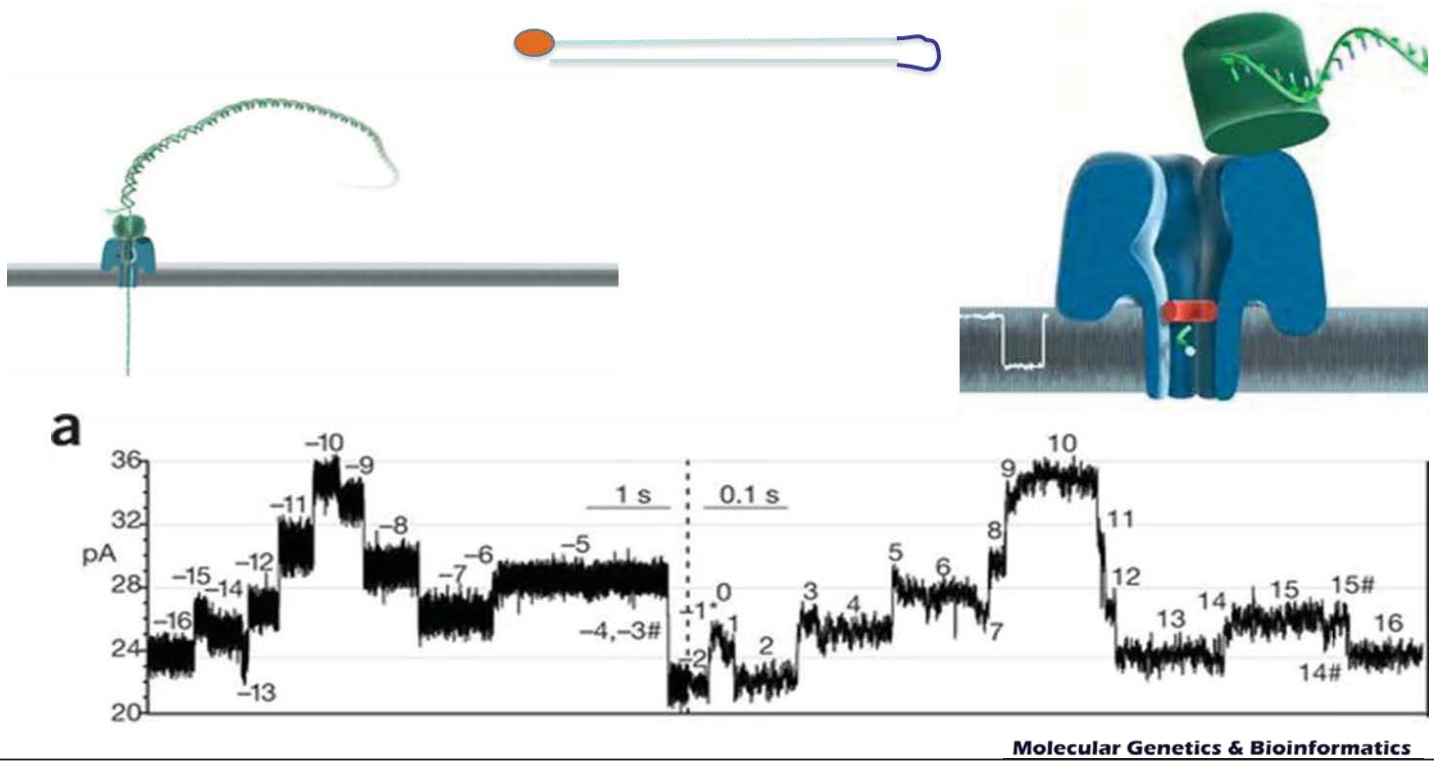
Electrically resistant Lipid bilayer



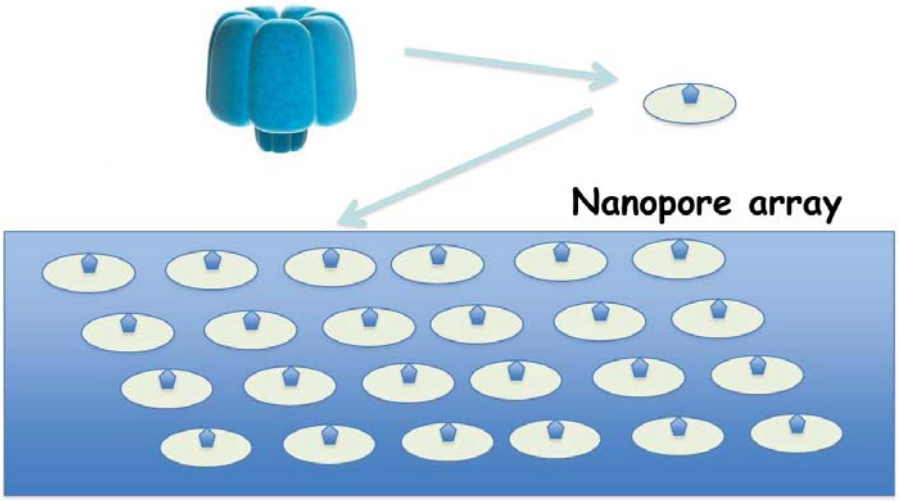
https://www.nanoporetech.com/technology/the-gridion-system/movie-an-introduction-to-the-gridion-system

SUN Lab
Molecular Genetics & Bioinformatics

Oxford Nanopore – new view on sequencing



Oxford Nanopore – new concepts



MinION



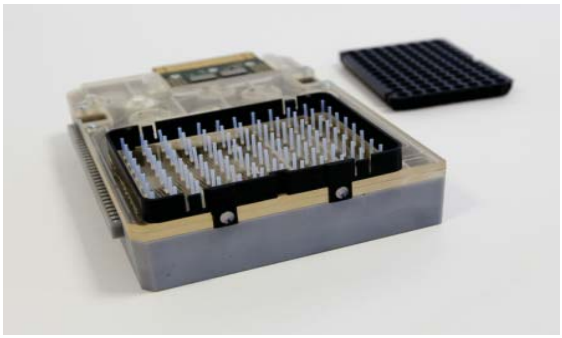
GridION



Oxford Nanopore – applications



- DNA sequencing
- Protein detection
- Protein DNA interaction
- Small molecule detection

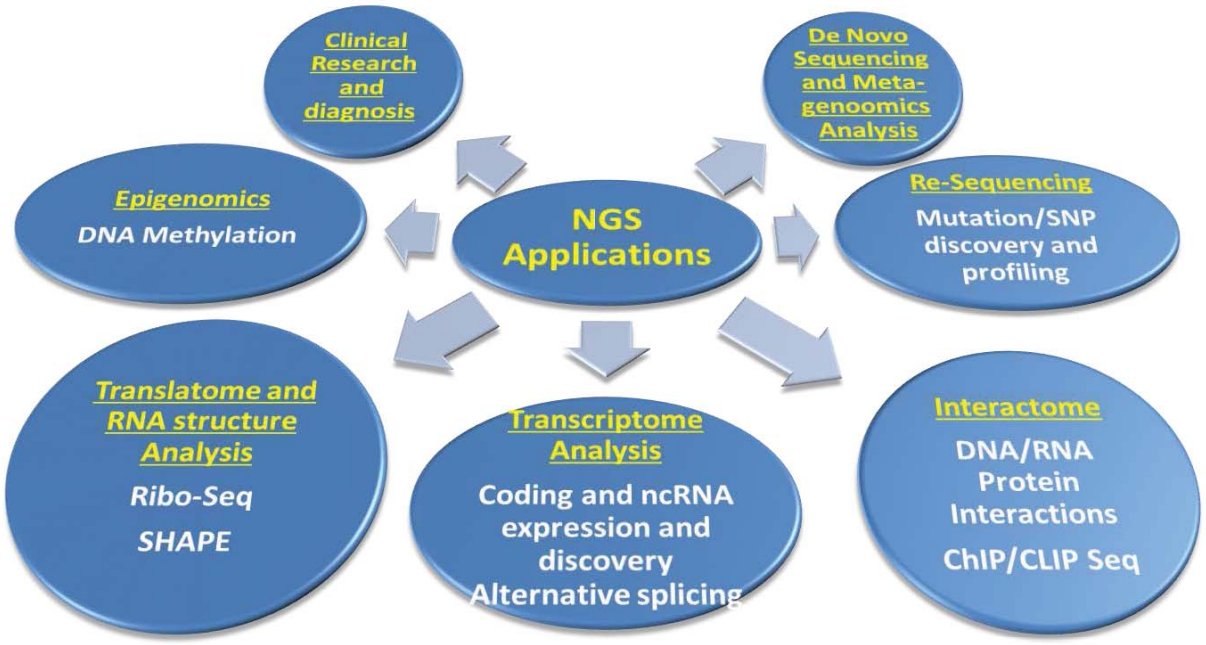


- 96 well plates for 96 samples
- Controlled time of sequencing



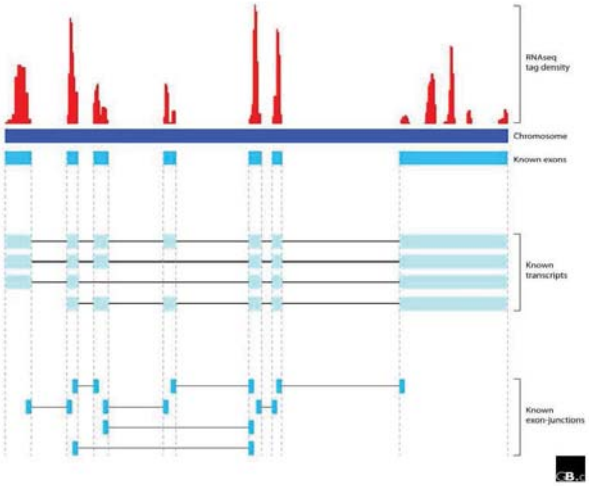
SUNLab
Molecular Genetics & Bioinformatics

Applications of NGS technologies



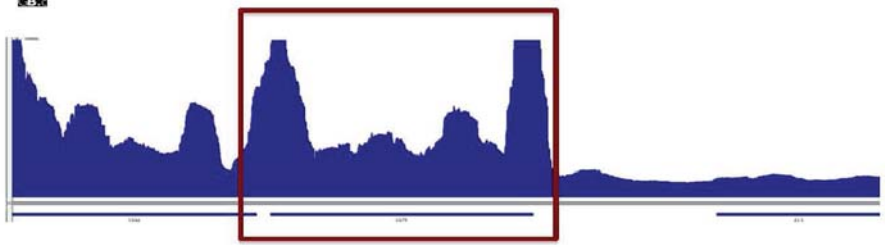
SUNLab
Molecular Genetics & Bioinformatics

RNA-Seq Results



Elucidate gene models

Quantify gene expression



Advantages of RNA-Seq

RNAseq

* Non targeted transcript detection
* No need of reference genome
* Strand specificity
* Find novels splicing sites
* Larger dynamic range
* Detects expression and SNVs
* Detects rare transcripts

....

microarrays

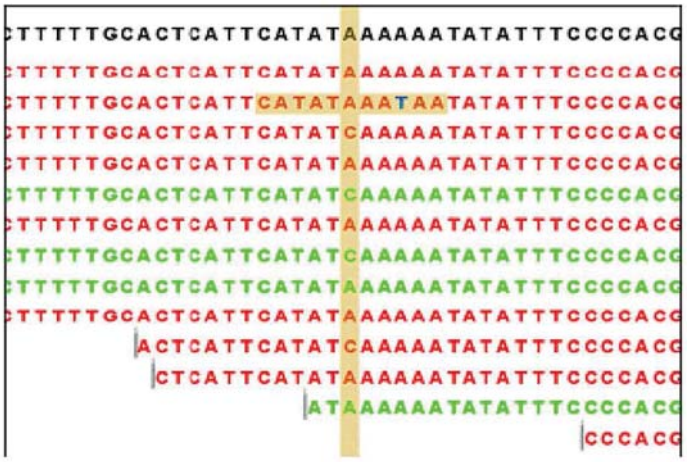
* Restricted to probes on array
* Needs genome knowledge
* Normally, not strand specific
* Exon arrays difficult to use
* Smaller dynamic range
* Does not provide sequence info
* Rare transcripts difficult

....

and.... are there any disadvantages?????

DNA re-sequencing





- Whole Genome Sequencing (WGS)
- Whole Exome Sequencing (WES)
- ChIP Sequencing
- ……
- …..

SUNLab
Molecular Genetics & Bioinformatics

What type of sequencing platforms should I choose?

-Application-dependent

-long vs. short read

-depth

-turn-around time

-……..

SUNLab
Molecular Genetics & Bioinformatics

個人履歷 (*Curriculum vitae*)

湯硯安 (Yen-An Tang)

國立成功大學 分子醫學研究所

ADDRESS: 704-56 台南市北區勝利路 367 號 4 樓

PHONE: +886- 62353535 ext.3645

E-MAIL: tangya1982@gmail.com

WEBSITE: tw.linkedin.com/pub/yen-an-tang/a2/447/b81/

■ 基本資料 / Personal Information

1. **國籍**/Nationality **台灣**/Taiwanese
2. **性別**/Sex **男**/Male
3. **出生日期**/Date of birth 01 Aug 1982

■ 主要學、經歷 / Education and Experiences

1. 2018/07-present **博士後研究員 /國立成功大學 基因體醫學中心**

   Postdoctoral research fellow, Center for Genomic Medicine, National Cheng Kung University, Taiwan

2. 2015/01-2018/06 **博士後研究員 / 新加坡基因體研究中心**

   Postdoctoral research fellow, Cancer Therapeutics & Stratified Oncology, Genome Institute of Singapore (GIS), A*STAR

3. 2013/01-2014/12 **博士後研究員 / 國立成功大學 藥理所**

   Postdoctoral research fellow, Department of Pharmacology, National Cheng Kung University, Taiwan

4. 2009/09-2012/12 **博士 / 國立成功大學 基礎醫學研究所**

   PhD program, Institute of Basic Medical Sciences, National Cheng Kung University, Taiwan

5. 2008/10-2009/08 **研究助理 / 國立成功大學 藥理所**

   Research assistant, Department of Pharmacology, National Cheng Kung University, Taiwan

6. 2005/09-2007/06 **碩士 / 國立清華大學 生物科技研究所**

   Master program, Institute of Biotechnology, National Tsing Hua University, Taiwan

7. 2000/09-2004/06 **學士 / 國立臺灣師範大學 生命科學系**

   Bachelor program, Department of Life Sciences, National Taiwan Normal University, Taiwan

■ 領域專長 / Fields of specialty

1.	Data analysis	Whole Exome sequencing; Microarray analysis; RNA-seq; ChIP-seq; Metagenomics; SPSS; Ingenuity Pathway Analysis (IPA); MetaCore software; Gene Set Enrichment Analysis (GSEA); Gene Spring software
2.	Cancer biology	Cancer epigenetics; Cancer metastasis; Cancer stem cells; Tumor-associated fibroblast; Anti-cancer therapeutics
3.	Biochemistry	Protein phosphorylation and ubiquitination; Nucleotide excision repair;
4.	Animal model	Orthotopic transplant model of breast and colorectal cancers; Patient-derived xenografts (PDXs) model; Subcutaneous xenograft model; Tail vein injection metastasis model

■ 論文著作 / Selected Publications:

1. **Yen-An Tang**, Y. Chen, Y. Bao, S. Mahara, S.M.J.M.Yatim, G. Oguz, P. L. Lee, M. Feng, Y. Cai, E. Y. Tan, S. S. Fong, Z. Yang, P. Lan, X. Wu, Q. Yu*. (**2018**) Hypoxic tumor microenvironment activates GLI2 via HIF-1α and TGF-β2 to promote chemoresistance in colorectal cancer. ***PNAS***. 115: E5990-E5999. (SCI, IF=9.504)

2. J. Jen, **Yen-An Tang**, Y.-H. Lu, C.-C. Lin, W.-W. Lai, Y.-C. Wang*. (**2017**). Oct4 transcriptionally regulates the expression of long non-coding RNAs NEAT1 and MALAT1 to promote lung cancer progression. ***Mol. Cancer***, 16:104. (SCI, IF=7.776)

3. **Yen-An Tang**, C.-H. Chen, H. S. Sun, C.-P. Cheng, V. S. Tseng, H.-S. Hsu, W.-C. Su, W.-W. Lai, Y.-C. Wang*. (**2015**). Global Oct4 target gene analysis reveals novel downstream PTEN and TNC genes required for drug-resistance and metastasis in lung cancer. ***Nucleic Acids Res.***, 43:1593-1608. (SCI, IF=11.561)

4. R.-C. Tseng, J.-M. Chang, J.-H. Chen, W.-R. Huang, **Yen-An Tang**, I.-Y. Kuo, J.-J. Yan, W.-W. Lai, Y.-C. Wang*. (**2015**). Deregulation of SLIT2-mediated cdc42 activity is associated with esophageal cancer metastasis and poor prognosis. ***J. Thorac. Oncol.***, 10:189-198. (SCI, IF=10.336)

5. Y.-C. Yang, **Yen-An Tang (co-first)**, J.-M. Shieh, R.-K. Lin, H.-S. Hsu, Y.-C. Wang*. (**2014**). DNMT3B overexpression by deregulation of FOXO3a-mediated transcription repression and MDM2 overexpression in lung cancer. ***J. Thorac. Oncol.***, 9:1305-1315. (SCI, IF=10.336)

6. **Yen-An Tang**, R.-K. Lin, Y.-T. Tsai, H.-S. Hsu, Y.-C. Yang, C.-Y. Chen, Y.-C. Wang*. (**2012**). MDM2 overexpression deregulates the transcriptional control of RB/E2F leading to DNA methyltransferase 3A overexpression in lung cancer. ***Clin. Cancer Res.***, 18:4325-4333. (SCI, IF=10.199)

Exome sequencing data analysis

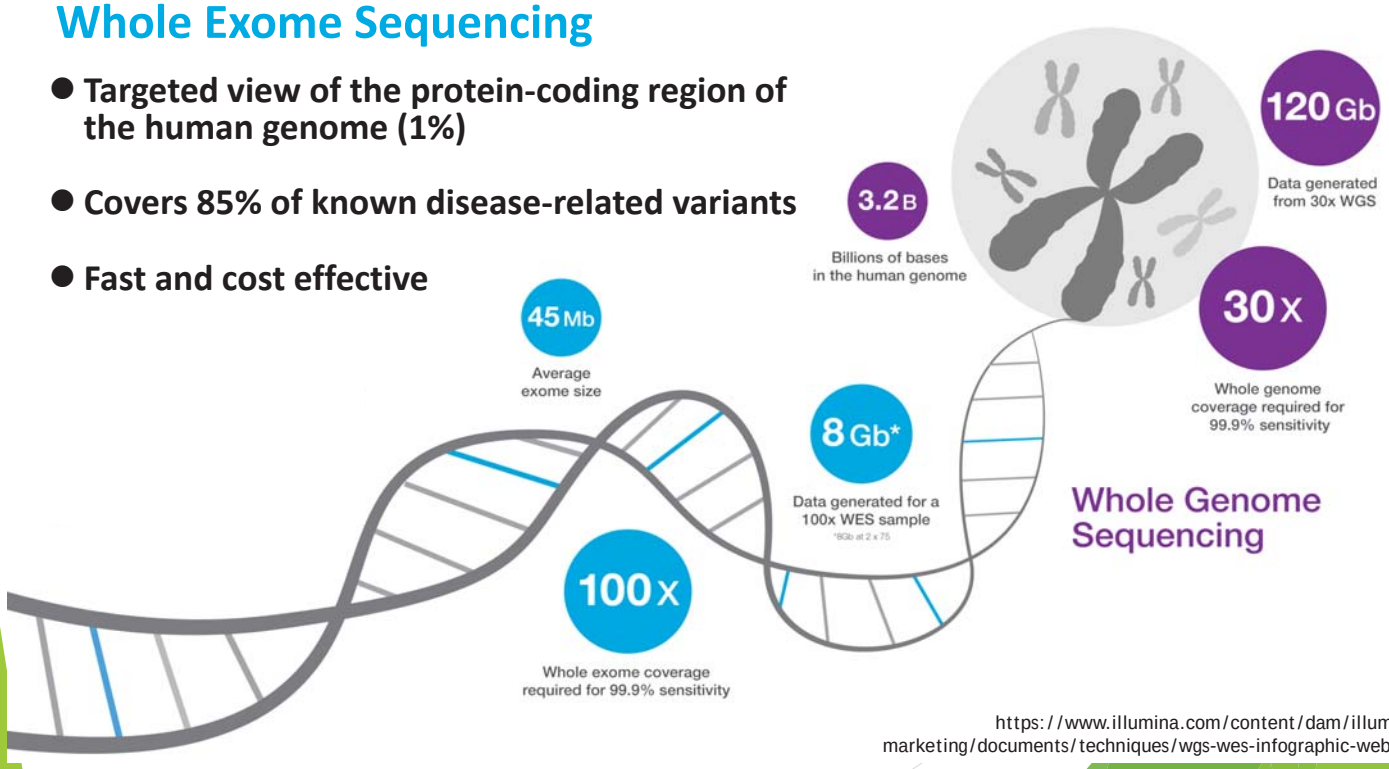
Yen An Tang, PhD

CENTER FOR GENOMIC MEDICINE,
NATIONAL CHENG KUNG UNIVERSITY

2019/07/10

Whole Exome Sequencing

- Targeted view of the protein-coding region of the human genome (1%)
- Covers 85% of known disease-related variants
- Fast and cost effective



https://www.illumina.com/content/dam/illumina-marketing/documents/techniques/wgs-wes-infographic-web.pdf

Outline

WES–Variant Calling pipeline

FASTQ

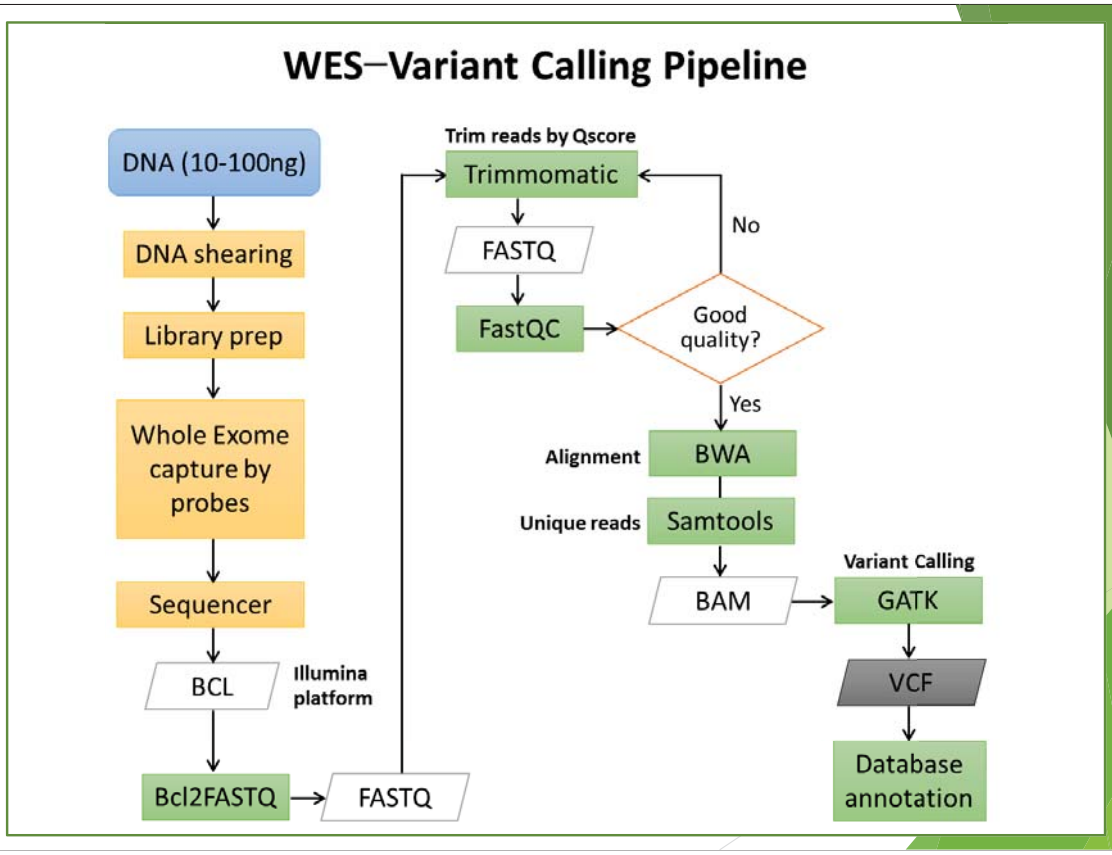
Trimming

QC

Alignment

Variant Calling

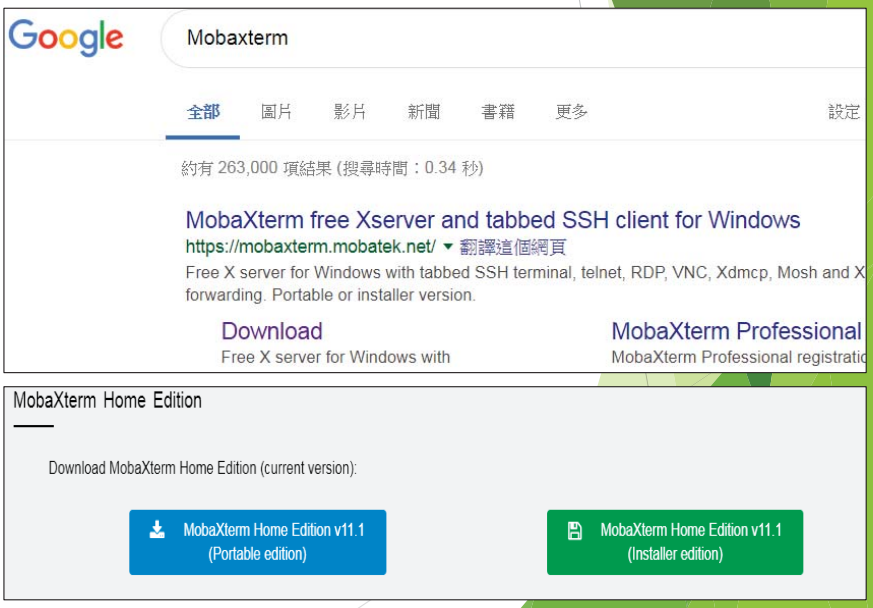
Visualization



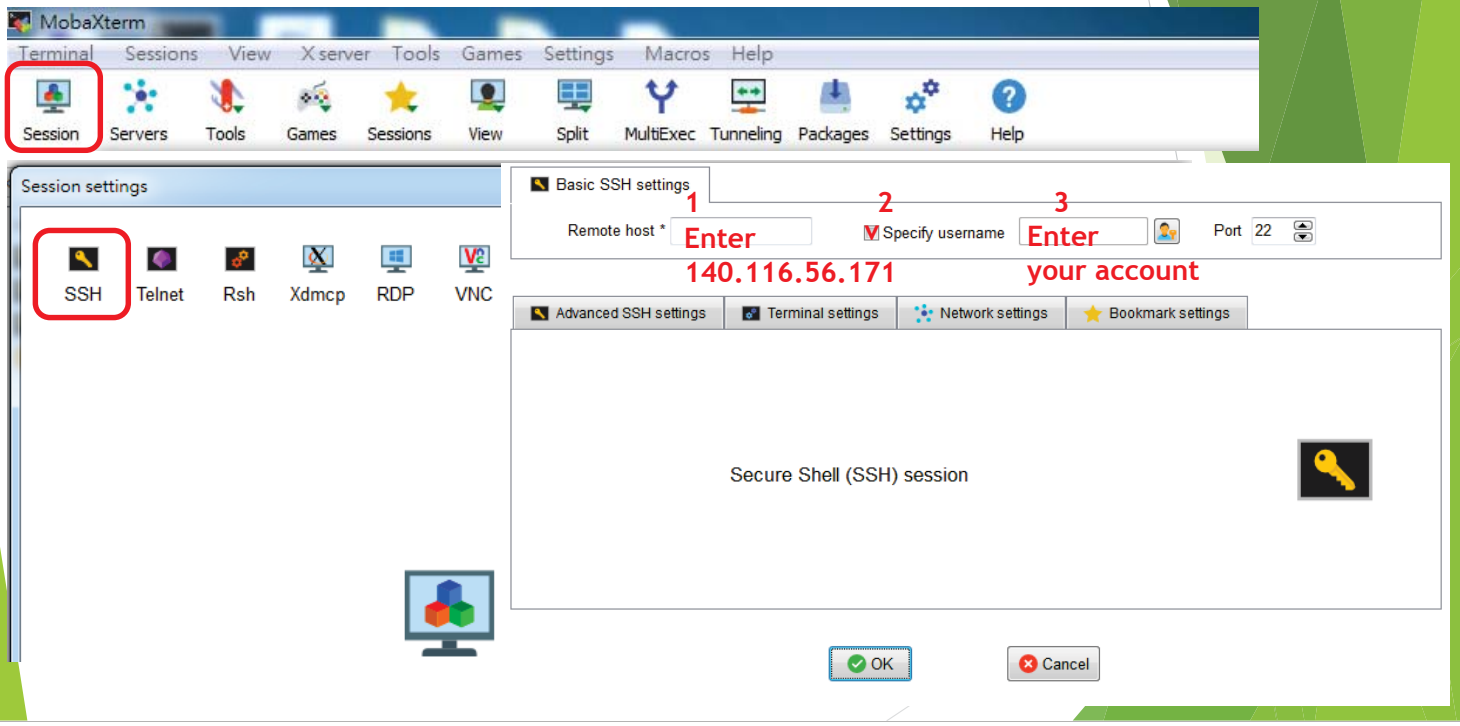
Linux Interface

- Windows: PuTTY/PieTTY
- Mac: Terminal

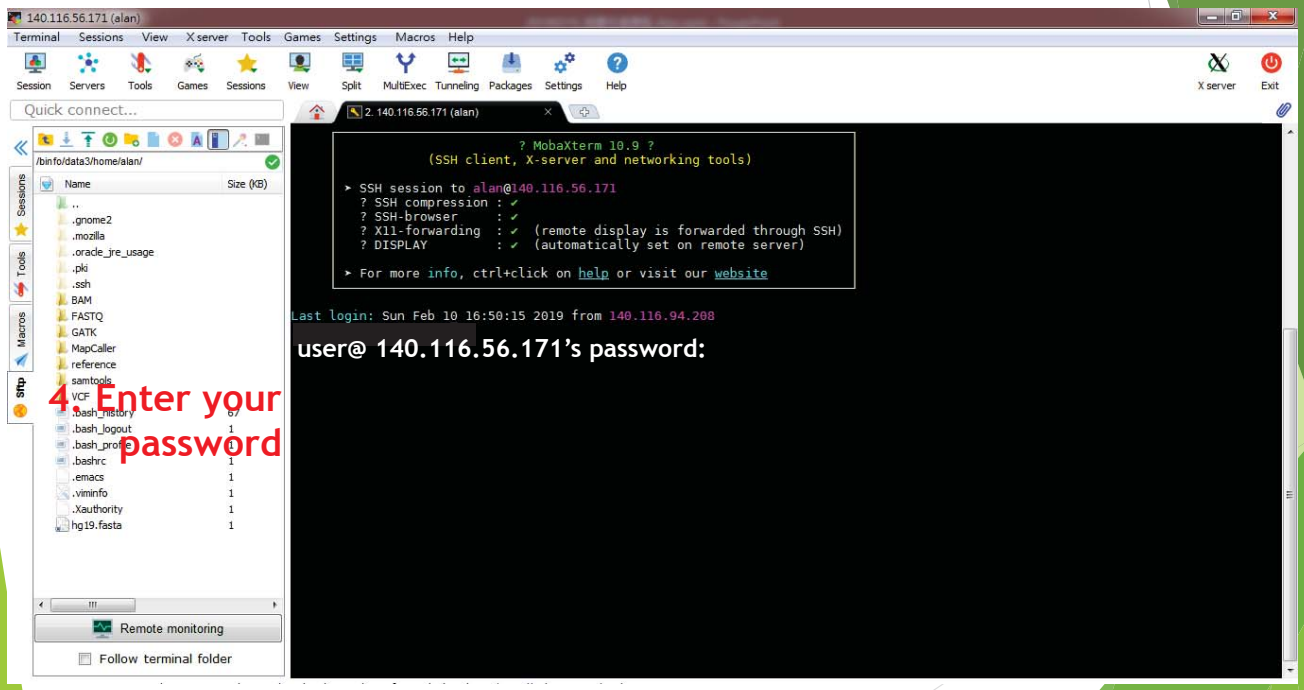
- Xwindow: MobaXterm
- VNC: RealVNC



MobaXterm



MobaXterm



4. Enter your password

Basic command

- ls (list)
- ls –l (list)
- cd (change directory)
- mkdir (make directory)
- rm (remove)
- cp (copy)
- mv (move)
- cat/zcat (print/print compressed file)
- less
- help

1 Sequencer output Format: FATSQ

Example:

@A00361:16:H2YVGDSXX:3:1101:1235 1:N:0:CTGAAGCT+GGCTCTGA
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

- Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.
- Line 4 encodes the Phred Score(Quality Score) of Sequence.

1 Phred Score

Calculating Phred Quality Scores

Q scores are defined as a property that is logarithmically related to the base calling error probabilities (P)[2].

$$Q = -10 \log_{10} P$$

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Phred 33:

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

2 Trimming

- Purpose of trimming
  --Remove adapter/barcode
  --Remove low quality bases/reads

- Trimming tool： Trimmomatic or cutadapt

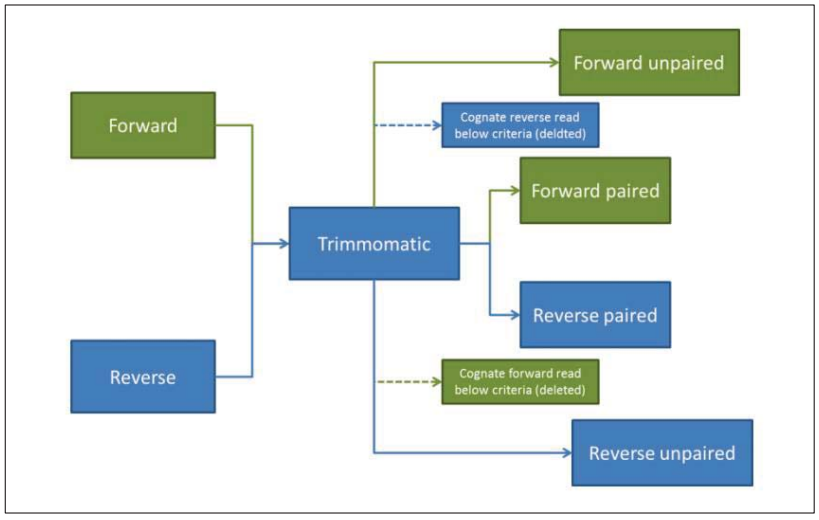
- Command:

java -jar /tools/Trimmomatic/trimmomatic_0.38/trimmomatic-0.38.jar PE -phred33 <input_R1.fastq> <input_R2.fastq> -baseout <output_trim> SLIDINGWINDOW:4:30 MINLEN:15

2 Trimming

- Command:

java -jar /tools/Trimmomatic/trimmomatic_0.38/trimmomatic-0.38.jar PE -phred33 <input_R1.fastq> <input_R2.fastq> -baseout <output_trim> SLIDINGWINDOW:4:30 MINLEN:15

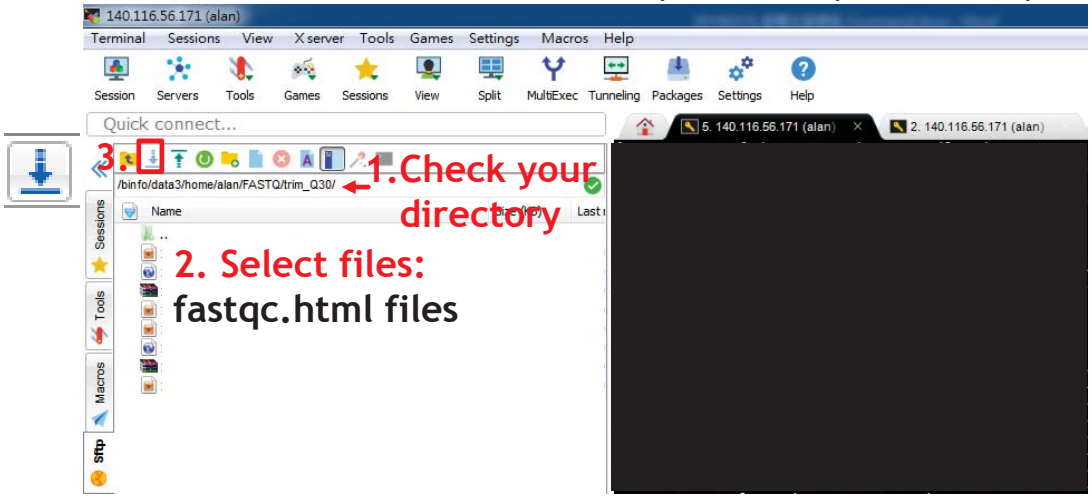


3 Quality Control: FastQC

- Command:

/tools/FastQC/fastqc <input 1> <input 2> <input 3> <input 4>

→ It will generate fastqc.html and fastqc.gz files
→ Download the files (SFTP) to local computer and open the fastqc.html file



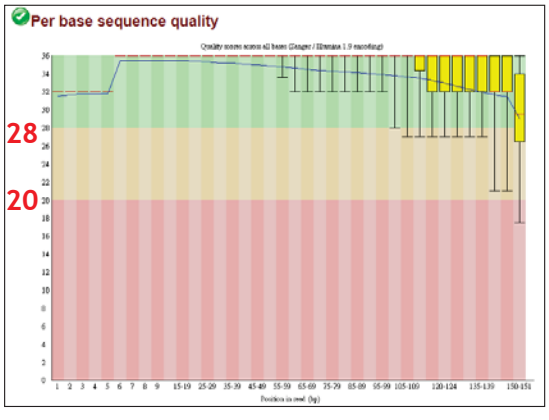
3 Quality Control: FastQC

- Command:

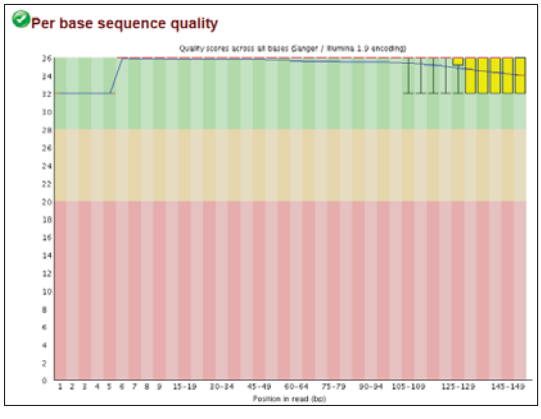
/tools/FastQC/fastqc <input 1> <input 2> <input 3> <input 4>

→ It will generate fastqc.html and fastqc.gz files
→ Download the files and open the fastqc.html file

11111-A_R1.fastqc.html

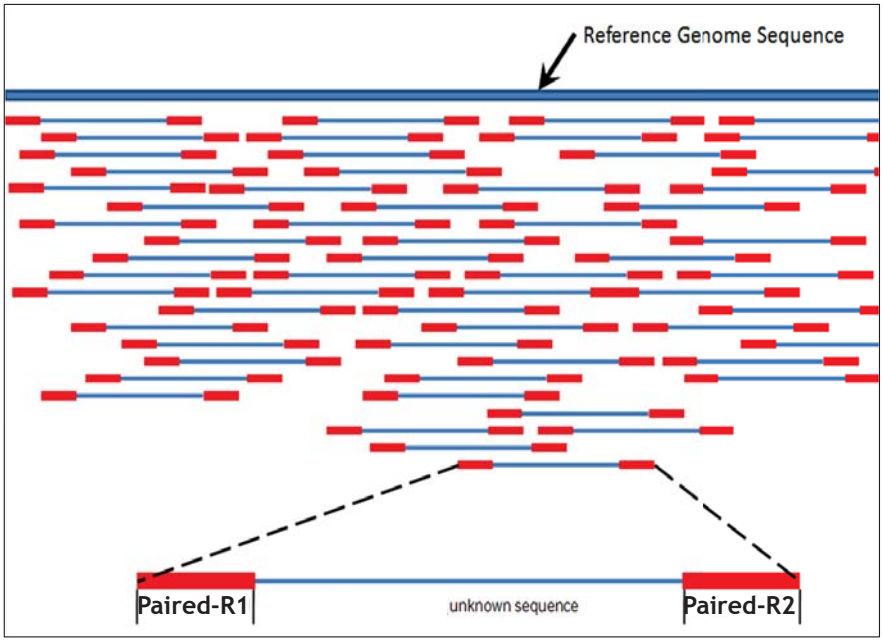


11111-A_R1_trim_q30.fastqc.html



4 Alignment

- Steps: 1. indexing
  2. mapping
  3. Sam→ BAM
  4. sorting
- Tools： BWA/Bowtie
  Samtools



4 Alignment: BWA

Indexing:

Command: bwa index –p <prefix> <ref.fa>

/tools/bwa-0.7.17/bwa index –p hg19 hg19.fasta

```
Oct 29 15:38 hg19.fasta
Oct 29 15:38 hg19.fasta.amb
Oct 29 15:39 hg19.fasta.ann
Oct 29 15:39 hg19.fasta.bwt
Oct 29 15:39 hg19.fasta.fai
Oct 29 15:39 hg19.fasta.pac
Oct 29 15:40 hg19.fasta.sa
```

4 Alignment: BWA

Mapping: BWA

- Three algorithms (module name)
  1. BWA-backtrack (bwa aln)
     - reads up to 100pb
  2. BWA-SW (bwa bwasw)
     - reads from 70bp~1Mbp
  3. BWA-MEM (bwa mem)
     - reads from 70bp~1Mbp

- BWA-MEM is faster and more accurate than BWA-SW

http://bio-bwa.sourceforge.net/bwa.shtml

4 Alignment: BWA

Mapping:

Command: /tools/bwa-0.7.6a/bwa mem <reference_prefix> <trim_R1.fastq> <trim_R2.fastq > > output.sam

4 Alignment: BWA

SAM format:

Command: cat 11111-A.chr17.sam | less

```
1                                   2     3       4          5     6
NB551359:21:HC2KVAFXY:1:11101:1004:12221      83    chr17   41197197      60    151M    =     41197175      -173    CC
CAAGGACTATTCTGACTTTAAGTCACATAATCGATCCCAAGCACTCTCCTTCCATTGAAGGGTCTGACTCTCTGCCTTTGTGAACACAGGGTTTTAGAGAAGTAAACTTAGGGAAACCAGCTATTCTCTT
GAGGCCAAGCCACTCTGTN AEEEA/AEEEEAEEEAEEEEEEAEEEEEE<EEEEEEEEEEEEEEEEEEEEEEEEEAEEEEE<EEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
EEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAAAA# NM:i:1  MD:Z:150C0      AS:i:150        XS:i:0  RG:Z:NextSeq550 2019
```

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{31}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,2^{31}-1]	Position of the mate/next read
9	TLEN	Int	[-2^{31}+1,2^{31}-1]	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

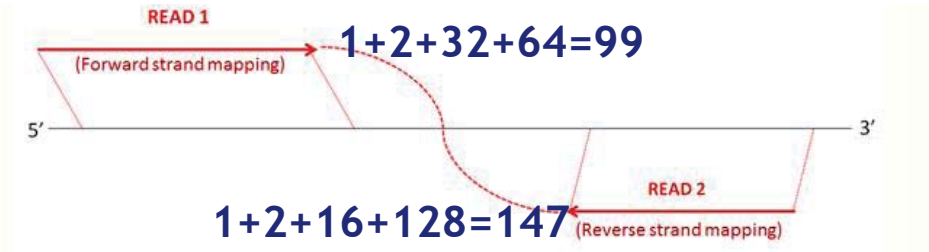
4 Alignment: BWA

SAM format:

Flag

FLAG: Combination of bitwise FLAGs.[7] Each bit is explained in the following table:

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment



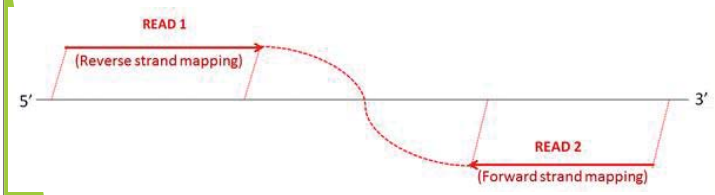
4 Alignment: BWA

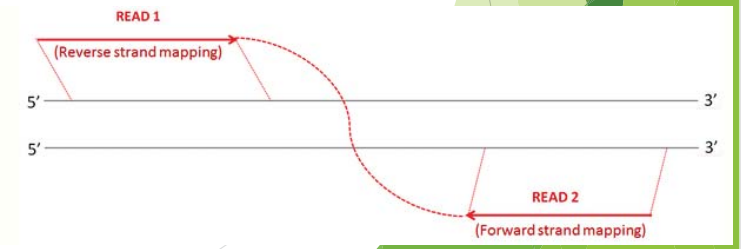
SAM format:

Flag

FLAG: Combination of bitwise FLAGs.[7] Each bit is explained in the following table:

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment





4 Alignment: BWA

SAM format:

CIGAR

Col	Field
1	QNAME
2	FLAG
3	RNAME
4	POS
5	MAPQ
6	CIGAR
7	RNEXT
8	PNEXT
9	TLEN
10	SEQ
11	QUAL

CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

Op	BAM	Description	Consumes query	Consumes reference
M	0	alignment match (can be a sequence match or mismatch)	yes	yes
I	1	insertion to the reference	yes	no
D	2	deletion from the reference	no	yes
N	3	skipped region from the reference	no	yes
S	4	soft clipping (clipped sequences present in SEQ)	yes	no
H	5	hard clipping (clipped sequences NOT present in SEQ)	no	no
P	6	padding (silent deletion from padded reference)	no	no
=	7	sequence match	yes	yes
X	8	sequence mismatch	yes	yes

- "Consumes query" and "consumes reference" indicate whether the CIGAR operation causes the alignment to step along the query sequence and the reference sequence respectively.

For example: 151M or 37M1D106M

4 Alignment: Samtools

- Steps
  1. Convert Sam to Bam
     - Bam is compressed version of Sam.
  2. Fixmate
     - Remove secondary alignments.
  3. Sort
  4. Index

4 Alignment: Samtools

Commands

1. Convert Sam to Bam
   - tools/samtools-1.2/samtools view -b <input.sam> -o <output.bam>
2. Fixmate
   - tools/samtools-1.2/samtools fixmate –r <input: bam> <output: fixmate.bam>
3. Sort
   - tools/samtools-1.2/samtools sort <input: fixmate.bam> <output: fixmate.sorted.bam>
4. Index
   - tools/samtools-1.2/samtools index <input: fixmate.sorted.bam>

5 Variant Calling: GATK-HaplotypeCaller

Call germline SNPs and indels via local re-assembly of haplotypes. The HaplotypeCaller is capable of calling SNPs and indels simultaneously via local de-novo assembly of haplotypes in an active region.

- https://software.broadinstitute.org/gatk/documentation/tooldocs/current/org_broadinstitute_hellbender_tools_walkers_haplotypecaller_HaplotypeCaller.php
- https://software.broadinstitute.org/gatk/best-practices/workflow?id=11145

Command:

/tools/gatk-4.1.0.0/gatk --java-options "-Xmx80g“ HaplotypeCaller -R <reference> -I <input.bam> -O <output.vcf.gz> --intervals chr17 --native-pair-hmm-threads 24

5 Variant Calling: GATK-HaplotypeCaller

VCF format:

VCF (Variant Call Format) is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome.

##Header: Version, Filter, Format, Info, Contig

```
##fileformat=VCFv4.2
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order l
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mat
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as def

##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as l
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=ExcessHet,Number=1,Type=Float,Description="Phred-scaled p-value for exact test of excess heterozy
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect stra
```

5 Variant Calling: GATK-HaplotypeCaller

VCF format:

#CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, FORMAT, NAME

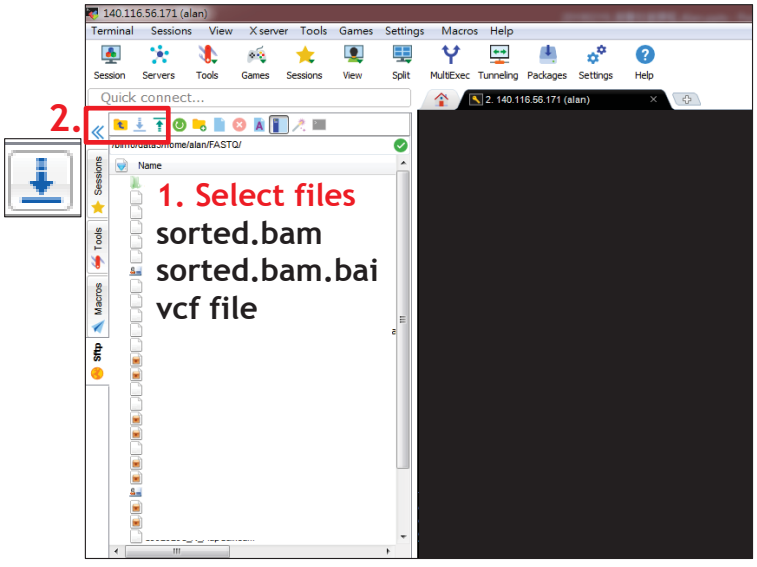
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	11111-A_R
chr17	6115	.	G	C	5081.03	.	AC=2;AF=1.00;AN=2;DP=184;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=27.61;SOR=0.932	GT:AD:
chr17	6157	.	A	G	3873.03	.	AC=2;AF=1.00;AN=2;DP=140;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=27.66;SOR=0.980	GT:AD:
chr17	7242	.	G	T	37.28	.	AC=2;AF=1.00;AN=2;DP=2;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=60.00;QD=18.64;SOR=0.693	GT:AD:
chr17	11743	.	A	G	270.98	.	AC=2;AF=1.00;AN=2;DP=7;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=25.36;SOR=4.174	GT:AD:DP:GQ:PL
chr17	11830	.	A	G	1040.03	.	AC=2;AF=1.00;AN=2;DP=33;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=31.52;SOR=2.517	GT:AD:
chr17	13905	.	G	A	2344.03	.	AC=2;AF=1.00;AN=2;DP=84;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=27.91;SOR=0.954	GT:AD:
chr17	14008	.	G	C	2285.03	.	AC=2;AF=1.00;AN=2;DP=82;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=27.87;SOR=1.220	GT:AD:
chr17	15463	.	C	T	76.80	.	AC=2;AF=1.00;AN=2;DP=3;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=60.00;QD=25.60;SOR=1.179	GT:AD:
chr17	19549	.	C	G	60.80	.	AC=2;AF=1.00;AN=2;DP=3;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=59.02;QD=20.27;SOR=1.179	GT:AD:
chr17	22211	.	G	A	898.03	.	AC=2;AF=1.00;AN=2;DP=25;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=28.73;SOR=5.058	GT:AD:
chr17	32217	.	C	G	37.28	.	AC=2;AF=1.00;AN=2;DP=2;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=60.00;QD=18.64;SOR=0.693	GT:AD:
chr17	32249	.	C	G	37.28	.	AC=2;AF=1.00;AN=2;DP=2;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=60.00;QD=18.64;SOR=0.693	GT:AD:
chr17	32309	.	G	C	78.28	.	AC=2;AF=1.00;AN=2;DP=2;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=60.00;QD=30.97;SOR=0.693	GT:AD:
chr17	32311	.	A	G	78.28	.	AC=2;AF=1.00;AN=2;DP=2;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=60.00;QD=27.24;SOR=0.693	GT:AD:
chr17	37234	.	T	C	37.28	.	AC=2;AF=1.00;AN=2;DP=2;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=60.00;QD=18.64;SOR=0.693	GT:AD:
chr17	46107	.	A	G	37.28	.	AC=2;AF=1.00;AN=2;DP=2;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=60.00;QD=18.64;SOR=0.693	GT:AD:
chr17	50170	.	A	G	37.28	.	AC=2;AF=1.00;AN=2;DP=2;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=60.00;QD=18.64;SOR=0.693	GT:AD:

6 Download data from Linux

download Bam, Bam.Bai, and VCF file to local computer:

→ Command: gunzip <vcf.gz>

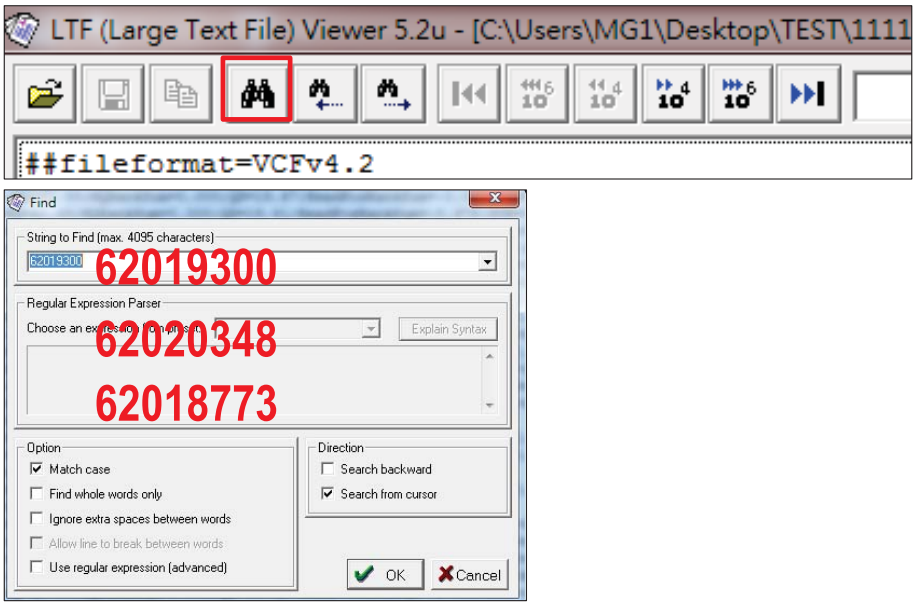
→ Use the SFTP (left window) to download the files



7 View of VCF file

OPEN VCF file using LTFViewr5u.exe or EXCEL (lesser than 10M)

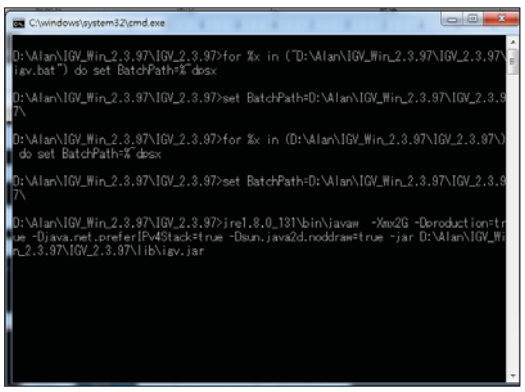


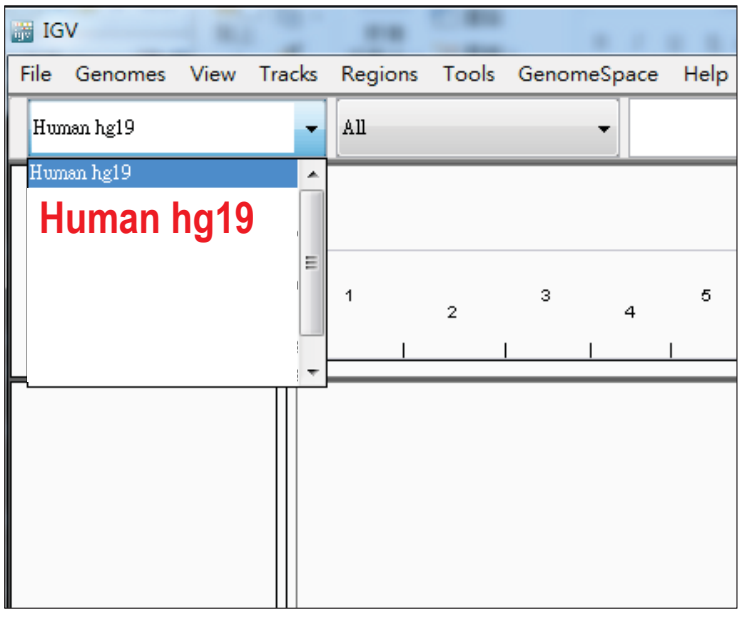


8 Visualization of reads: BAM file

OPEN Bam file using IGV software

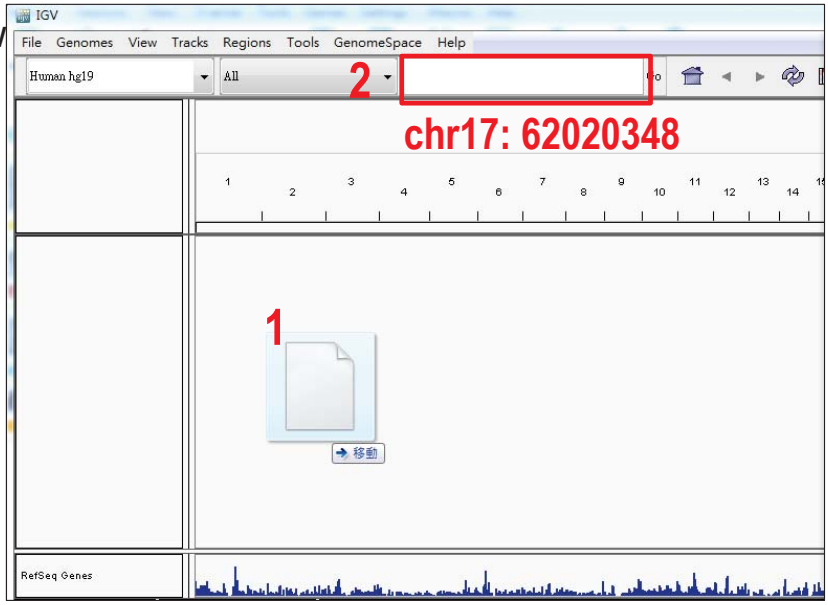






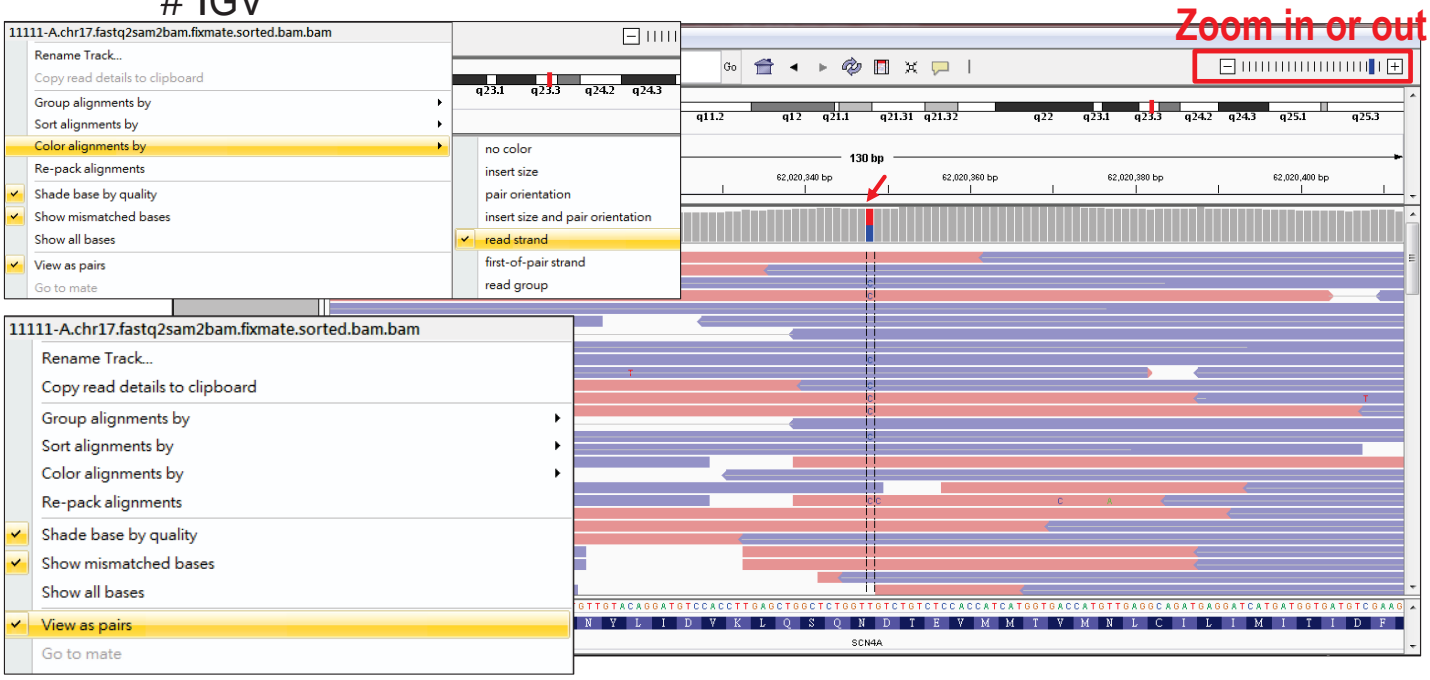
8 Visualization of reads: BAM file

1. Drag your Bam file into IGV window
2. Key in the chromosome and coordinate → Enter



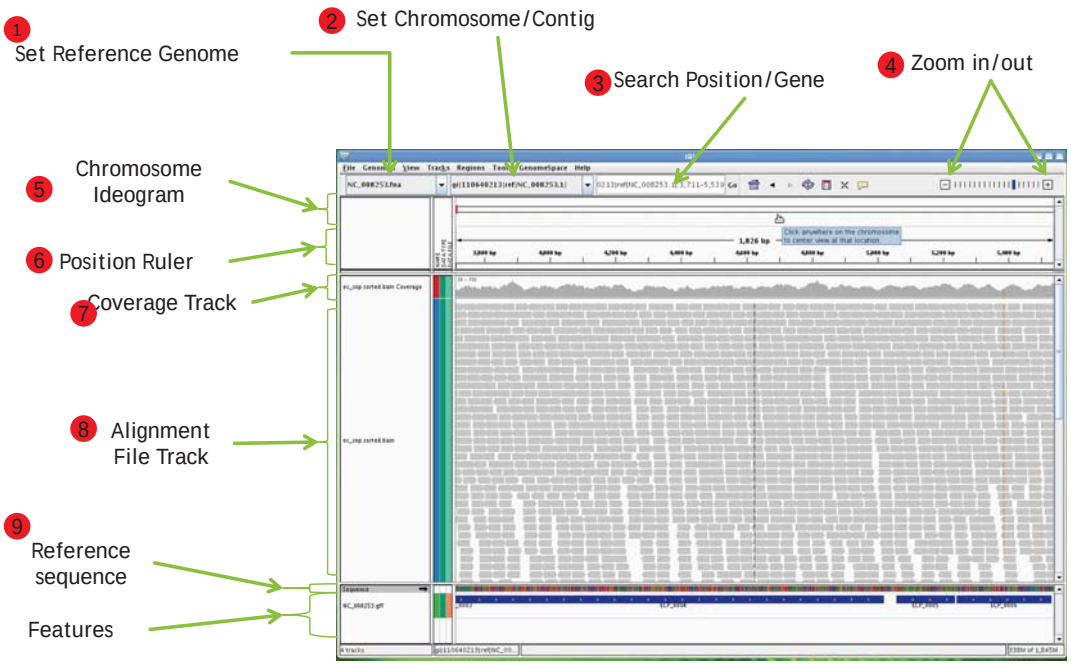
8 Visualization of reads: BAM file

IGV



8 IGV

http://software.broadinstitute.org/software/igv/regionsofinterest



THANKS

Commands

#Create directory

mkdir trim_Q30

ls

1#Trimming ~1min5sec

qsub ./qsub/trimming.sh

qstat

[command]

java -jar /share/app/Trimmomatic/Trimmomatic-0.38/trimmomatic-0.38.jar PE
-phred33 11111-A.chr17.r1.fastq.gz 11111-A.chr17.r2.fastq.gz
-baseout ./trim_Q30/11111-A.chr17_trim.fastq.gz SLIDINGWINDOW:4:30 MINLEN:15

2#FASTQC ~1min30sec

cd trim_Q30

qsub ../qsub/fastqc.sh

qstat

[command]

/share/app/FastQC/fastqc 11111-A.chr17_trim_1P.fastq.gz
11111-A.chr17_trim_2P.fastq.gz

3#Mapping ~1min30sec

qsub ../qsub/mapping.sh

qstat

[command]

/share/app/bwa-0.7.6a/bwa mem -M -t 12
/binfo/data3/home/GATK_training/reference/hg19.fasta -R
"@RG\tID:NextSeq550_2019\tSM:11111-A_R\tPL:NextSeq\tPI:550"
11111-A.chr17_trim_1P.fastq.gz 11111-A.chr17_trim_2P.fastq.gz >
11111-A.chr17.sam

4#Convert SAM to BAM ~2min30sec

qsub ../qsub/sam2bam.sh
qstat

[command]
date; echo ‘convert’; /share/app/samtools-1.2/samtools view -b 11111-A.chr17.sam -o 11111-A.chr17.bam; date; echo ‘fixmate’; /share/app/samtools-1.2/samtools fixmate -r 11111-A.chr17.bam 11111-A.chr17.fixmate.bam; date; echo ‘sort’; /share/app/samtools-1.2/samtools sort 11111-A.chr17.fixmate.bam 11111-A.chr17.fixmate.sorted; date; echo ‘index’; /share/app/samtools-1.2/samtools index 11111-A.chr17.fixmate.sorted.bam; date

5#GATK-HaplotypeCaller ~3min30sec
qsub ../qsub/gatk.sh
qstat

[command]
date;/share/app/GATK/jre1.8.0_151/bin/java -jar /share/app/GATK/gatk-4.1.1.0/gatk-package-4.1.1.0-local.jar HaplotypeCaller -R /binfo/data3/home/GATK_training/reference/hg19.fasta -I 11111-A.chr17.fixmate.sorted.bam -O 11111-A.chr17.vcf.gz --intervals chr17 --native-pair-hmm-threads 24 > 11111-A_HaplotypeCaller.log 2>&1;date

6#Decompress vcf.gz
gunzip 11111-A.chr17.vcf.gz

Jun 24, 2019

Curriculum Vita

Shuen-Lin Jeng, Ph.D.
Associate Professor
Department of Statistics
National Cheng Kung University
Tainan, Taiwan

Office Phone: 06-2757575-53640
Home Phone: 06-3507543
Mobile Phone: 0952439106
Email: sljeng@mail.ncku.edu.tw
www.stat.ncku.edu.tw/members/teacher.asp

I. DEGREES HELD:

Ph.D. Iowa State University, 1998, Statistics
M.S. Fu-Jen University, 1989, Applied Mathematics
B.S. National Cheng Kung University, 1987, Mathematics

II. EXPERIENCE:

- Associate Professor
  Department of Statistics, National Cheng Kung University, 02/2007-Now
- Administrative Group Leader
  Center for Micro/Nano Science and Technology(CMNST), National Cheng Kung University, 02/2015-07/2016.
- Visiting scholar, School of Mathematics, Beijing Institute of Technology, 03/07/2012-18/07/2012. Visited Professor Yubin Tian.
- Visiting scholar, Beijing Institute of Genomics, Chinese Academy of Science, 08/10/2011-04/11/2011. Visited Professor Chung-I Wu.
- Visiting scholar, Vanderbilt Center for Quantitative Sciences, Vanderbilt University, 27/03/2011-07/04/2011. Visited Professor Yu Shyr.
- Visiting scholar, industrial and systems engineering, Georgia Institute of Technology, 01/2005-09/2005. Visited Professor C. F. Jeff Wu.
- **Special interests in reliability data analysis, data mining, statistical computing and bioinformatics.**

III. AWARDS AND HONORS:

- Distinguished Student Counseling Award, NCKU, 2017.
- Distinguished Teaching Award, NCKU, 2014.
- Excellent Service Award, Department of Statistics, NCKU, 2012.
- Excellent Teaching Award, NCKU, 2011.
- Excellent Research Award, College of Management, NCKU, 2011.
- Elected Member of the International Statistical Institute, 2004.

Exome sequencing data analysis: structural variant identification and analysis

cn.MOPS:

mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate

Klambauer, et al.
Nucleic Acids Research, 2012, 1–14

INTRODUCTION

- Copy number analysis by NGS
  1. Estimation of integer copy numbers from NGS data is more accurate for large copy numbers
  2. Breakpoints of copy number regions can be determined more precisely
  3. Allele-specific copy numbers may be estimated for observed alleles

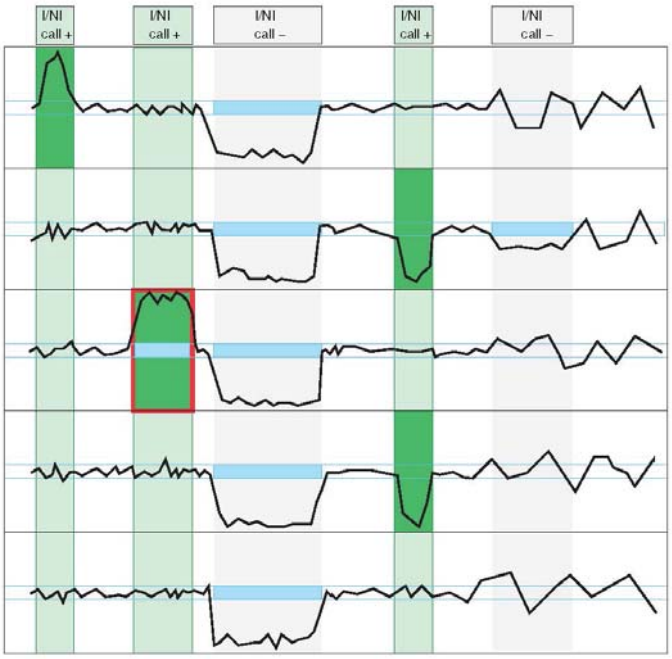
INTRODUCTION

- Existing methods for estimating DNA copy numbers in NGS data
  1. Approaches detecting read count deviations
  2. Reference-based approaches
- High FDR
  1. Read count variations along the chromosome, especially if no references are used
  2. Variations in read counts (noisy counts) across samples that may occur even for constant copy numbers

INTRODUCTION: cn.MOPS

- Contrast to most previous methods
  1. Provides integer copy numbers
  2. Estimates variations in read counts across samples
  3. Uses these estimates for CNV calling, thereby keeping the FDR low
- The novelty of cn.MOPS
  1. Modeling across samples

CNV among 5 samples



METHODS

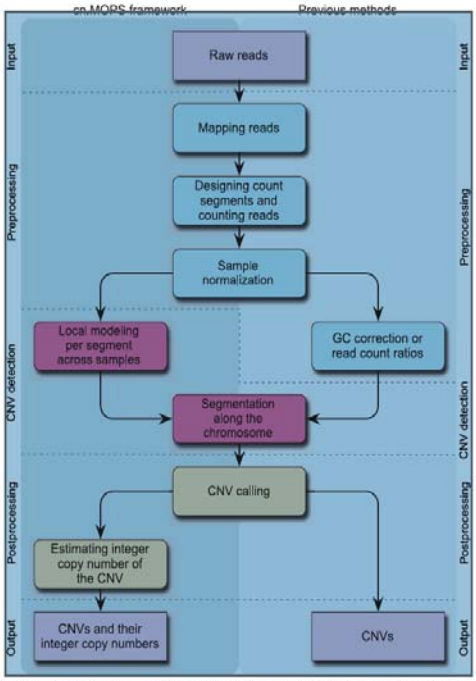
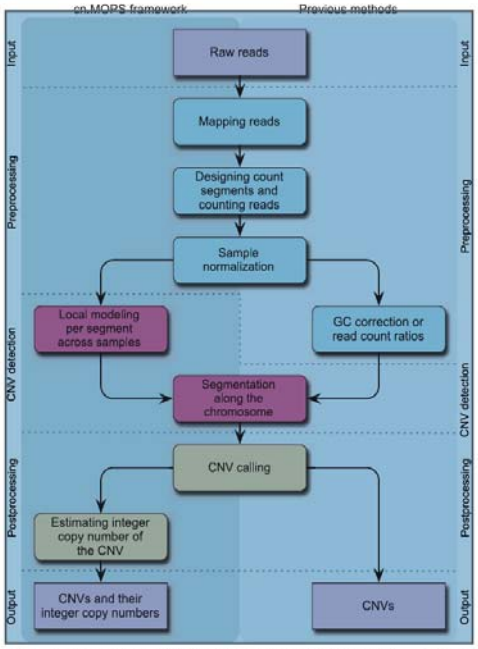


Figure 1. The processing pipelines for CNV detection in NGS data. Left column: modeling across samples and integer copy number estimation are unique to cn.MOPS. Right column: either GC correction [class (a) methods] or read count ratios [class (b) methods] are required for previous pipelines.

METHODS

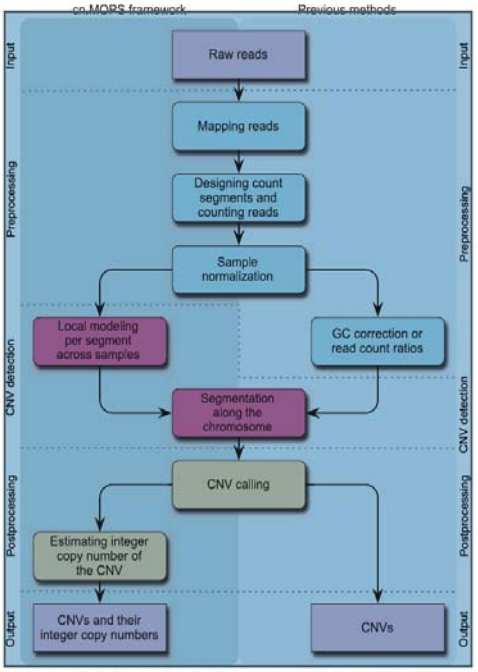


gure 1. The processing pipelines for CNV detection in NGS data. Left column: modeling across samples and integer copy number estimation are ique to cn.MOPS. Right column: either GC correction [class (a) methods] or read count ratios [class (b) methods] are required for previous elines.

Read mapping and segment read counts

- The parameters of the mapping method should be adjusted to minimize the number of false positives without generating too many false negatives.
- Expected sequencing errors
- They mapped the reads by Bowtie (25) for paired reads, allowed two mismatches and mapped to one random best mapping position

METHODS

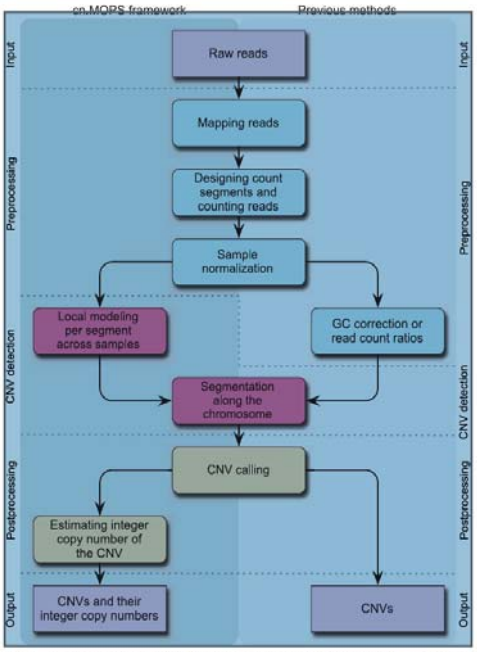


gure 1. The processing pipelines for CNV detection in NGS data. Left column: modeling across samples and integer copy number estimation are ique to cn.MOPS. Right column: either GC correction [class (a) methods] or read count ratios [class (b) methods] are required for previous elines.

Read mapping and segment read counts

- segments must be defined in which the reads are counted
- cn.MOPS also uses equally sized segments by default, equal size is not strictly required
- segmentation along the chromosome is based on the expected copy number, which is independent of the segment length

METHODS

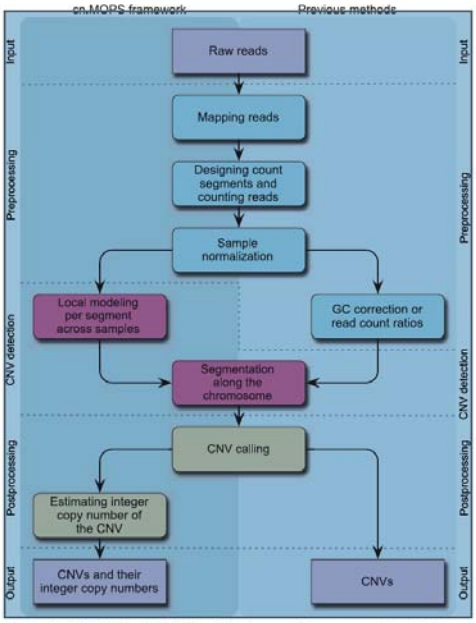


ure 1. The processing pipelines for CNV detection in NGS data. Left column: modeling across samples and integer copy number estimation are ique to cn.MOPS. Right column: either GC correction [class (a) methods] or read count ratios [class (b) methods] are required for segmen elines.

Sample normalization and GC correction

- apply a segmentation algorithm to (log) read counts along the chromosome
- (log) read counts must be normalized to be comparable between different genomic loci
- numbers of reads within segments depend on their GC content
- corrects the read counts of one sample by the number of mappable reads of the sample

METHODS



ure 1. The processing pipelines for CNV detection in NGS data. Left column: modeling across samples and integer copy number estimation are ique to cn.MOPS. Right column: either GC correction [class (a) methods] or read count ratios [class (b) methods] are required for previous elines.

Sample normalization and GC correction

data of HapMap individuals from the 1000 Genomes Project. Tested read counts of 25 kbp segments for being Poisson distributed with and without sample normalization.

Without sample normalization => **reject for 92%**
Using normalization => **rejected for only 2%**

The mixture of Poissons model

- Not affected by read count variations along the chromosome
- The model incorporates the linear dependency between average read counts in segments and copy numbers
- Dirichlet prior on the mixture components prefers a constant copy number of 2 for all samples.
- if the data drive the posterior away from this prior, the segment receives a high informative/non-informative call (I/NI call)

(I/NI call) =>detects variation across samples

The mixture of Poissons model

- cn.MOPS is a generative probabilistic model that explains the observed read counts by copy numbers and by measurement variations.
- Assumes that the read counts x in a segment are distributed across samples according to a mixture of Poissons, in which each mixture component corresponds to specific copy number and the Poisson parameter reflects the noise:

$$p(x) = \sum_{i=0}^{n} \alpha_i \, \mathrm{P}(x; \tfrac{i}{2}\lambda) \; . \qquad (1)$$

The mixture of Poissons model

$$p(x) = \sum_{i=0}^{n} \alpha_i \, \mathrm{P}(x; \tfrac{i}{2}\lambda) \; . \qquad (1)$$

- α_i is the percentage of samples with copy number i for $0 \leq i \leq n$
- λ is the mean read count for copy number 2
- P is the Poisson distribution

$$\mathrm{P}(x; \beta) = \frac{1}{x!} \, e^{-\beta} \, \beta^{x} \; . \qquad (2)$$

The mixture of Poissons model

- For copy number $i \geq 1$, the mean read count is thus $\beta = \frac{i}{2}\lambda$
- For copy number I = 0, we assume a Poisson distribution with parameter $\beta = \frac{\varepsilon}{2}\lambda$
- The robustness is due to the fact that copy number zero can be detected with a broad range of e values.

The mixture of Poissons model

- Prior probability that a read count stems from copy number i

  => $p(i) = \alpha_i$
- The likelihood that a read count x is produced by the i-th mixture component is

  => $P(x|i)=P(x; \frac{i}{2}\lambda)$
- Bayes' formula can be used to compute the posterior

  => $p(i \mid x)$

  =>the probability that read count x is caused by the i-th component corresponding to copy number i

The mixture of Poissons model

- Model Selection by an EM Algorithm and Dirichlet Prior

➢Suppose that read counts $\{x_1, \ldots, x_n\}$ have been observed for N samples in a given segment.

➢fitting a model that best explains the training data $\{x_1, \ldots, x_n\}$

The mixture of Poissons model

- Bayesian framework

=> α and λ are considered as random variables

=> p(x) in Equation (1) becomes a conditional probability p(x | α, λ)

=> we assume that, for the prior distribution, the parameters α and λ are independent [p(α,λ) = p(α)p(λ)].

=> the parameter posterior is

$$p(\boldsymbol{\alpha}, \lambda \mid x) = \frac{p(x \mid \boldsymbol{\alpha}, \lambda)\, p(\boldsymbol{\alpha})\, p(\lambda)}{\int p(x \mid \boldsymbol{\alpha}, \lambda)\, p(\boldsymbol{\alpha})\, p(\lambda)\, d\boldsymbol{\alpha}\, d\lambda} . \qquad (3)$$

The mixture of Poissons model

- Dirichlet prior p(α) on $\alpha = \{\alpha_0, \alpha_1, \ldots, \alpha_n\}$
- include the prior knowledge that almost all locations have copy number 2 for all samples, which is the null hypothesis of constant copy number 2

$$p(\alpha) = \frac{1}{\mathrm{B}(\gamma)} \prod_{i=0}^{n} \alpha_i^{\gamma_i - 1} \qquad (4)$$

$\gamma = \{\gamma_0, \gamma_1, \ldots, \gamma_n\}$

The mixture of Poissons model

- Setting $\gamma_2 >> \gamma_i \geq 1$, for $(i \neq 2)$, we ensure that vectors α with a large value for α_2 are the most likely to be drawn.

  (α_2 is the percentage of samples having copy number 2)
- Each component i of the Dirichlet distribution p(α) is distributed according to a beta distribution with the mode $(\gamma_i - 1)(\gamma_s - n)$

$$\gamma_s = \sum_{i=0}^{n} \gamma_i .$$

The mixture of Poissons model

- The prior on λ, choose a uniform distribution on a sufficiently large interval with left endpoint 0
- EM algorithm minimizes an upper bound on the negative log-posterior of the parameters α and λ

$$\hat{\alpha}_{ik} = \frac{\alpha_i^{\text{old}} \, \mathrm{P}(x_k; \frac{i}{2}\lambda^{\text{old}})}{p(x_k \mid \boldsymbol{\alpha}^{\text{old}}, \lambda^{\text{old}})} , \qquad (5)$$

$$\alpha_i^{\text{new}} = \frac{\frac{1}{N}\sum_{k=1}^{N} \hat{\alpha}_{ik} + \frac{1}{N}(\gamma_i - 1)}{1 + \frac{1}{N}(\gamma_s - n)} , \qquad (6)$$

$$\lambda^{\text{new}} = \frac{\frac{1}{N}\sum_{k=1}^{N} x_k}{\sum_{i=0}^{n} \left(\frac{1}{N} \frac{i}{2} \sum_{k=1}^{N} \hat{\alpha}_{ik} \right)} . \qquad (7)$$

The mixture of Poissons model

- $\hat{\alpha}_{ik}$ is an estimate (the E-step of the EM algorithm) of the posterior $\alpha_{ik} = p(i \mid x_k, \alpha, \lambda)$ using current estimations α^{old} and λ^{old} of the parameters.
- Simplify the hyperparameter vector c to one intuitively interpretable hyperparameter G by setting $\gamma_i = 1$ for $i \neq 2$ and $\gamma_2 = 1+G$

=> ensures that $\alpha_2 > G/(G+N)$ in the α_i update Equation (6).

I/NI Call: Information Gain of Posterior over Prior

- Bayesian framework

=> define an informative/non-informative (I/NI) call analogous to that of the FARMS algorithm

- The I/NI call measures the information gain of the posterior compared to its prior distribution $p(\alpha)$, which represents the null hypothesis that all samples have copy number 2
- I/NI call measures the tendency to reject the null hypothesis based on the observed data

I/NI Call: Information Gain of Posterior over Prior

- Define the I/NI call as a weighted distance between the posterior's mode and the prior's mode (0, 0, 1, 0, . . . ,0)

$$\mathrm{I/NI}(\alpha) = \sum_{i=0}^{n} \alpha_i \, |\log(i/2)| = \sum_{i=0}^{n} \frac{1}{N} \sum_{k=1}^{N} \alpha_{ik} \, |\log(i/2)|, \quad (8)$$

$$\alpha_i = \frac{1}{N} \sum_{k=1}^{N} \alpha_{ik}$$

- 'log(0/2)' must be understood as log(ε/2)
- The more copy numbers differ from 2, the higher is the I/NI call

I/NI Call: Information Gain of Posterior over Prior

- The I/NI call can be understood as the sum of individual I/NI calls, I/NI(α_k) , the contribution of the k-th sample to the I/NI call:

$$\mathrm{I/NI}(\boldsymbol{\alpha}) = \frac{1}{N} \sum_{k=1}^{N} \sum_{i=0}^{n} \alpha_{ik} |\log(i/2)| = \frac{1}{N} \sum_{k=1}^{N} \mathrm{I/NI}(\boldsymbol{\alpha}_k), \quad (9)$$

$\alpha_k = \{\alpha_{0k}, \alpha_{1k}, \alpha_{2k}, \ldots, \alpha_{nk}\}$ the vector of posteriors for the read count x_k

Segmentation and CNV call

- cn.MOPS detects CNVs by segmenting the chromosomes of individuals based on their individual I/NI calls, joining genomically adjacent I/NI calls that show the same copy numbers.
- The I/NI call defined in Equation (9) does not distinguish between losses and gains with the same fold change.
- They define the signed individual I/NI call as the expected log fold change:

$$\mathrm{sI/NI}(\boldsymbol{\alpha}_k) \;=\; \sum_{i=0}^{n} \alpha_{ik} \;\log(i/2)\;. \qquad (\mathbf{10})$$

Segmentation and CNV call

- The absolute value of the signed I/NI call $|\mathrm{sI/NI}(\alpha_k)|$ is not exactly the I/NI call $\mathrm{I/NI}(\alpha_k)$, but the two values are always very close
- cn.MOPS applies either its own algorithm 'fastseg' or, "the circular binary segmentation algorithm [DNAcopy (32)]" to $\mathrm{sI/NI}(\alpha_k)$ along the chromosome.
- The segmentation algorithm joins consecutive segments with large or small expected fold changes to make a candidate segment.
- Supplies candidate segments that show variations along the chromosome and also across samples indicated by the signed individual I/NI calls

Segmentation and CNV call

- sI/NI(α_k) over the segment is at least 0.6 ≈ $\log_2(3/2)$ for gains or at most -1 = $\log_2(1/2)$

  => candidate segment is called a CNV segment

- This CNV call incorporates two calls:
  1. an I/NI call across samples
  2. a segment call along the chromosome

  => if consecutive segments obtain an I/NI call, they are joined by the segmentation algorithm

Integer copy number estimation (final step)

- Concerned with estimating the integer copy numbers of the CNVs

=>cn.MOPS automatically supplies posterior estimates of the integer copy numbers for each segment

Exome sequencing data analysis: structural variant identification and analysis

cn.MOPS:

mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate

Workshop Notes

上課老師：鄭順林 副教授
助教：廖鈺茹（碩二）
盧配天（研究助理）
國立成功大學統計系

程式環境與執行指令

- 程式環境 (140.116.56.169 server)
  1. R-3.4.1 (以上的版本) https://www.r-project.org/
  2. cn.mops-1.28
     #source("http://www.bioconductor.org/biocLite.R")
     #biocLite("cn.mops")

- 執行指令
  1. 使用pietty 連結到 140.116.56.169 linux server
     執行 cnmops_process_workflow.txt 的指令。
     Round 1: 畫圖和分析。
     Round 2: 修改圖形範圍。
  2. 使用filezilla 連結到140.116.56.169 下載執行結果

cnmops_process_workflow.txt

#Round1: read raw bam files, draw data plots and analize data
Rscript /Pathxxxxx/pop_analysis_Rscript.r /Pathxxxxx/pop_analysis_para_input_example.txt

Rscript /Pathxxxxx/new_vs_pop_analysis_Rscript.r /Pathxxxxx/new_vs_pop_analysis_para_input_example.txt

#Round2: use .Rdata from Rounnd1 and draw the modified data plots
Rscript /Pathxxxxx/pop_analysis_Rscript.r /Pathxxxxx/pop_analysis_para_round2_input_example.txt

Rscript /Pathxxxx/new_vs_pop_analysis_Rscript.r /Pathxxxx/new_vs_pop_analysis_para_round2_input_example.txt

Round 1:
Population Analysis (Reference samples)
pop_analysis_para_input_example.txt

Round	1
RData.base.path	/RdataPathxxxxx/CGM_cnMOPS_Analysis.Rdata
bedfile.path	/bioinformatics/sljeng/hg19_100k.bed
reference.bam.path.v	/bioinformatics/sljeng/A1.bam,/bioinformatics/sljeng/A2.bam, ….
reference.output.names.v	A1.bam,A2.bam,A3.bam,A4.bam,Q1.bam,Q2.bam
ploidy.v	2,1,1,2,2,1
parallel	6
reference.output.name.first.part.v	Pop6
save.path	/SavePathxxxxx/cnMOPS/

Round 1:
Population Analysis (Reference samples)
pop_analysis_para_input_example.txt

scatter.onepage	TRUE
cor.threshold	0.7
cor.xylim.v	0,500
count.ylim.v	0,800
hist.xlim.v	0,800
graph.focus	TRUE
range.no.100K	7
output.RData.name	cnMOPS_Analysis_reference.Rdata

Round 1:
New Sample VS Population Analysis
new_vs_pop_analysis_para_input_example.txt

Round	1
RData.base.path	/SavePathxxxxx/cnMOPS/cnMOPS_Analysis_reference.Rdata
new.sample.bam.path	/bioinformatics/sljeng/Q3.bam,/bioinformatics/sljeng/Q4.bam
parallel	2
new.ploidy.v	1,2
new.sample.output.file.name.first.part.v	Pop6_vs_Q3_Male,Pop6_vs_Q4_Female
new.sample.output.names.v	Q3_Male,Q4_Female
save.path	/SavePathxxxxx/cnMOPS/

Round 1:
New Sample VS Population Analysis
new_vs_pop_analysis_para_input_example.txt

scatter.onepage	TRUE
cor.threshold	0.8
cor.xylim.v	0,400
count.ylim.v	0,800
hist.xlim.v	0,800
graph.focus	TRUE
range.no.100K	6
sign.ploidy	TRUE
output.RData.name	cnMOPS_Analysis_new.Rdata

Round 2: 觀察資料圖形然後進行修改圖形範圍
Population Analysis (Reference samples)
pop_analysis_para_round2_input_example.txt

Round	2
RData.base.path	/SavePathxxxxx/cnMOPS/cnMOPS_Analysis_reference.Rdata
scatter.onepage	TRUE
cor.threshold	0.8
cor.xylim.v	0,800
count.ylim.v	0,800
hist.xlim.v	0,800
output.RData.name	cnMOPS_Analysis_reference2.Rdata

Round 2:觀察資料圖形然後進行修改圖形範圍
New Sample VS Population Analysis
new_vs_pop_analysis_para_round2_input_example.txt

Round	2
RData.base.path	/SavePathxxxxx/cnMOPS/cnMOPS_Analysis_new.Rdata
scatter.onepage	TRUE
cor.threshold	0.8
cor.xylim.v	0,800
count.ylim.v	0,800
hist.xlim.v	0,800
graph.focus	TRUE
output.RData.name	cnMOPS_Analysis_new.Rdata

‘cn.mops’ process workflow description

June 25, 2019

Description:

This is a readme to introduce the ‘cn.mops’ process workflow, which includes parameters and the results of “pop_analysis_Rscript.r” and “new_vs_pop_analysis_Rscript.r”.

Package: cn.mops

Package version: 1.28.0

Install package method:

R version ≥3.5

```
if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager")
BiocManager::install("cn.mops", version = "3.8")
library(cn.mops)
```

2.12 ≤ R version <3.5

```
source("http://www.bioconductor.org/biocLite.R")
biocLite("cn.mops")
library(cn.mops)
```

1. Version of this process

- The following process were tested on the machine below at 2019/06/24.
- Linux version 2.6.18-348.1.1.el5 (mockbuild@builder10.centos.org) (gcc version 4.1.2 20080704 (Red Hat 4.1.2-54)) #1 SMP Tue Jan 22 16:19:19 EST 2013
- R version 3.4.1
- cn.MOPS version 1.28.0

2. Setting the parameters in “pop_analysis_para_input_example.txt”, “new_vs_pop_analysis_para_input_example.txt”, “pop_analysis_para_round2_input_example.txt”, “new_vs_pop_analysis_para_round2_input_example.txt”

Suppose that all the paths are “/bioinformatics/sljeng”, and referred to as PATH

Parameters	Example	Explanation
Round	1	1: running the process for the first time, 2: others
RData.base.path	PATH/Analysis.Rdata	RData base path and name
bedfile.path	PATH/hg19_100k.bed	Bed file path and bed file name
reference.bam.path.v	PATH/A1.bam,PATH/A2.bam	All reference bam file paths and bam file names you want to input
reference.output.names.v	A1.bam,A2.bam	All reference bam file names you want to output
ploidy.v	1,2	Sex: M:1, F:2
parallel	4	How many parallel processing you want to run
reference.output.name.first.part.v	Pop6	The first part of output file names
save.path	PATH/cnMOPS/	The path for output files
scatter.onepage	TRUE	Whether to draw all scatter plots in one page
cor.threshold	0.7	Give the graph a red title warning if the correlation is smaller than the cor.threshold
count.ylim.v	0,800	Set the y axis range for the position-count plots
cor.xylim.v	0,500	Set the x and y axis range for the correlation plots
hist.xlim.v	0,800	Set the x axis range for the histogram plots
graph.focus	TRUE	Whether to draw the new sample in different color
range.no.100K	7	Set range.no.100K as number of the ± range by 100K around cnvs/cnvr
output.RData.name	Analysis_reference.Rdata	The output RData base path and name
new.sample.bam.path.v	PATH/Q3.bam,PATH/Q4.bam	All new samples bam file paths and bam file names you want to input
new.ploidy.v	1.2	Sex: M:1, F:2
new.sample.output.file.name.first.part.v	Pop6_vs_Q3,Pop6_vs_Q4	The first part of output file names
new.sample.output.names.v	Q3.bam,Q4.bam	All new samples bam file names you want to output
sign.ploidy	TRUE	Whether to draw the same ploidy in the same color

3. Running the linux command in “process_analysis_comments_Rscript_example.txt”

Two rounds

- Round1: Read raw bam files, draw data plots and analyze data
- Round2: Observe the data graph and then modify the graphic range, use the .Rdata from Round1 and draw data plots again

For Example (output folders):

- Population: folder “Population-Samples-Analysis”, which includes six pdf files, four jpeg files, two csv files, and after running round2 you will get one more folder.

  six pdf files:

Example	Explanation
Pop4-Position-Count-Original.pdf	Position vs Coverage−depth (original)
Pop4-Position-Count-Normalized.pdf	Position vs Coverage−depth (normalized)
Pop4-Position-Count-Smoothed-Original.pdf	Position vs Coverage−depth (smoothed original)
Pop4-Position-Count-Smoothed-Normalized.pdf	Position vs Coverage−depth (smoothed normalized)
Pop4-cnvs_700K_Significant-CopyNumber-plot_normalized.pdf	Significant Copy Number Plot_cnvs
Pop4-cnvr_700K_Significant-CopyNumber-plot_normalized.pdf	700 K_Significant Copy Number Plot_cnvr_Normalized

  four jpeg files:

Example	Explanation
Pop4-focus-1p-Correlation-Original.jpeg	Correlation of Coverage-depth (Original)
Pop4-focus-1p-Correlation-Normalized.jpeg	Correlation of Coverage-depth (Min1 & Normalized)
Pop4-focus-Histogram-Original.jpeg	Histograms of Coverage-depth (Original)
Pop4-focus-Histogram-Normalized.jpeg	Histograms of Coverage-depth (Min1 & Normalized)

  two csv files:

Example	Explanation
Pop4-cnvs_min1_cnmops.csv	The locations of cnvs
Pop4-reorder-cnvr_min1_cnmops_order.csv	The locations of cnvr

  one folder:

  Round2

- New sample: folder “2019-06-24-Q1-Combined”, which includes five pdf files, four jpeg files, three csv files, and after running round2 you will get one more folder.

  five pdf files:

Example	Explanation
Pop4_vs_Q1-Position-Count-Original.pdf	Position vs Coverage−depth, Mean, Q1 added (original)
Pop4_vs_Q1-Position-Count-Smoothed-Original.pdf	Position vs Coverage−depth, Mean, Q1 added (smoothed original)
Pop4_vs_Q1-Position-Count-Normalized.pdf	Position vs Coverage−depth, Mean, Q1 added (normalized)
Pop4_vs_Q1-Position-Count-Smoothed-Original.pdf	Position vs Coverage−depth, Mean, Q1 added (smoothed normalized)
Pop4_vs_Q1_case_control_ploidy_cnvs_600K_Sign-CN-plot_norm.pdf	Significant Copy Number Plot_cnvs

  four jpeg files:

Example	Explanation
Pop4_vs_Q1-focus-1p-Correlation-Original.jpeg	Correlation of Coverage-depth, Mean, Q1 added (Original)
Pop4_vs_Q1-focus-1p-Correlation-Normalized.jpeg	Correlation of Coverage-depth, Mean, Q1 added (Min1 & Normailized)
Pop4_vs_Q1-focus-Histogram-Original.jpeg	Histograms of Coverage-depth, Mean, Q1 added (Original)
Pop4_vs_Q1-focus-Histogram-Normalized.jpeg	Histograms of Coverage-depth, Mean, Q1 added (Min1 & Normailized)

  three csv files:

Example	Explanation
Pop4_vs_Q1-original_count.csv	Counts of population and new sample
Pop4_vs_Q1-norm_count.csv	Counts of population and new sample after normalization
Pop4_vs_Q1-cnvs_min1_cnmops.csv	Do case-control to find the location of cnvs

  one folder:

  Round2