Exome sequencing data analysis

全外顯子組定序數據分析

Yen An Tang, PhD

CENTER FOR GENOMIC MEDICINE,
NATIONAL CHENG KUNG UNIVERSITY

2020/7/8

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

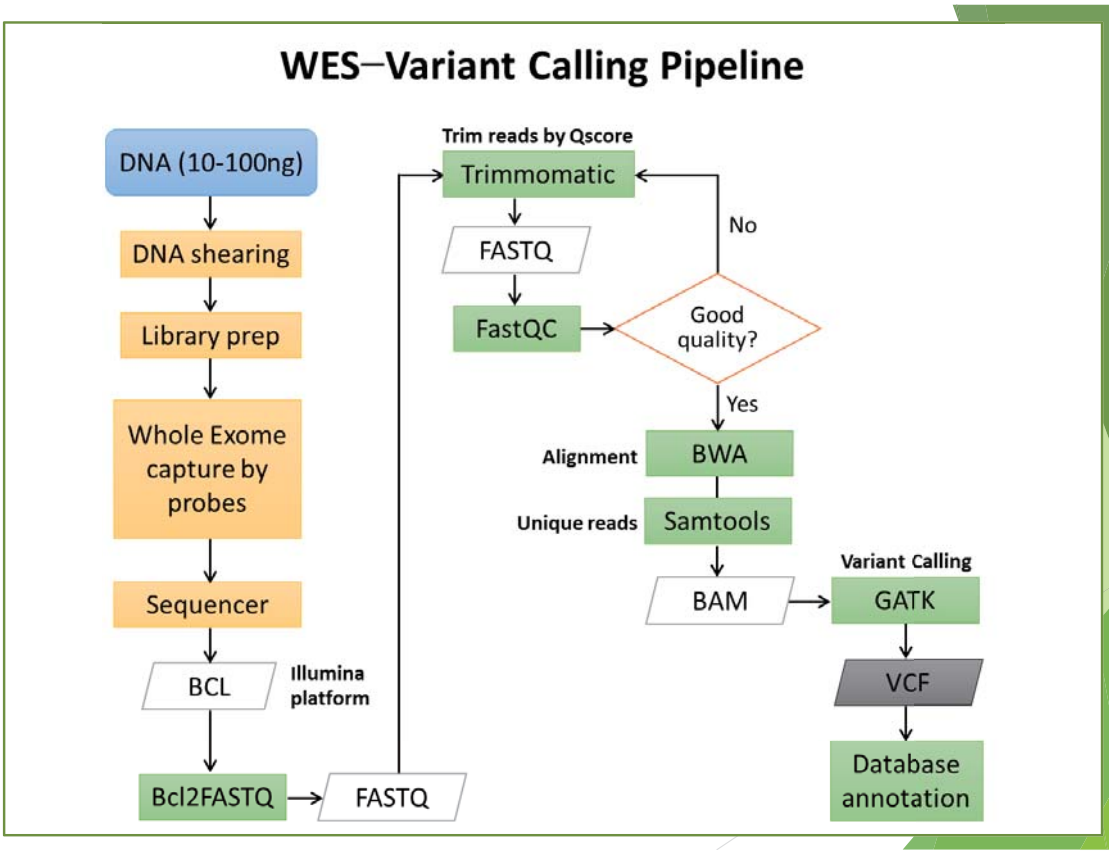
Trimming

QC

Alignment

Variant Calling

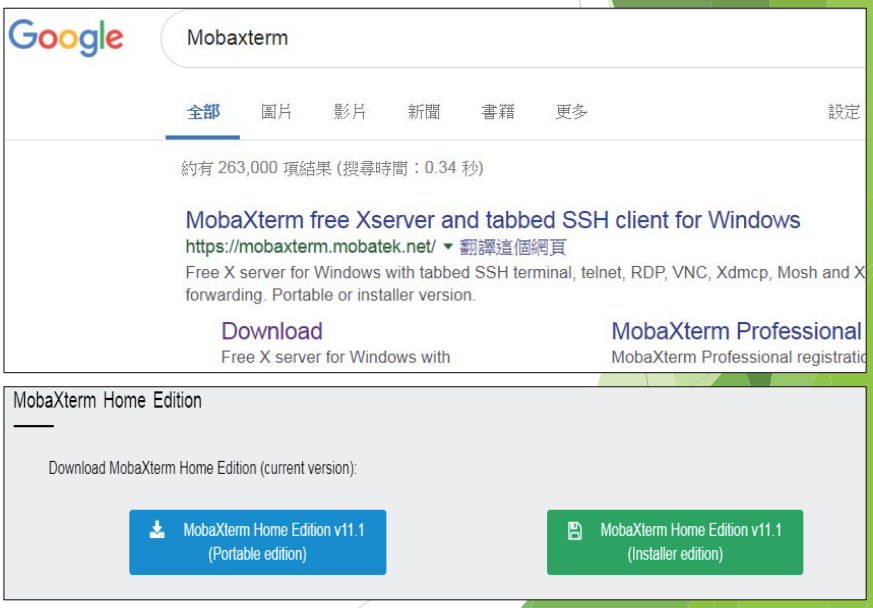
Visualization



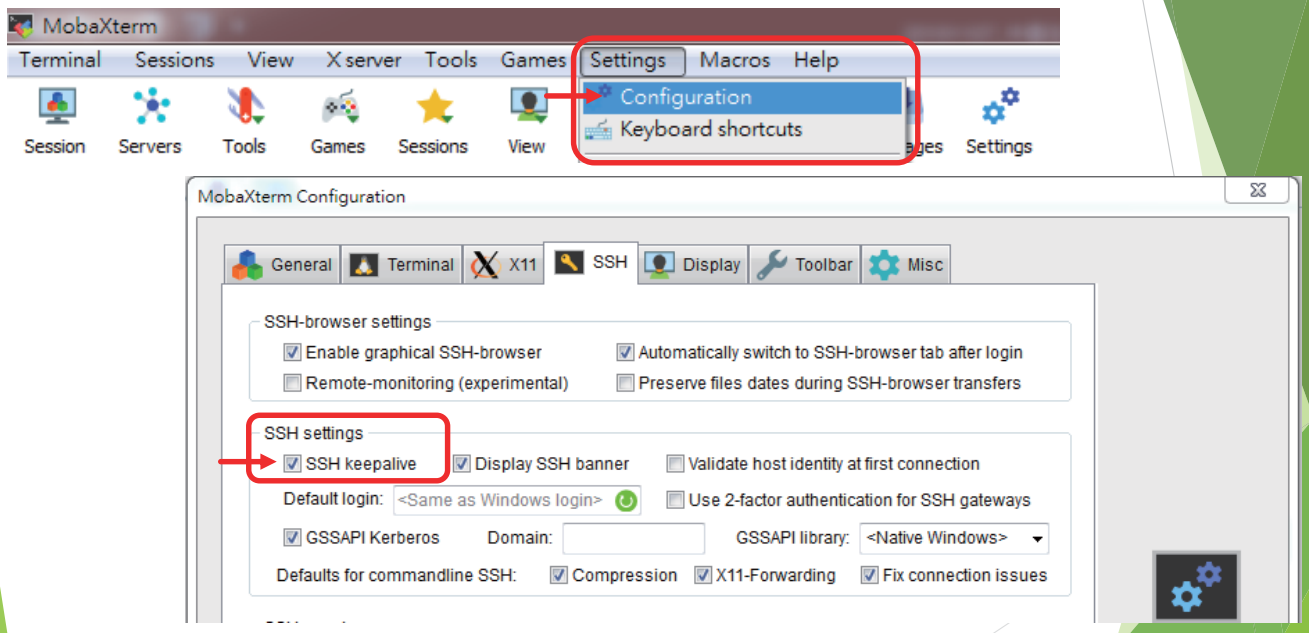
Linux Interface

- Windows: PuTTY/PieTTY
- Mac: Terminal

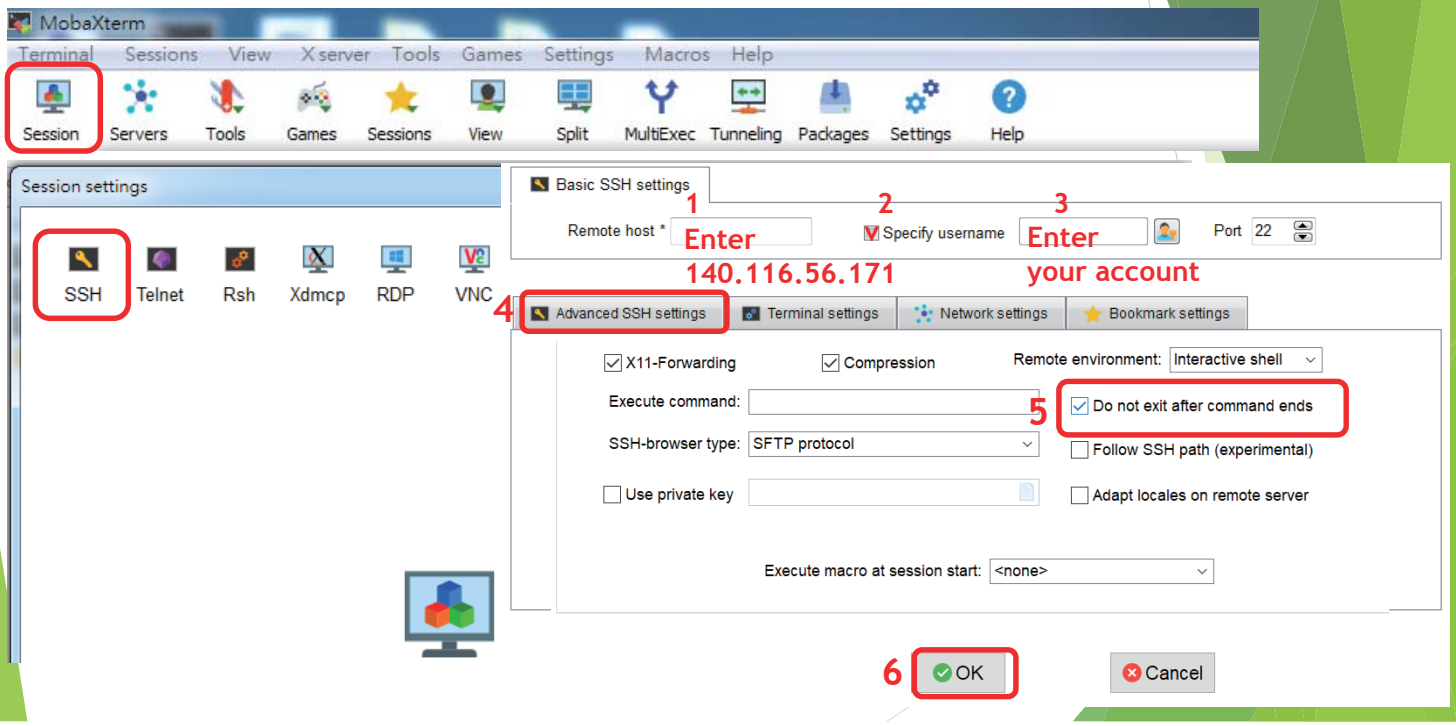
- Xwindow: MobaXterm
- VNC: RealVNC



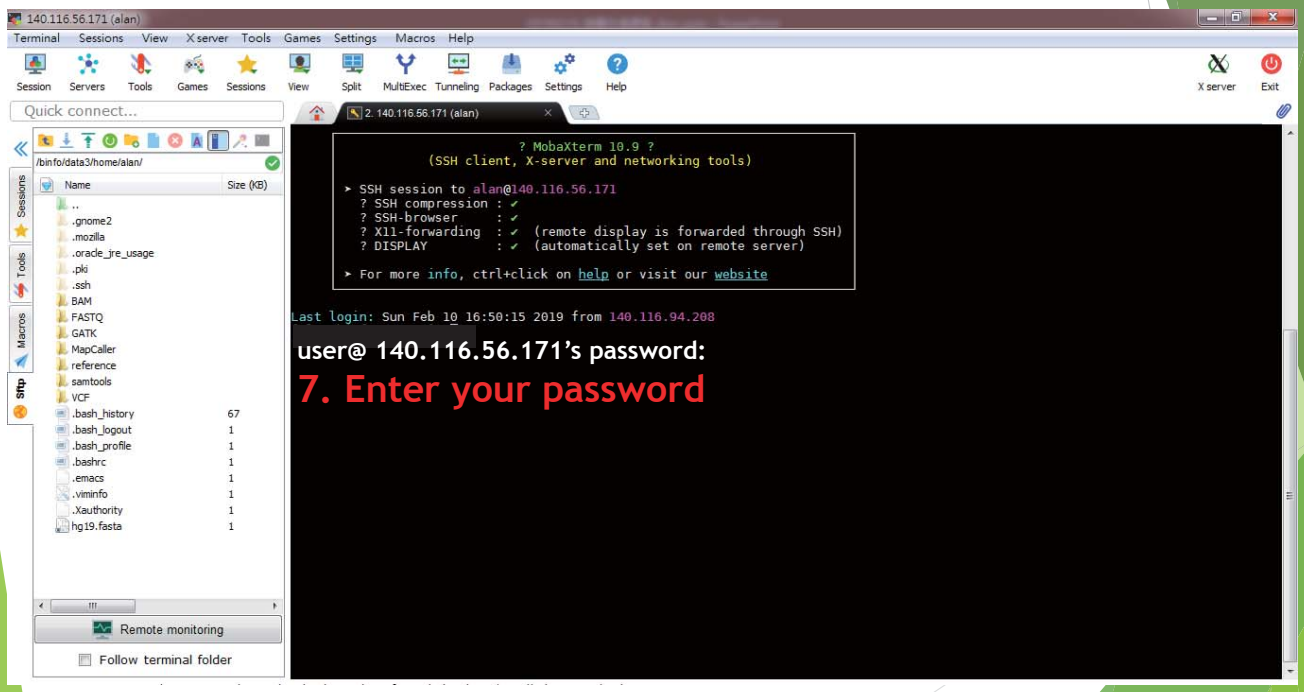
MobaXterm



MobaXterm



MobaXterm



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

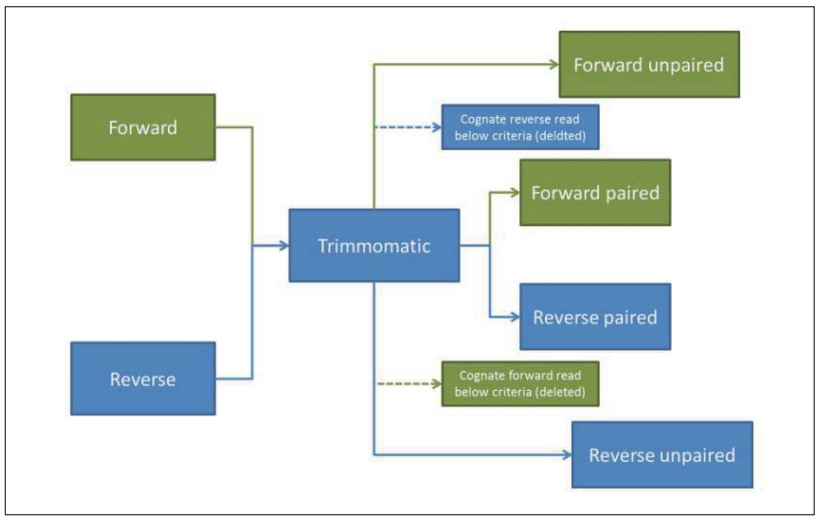
- Command:

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

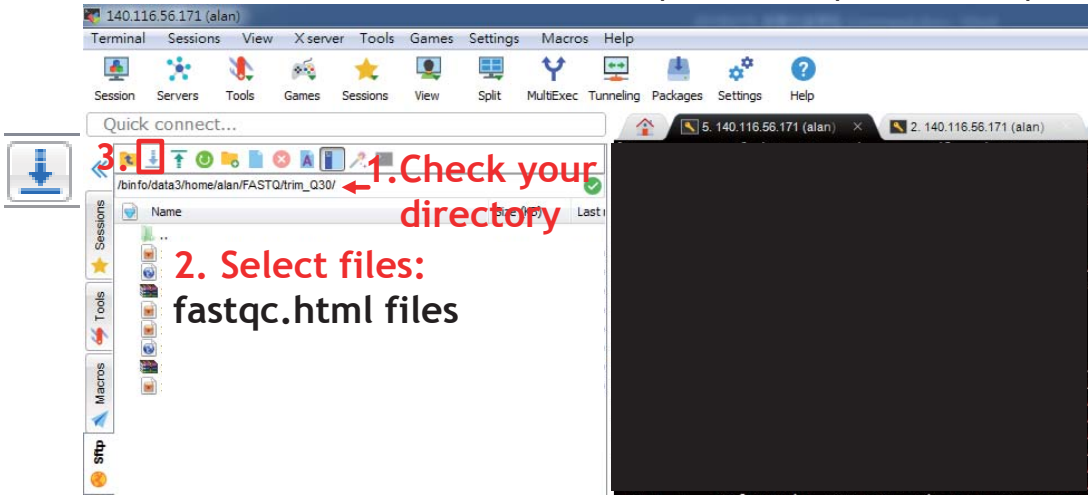


3 Quality Control: FastQC

- 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



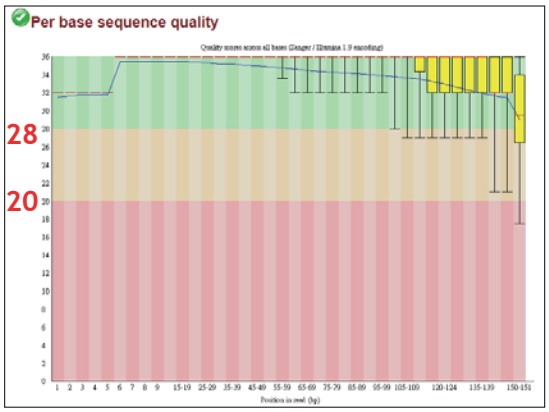
3 Quality Control: FastQC

- Command:

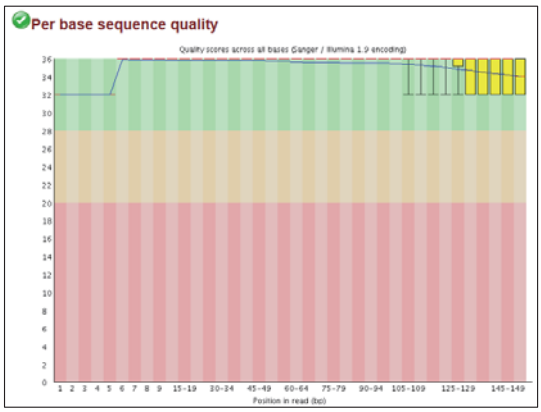
/tools/FastQC/fastqc <input 1> <input 2> <input 3> <input 4>

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

11111-A_R1.fastqc.html



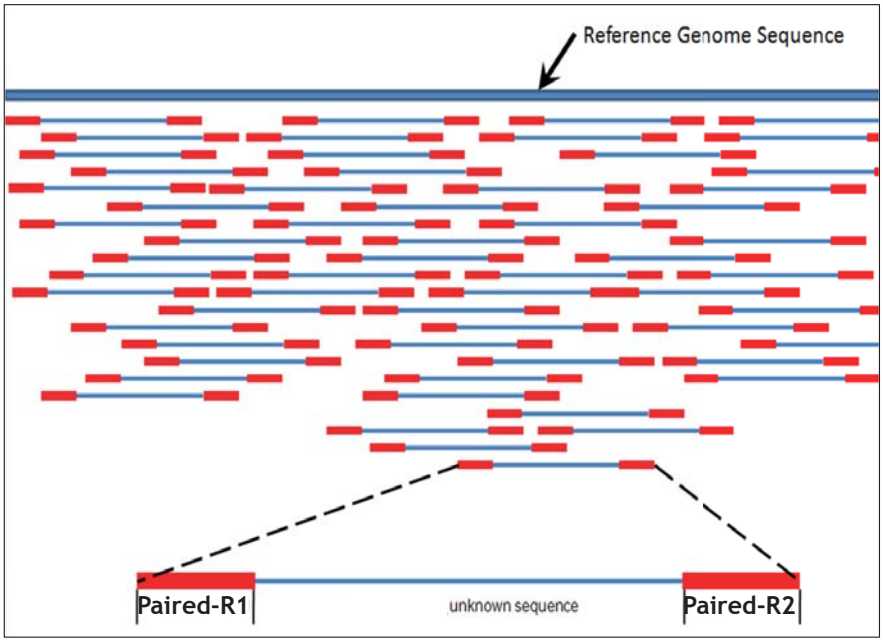
11111-A_R1_trim_q30.fastqc.html



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
EEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAAAA# NM:i:1  MD:Z:150C0      AS:i:150        XS:i:0  RG:Z:NextSeq550 2019
```

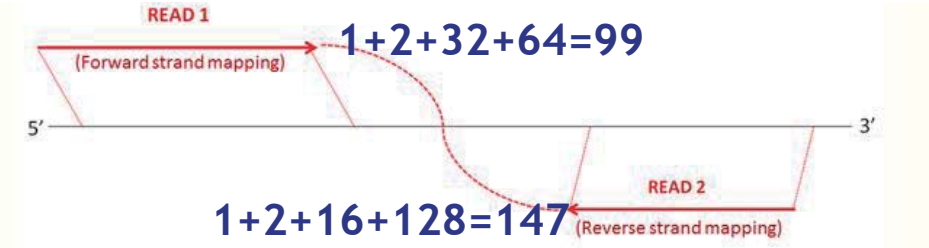
Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{31}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String	*|([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

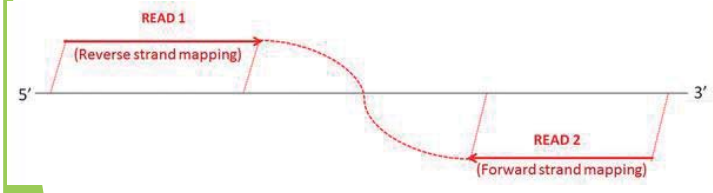


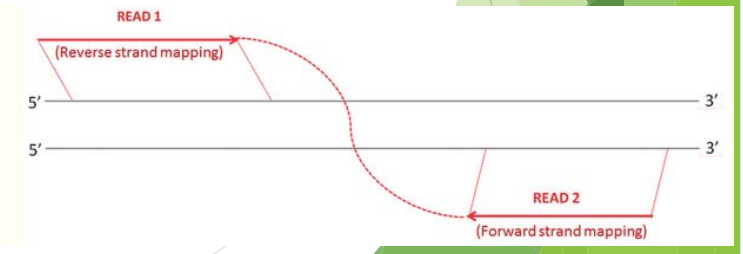
4 Alignment: BWA

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 or 30S90M

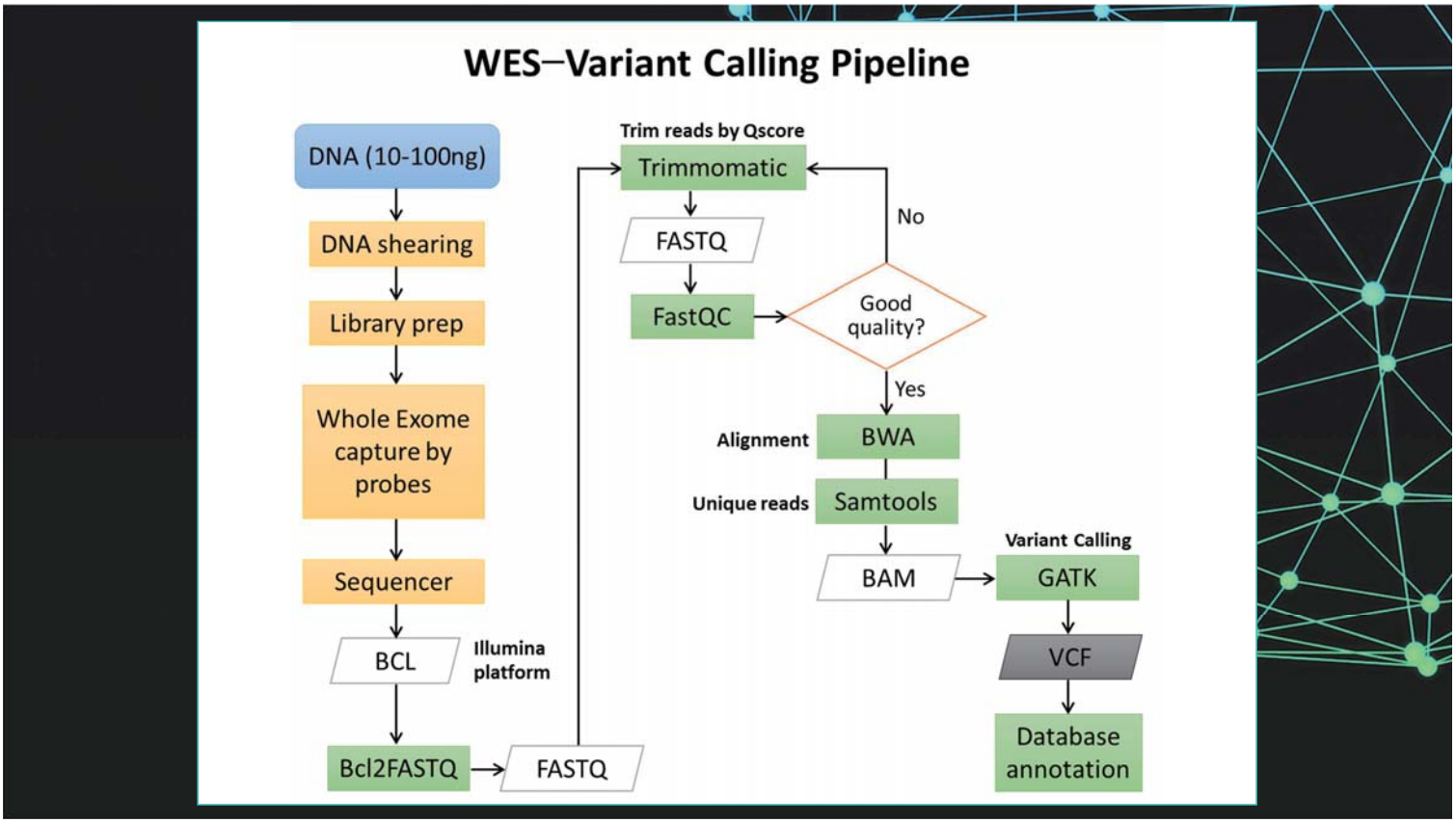
4 Alignment: Samtools

- 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

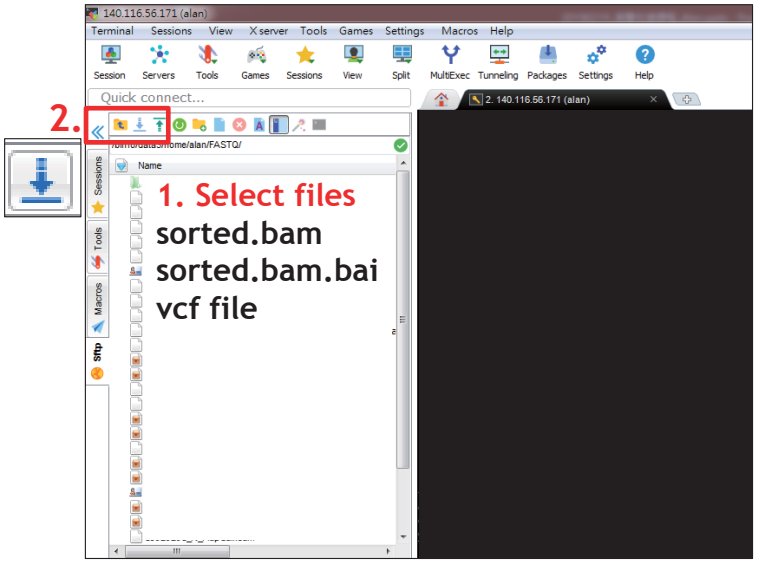
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	11111-A_R
chr17	6115	.	G	C	5081.03	.	AC=2;AF=1.00;AN=2;DP=184;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=27.61;SOR=0.932		GT:AD:
chr17	6157	.	A	G	3873.03	.	AC=2;AF=1.00;AN=2;DP=140;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=27.66;SOR=0.980		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>

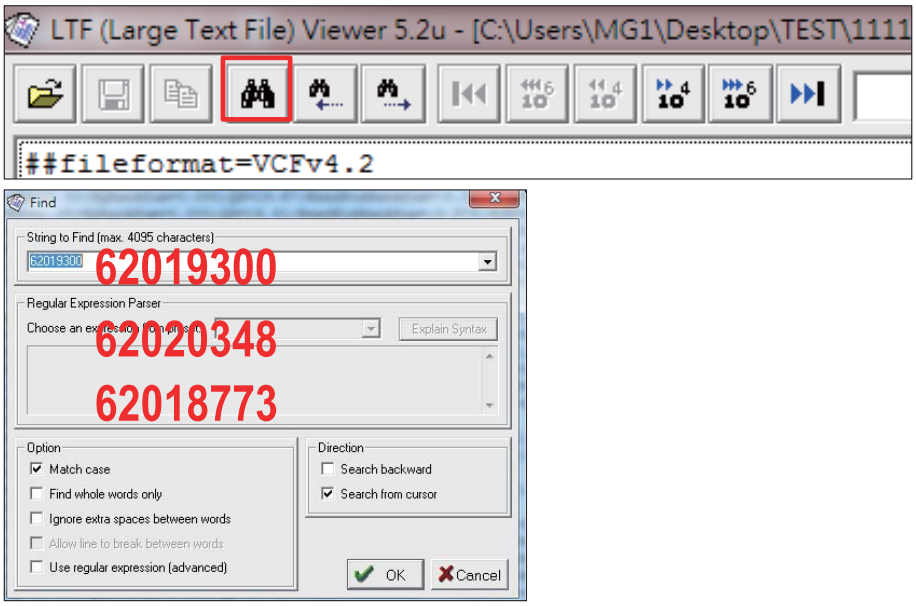
→ Use the SFTP (left window) to download the files



7 View of VCF file

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

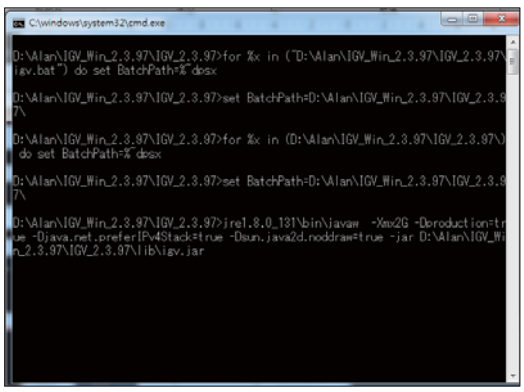


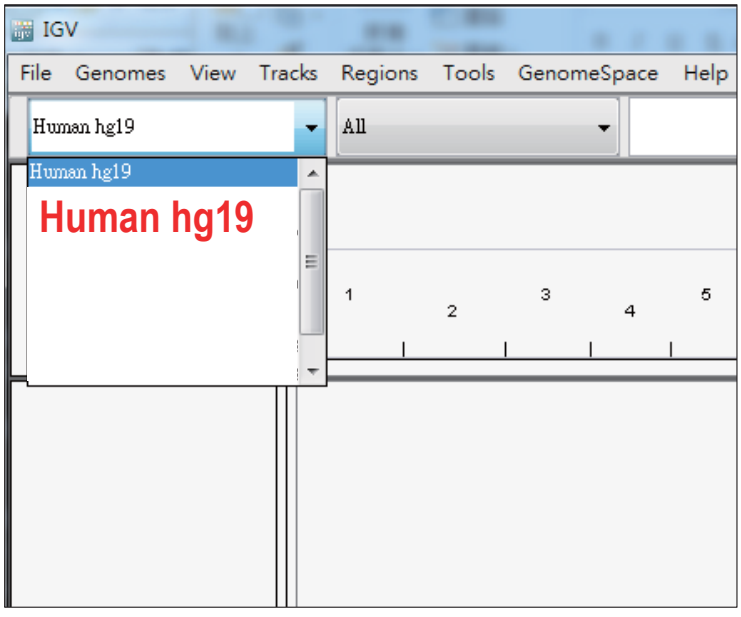


8 Visualization of reads: BAM file

OPEN Bam file using IGV software

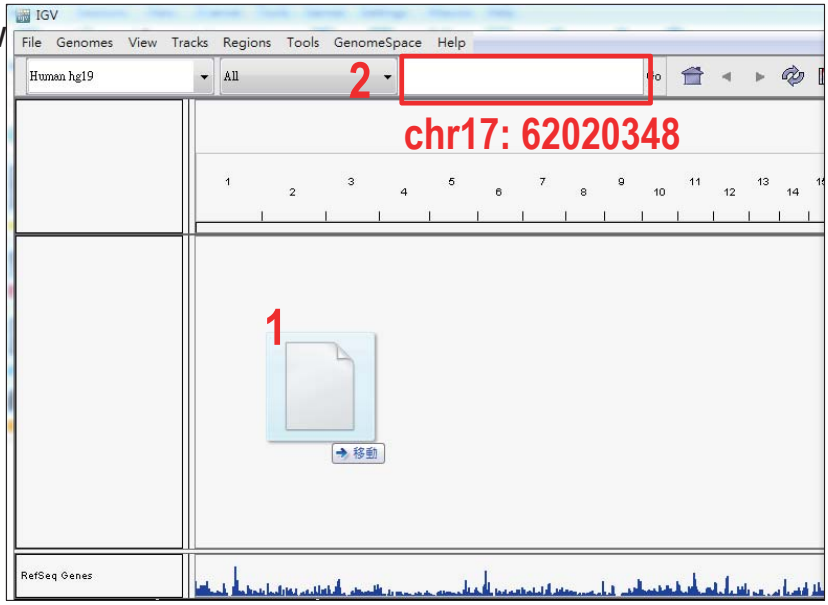






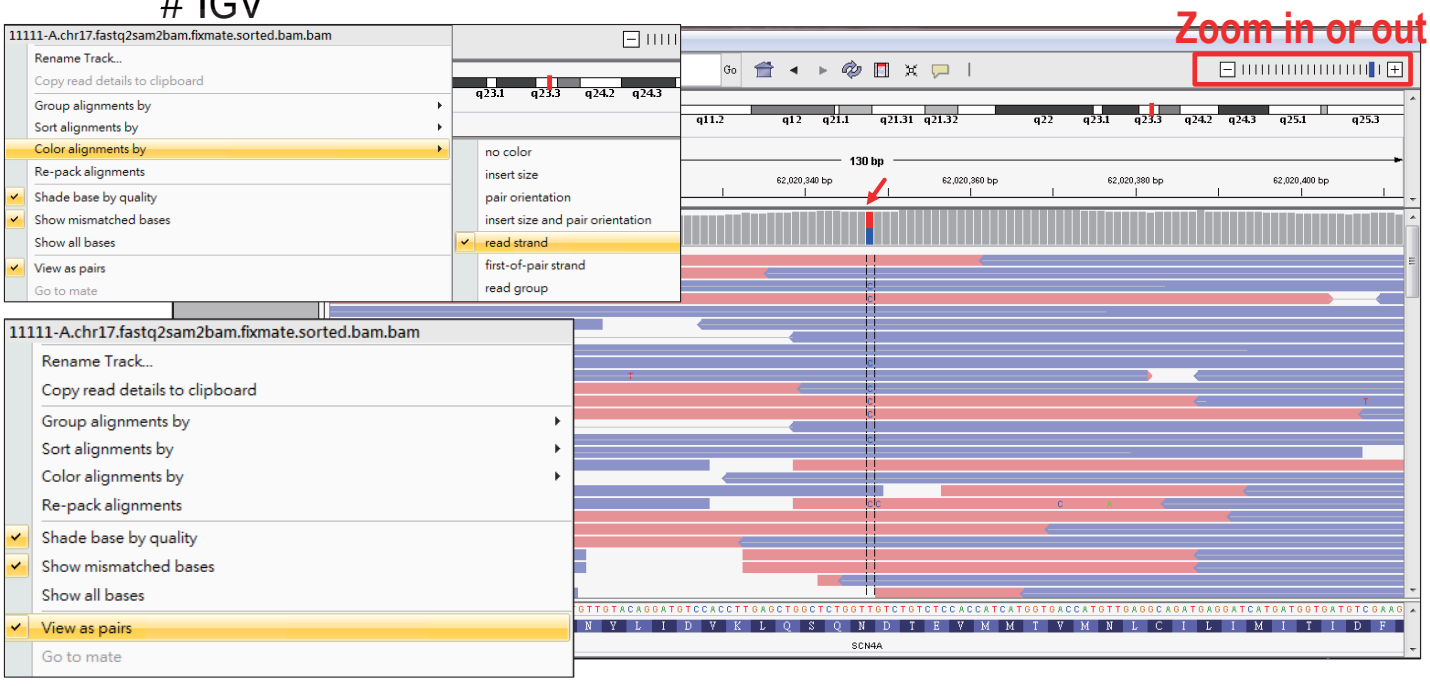
8 Visualization of reads: BAM file

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



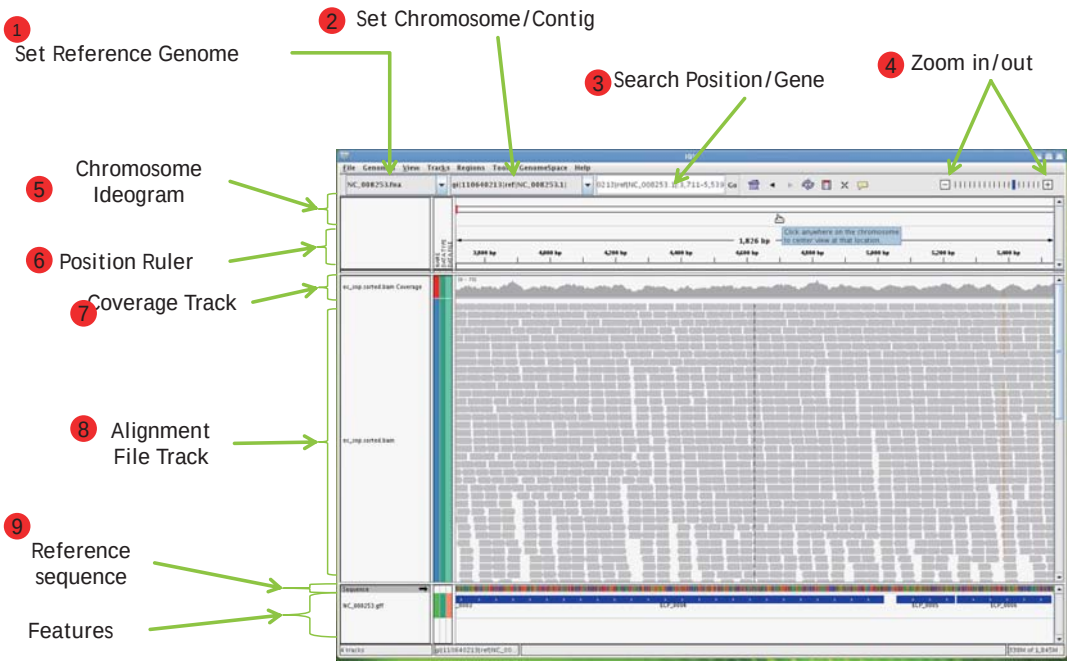
8 Visualization of reads: BAM file

IGV



8 IGV

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



THANKS

WES (Whole Exome Seq) - GATK (Genome Analysis Toolkit)

Commands

【Step1】Create directory

mkdir trim_Q30

ls

【Step2】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

【Step3】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

【Step4】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

【Step5】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 △1

1111-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

【Step6】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

【Step7】Decompress vcf.gz

gunzip 11111-A.chr17.vcf.gz

Exome-Seq Data Analysis
全外顯子組定序數據分析

Introduction of copy number variation (CNV) and CNV Detection

Yen-An Tang, Ph.D

CENTER FOR GENOMIC MEDICINE,
NATIONAL CHENG KUNG UNIVERSITY

2020/7/8

CONTENTS

01 Introduction of copy number variation (CNV)

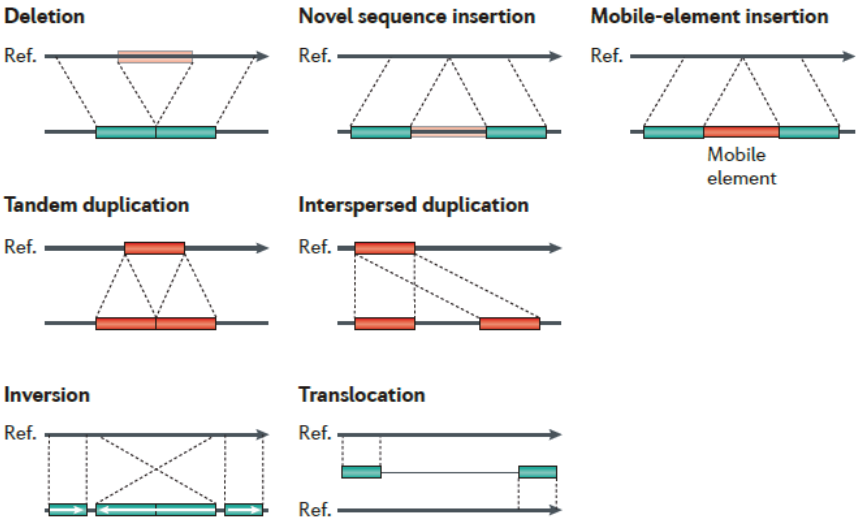
02 Chromosomal structure associated diseases

03 CNV databases and tools

04 Hands on: CNVkit

Chapter 1. Introduction of copy number variation (CNV)

Currently, >50 bp is used as an operational demarcation between small insertion/deletion (indels) and copy number variants (CNVs).



NATURE REVIEWS | GENETICS VOLUME 12 | MAY 2011 |

Major mechanisms underlying human genomic rearrangements and CNV formation

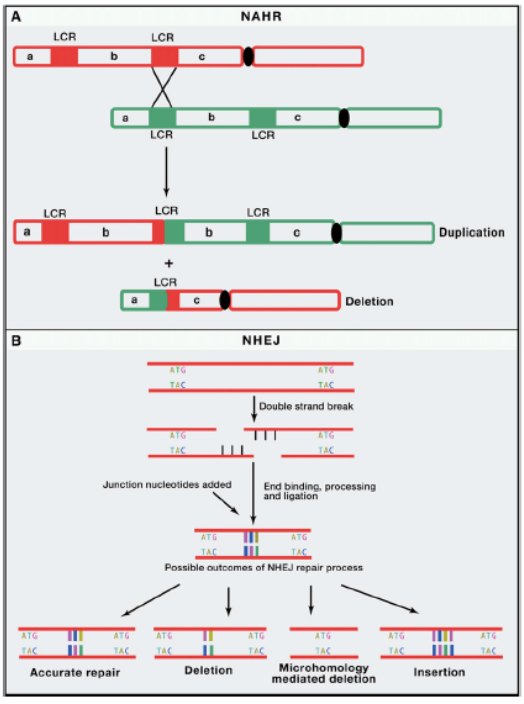
(a) Non-Allelic Homologous Recombination (NAHR) between repeat sequences (LCRs/SDs, Alu, or L1 elements);

Meiosis; Large deletion/insertion

(b) Non-Homologous End-Joining (NHEJ), recombination repair of double strand break;

(c) Fork Stalling and Template Switching (FoSTeS);

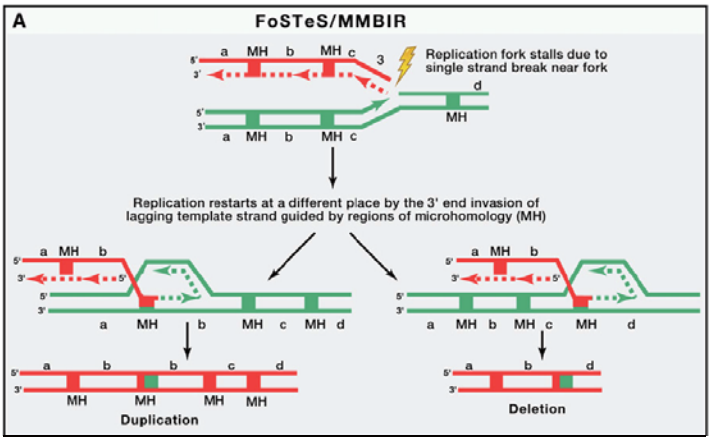
(d) L1 Retrotransposition.

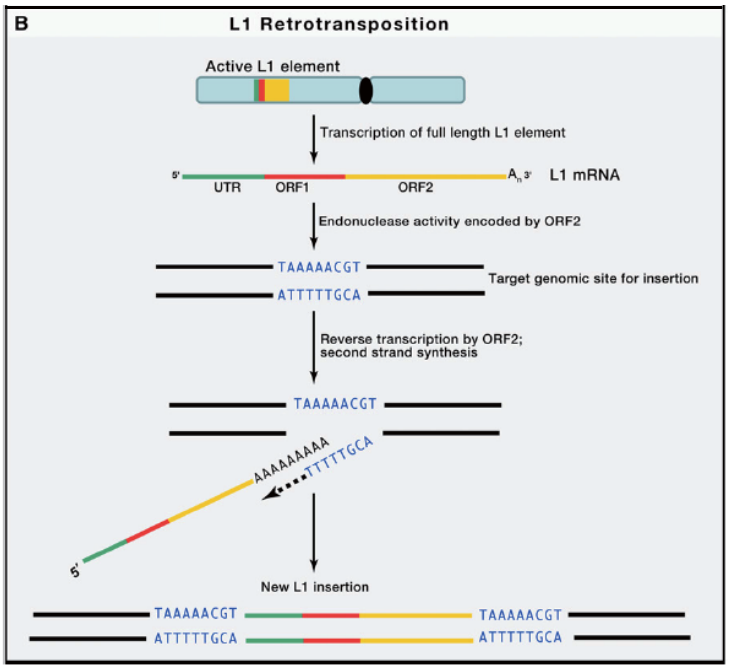


Malhotra and Sebat, 2012, *Cell*. 4:1223-1241

Zhang et. al., 2009, *Annu. Rev. Genomics Hum. Genet.*10:451–81

Major mechanisms underlying human genomic rearrangements and CNV formation



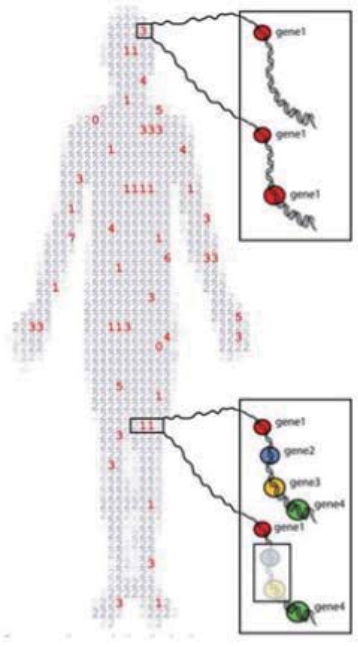


Replication-based mechanisms:
Fork Stalling and Template Switching (FoSTeS)
Microhomology-mediated break-induced replication (MMBIR)

MH (microhomology): short regions of DNA sequence homology

Malhotra and Sebat, 2012, *Cell*. 4:1223-1241

Chapter 2. Chromosomal structure associated diseases



- **Copy number gain (CN >2):** Increase of copy number due to genomic rearrangement like insertion/duplication
- **Copy number loss (CN <2):** Decrease of copy number due to deleterious genomic rearrangement
- **Copy number aberration (CNA):** Refers to CNV particularly when the events are associated with disease or phenotype

Disorder name	Chromosome	Chromosome size (Mb)	Deletion size (Mb)	% chromosome loss
1p36 deletion syndrome	1p36	247	1.5-10	0.6-4
Wolf-Hirschhorn syndrome	4p	191	1.4-8.4	0.7-4.4
Cri du chat syndrome	5p	181	10.5	5.8
Jacobsen syndrome	11q	134	4.1-12.8	3.1-10
Prader-Willi syndrome	15q	100	5-6	5-6
Angelman syndrome	15q	100	3.5	3.5
18p deletion syndrome	18p	76	3.8-30	5-40

CNVs plays key roles in neuropsychiatric disorders

nature Vol 466|15 July 2010|doi:10.1038/nature09146

LETTERS

Functional impact of global rare copy number variation in autism spectrum disorders

Leading Edge Review Cell

CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics

Cell *148*, March 16, 2012

Human Molecular Genetics, 2014 1–7
doi:10.1093/hmg/ddu379

***De novo* CNVs in bipolar affective disorder and schizophrenia**

Article
Cell Reports

Comparative Analyses of Copy-Number Variation in Autism Spectrum Disorder and Schizophrenia Reveal Etiological Overlap and Biological Insights

Cell Reports *24*, 2838–2856, September 11, 2018

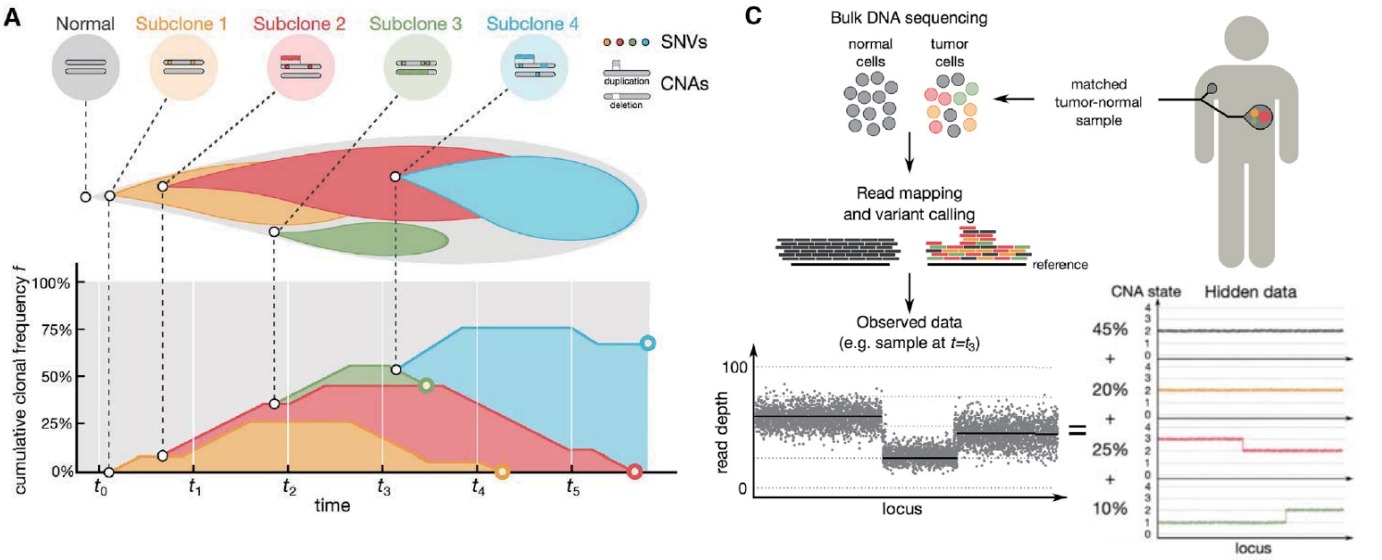
JAMA Psychiatry | Original Investigation

Analysis of Intellectual Disability Copy Number Variants for Association With Schizophrenia

JAMA Psychiatry September 2016 Volume 73, Number 9

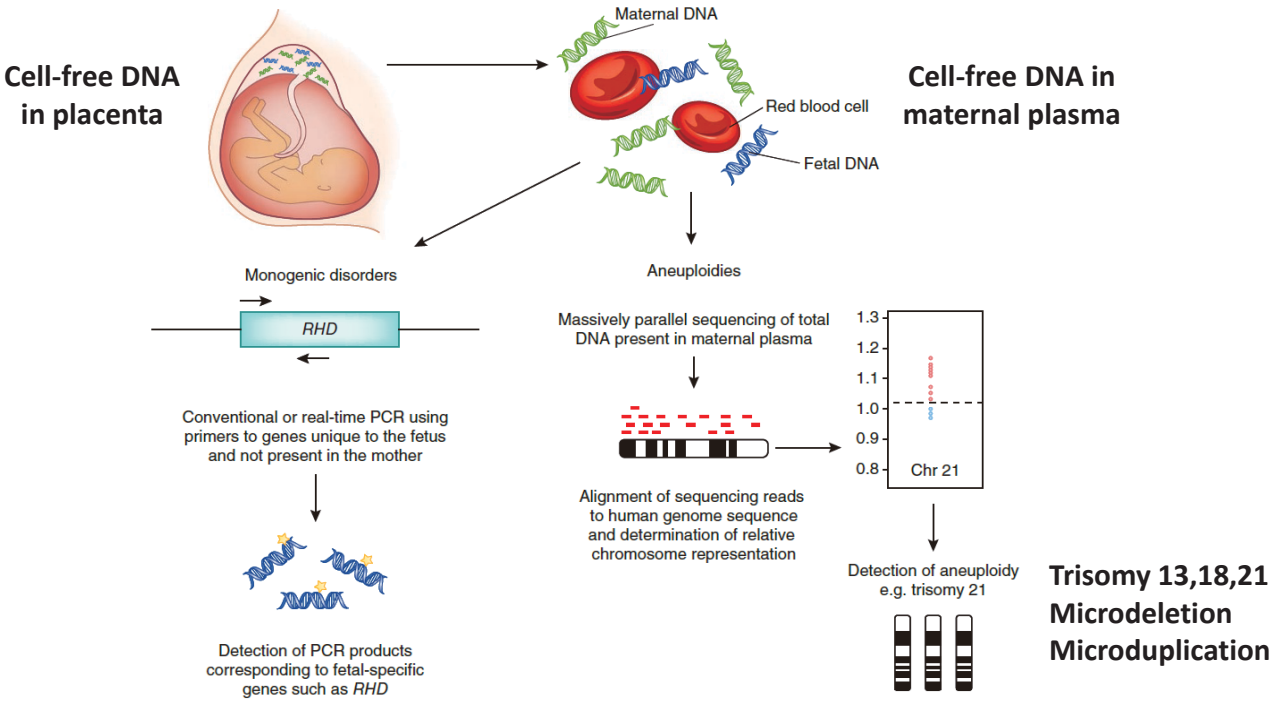
CNVs detection in somatic events

- **Clonal evolution of mutations and clonal heterogeneity in tumors**



Fischer et al., *Cell Reports* 7, 1740–1752, 2014

Non-Invasive Prenatal Testing (NIPT)



NATURE MEDICINE VOLUME 18 | NUMBER 7 | JULY 2012

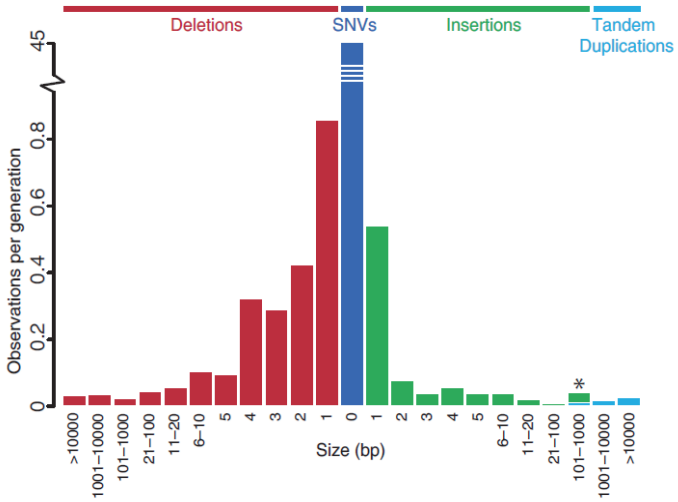
Comparison of SNVs and CNVs in population

1. **On average, each person is found to carry approximately 250–300 loss-of-function SNVs/indels in annotated genes and 24–30 variants previously implicated in rare disorders through ClinVar.** (1000 Genomes Project, *Nature*, 2010; 1000 Genomes Project, *Nature*, 2015)

2. **Approximate 50–100 new SNVs/indels occur each generation.** (Kloosterman et al., *Genome Research*, 2015; Lappalainen et al., *Cell*, 2019)

3. **The mutation rate for each structural variation is estimated about 0.16–0.226 per generation.** (Malhotra and Sebat, *Cell*, 2012; 1000 Genomes Project, *Nature*, 2015; Kloosterman et al., *Genome Research*, 2015;)

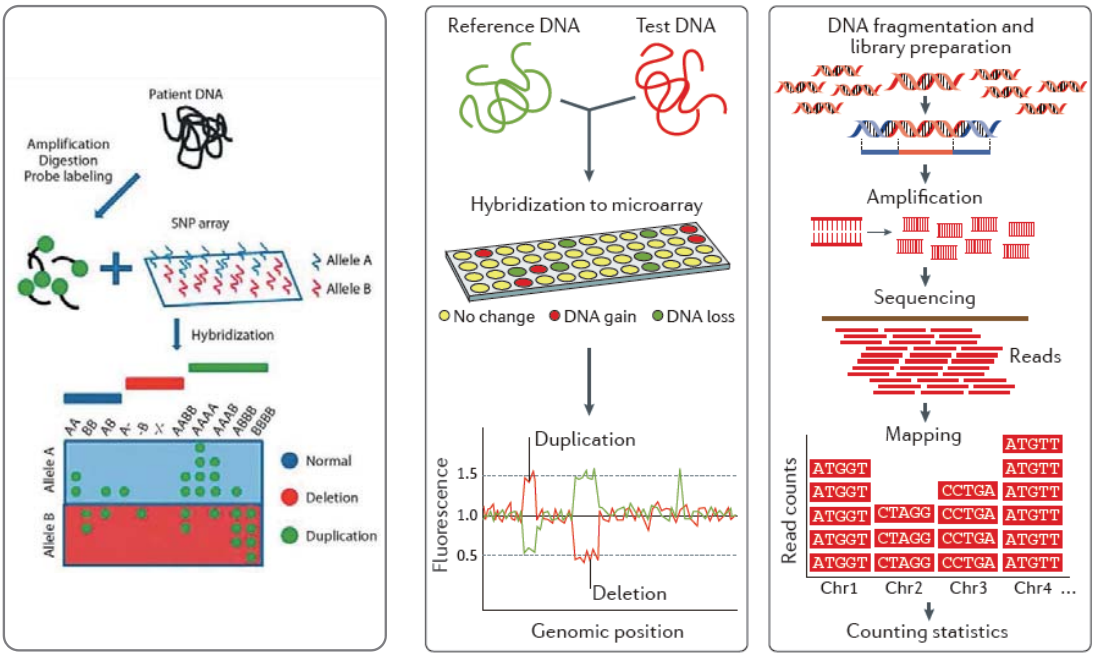
4. **The overall rate of genomic change by CNV (as measured in nucleotides) is high, on the order of 1,000 bp per generation, and the functional impact per site is large.** (Malhotra and Sebat, *Cell*, 2012; 1000 Genomes Project, *Nature*, 2015; Birney and Soranzo, *Nature*, 2015)



Kloosterman et al., *Genome Research*, 25:792–801,2015

The techniques for CNV detection

SNP array **aCGH** **NGS**



https://www.contemporarypediatrics.com/

NATURE REVIEWS | GENETICS VOLUME 17 | OCTOBER 2016

The limitation and strategy of CNV detection using NGS

Table 1. The Landscape of Human Genome Variation

Variant Class	Subclass, Other Term(s)	Size	Num. / genome (Illumina WGS)	Arrays	Short-Read WGS	Long-Read WGS	WES
Single Nucleotide Variation (SNV)	point mutation; substitution	1 bp	3.5×10^6	XX	XXX	XX	XXX
Small Insertion/Deletion Variation (indel)	insertion; deletion; complex indel	1-49 bp	4.5×10^5	XX	XX	X	XX
Structural Variation (SV) copy number variation (CNV)	deletion	≥50 bp	5,000	X	XX	XXX	XX
	duplication (tandem, interspersed)		1,000	X	XX	XXX	X
	multi-allelic CNV; tandem repeats		450	X	XX	XX	-
insertion	novel, templated or repeat insertion		1,500	-	X	XXX	-
balanced rearrangement	inversion		40	-	XX	XXX	-
	reciprocal translocation	inter-chrom	0.001	-	XX	XXX	-
complex genomic rearrangement	complex SV; chromothripsis	>1 mb	0.01	-	XX	XXX	-
extremely large copy number variant	aneuploidy; chrom. abnormality	>1 mb	0.01	XXX	XXX	XXX	XXX
retrogene insertion	retroduplication; retrocopy	gene coding length	10	-	XX	XXX	-
mobile element insertion (MEI)	SINE; LINE; SVA	0.3-7 kb	2,000	-	X	XXX	-

"–" indicates a near-complete inability to detect that variant class, XXX indicates that the technology is highly effective.

Lappalainen et al., 2019, *Cell,* 177:70-84

The advantages and disadvantages of long read sequencing

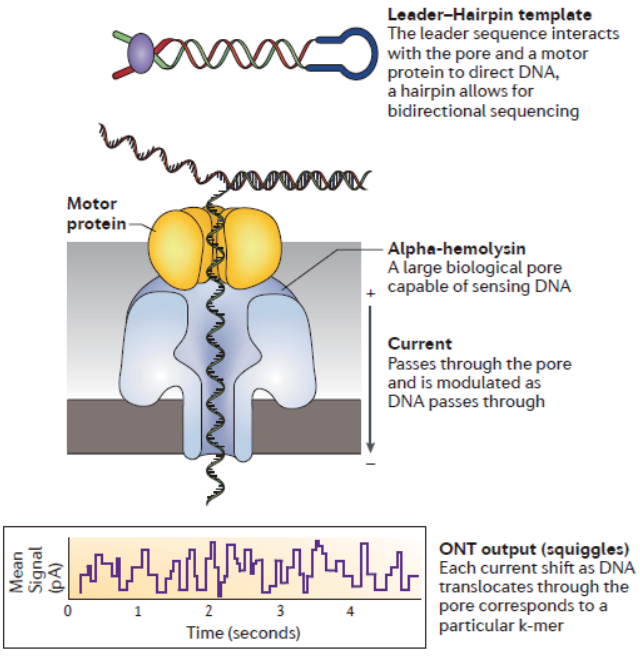
- ✓ **Balanced reciprocal translocation (BRT)**
- ✓ **Unbalanced translocations**
- ✓ **Uniparental disomy (UPD) chromosome**

Platform	Read length (bp)	Throughput	Reads	Runtime	Error profile
Single-molecule real-time long reads					
Pacific BioSciences RS II	~20 Kb	500 Mb–1 Gb*	~55,000*	4 h*	13% single pass, ≤1% circular consensus read, indel‡
Oxford Nanopore MK 1 MinION	Up to 200 Kb[159]	Up to 1.5 Gb[159]	>100,000 (REF. 159)	Up to 48 h[160]	~12%, indel[159]
Synthetic long reads					
10X Genomics	Up to 100 Kb synthetic length*	See HiSeq 2500	See HiSeq 2500	See HiSeq 2500	See HiSeq 2500 (possible barcoding and partitioning errors)

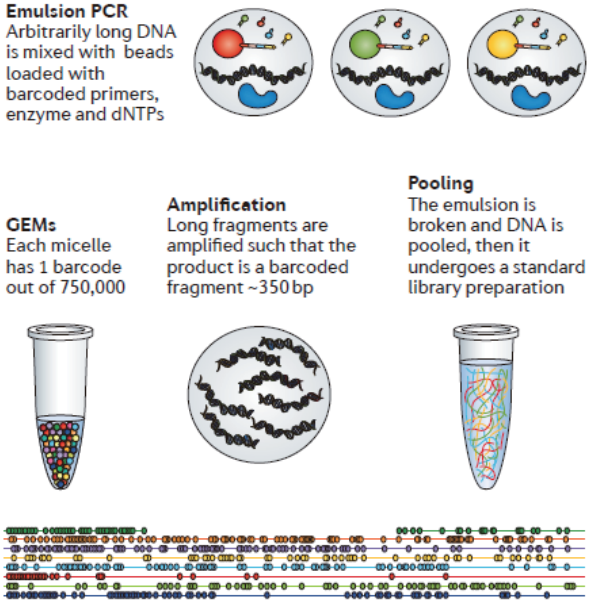
NATURE REVIEWS | GENETICS VOLUME 17 | JUNE 2016

The platforms of long read sequencing

Ab Oxford Nanopore Technologies



Bb 10X Genomics linked-read sequencing



NATURE REVIEWS | GENETICS VOLUME 17 | JUNE 2016

Chapter 3. CNV databases

a) **DGV: Database of Genomic Variants**
>22,300 genomes, microarray/NGS data, records: > 7.0 million entries

http://dgv.tcag.ca/dgv/app/home

b) **DECIPHER: DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources)**

33,959 clinical patients, microarray/NGS data, record: 36,224 CNVs with clinical indication https://decipher.sanger.ac.uk/

c) **ISCA Consortium: International Standards for Cytogenomic Arrays**

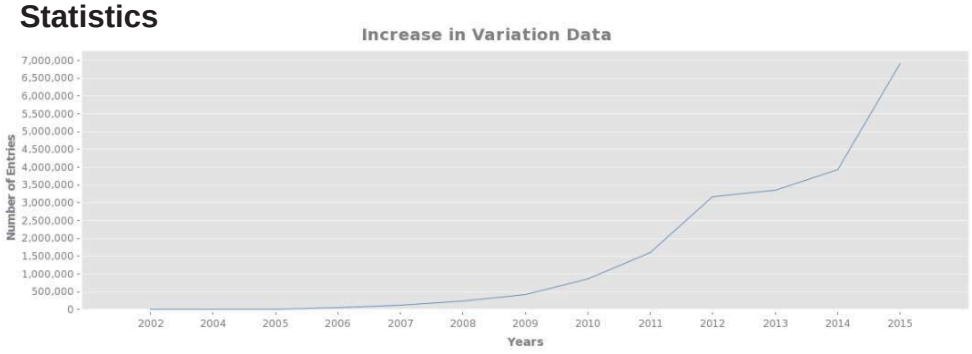
2007-2012: >200,000 clinical patients, aCGH and SNP array,
publicly available databases through NCBI (ClinGen) with curated clinical indication

http://dbsearch.clinicalgenome.org/search/

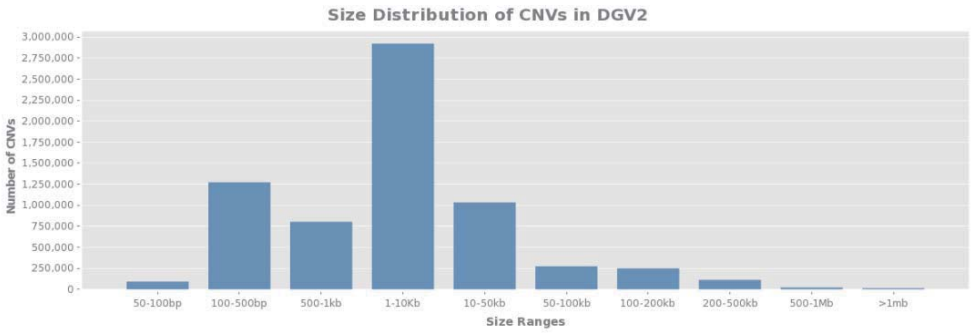
Database of Genomic Variants

A curated catalogue of human genomic structural variation

- **Statistics**



Content Growth
This graph shows the increase in published structural variation data that have been added to the database since its start in 2004; the numbers reflect the year of publication.



Size distribution (CNV)
The graph displays the size distribution of CNVs in the database.

http://dgv.tcag.ca/dgv/app/statistics?ref=GRCh37/hg19

CNVs detection tools for NGS data

- **ExomeCNV (2011)**
  - -Read counts & LOH
  - -cited 350 times
  - -somatic CNV only, WES only
  - https://cran.r-project.org/src/contrib/Archive/ExomeCNV/
- **CoNIFER (2012)**
  - -Read counts method
  - -cited 403 times
  - -somatic & germline CNV, WES only
  - http://conifer.sourceforge.net/
- **cn.MOPS (2012)**
  - -Read counts method
  - -cited 270 times
  - -somatic & germline CNV
  - -WES & WGS
  - https://bioconductor.org/packages/release/bioc/html/cn.mops.html
- **sCNAphase (2017)**
  - -Read counts & SNP
  - -somatic CNV, WGS only
  - https://github.com/Yves-CHEN/sCNAphase
- **GATK—ModelSegments (2018)**
  - -Read counts & SNP
  - -somatic CNV only
  - -WES and WGS
  - https://software.broadinstitute.org/gatk/
- **CNVkit (2016)**
  - -Read counts & SNP
  - -cited 254 times
  - -somatic & germline CNV
  - -WES, targeted-seq, WGS
  - https://cnvkit.readthedocs.io/en/latest/index.html

Chapter 4

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

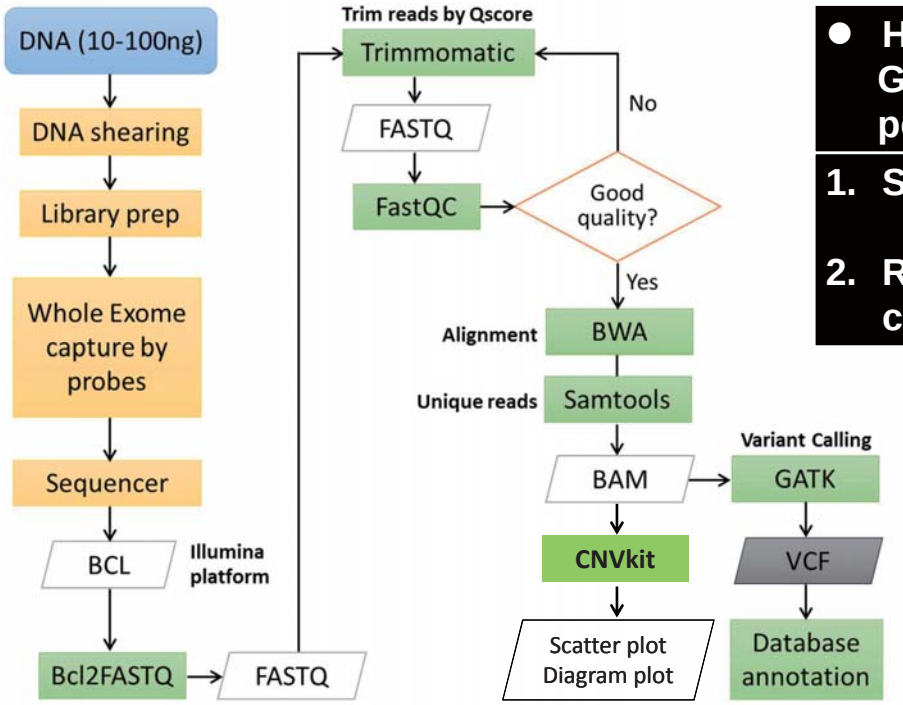
PLOS Computational Biology | DOI:10.1371/journal.pcbi.1004873 April 21, 2016

CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing

Eric Talevich[1,2,3], A. Hunter Shain[1,2,3], Thomas Botton[1,2,3], Boris C. Bastian[1,2,3]*

- **Python script**
- **Platform:** WES, Targeted sequencing, WGS
- **Methods:** reads counts/SNP
- Bias correction
- Copy number analysis of samples prepared with the same library preparation protocol and sequenced on the same platform

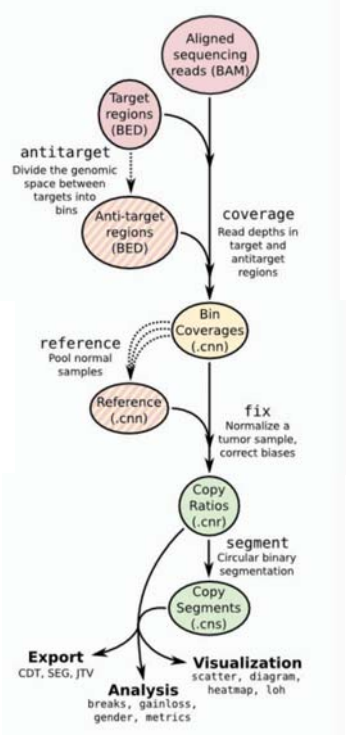
WES – CNV analytic pipeline



- **Hands on：**
  Germline DNA from peripheral whole blood

1. **Samples: 19030164**
   19030165
2. **References: nine normal controls**

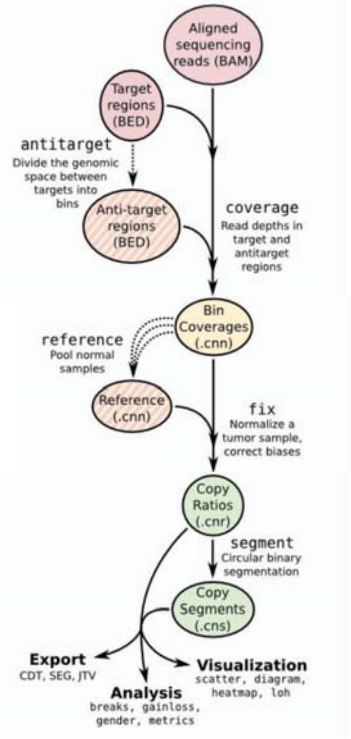
CNVkit – the flow of analytic pipeline



1. **Target** — Prepare a BED file of baited regions (on-target) and anti-target (off-target) region (only once for the same platform)
2. **Antitarget**
3. **Coverage** — For every samples (including reference)
4. **Reference** — Only for the samples serving as reference
5. **Fix** — Samples/reference (normalization)
6. **Segment** — Algorithms to call segment
7. **Plots** — Scatter and diagram plot

https://cnvkit.readthedocs.io/en/latest/index.html

CNVkit – preparation at the beginning



Bin size and resolution

Exons in the human genome have an average size of about 200bp. The target bin size default of 267 is chosen so that splitting larger exons will produce bins with a minimum size of 200. Since bins that contain fewer reads result in a noisier copy number signal, this approach ensures the "noisiness" of the bins produced by splitting larger exons will be no worse than average.

autobin

[cnvkit.py autobin *.bam -t my_targets.bed -b -g access.hg19.bed]

-b --bp-per-bin Desired average number of sequencing read bases mapped to each bin

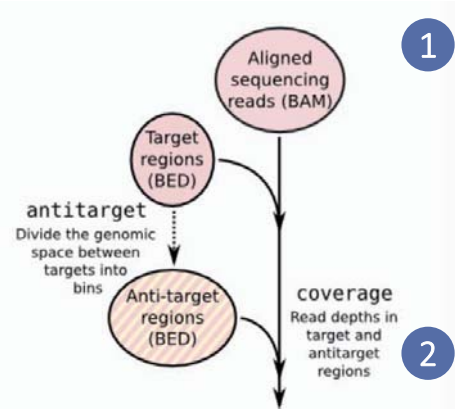
In practice we see good results with an average of 200–300 reads per bin

150base*250reads≒38000 (base coverage)

cnvkit.py autobin 19030165_F0x100.bam -t Exome_hg19_capture.bed -b 38000 -g access-5k-mappable.hg19.bed

	Depth	Bin size
Target:	87.899	432
Antitarget:	1.021	37208

CNVkit – Target & Antitarget



1 **Target: Prepare a BED file of baited regions for use with CNVkit**

[cnvkit.py target my_baits.bed --split -o my_targets.bed]

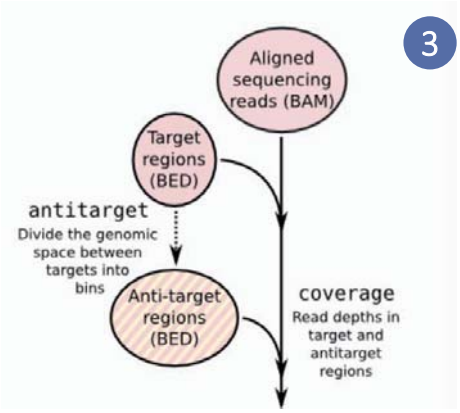
cnvkit.py target Exome_hg19_capture.bed --split --average-size 432 -o Exome_hg19_capture_targets_432.split.bed

2 **Antitarget: Given a "target" BED file that lists the chromosomal coordinates of the tiled regions, derive a BED file off-target regions**

[cnvkit.py antitarget my_targets.bed -g data/access-5kb-appable.hg19.bed -o my_antitargets.bed]

cnvkit.py antitarget Exome_hg19_capture_targets_432.split.bed -g access-5k-mappable.hg19.bed -o Exome_hg19_capture_targets_432.auto.split.antitarget.bed

CNVkit – Coverage



3 **Coverage: Calculate coverage in the given regions from BAM read depths for each sample**

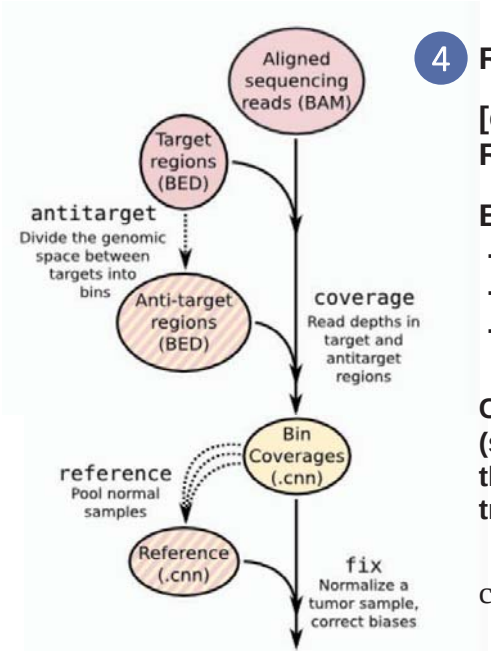
(1) [cnvkit.py coverage Sample.bam targets.bed -o Sample.targetcoverage.cnn]

(2) [cnvkit.py coverage Sample.bam antitargets.bed -o Sample.antitargetcoverage.cnn]

cnvkit.py coverage 19030165_F0x100.bam ../Exome_hg19_capture_targets_432.split.bed -o 19030165_F0x100_432.targetcoverage.cnn

cnvkit.py coverage 19030165_F0x100.bam ../Exome_hg19_capture_targets_432.auto.split.antitarget.bed -o 19030165_F0x100_432.antitargetcoverage.cnn

CNVkit – Reference



4 **Reference: Combine all normal samples into a pooled reference**

[cnvkit.py reference *coverage.cnn -f ucsc.hg19.fa -o Reference.cnn]

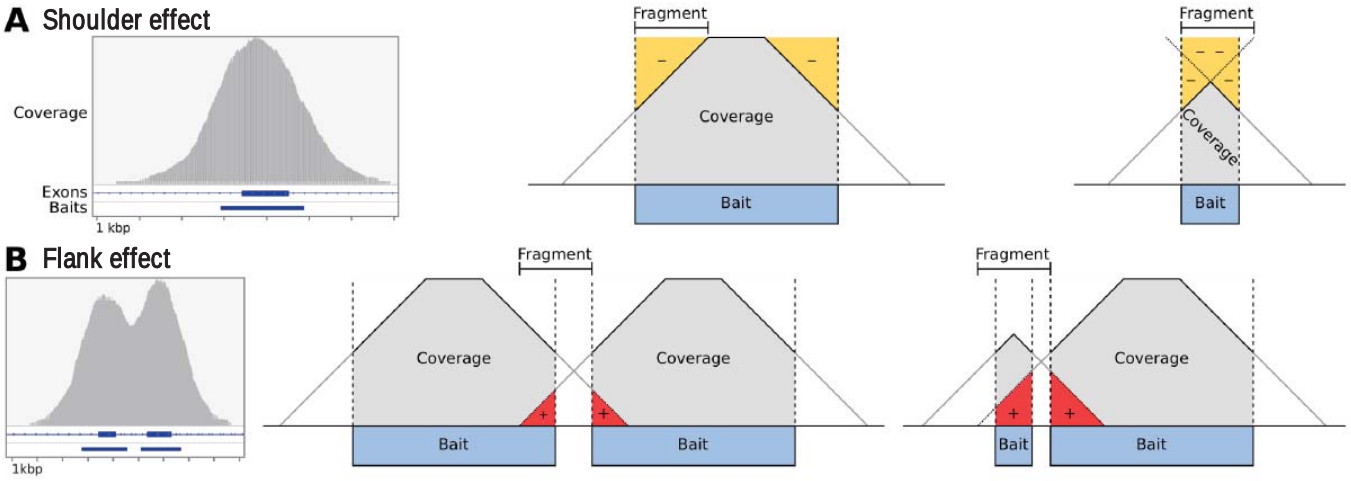
Bias correction:
--gc, GC correction
--edge, edge-effect correction
--rmask, RepeatMasker correction

CNVkit calculates the association between each bin's GC content (stored in the reference) and observed read depth, fits a trendline through the bin read depths ordered by GC value, and subtracts this trend from the original read depths.

cnvkit.py reference *.cnn -f hg19.fasta -o CNVKit_9sample_432.cnn

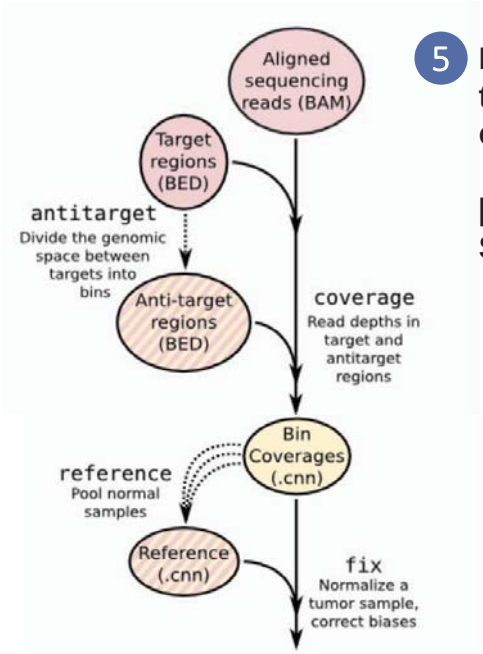
PLOS COMPUTATIONAL BIOLOGY
PLOS Computational Biology | DOI:10.1371/journal.pcbi.1004873 April 21, 2016

Baited region size and spacing affect read depth systematically



•Within the baited region, read depth is lower at the "shoulders" where sequence fragments are not completely captured.
•Just outside the baited region, in the "flanks", read depth is elevated to nearly that of the adjacent baited sites due to the same effect. If two targets are very close together, the sequence fragments captured for one target can increase the read depth in the adjacent target.

CNVkit – Fix



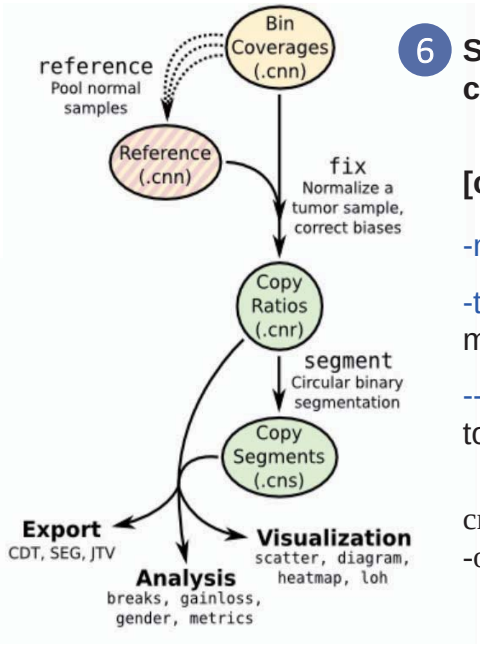
5 **Fix: Combine the uncorrected target and antitarget coverage tables (.cnn) and correct for biases in regional coverage and GC content, according to the given reference**

[cnvkit.py fix Sample.targetcoverage.cnn Sample.antitargetcoverage.cnn Reference.cnn -o Sample.cnr]

Bias correction:
--no gc, skip GC correction
--no edge, skip edge-effect correction
--no rmask, skip RepeatMasker correction

cnvkit.py fix 19030165_F0x100_432.targetcoverage.cnn 19030165_F0x100_432.antitargetcoverage.cnn CNVKit_9sample_432.cnn -o 19030165_F0x100_CNVKit_432_fixRef.cnr

CNVkit – Segment



6 **Segment: Infer discrete copy number segments from the given coverage table**

[cnvkit.py segment Sample.cnr -o Sample.cns]

-m {cbs,flasso,haar,none,hmm,hmm-tumor,hmm-germline},

-t, --threshold Significance threshold (p-value or FDR, depending on method) to accept breakpoints during segmentation.

--drop-low-coverage Drop very-low-coverage bins before segmentation to avoid false-positive deletions in poor-quality tumor samples.

cnvkit.py segment -m cbs 19030165_F0x100_CNVKit_432_fixRef.cnr -o 19030165_F0x100_CNVKit_432_fixRef.cns

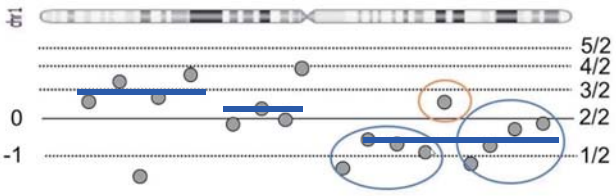
Circular binary segmentation for the analysis of array-based DNA copy number data

Biostatistics (2004), **5**, 4, *pp.* 557–572

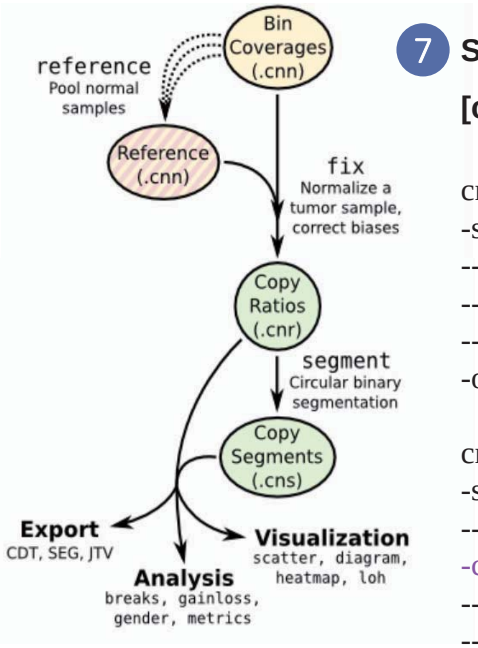
- **The CBS is a well-established segmentation method and many CNV detection tools use it for identifying CNV segments.**
- **The statistic for testing the null hypothesis of no change against the alternative of exactly one change at an unknown location.**
- **Split the chromosomes into regions of equal copy number that accounts for the noise in the data.**
- **The binary segmentation procedure applies the test recursively until no more changes are detected in any of the segments obtained from the change-points already found.**

Identify the "change point"

- Recursively divide up the genome until identify segments that have probe distribution different than neighbors



CNVkit – Visualization



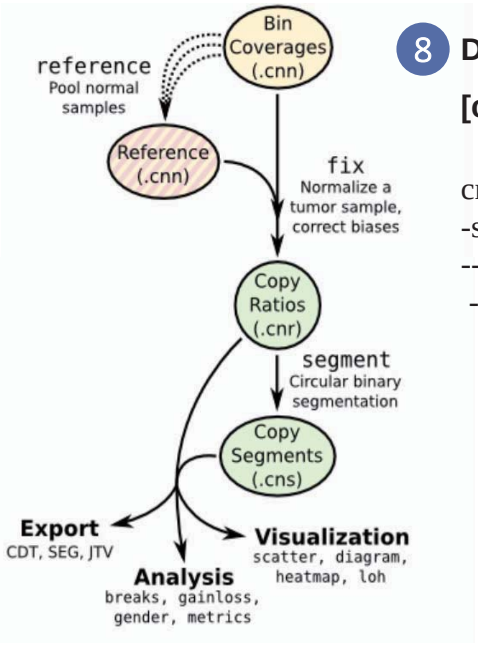
7 **Scatter plot:**

[cnvkit.py scatter Sample.cnr -s Sample.cns]

cnvkit.py scatter 19030165_F0x100_CNVKit_432_fixRef.cnr
-s 19030165_F0x100_CNVKit_432_fixRef_Seg_cbs.cns
--title 19030165_F0x100_432_cbs
--y-max 2
--y-min -2
-o 19030165_CNVKit_432_fixRef_Seg_cbs.pdf

cnvkit.py scatter 19030165_F0x100_CNVKit_432_fixRef.cnr
-s 19030165_F0x100_CNVKit_432_fixRef_Seg_cbs.cns
--title 19030165_F0x100_432_cbs_chr22
-c chr22
--y-max 2
--y-min -2
-o 19030165_CNVKit_432_fixRef_Seg_cbs_chr22.pdf

CNVkit – Visualization

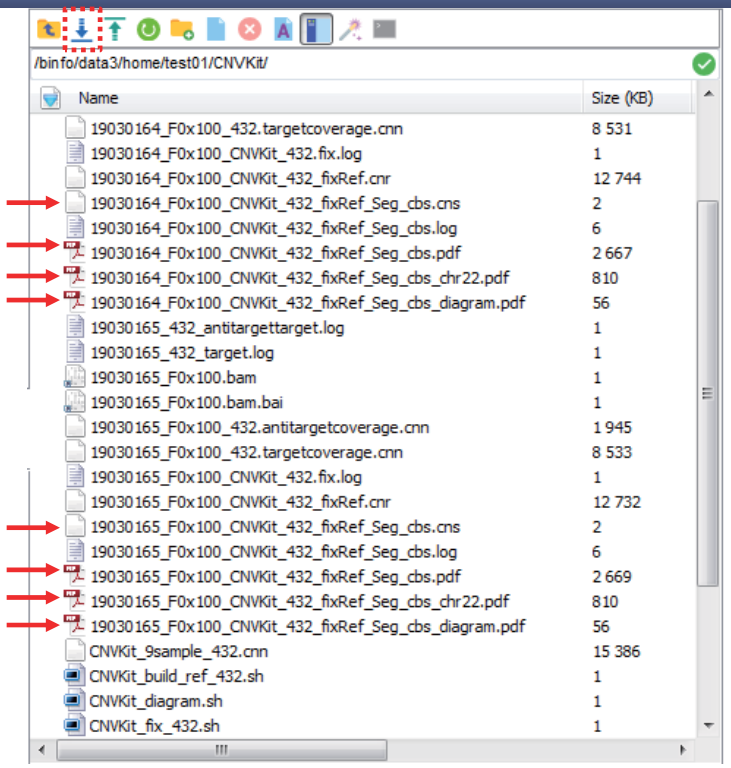


8 **Diagram plot:**

[cnvkit.py diagram -s Sample.cns]

cnvkit.py diagram
-s 19030165_F0x100_CNVKit_432_fixRef_Seg_cbs.cns
--title 19030165_F0x100_432_cbs
-o 19030165_CNVKit_432_fixRef_Seg_cbs_diagram.pdf

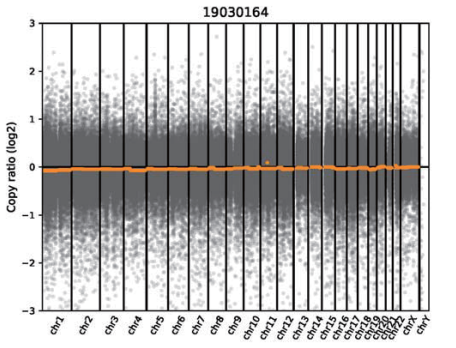
CNVkit – Visualization

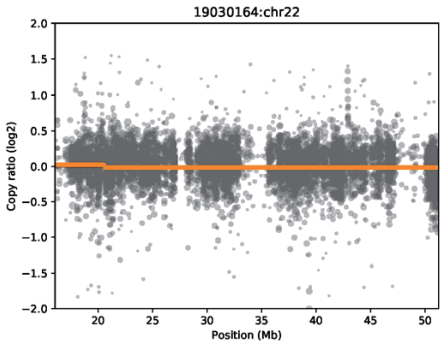


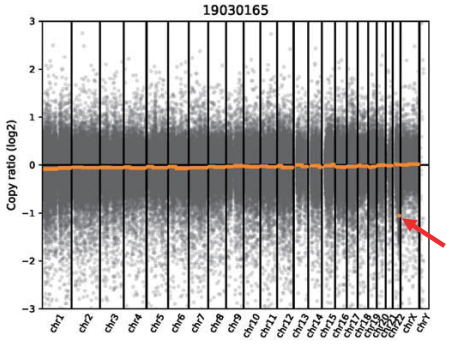
Download 8 files:
→ 2 for .cns file
→ 6 for .pdf file

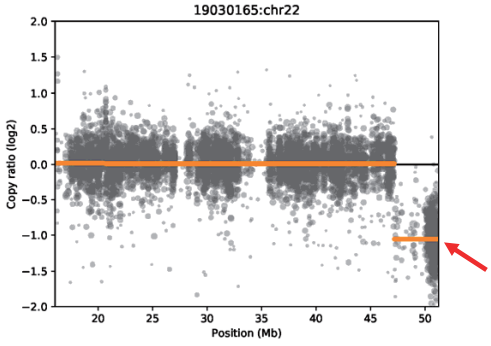
CNVkit – Visualization

Scatter plot:









Copy ratio (log2)

Ratio= sample/ref

CN=2/2=1 → 2^{0}

CN=1/2=0.5 → 2^{-1}

CN=3/2=1.5 → $2^{0.58}$

CNVkit – Visualization

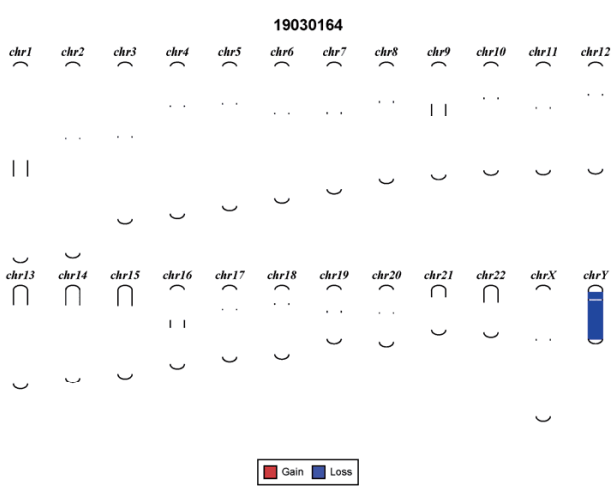
19030165_F0x100_CNVKit_432_fixRef_Seg_cbs.cns

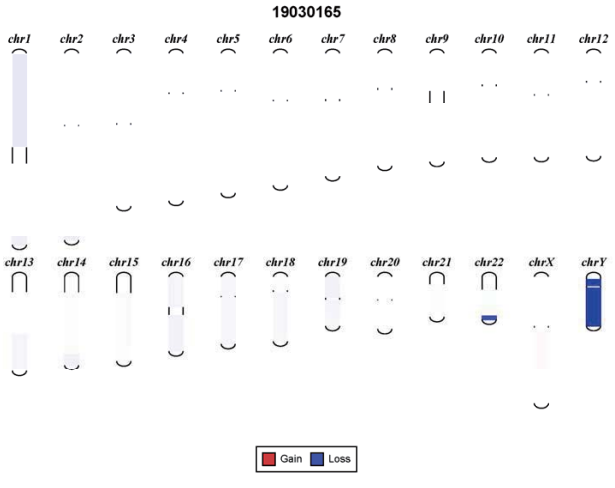
	A	B	C	D	E	F	G	H
1	chromosome	start	end	gene	log2	depth	probes	weight
38	chr19	110678	24364012	-	-0.05434	96.6478	6465	3539.08
39	chr19	28263182	59118483	-	-0.03864	95.8183	6915	3853.68
40	chr20	68326	26256532	-	-0.00716	84.8299	1961	998.111
41	chr20	29420069	62904970	-	-0.01865	93.7371	4010	2108.88
42	chr21	9411693	48119395	-	-0.0213	86.5959	2469	1250.45
43	chr22	16258185	22599772	-	-0.02319	93.4494	893	461.042
44	chr22	22842063	47193517	-	-0.00228	105.301	3420	1808.32
45	chr22	47194017	51244066	-	-1.08323	55.9293	556	299.439
46	chrX	60500	58488398	-	-0.02549	116.858	3814	1885.52
47	chrX	62042080	1.55E+08	-	-0.00971	111.656	5274	2664.32
48	chrY	2735457	9236561	-	-5.75696	4.41143	66	33.0407
49	chrY	9304587	10058894	-	-16.4475	0.21957	18	9.53852
50	chrY	13105053	59099792	-	-10.9562	1.07128	129	60.9771

A weight is assigned to each remaining bin depending on:
1.The size of the bin;
2.The deviation of the bin's log2 value in the reference from 0;
3.The "standard deviation" of the bin in the reference.

CNVkit – Visualization

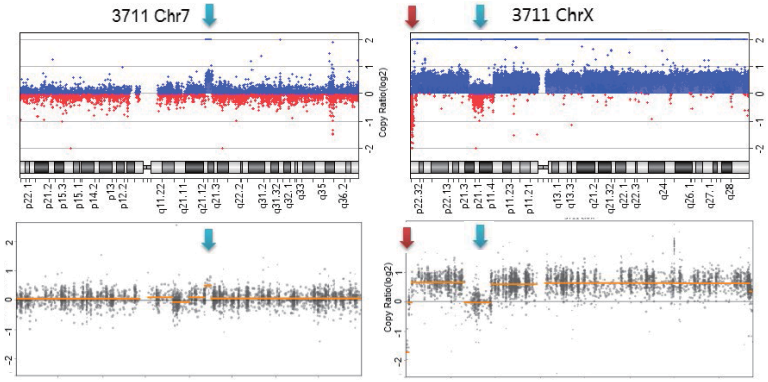
Diagram plot:





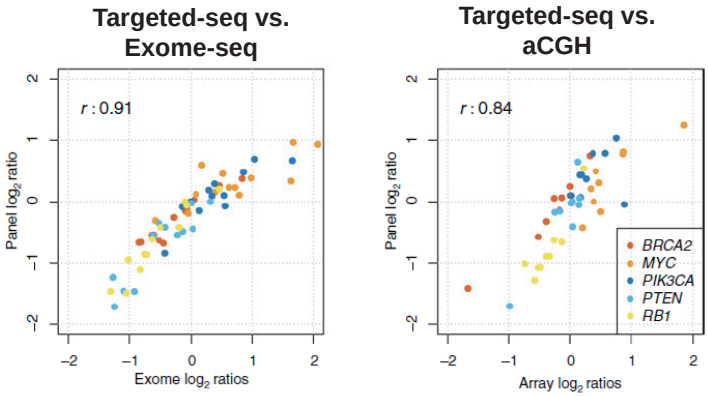
CNVkit copy ratios agree with aCGH in clinical samples

Germline DNA, Inherited disease



Yao et al., *Mol Genet Genomic Med*, 2019

Prostate tumor DNA



Target panel: 113 genes, covered 564 kb of the human genome. CNVkit was performed using targeted-seq and exome-seq data.

Seed et al., *Clinical Cancer Research*, 2017

成大基因體與生物資訊研究平台服務

壹、儀器設備說明 (規格)：

本平台係基於叢集架構，並搭配圖形處理器(GPU)與大量儲存設備建構之運算主機，利用其適合單一且大量重複計算的特性，提供一個具高速運算效能的基因體資訊演算環境，加速相關研究之進行。

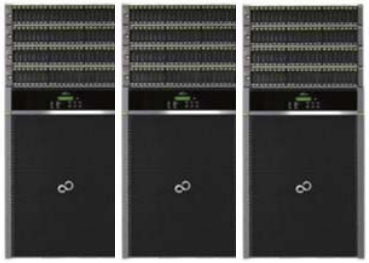
本平台包含：

- 高效能運算叢集系統：基因體資訊演算法研發、非模式物種基因體解析系統
- NGS商用分析軟體

一、硬體規格

■ 運算節點*8

1. Fujitsu PRIMERGY RX350 S8
2. Coprocessor - NVIDIA® Tesla™ K20C GPGPU
3. RAM: 512 GB ~768GB DDR3/每節點





二、軟體規格

■ 作業系統: Red Hat Enterprise Linux Server release 6.4 (Santiago)

■ 應用套件

1	BLAST	8	GATK	15	Ray	22	Velvet
2	Bowtie	9	Glimmer	16	SAMtools	23	R
3	BWA	10	InterProScan	17	SRAtoolkit	24	Python
4	Cufflinks	11	MACS	18	TopHat	25	Perl
5	DiscoveryStudio	12	MAQ	19	Trinity	26	Java
6	FastQC	13	mpiBLAST	20	tRNAcan	27	CNVkit
7	FASTX-toolkit	14	PartekFlow	21	tRNA-scanSE		

成大基因體與生物資訊研究平台服務

貳、服務項目：

1. 提供上述應用套件與其高速運算效能，滿足 研究需求之執行，如：次世代定序 (NGS) 分 析、非模式物種基因體 分析、 基因體註解、蛋白質結構 、組裝皆可提供服務。
2. 提供商用軟體，如 DiscoveryStudio、CLC Genomics Workbench，採預約付費使用。
   高效能運算叢集系統：基因體資訊演算法研發、非模式物種基因體解析系統

參、申請與使用應注意事項：

1. 申請時需說明使用目的、來源IP。
2. 設備使用方式：由使用者在自備電腦或終端機，以SSH登入/操作機器。
3. 每帳號空間最多為1.5TB(包含使用者上傳與產生的檔案)，每季統計該帳號空間使用情況，並通知清理空間。
4. 使用本系統所產生的資料於成果發表時，不要需要將服務提供者列為共同作者，但須於文章致謝欄中載明。

肆、使用時段與預約：

1. 商用軟體：開放24小時線上預約系統，自己掌握可用時段。
2. 高效能運算叢集系統：開放24小時以Queue排程指令執行。

伍、收費辦法及標準：

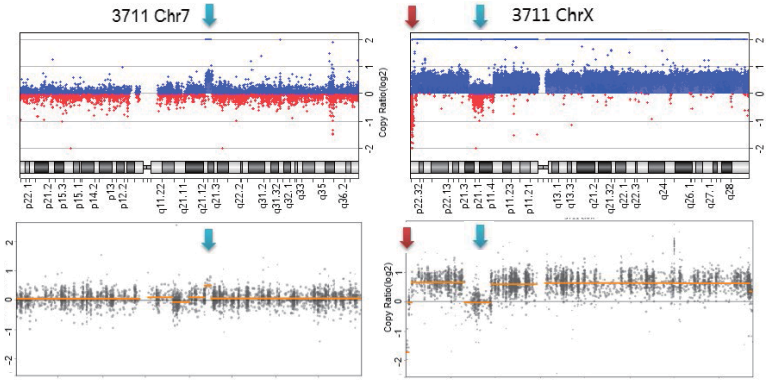
1. 商用軟體:每時段400元(採預付制)
2. 高效能運算叢集系統使用費: 每帳號8,000元/年； 5,000元/半年

陸、儀器聯絡與管理人：第一管理者：陳小姐 em61330@email.ncku.edu.tw

Thank you

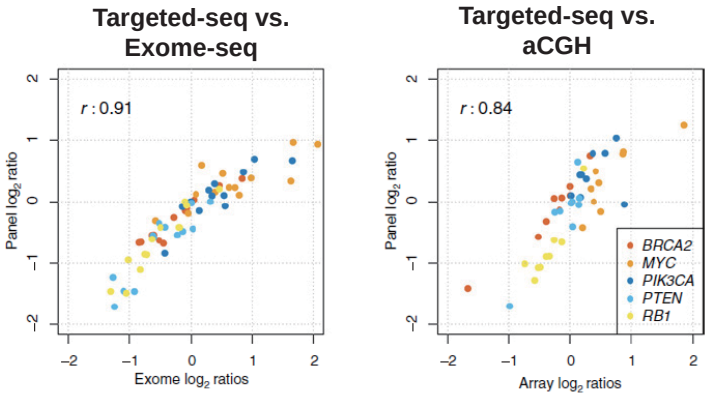
CNVkit copy ratios agree with aCGH in clinical samples

Germline DNA, Inherited disease



Yao et al., *Mol Genet Genomic Med*, 2019

Prostate tumor DNA



Target panel: 113 genes, covered 564 kb of the human genome. CNVkit was performed using targeted-seq and exome-seq data.

Seed et al., *Clinical Cancer Research*, 2017

Tumor analysis

- **To analyze a cohort sequenced on a single platform, CNVkit recommends combining all normal samples into a pooled reference, even if matched tumor-normal pairs were sequenced – our benchmarking showed that a pooled reference performed slightly better than constructing a separate reference for each matched tumor-normal pair.**

- **Furthermore, even matched normals from a cohort sequenced together can exhibit distinctly different copy number biases (Plagnol et al. 2012 and Backenroth et al. 2014); reusing a pooled reference across the cohort provides some consistency to help diagnose such issues.**

Adjusting copy ratios and segments for normal cell contamination

[cnvkit.py call -m clonal Sample.cns --purity 0.6 -o Sample.call.cns]

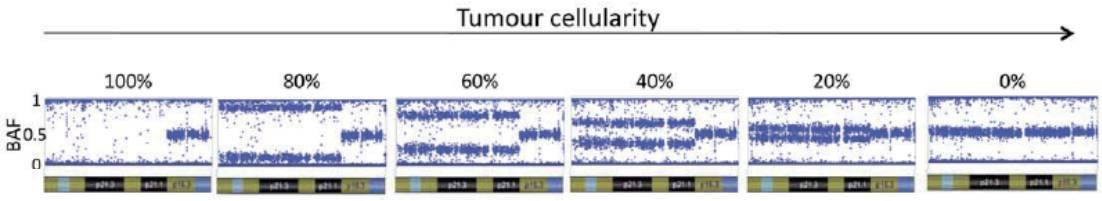
Inferring tumor purity and subclonal population fractions from sequencing: CNVkit provides wrappers for exporting .cns segments to THetA2's input format and importing THetA2's result file as CNVkit's segmented .cns files.

If SNV calls are available in VCF format, use the -v/--vcf option in the call and scatter commands to calculate or plot b-allele frequencies alongside each segment's total copy number or log2 ratio.

https://cnvkit.readthedocs.io/en/latest/heterogeneity.html

CNVkit – SNV b-allele frequencies (BAF)

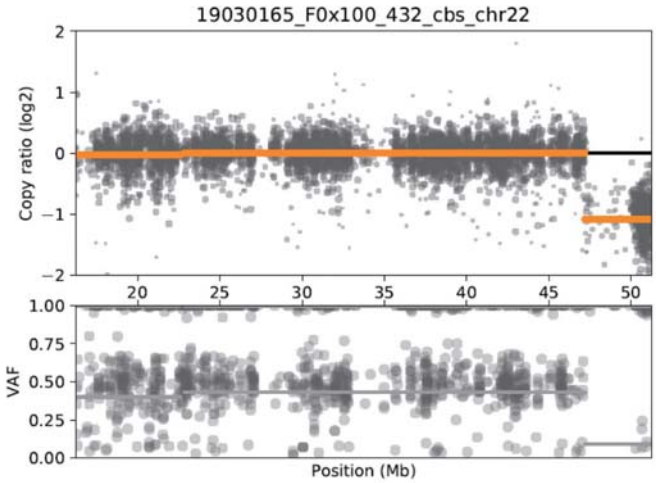
Genotype	BAF
A/A	0
A/B	0.5
B/B	1

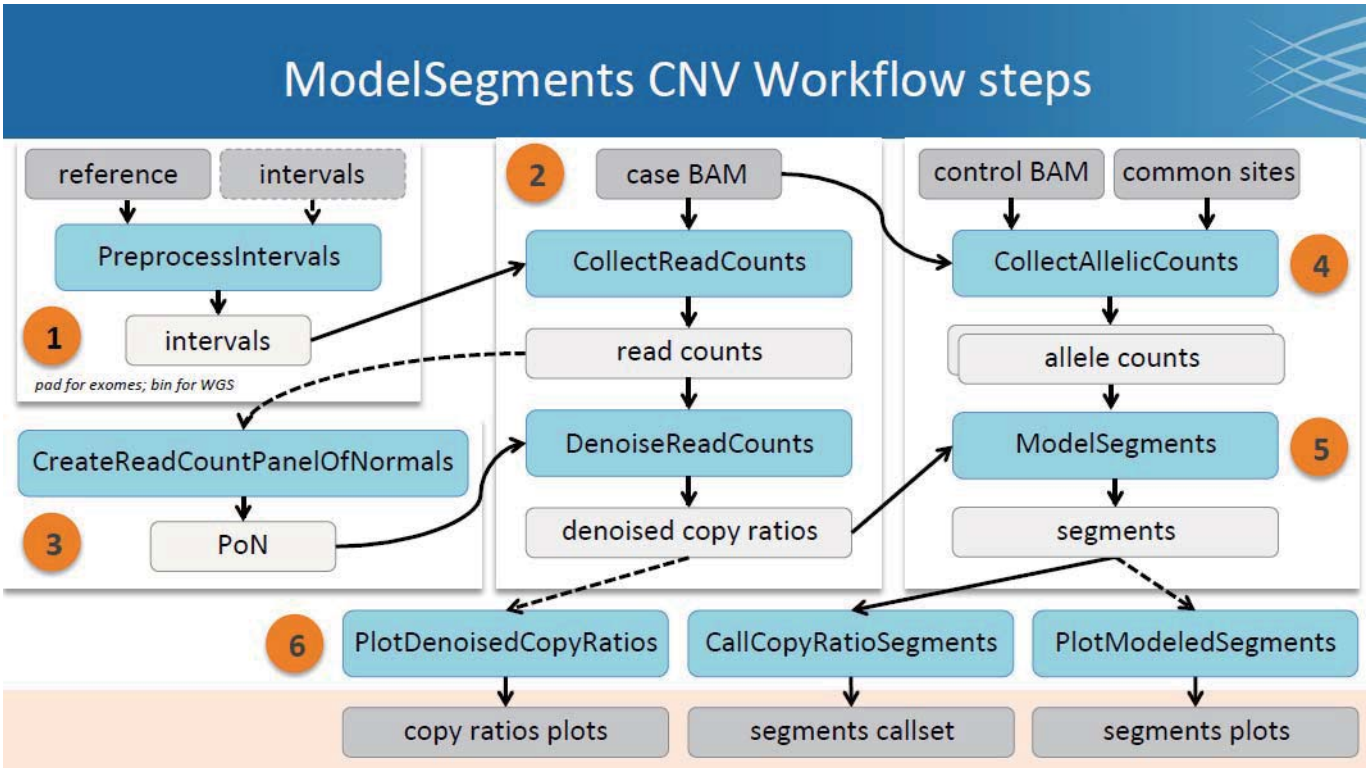


[cnvkit.py segment Sample.cnr --vcf -o Sample_vcf.cns]

cnvkit.py segment -m cbs
19030165_F0x100_CNVKit_432_fixRef.cnr
--vcf 19030165.vcf
-o 19030165_F0x100_CNVKit_432_fixRef_vcf.cns

[cnvkit.py scatter Sample.cnr -s Sample_vcf.cns --vcf]





https://software.broadinstitute.org/gatk/documentation/tooldocs/4.1.4.0/org_broadinstitute_hellbender_tools_copynumber_ModelSegments.php

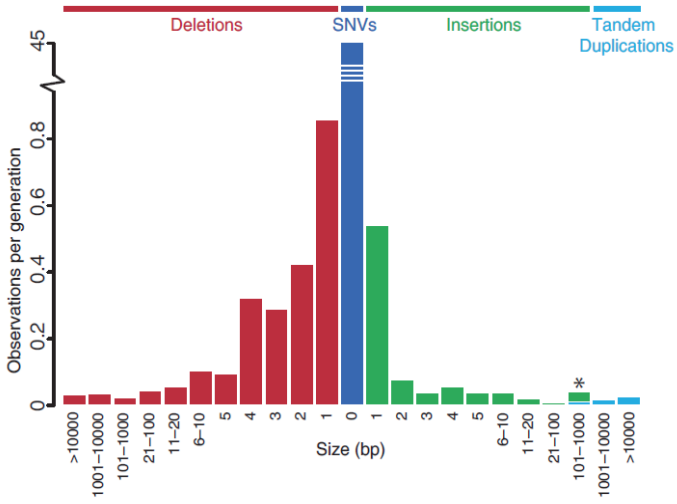
Comparison of SNVs and CNVs in population

1. **On average, each person is found to carry approximately 250–300 loss-of-function SNVs/indels in annotated genes and 24–30 variants previously implicated in rare disorders through ClinVar.** (1000 Genomes Project, *Nature*, 2010; 1000 Genomes Project, *Nature*, 2015)

2. **Approximate 50–100 new SNVs/indels occur each generation.** (Kloosterman et al., *Genome Research*, 2015; Lappalainen et al., *Cell*, 2019)

3. **The mutation rate for each structural variation is estimated about 0.16–0.226 per generation.** (Malhotra and Sebat, *Cell*, 2012; 1000 Genomes Project, *Nature*, 2015; Kloosterman et al., *Genome Res*earch, 2015;)

4. **The overall rate of genomic change by CNV (as measured in nucleotides) is high, on the order of 1,000 bp per generation, and the functional impact per site is large.** (Malhotra and Sebat, *Cell*, 2012; 1000 Genomes Project, *Nature*, 2015; Birney and Soranzo, *Nature*, 2015)



Kloosterman et al., *Genome Research*, 25:792–801,2015

CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics

Dheeraj Malhotra[1,2] and Jonathan Sebat[1,2,3,4,*]
[1]Beyster Center for Genomics of Psychiatric Diseases
[2]Department of Psychiatry
[3]Department of Cellular Molecular and Molecular Medicine
[4]Institute for Genomic Medicine
University of California, San Diego, La Jolla, CA 1020103, USA
*Correspondence: jsebat@ucsd.edu
DOI 10.1016/j.cell.2012.02.039

On average, there are > 1000 CNVs in the genome, accounting for 4 million base pairs of genomic difference (Conrad et al., 2010; Mills et al., 2011).

The rate of nucleotide substitutions genome-wide is estimated at 30–100 per generation (Conrad et al., 2011) and ~1 per exome.

Thus, we estimate that the rate of the multiple classes of structural mutation combined is 0.07–0.12 per generation.

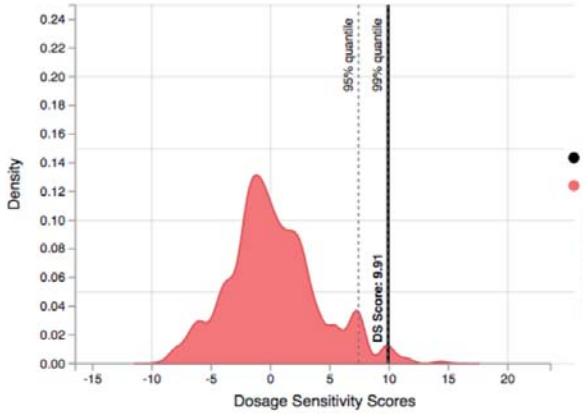
An integrated map of structural variation in 2,504 human genomes

A list of authors and their affiliations appears at the end of the paper.

We estimate that individuals harbour a median of 18.4 Mbp of SVs per diploid genome, an excess contributed to a large extent by mCNVs (11.3 Mbp) and biallelic deletions (5.6Mbp; Table 1).

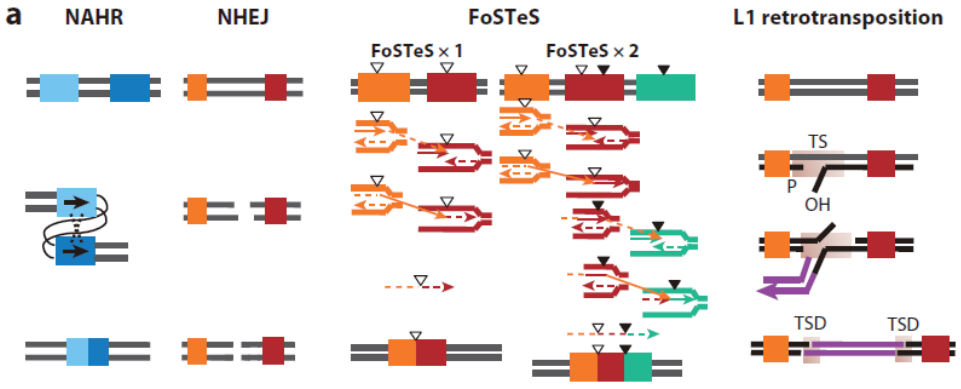
Dosage-sensitivity scores and sampling probabilities: We now automatically calculate dosage sensitivity (DS) scores for all CNVs deposited to DECIPHER. DS scores provide an estimate of the likely pathogenicity of the variant based on log-odds scores of predicted haploinsufficient genes within a copy-number variant. Higher scores indicate that the CNV is likely to be pathogenic while low or negative scores could be indicative of non-pathogenicity.

▫ **Dosage-sensitivity scores and sampling probabilities:** We now automatically calculate dosage sensitivity (DS) scores for all CNVs deposited to DECIPHER. DS scores provide an estimate of the likely pathogenicity of the variant based on log-odds scores of predicted haploinsufficient genes within a copy-number variant. Higher scores indicate that the CNV is likely to be pathogenic while low or negative scores could be indicative of non-pathogenicity.



- Patient: The dosage sensitivity score for the patient
- Control Population: Dosage sensitivity (DS) scores in a general population sample. The WTCCC dataset was used to calculate dosage sensitivity scores for all copy-number variants reported in the dataset and then plotting the highest DS scores for each sample for a deletion (red) and a duplication (blue).

Major mechanisms underlying human genomic rearrangements and CNV formation



b	NAHR	NHEJ	FoSTeS	Retrotransposition
Structural variation type	dup, del, inv	dup, del	dup, del, inv, complex	ins
Homology flanking breakpoint (before rearrangement)?	Yes (LCR/SD, *Alu*, L1, or pseudogene)	No	No	No
Breakpoint	Inside homology	Addition or deletion of basepairs, or microhomology	Microhomology	No specification
Sequence undergoing SV	Any	Any	Any	Transcribed sequences

Zhang et. al., 2009, Annu. Rev. Genomics Hum. Genet.10:451–81

The limitation of CNV detection using short-reads NGS

Table 1. The Landscape of Human Genome Variation

Variant Class		Subclass, Other Term(s)	Size	Num. / genome (Illumina WGS)	Arrays	Short-Read WGS	Long-Read WGS
Single Nucleotide Variation (SNV)		point mutation; substitution	1 bp	3.5×10^6	XX	XXX	XX
Small Insertion/Deletion Variation (indel)		insertion; deletion; complex indel	1-49 bp	4.5×10^5	XX	XX	X
Structural Variation (SV)	copy number variation (CNV)	deletion	≥50 bp	5,000	X	XX	XXX
		duplication (tandem, interspersed)		1,000	X	XX	XXX
		multi-allelic CNV; tandem repeats		450	X	XX	XX
	insertion	novel, templated or repeat insertion		1,500	-	X	XXX
	balanced rearrangement	inversion		40	-	XX	XXX
		reciprocal translocation	inter-chrom	0.001	-	XX	XXX
	complex genomic rearrangement	complex SV; chromothripsis	>1 mb	0.01	-	XX	XXX
	extremely large copy number variant	aneuploidy; chrom. abnormality	>1 mb	0.01	XXX	XXX	XXX
	retrogene insertion	retroduplication; retrocopy	gene coding length	10	-	XX	XXX
	mobile element insertion (MEI)	SINE; LINE; SVA	0.3-7 kb	2,000	-	X	XXX
Tandem Repeat Variation	short tandem repeat (STR)	microsatellite; simple sequence repeat	1-6 bp (repeat unit)	1×10^5	-	X	XXX
	variable number tandem repeat (VNTR)	minisatellite	7-49 bp (repeat unit)	unknown	-	X	XX
	centromeric & heterochromatic repeats	satellite DNA (α, β, 1-3)	various	unknown	-	-	XX

"–" indicates a near-complete inability to detect that variant class, XXX indicates that the technology is highly effective.

Lappalainen et al., 2019, Cell. 177:70-84

Table 1. Comparison of Genetic Methodologies

	Sanger sequencing	NGS		
		Targeted NGS	WES	WGS
Expected detection range	Small to medium indels (<300 bp) Introns Promoters	Small to medium indels (<300 bp) Introns Promoters	Small to medium indels (<300 bp) Introns Promoters	Small to large indels Introns Promoters
Candidates for application	Detection of known mutations and research into new causative genes	Analysis of known causative genes	Identification of new causative genes	Studies of population genomics
Advantages	Traditional and practiced methodology Cost-effective analysis for	Proper option in clinical application Customizable for different sample types Relatively low cost	Identification of SNVs, indels, SVs, and CNVs in coding regions Detection of mutations in unsuspected genes Moderate cost between targeted NGS and WGS	Identification of SNVs, indels, SVs, and CNVs in coding and non-coding regions PCR not required Obtaining complete genetic information
Limitations	Variable costs depending on gene size and exon number False negatives if the wrong area is analyzed False positive interpretation for benign polymorphisms	Restricted analysis range Biases of library preparation Heterogeneous coverage Difficult identification of CNVs	Heterogeneous coverage High risk of incidental genetic findings	High cost Greater effort in data interpretation

Endocrinol Metab. 2019 Mar;34(1):23-28.
https://doi.org/10.3803/EnM.2019.34.1.23
Copyright © 2019 Korean Endocrine Society
A in pseudogenes and GC-rich genes

NGS, next-generation sequencing; WES, whole exome sequencing; WGS, whole genome sequencing; indels, insertions or deletions; GC, guanine-cytosine; SNV, single nucleotide variant; SV, structural variant; CNV, copy number variant; PCR, polymerase chain reaction.

Endocrinol Metab. 2019 Mar;34(1):23-28.
https://doi.org/10.3803/EnM.2019.34.1.23
Copyright © 2019 Korean Endocrine Society

TABLE Close

Examples of Microdeletion Syndromes

Syndrome	Chromosomal Deletion	Description
Alagille syndrome	20p.12	Cholestasis, bile duct paucity, cardiac anomalies, pulmonary artery stenosis, butterfly vertebrae, posterior embryotoxon of the eye
Angelman syndrome	Maternal chromosome at 15q11	Seizures, puppet-like ataxia, frequent laughter, hand flapping, severe intellectual disability
DiGeorge syndrome (DiGeorge anomaly, velocardiofacial syndrome, pharyngeal pouch syndrome, thymic aplasia)	22q11.21	Hypoplasia or lack of thymus and parathyroids, cardiac anomalies, cleft palate, intellectual disability, psychiatric problems
Langer-Giedion syndrome (trichorhinophalangeal syndrome type II)	8q24.1	Exostosis, cone epiphyses, sparse hair, bulbous nose, hearing loss, intellectual disability
Miller-Dieker syndrome	17p13.3	Lissencephaly; short, upturned nose; severe growth retardation; seizures; severe intellectual disability
Prader-Willi syndrome	Paternal chromosome at 15q11	In infancy: Hypotonia, poor feeding, failure to thrive In childhood and adolescence: Obesity, hypogonadism, small hands and feet, intellectual disability, obsessive-compulsive behaviors

https://www.msdmanuals.com/professional/pediatrics/chromosome-and-gene-anomalies/microdeletion-and-microduplication-syndromes#

ISCA Database Search

The ISCA Database Search is currently utilizing the following databases: ISCA/NCBI: November 6, 2012, OMIM: December 6, 2012, UCSC: January 3, 2013. Please contact us at clingen@clinicalgenome.org if you have any questions.

Enter search parameters

Genome Build: GRCh37/hg19

Location/Gene: chr3:33183884-33900000 View in ...

e.g. chrX:70367400-70384667 or NLGN3 or Xq13.1

Type: Any

Gender: Any

Phenotype(s): Select or enter HPO term

Search ISCA Database | Summary Table of Genes

Download as Excel

Click on headers to sort

Variant ID	Location	Size (Kb)	Type	Phenotypes	Inheritance	Gender	Interpretation
nssv577899	chr3:32363874-36817097	4,453.22	loss	• Developmental Delay and additional significant developmental and morphological phenotypes referred for genetic testing			Pathogenic

•Last updated: January 3, 2013Database updates

- ISCA/NCBI: November 6, 2012
- OMIM: December 6, 2012
- UCSC: January 3, 2013

ISCA Consortium and Public Database

International Standards for Cytogenomic Arrays (ISCA) Consortium is a rapidly growing group of clinical cytogenetics and molecular genetics laboratories committed to improving quality of patient care related to clinical genetic testing using new molecular cytogenetic technologies including array comparative genomic hybridization (aCGH) and quantitative SNP analysis by microarrays or bead chip technology.

http://dbsearch.clinicalgenome.org/search/

http://www.iscaconsortium.org/learn-about-isca/

ClinVar and ClinGen

ClinVar partners with ClinGen to advance knowledge connecting human variation to human health. Although both support centralization of information to improve access, there are some differences worth noting. This page summarizes functions specific to ClinGen, with links to more information on ClinGen's web site

ClinVar is like PubMed for clinical variants

•It is a long term internally funded government archive.
•It collects data about genotype-phenotype relationships from many contributors.
•Each contribution is intellectually attributed to the author, not to ClinVar.
•ClinVar sets the editorial and structural standards for the data, but the intellectual content comes from the author.
•Authors in ClinVar may provide data that disagrees with each other.
•ClinVar contains contributions at many levels: basic reports, as well as authoritative reviews or guidelines. These different levels of report are explicitly flagged and clearly distinguished from each other.

ClinGen is like a scientific journal that publishes reviews

•It may have various funding sources and lifetimes.
•It reviews data about genotype-phenotype relationships, from ClinVar and other sources.
•It has an editorial board that reviews credentials of data sources.
•It publishes reports about specific genotype-disease association topics.
•Each report represents the intellectual effort of the authors and the "journal" staff.
•The report may be available directly from the website in various forms.
•The report will be deposited in ClinVar in a standard form for long-term archiving and for uniform access by users.

ClinGen
Clinical Genome Resource

Get Started | About Us | Curation Activities | Working Groups | Expert Panels | Documents & Annoucements | Tools

Explore the clinical relevance of genes & variants

ClinGen is a National Institutes of Health (NIH)-funded resource dedicated to building a central resource that defines the clinical relevance of genes and variants for use in precision medicine and research.

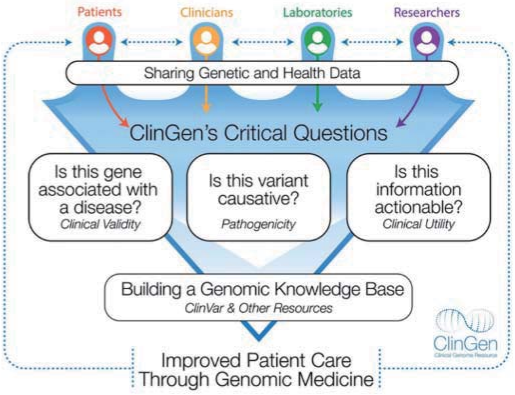
ClinGen - Clinical Genome Resource

While knowledge in the field of human genetics has greatly increased since the time of the Human Genome Project, we are still learning all of the ways in which changes in our DNA contribute to human health and disease. The Clinical Genome Resource, or ClinGen, is a National Institutes of Health funded initiative to increase the community's knowledge about the relationship between genes and health. We are dedicated to building a knowledge base that defines the clinical relevance of genes and variants for use in precision medicine and research. We do this by first encouraging the sharing of genetic and health data by our key stakeholder groups: patients, clinicians, laboratories, and researchers.

We then use this data to answer a number of key curation questions:

- Is this gene associated with a disease, and by which mechanisms do variation cause this disease?
- Is this variant causative?
- Will this information affect medical management?

Once we answer these questions via our various curation efforts, we make this information publicly available, building a genomic knowledge base with the goal of improving patient care through genomic medicine.



https://clinicalgenome.org/

ClinGen
Clinical Genome Resource

ClinGen Dosage Sensitivity Map

The Clinical Genome Resource (ClinGen) consortium is curating genes and regions of the genome to assess whether there is evidence to support that these genes/regions are dosage sensitive and should be targeted on a cytogenomic array.

All data are shown in GRCh37 and GRCh38 coordinates.

Search By Gene Name

Symbol:

Or click on the following examples: ZEB2, PTEN, MAPT

Go

Search By Location (GRCh37)

Location:

example: chr2:44,000,000-45,500,000, 2p21-2p16.2

Go

Links

- ClinGen Home Page
- Help with this site
- FAQ
- Contact Us
- Pathogenic CNV regions
- Curation of the ACMG 59 Genes
- FTP

Gene/Region Curation Stats

Review Complete	1,474
Under Primary Review	107
Under Secondary Review	21
Under Group Review	23
Awaiting Review	42,318

https://www.ncbi.nlm.nih.gov/projects/dbvar/clingen/index.shtml

Submitted location: chr3:33,183,884-33,900,000

"N/A" indicates that this gene has not yet been evaluated

Gene Issues (10) | Region Issues (0)

Location search results

Gene Symbol	Haploinsufficiency
CRTAP	30: Gene associa autosomal recess
SUSD5	N/A
FBXL2	N/A
UBP1	N/A
RNU7-110P	N/A
RNA5SP128	N/A
CLASP2	N/A
COX6CP10	N/A
SDAD1P3	N/A

CRTAP

Curation Status: Complete

id: ISCA-4683

Date last evaluated: 2016-08-22

Issue Type: ClinGen Gene Curation

Gene type: protein-coding

Entrez Gene: https://www.ncbi.nlm.nih.gov/gene/10491

OMIM: https://omim.org/entry/610854

ClinGen Haploinsufficiency Score: Gene associated with autosomal recessive phenotype

ClinGen Triplosensitivity Score: 0

3p26 3p24.1 3p2

Location Informati

3p22.3

GRCh37/hg19 chr3

View: NCBI | Ensem

GRCh38/hg38 chr3

View: NCBI | Ensem

Genome View | Evidence for Haploinsufficiency Phenotypes | Evidence for Triplosensitive

Haploinsufficiency score: Gene associated with autosomal recessive phenotype
Strength of Evidence (disclaimer): Gene associated with autosomal recessive phenotype
Haploinsufficiency Phenotype: OSTEOGENESIS IMPERFECTA, TYPE VII; OI7

CRTAP

Curation Status: Complete

Gene Information

- id: ISCA-4683
- Date last evaluated: 2016-08-22
- Issue Type: ClinGen Gene Curation
- Gene type: protein-coding
- Entrez Gene: https://www.ncbi.nlm.nih.gov/gene/10491
- OMIM: https://omim.org/entry/610854

Location Information

- 3p22.3
- GRCh37/hg19 chr3: 33,155,450-33,189,265

Evidence for Loss Phenotypes

- Loss of function score: Gene associated with autosomal recessive phenotype
- Strength of Evidence: Gene associated with autosomal recessive phenotype
- Haploinsufficiency phenotype: OSTEOGENESIS IMPERFECTA, TYPE VII; OI7

Evidence for Triplosenstive Phenotype

- Triplosensitivity score: 0
- Strength of Evidence: No evidence for dosage pathogenicity

Print Full Report

Genome View | Evidence for Haploinsufficiency Phenotypes | Evidence for Triplosensitive Phenotypes

Haploinsufficiency score: 3
Strength of Evidence (disclaimer): Sufficient evidence for dosage pathogenicity
Haploinsufficiency Phenotype: OSTEOGENESIS IMPERFECTA, TYPE I; OI1

Evidence for haploinsufficiency phenotype

PubMed ID	Description
9067755	Korkko et al. (1997) analyzed COL1A1 in 8 patients with OI type I and found several mutations, including a ARG963TER mutation in 2 unrelated affected individuals and a 1 bp deletion resulting in a premature stop codon.
9016532	Dalgleish (1997) describes a mutation database for COL1A1 and COL1A2, including 6 frameshift mutations in COL1A1 and 7 deletions in COL1A1.

Haploinsufficiency phenotype comments:

Mutations are associated with Osteogenesis imperfecta I-IV. There is a wide phenotypic spectrum.

Copy number variation data analysis

command

【Step1】sample1 bin coverage ~10min

cd ~/CNVKit

qsub CNVKit_sample1_target_antitarget.sh

qstat

command

- /share/app/Python-3.5.7_INSTALL/bin/cnvkit.py△coverage△./CNVKit/19030164_F0x100.bam△./CNVKit/hg19_20180808_CDS_wa_UTR_noNR_and_MedExome.merge_432.split.bed△-o△./CNVKit/19030164_F0x100_432.targetcoverage.cnn△>△./CNVKit/19030164_432_target.log△2>&1
- /share/app/Python-3.5.7_INSTALL/bin/cnvkit.py△coverage△./CNVKit/19030164_F0x100.bam△./CNVKit/hg19_20180808_CDS_wa_UTR_noNR_and_MedExome.merge_432.split.antitarget.bed△-o△./CNVKit/19030164_F0x100_432.antitargetcoverage.cnn△>△./CNVKit/19030164_432_antitargettarget.log△2>&1

【Step2】sample2 bin coverage ~10min

qsub CNVKit_sample2_target_antitarget.sh

qstat

command

- /share/app/Python-3.5.7_INSTALL/bin/cnvkit.py△coverage△./CNVKit/19030165_F0x100.bam△./CNVKit/hg19_20180808_CDS_wa_UTR_noNR_and_MedExome.merge_432.split.bed△-o△./CNVKit/19030165_F0x100_432.targetcoverage.cnn△>△./CNVKit/19030165_432_target.log△2>&1
- /share/app/Python-3.5.7_INSTALL/bin/cnvkit.py△coverage△./CNVKit/19030165_F0x100.bam△./CNVKit/hg19_20180808_CDS_wa_UTR_noNR_and_MedExome.merge_432.split.antitarget.bed△-o△./CNVKit/19030165_F0x100_432.antitargetcoverage.cnn△>△./CNVKit/19030165_432_antitargettarget.log△2>&1

【Step3】build reference ~8min

qsub CNVKit_build_ref_432.sh

qstat

command

- /share/app/Python-3.5.7_INSTALL/bin/cnvkit.py△reference△./CNVKit/ref_432/*.cnn△-f△/binfo/data3/home/CNVKit_training/reference/hg19.fasta△-o△./CNVKit/CNVKit_9sample_432.cnn△>△./CNVKit/CNVKit_ref_432.log△2>&1

【Step4】fix ~2min

qsub CNVKit_fix_432.sh

qstat

command

- /share/app/Python-3.5.7_INSTALL/bin/cnvkit.py△fix△./CNVKit/19030164_F0x100_432.targetcoverage.cnn△./CNVKit/19030164_F0x100_432.antitargetcoverage.cnn△./CNVKit/CNVKit_9sample_432.cnn△-o△./CNVKit/19030164_F0x100_CNVKit_432_fixRef.cnr△>△./CNVKit/19030164_F0x100_CNVKit_432.fix.log△2>&1

- /share/app/Python-3.5.7_INSTALL/bin/cnvkit.py△fix△./CNVKit/19030165_F0x100_432.targetcoverage.cnn△./CNVKit/19030165_F0x100_432.antitargetcoverage.cnn△./CNVKit/CNVKit_9sample_432.cnn△-o△./CNVKit/19030165_F0x100_CNVKit_432_fixRef.cnr△>△./CNVKit/19030165_F0x100_CNVKit_432.fix.log△2>&1

【Step5】Segment ~2min

qsub CNVKit_segment_432.sh

qstat

command

- /share/app/Python-3.5.7_INSTALL/bin/cnvkit.py△segment△-m△cbs△-o△./CNVKit/19030164_F0x100_CNVKit_432_fixRef_Seg_cbs.cns△./CNVKit/19030164_F0x100_CNVKit_432_fixRef.cnr△>△./CNVKit/19030164_F0x100_CNVKit_432_fixRef_Seg_cbs.log△2>&1

- /share/app/Python-3.5.7_INSTALL/bin/cnvkit.py△segment△-m△cbs△-o△./CNVKit/19030165_F0x100_CNVKit_432_fixRef_Seg_cbs.cns△./CNVKit/19030165_F0x100_CNVKit_432_fixRef.cnr△>△./CNVKit/19030165_F0x100_CNVKit_432_fixRef_Seg_cbs.log△2>&1

【Step6】scatter ~2min

qsub CNVKit_scatter.sh

qstat

command

- /share/app/Python-3.5.7_INSTALL/bin/cnvkit.py△scatter△./CNVKit/19030164_F0x100_CNVKit_432_fixRef.cnr△-s△./CNVKit/19030164_F0x100_CNVKit_432_fixRef_Seg_cbs.cns△--title△19030164_F0x100_432_cbs△--y-max△2△--y-min△-2△-o△./CNVKit/19030164_F0x100_CNVKit_432_fixRef_Seg_cbs.pdf

- /share/app/Python-3.5.7_INSTALL/bin/cnvkit.py△scatter△./CNVKit/19030164_F0x100_CNVKit_432_fixRef.cnr△-s△./CNVKit/19030164_F0x100_CNVKit_432_fixRef_Seg_cbs.cns△--title△19030164_F0x100_432_cbs_chr22△-c△chr22△--y-max△2△--y-min△-2△-o△./CNVKit/19030164_F0x100_CNVKit_432_fixRef_Seg_cbs_chr22.pdf

■ /share/app/Python-3.5.7_INSTALL/bin/cnvkit.py△scatter ./CNVKit/19030165_F0x100_CNVKit_432_fixRef.cnr△-s△./CNVKit/19030165_F0x100_CNVKit_432_fixRef_Seg_cbs.cns△--title△19030165_F0x100_432_cbs△--y-max△2△--y-min△-2△-o△./CNVKit/19030165_F0x100_CNVKit_432_fixRef_Seg_cbs.pdf

■ /share/app/Python-3.5.7_INSTALL/bin/cnvkit.py△scatter ./CNVKit/19030165_F0x100_CNVKit_432_fixRef.cnr△-s△./CNVKit/19030165_F0x100_CNVKit_432_fixRef_Seg_cbs.cns△--title△19030165_F0x100_432_cbs_chr22△-c△chr22△--y-max△2△--y-min△-2△-o△./CNVKit/19030165_F0x100_CNVKit_432_fixRef_Seg_cbs_chr22.pdf

【Step7】diagram ~1min

qsub CNVKit_diagram.sh

qstat

command

■ /share/app/Python-3.5.7_INSTALL/bin/cnvkit.py△diagram△-s△./CNVKit/19030164_F0x100_CNVKit_432_fixRef_Seg_cbs.cns△--title△19030164_F0x100_432_cbs△-o△./CNVKit/19030164_F0x100_CNVKit_432_fixRef_Seg_cbs_diagram.pdf

■ /share/app/Python-3.5.7_INSTALL/bin/cnvkit.py△diagram△-s△./CNVKit/19030165_F0x100_CNVKit_432_fixRef_Seg_cbs.cns△--title△19030165_F0x100_432_cbs△-o△./CNVKit/19030165_F0x100_CNVKit_432_fixRef_Seg_cbs_diagram.pdf