DETECTING CNV BY EXOME SEQUENCING

Fah Sathirapongsasuti

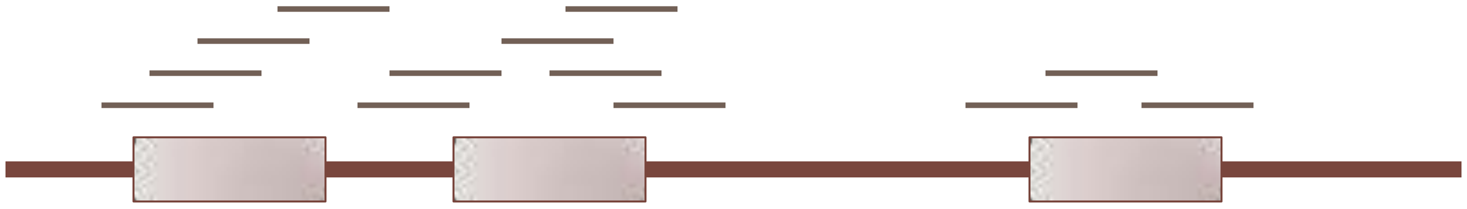
Biostatistics, HSPH

Exome Sequencing

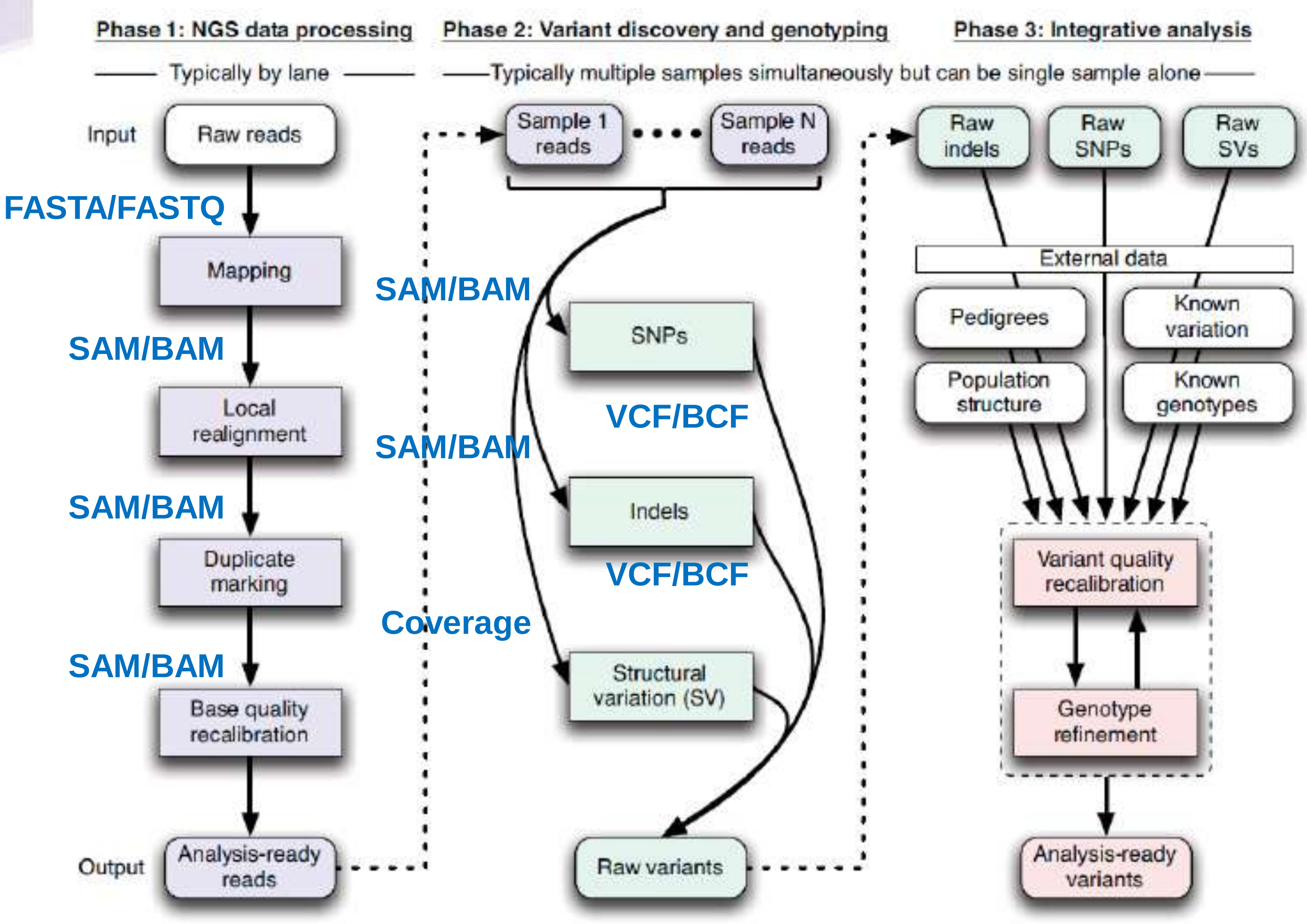
- Capturing protein coding portion of the genome

- ~85% of the disease-causing mutations occur in protein coding regions (exome)
- Exome constitutes 1% of the genome
- About 160,000-180,000 exons

- Time-saving and cost-effective

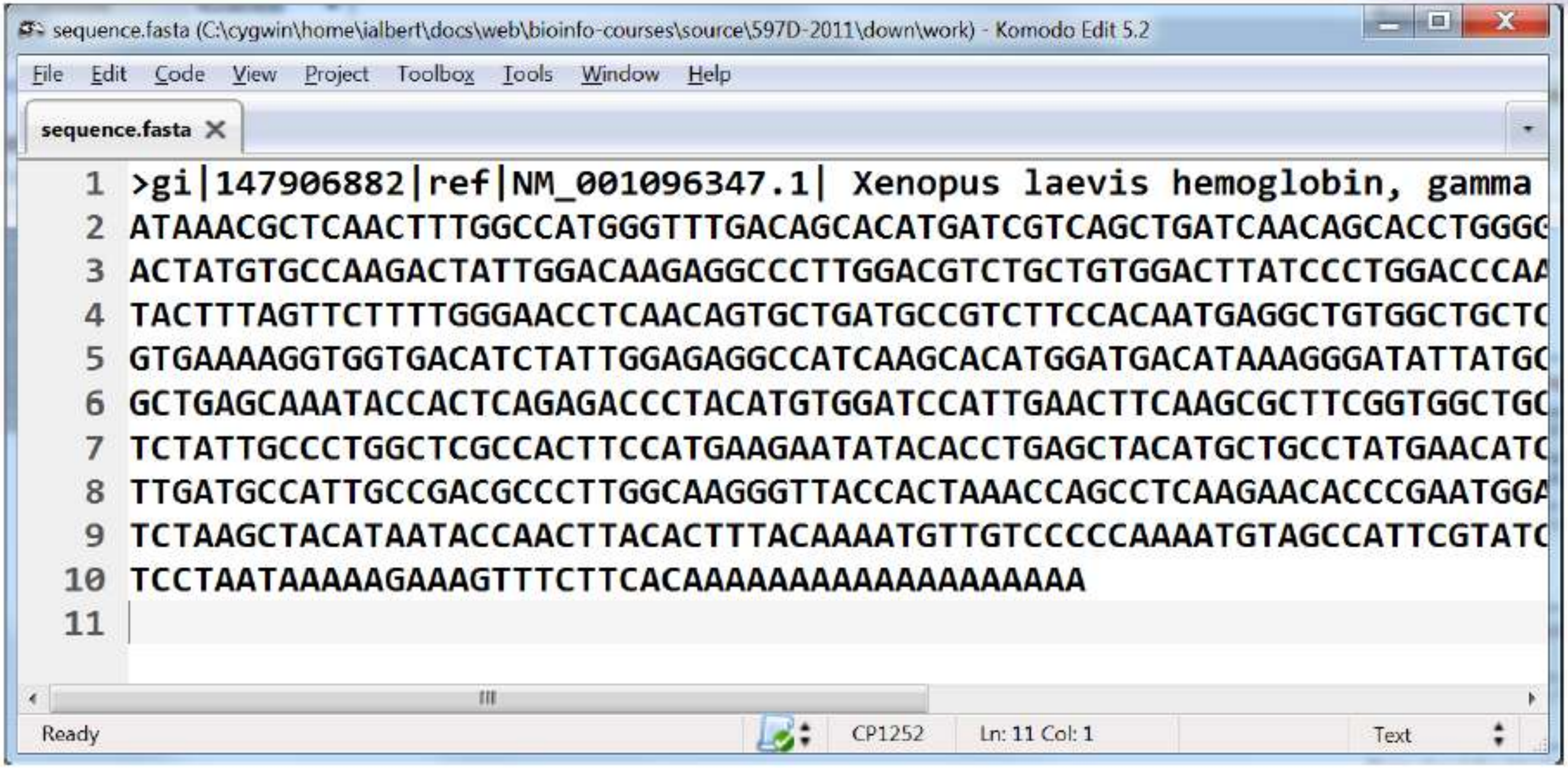


General Workflow



Source:Nature Genetics 43, 491–498 (2011)

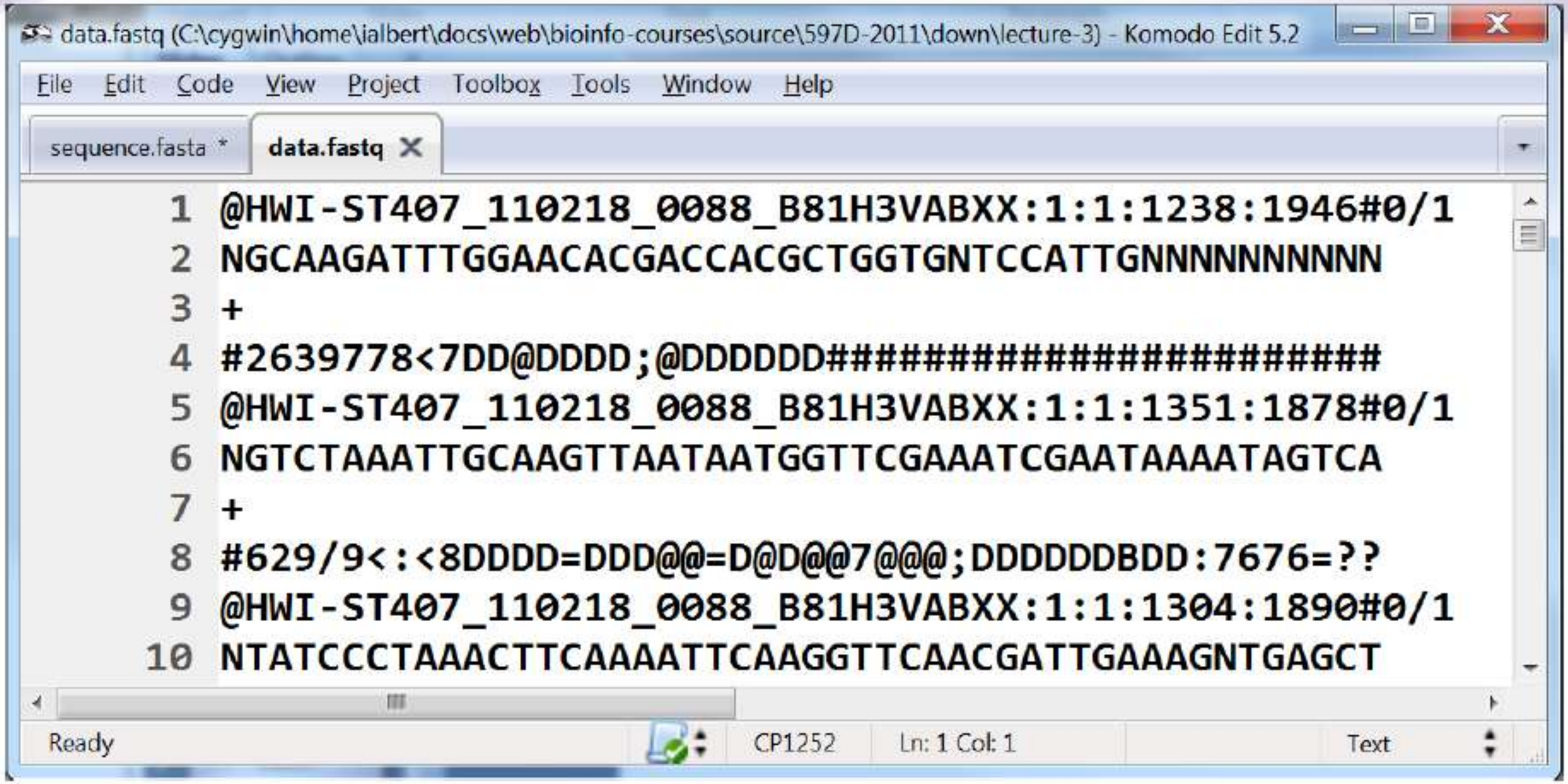
Fasta format



sequence.fasta (C:\cygwin\home\ialbert\docs\web\bioinfo-courses\source\597D-2011\down\work) - Komodo Edit 5.2
File
Edit
Code
View
Project
Toolbox
Tools
Window
Help
sequence.fasta
Ready
CP1252
Ln: 11 Col: 1
Text

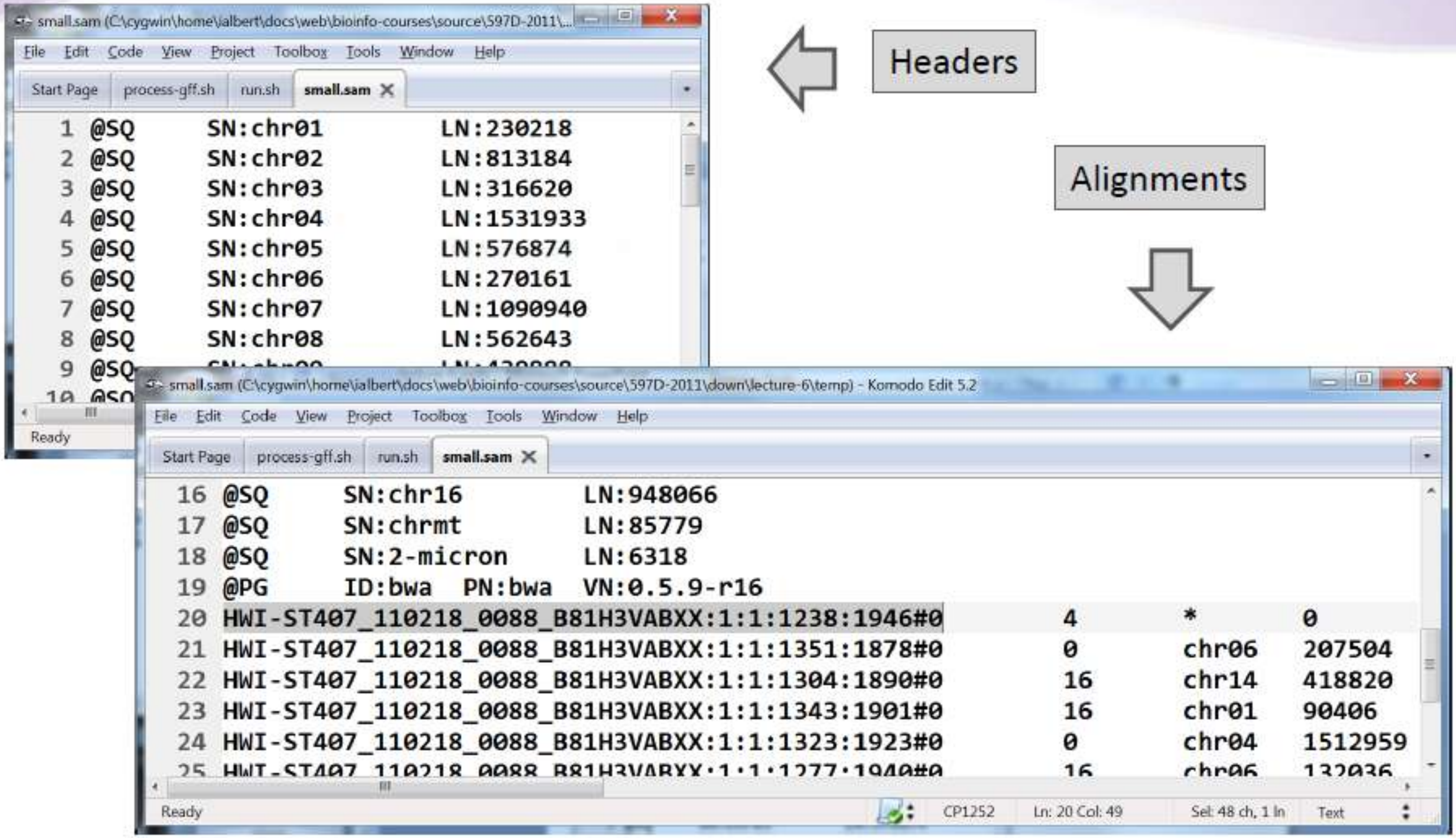
```
>gi|147906882|ref|NM_001096347.1| Xenopus laevis hemoglobin, gamma
ATAAACGCTCAACTTTGGCCATGGGTTTGACAGCACATGATCGTCAGCTGATCAACAGCACCTGGGG
ACTATGTGCCAAGACTATTGGACAAGAGGCCCTTGGACGTCTGCTGTGGACTTATCCCTGGACCCAA
TACTTTAGTTCTTTTGGGAACCTCAACAGTGCTGATGCCGTCTTCCACAATGAGGCTGTGGCTGCTC
GTGAAAAGGTGGTGACATCTATTGGAGAGGCCATCAAGCACATGGATGACATAAAGGGATATTATGC
GCTGAGCAAATACCACTCAGAGACCCTACATGTGGATCCATTGAACTTCAAGCGCTTCGGTGGCTGC
TCTATTGCCCTGGCTCGCCACTTCCATGAAGAATATACACCTGAGCTACATGCTGCCTATGAACATC
TTGATGCCATTGCCGACGCCCTTGGCAAGGGTTACCACTAAACCAGCCTCAAGAACACCCGAATGGA
TCTAAGCTACATAATACCAACTTACACTTTACAAAATGTTGTCCCCCAAAATGTAGCCATTCGTATC
TCCTAATAAAAAGAAAGTTTCTTCACAAAAAAAAAAAAAAAAAAA

```

Fastq format


data.fastq (C:\cygwin\home\ialbert\docs\web\bioinfo-courses\source\597D-2011\down\lecture-3) - Komodo Edit 5.2
File Edit Code View Project Toolbox Tools Window Help
sequence.fasta *
data.fastq
Ready
CP1252
Ln: 1 Col: 1
Text

```
@HWI-ST407_110218_0088_B81H3VABXX:1:1:1238:1946#0/1
NGCAAGATTTGGAACACGACCACGCTGGTGNTCCATTGNNNNNNNNNN
+
#2639778<7DD@DDDD;@DDDDDD#######################
@HWI-ST407_110218_0088_B81H3VABXX:1:1:1351:1878#0/1
NGTCTAAATTGCAAGTTAATAATGGTTCGAAATCGAATAAAATAGTCA
+
#629/9<:<8DDDD=DDD@@=D@D@@7@@@;DDDDDDBDD:7676=??
@HWI-ST407_110218_0088_B81H3VABXX:1:1:1304:1890#0/1
NTATCCCTAAACTTCAAAATTCAAGGTTCAACGATTGAAAGNTGAGCT
```

The structure of the SAM file


small.sam (C:\cygwin\home\ialbert\docs\web\bioinfo-courses\source\597D-2011\...
File Edit Code View Project Toolbox Tools Window Help
Start Page process-gff.sh run.sh small.sam
Ready
small.sam (C:\cygwin\home\ialbert\docs\web\bioinfo-courses\source\597D-2011\down\lecture-6\temp) - Komodo Edit 5.2
File Edit Code View Project Toolbox Tools Window Help
Start Page process-gff.sh run.sh small.sam
Ready CP1252 Ln: 20 Col: 49 Sel: 48 ch, 1 ln Text

Headers

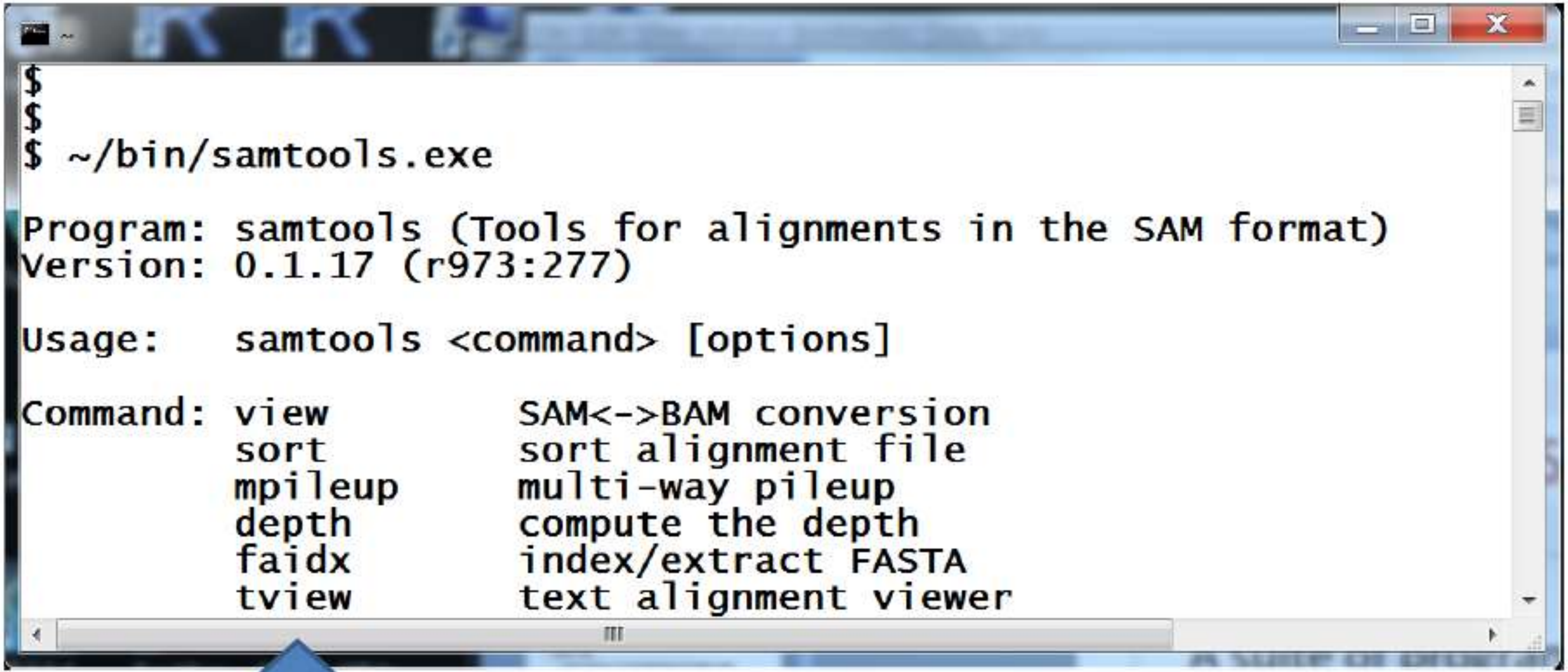
Alignments

```
@SQ     SN:chr01        LN:230218
@SQ     SN:chr02        LN:813184
@SQ     SN:chr03        LN:316620
@SQ     SN:chr04        LN:1531933
@SQ     SN:chr05        LN:576874
@SQ     SN:chr06        LN:270161
@SQ     SN:chr07        LN:1090940
@SQ     SN:chr08        LN:562643
@SQ     SN:chr09        LN:439888
@SQ
```

```
@SQ     SN:chr16        LN:948066
@SQ     SN:chrmt        LN:85779
@SQ     SN:2-micron     LN:6318
@PG     ID:bwa  PN:bwa  VN:0.5.9-r16
HWI-ST407_110218_0088_B81H3VABXX:1:1:1238:1946#0        4       *       0
HWI-ST407_110218_0088_B81H3VABXX:1:1:1351:1878#0        0       chr06   207504
HWI-ST407_110218_0088_B81H3VABXX:1:1:1304:1890#0        16      chr14   418820
HWI-ST407_110218_0088_B81H3VABXX:1:1:1343:1901#0        16      chr01   90406
HWI-ST407_110218_0088_B81H3VABXX:1:1:1323:1923#0        0       chr04   1512959
HWI-ST407_110218_0088_B81H3VABXX:1:1:1277:1940#0        16      chr06   132036
```

SAMtools

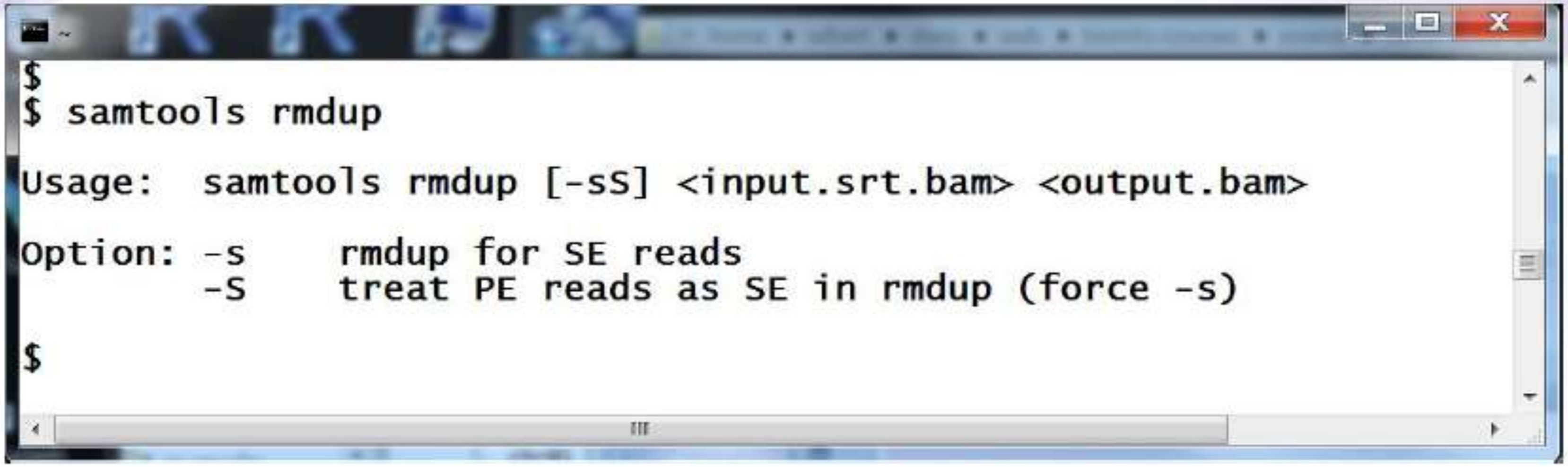
A suite of programs to manipulate and process SAM files



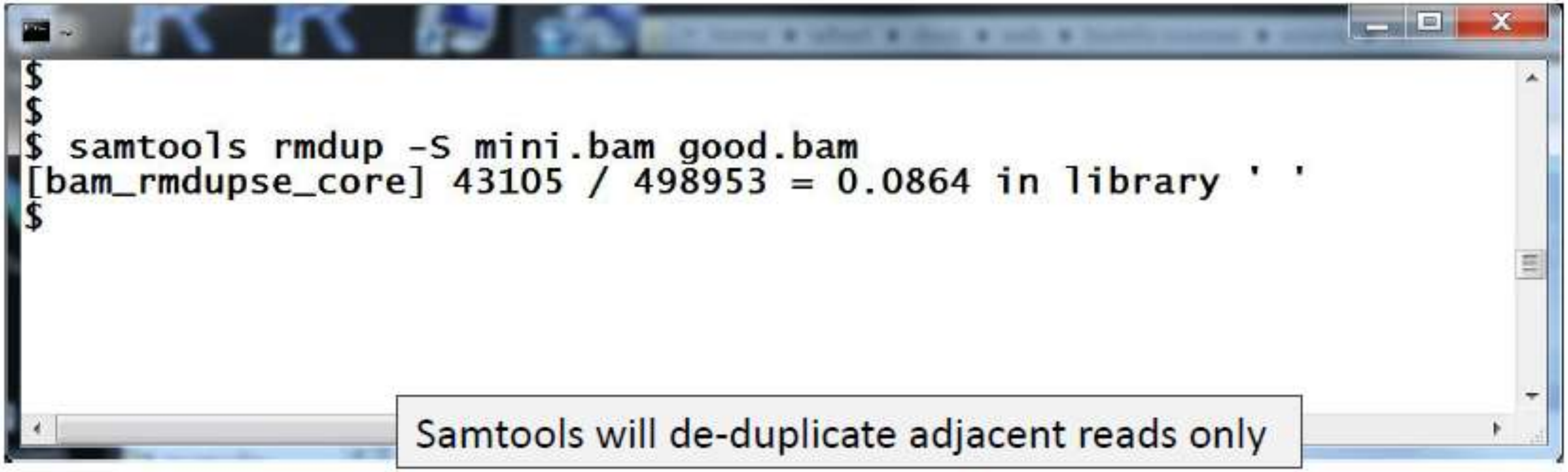
```
$
$
$ ~/bin/samtools.exe

Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.17 (r973:277)

Usage:   samtools <command> [options]

Command: view        SAM<->BAM conversion
         sort        sort alignment file
         mpileup     multi-way pileup
         depth       compute the depth
         faidx       index/extract FASTA
         tview       text alignment viewer
```

available actions with samtools

De-duplication with samtools



```
$
$ samtools rmdup

Usage:  samtools rmdup [-sS] <input.srt.bam> <output.bam>

Option: -s    rmdup for SE reads
        -S    treat PE reads as SE in rmdup (force -s)

$
```



```
$
$
$ samtools rmdup -S mini.bam good.bam
[bam_rmdupse_core] 43105 / 498953 = 0.0864 in library ' '
$
```

Samtools will de-duplicate adjacent reads only

Pileup

- Standard format for mapped data, position summaries

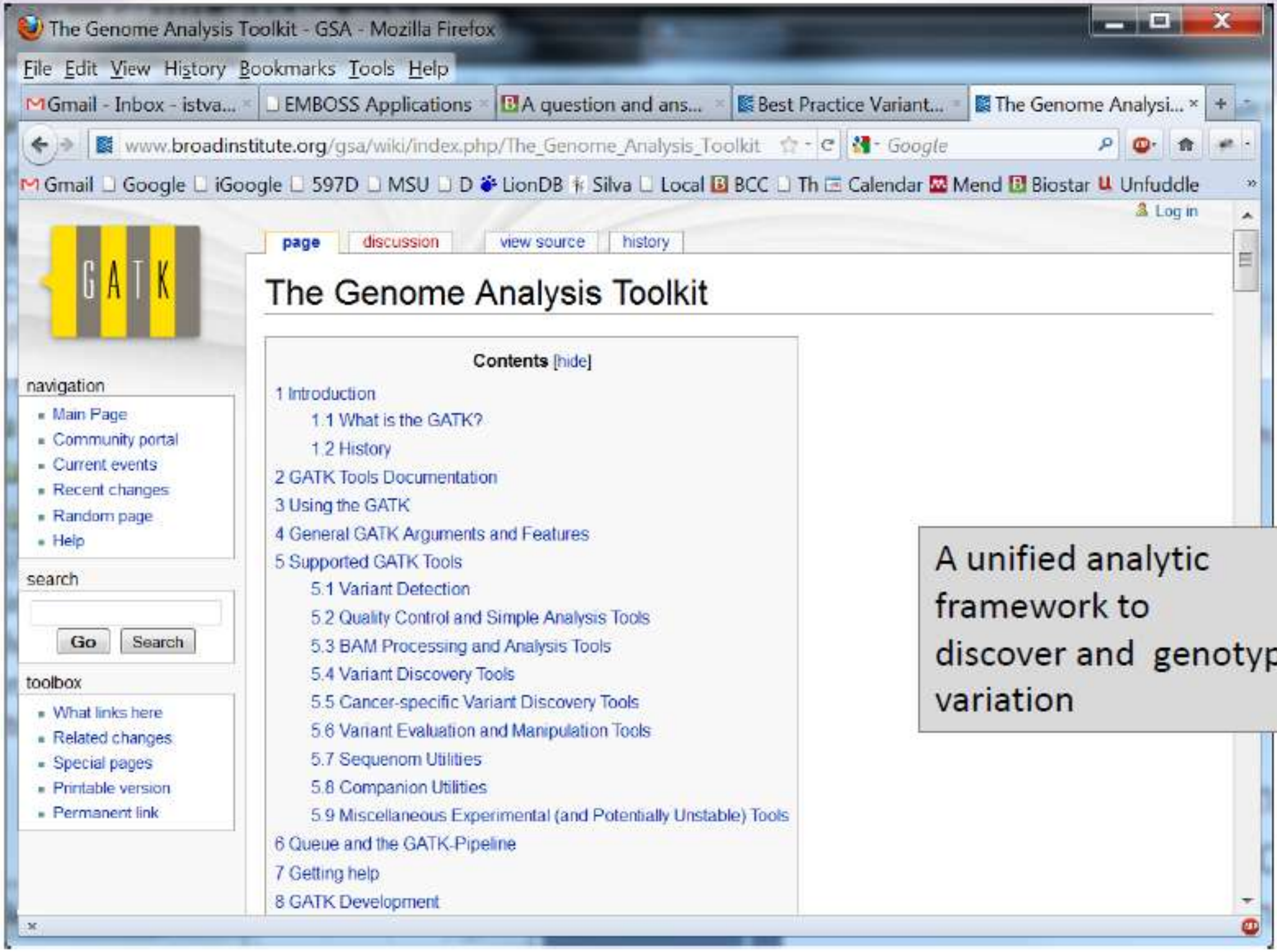
```
seq1 272 T 24  ,.$.....,,.,.,...,,,.,..^+. <<<+;<<<<<<<<<<<=<;<;7<&
seq1 273 T 23  ,.....,,.,.,...,,,.,..A     <<<;<<<<<<<<<3<=<<<;<<+
seq1 274 T 23  ,.$....,,.,.,...,,,.,...    7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23  ,$....,,.,.,...,,,.,...^l.  <+;9*<<<<<<<<<=<<:;<<<<
seq1 276 G 22  ...T,,.,.,...,,,.,....      33;+<<7=7<<7<&<<1;<<6<
seq1 277 T 22  ....,,.,.,.C.,,,.,..G.      +7<;<<<<<<<&<=<<:;<<&<
seq1 278 G 23  ....,,.,.,...,,,.,....^k.   %38*<<;<7<<7<=<<<;<<<<<
seq1 279 C 23  A..T,,.,.,...,,,.,.....     ;75&<<<<<<<<<=<<<9<<:<<
```

Seq.

Pos. Len. Alignment Quality

Ref.

Genome Analysis Toolkit



The Genome Analysis Toolkit

Contents [hide]

1 Introduction
- 1.1 What is the GATK?
- 1.2 History

2 GATK Tools Documentation

3 Using the GATK

4 General GATK Arguments and Features

5 Supported GATK Tools
- 5.1 Variant Detection
- 5.2 Quality Control and Simple Analysis Tools
- 5.3 BAM Processing and Analysis Tools
- 5.4 Variant Discovery Tools
- 5.5 Cancer-specific Variant Discovery Tools
- 5.6 Variant Evaluation and Manipulation Tools
- 5.7 Sequenom Utilities
- 5.8 Companion Utilities
- 5.9 Miscellaneous Experimental (and Potentially Unstable) Tools

6 Queue and the GATK-Pipeline

7 Getting help

8 GATK Development

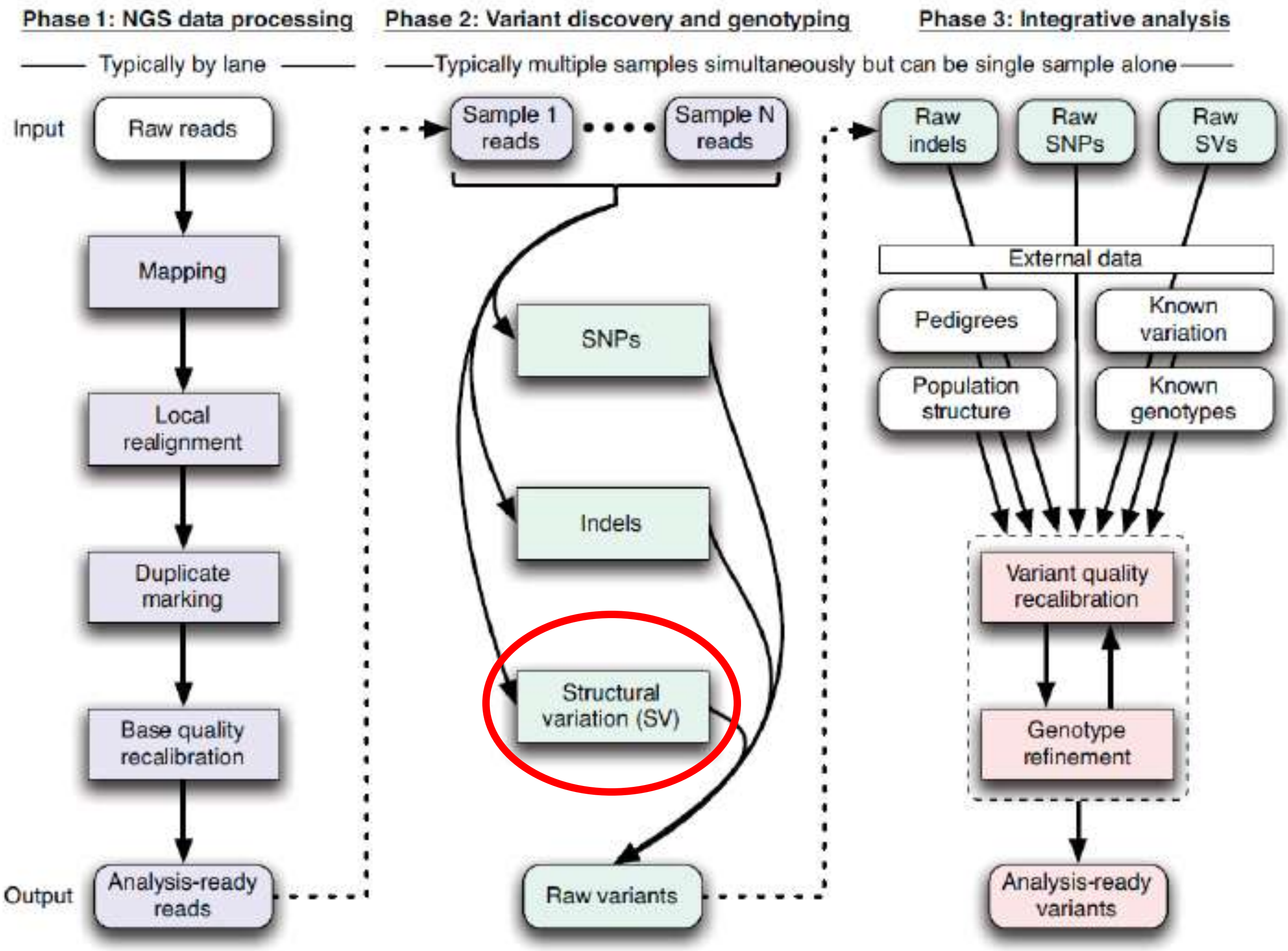
A unified analytic framework to discover and genotype variation

Variant Call Format

```
##format=PCFv1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
#CHROM  POS     ID        REF  ALT    QUAL FILTER  INFO                                  FORMAT       NA00001         NA00002
20      14370   rs6054257 G    A      29   0       NS=58;DP=258;AF=0.786;DB;H2           GT:GQ:DP:HQ  0|0:48:1:51,51  1|0:48:8:51,51
20      13330   .         T    A      3    q10     NS=55;DP=202;AF=0.024                 GT:GQ:DP:HQ  0|0:49:3:58,50  0|1:3:5:65,3
20      1110696 rs6040355 A    G,T    67   0       NS=55;DP=276;AF=0.421,0.579;AA=T;DB   GT:GQ:DP:HQ  1|2:21:6:23,27  2|1:2:0:18,2
20      10237   .         T    .      47   0       NS=57;DP=257;AA=T                     GT:GQ:DP:HQ  0|0:54:7:56,60  0|0:48:4:51,51
20      123456  microsat1 G    D4,IGA 50   0       NS=55;DP=250;AA=G                     GT:GQ:DP     0/1:35:4        0/2:17:2
```

```
##format=PCFv1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
#CHROM  POS     ID        REF  ALT    QUAL  FILTER  INFO
20      14370   rs6054257 G    A      29    0       NS=58;DP=258;AF=0.786;DB;H2
FORMAT       NA00001         NA00002
GT:GQ:DP:HQ  0|0:48:1:51,51  1|0:48:8:51,51
```

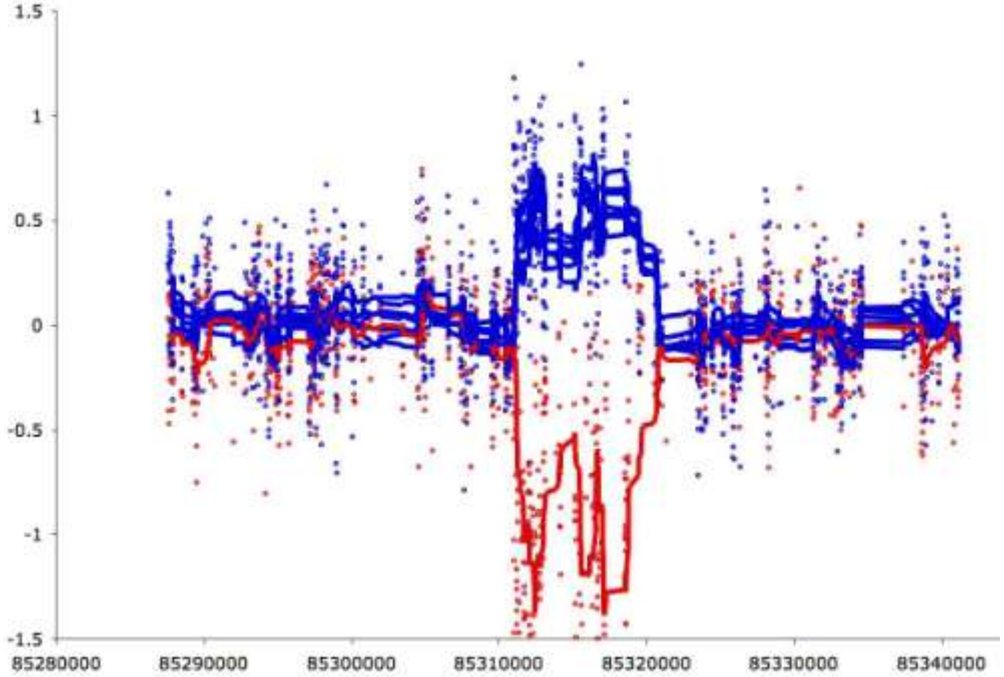
General Workflow



Source:Nature Genetics 43, 491–498 (2011)

Copy-Number Variation/Alteration

- CNV



Comparative **G**enomic **H**ybridisation

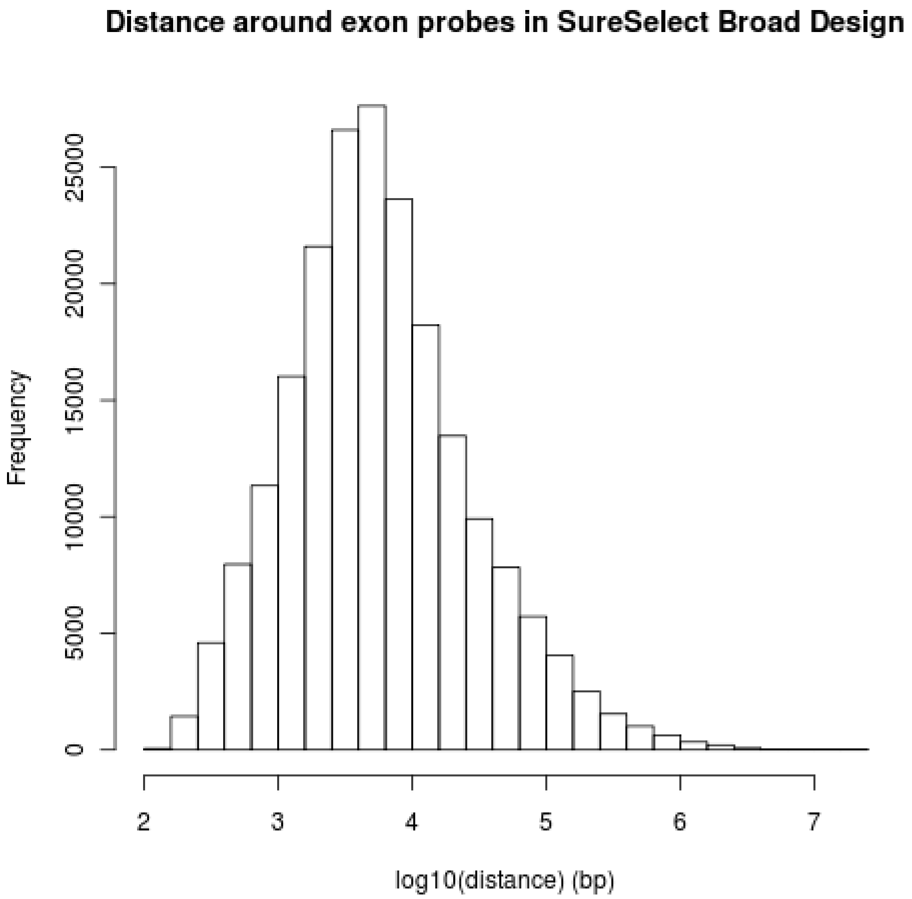
Blue lines: individuals with two copies.
Red line: individual with zero copy.

- gains and losses of chunks of DNA sequences
- Sizes:
  - 1kb-5Mb (Sanger's CNV Project)
  - Generally large chunks …
- Small gains/losses are called insertion/deletion (in-del)

CNV method specific for Exome Seq is needed

- All techniques were developed for whole genome sequencing or targeted sequencing of one continuous region.
- Two approaches:
  - Paired-End Methods (use insert size)
  - Depth of Coverage
- Challenges of Exome Sequencing:
  - **Discontinuous search space**
    - Paired-end methods won’t work
    - The only natural way to discretize the data is by exon
    - Resolution is limited by distance between exons
  - **Non-uniform distribution of reads**
    - Exon capture probes have different efficiency

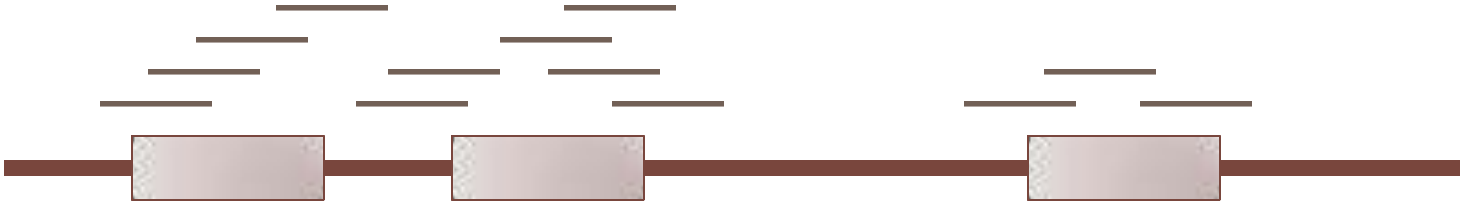
CNV Resolution is limited by exome probe design



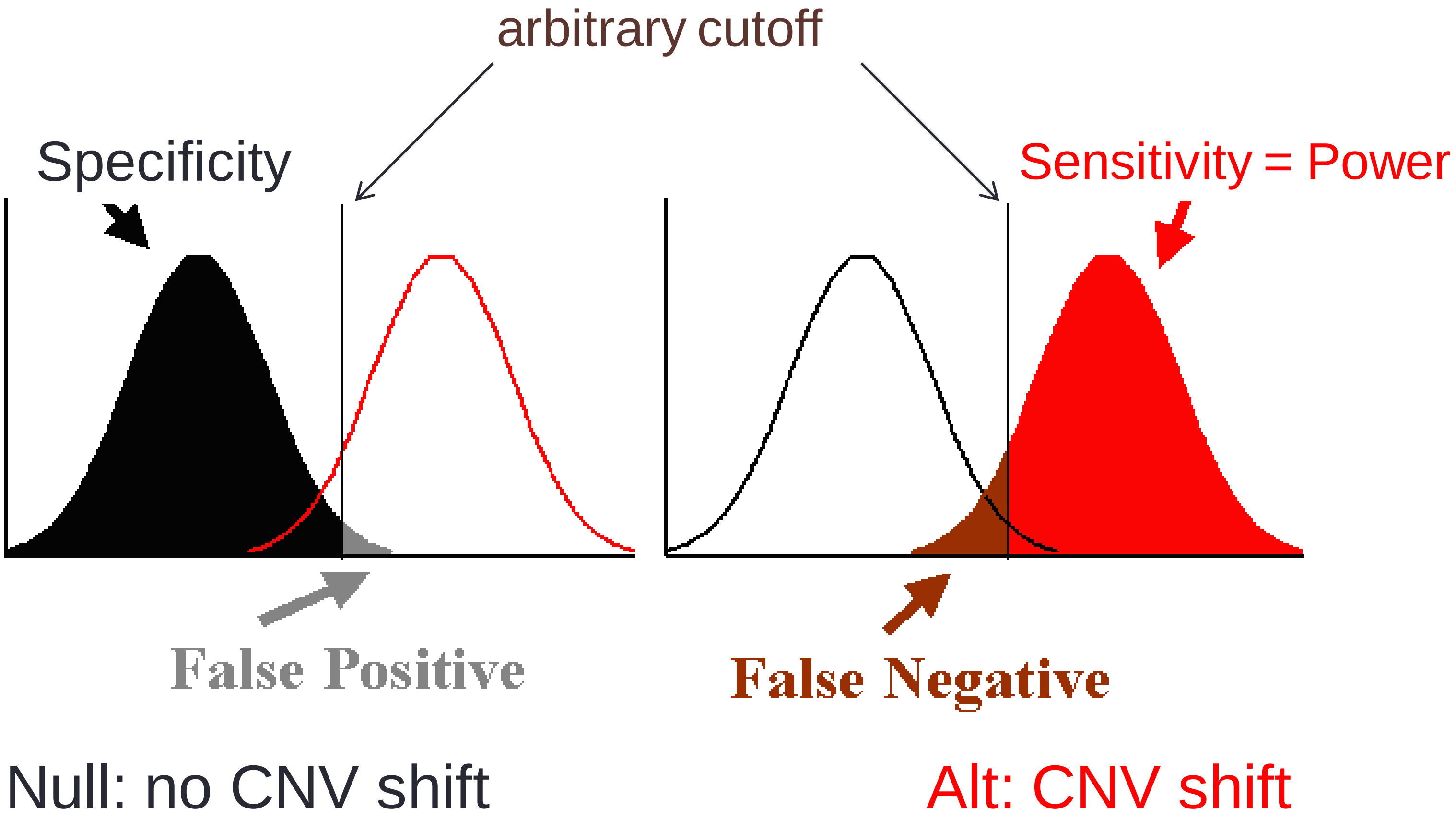
Min	1st Qu	Med	Mean	3rd Qu	Max
123	1,999	4,981	29,210	14,030	20,900,000

Depth of Coverage Approach

- Treat one exon as a unit (variable length)
- Measure depth of coverage (average coverage) per exon
- Key assumptions:
  - Number of reads over exons of certain size follows Poisson distribution
  - Average coverage is directly proportional to the number of reads; i.e.
    *average coverage = #reads * read length / exon length*



Using the ratio of depth-of-coverage to detect CNV



arbitrary cutoff

Specificity

Sensitivity = Power

False Positive

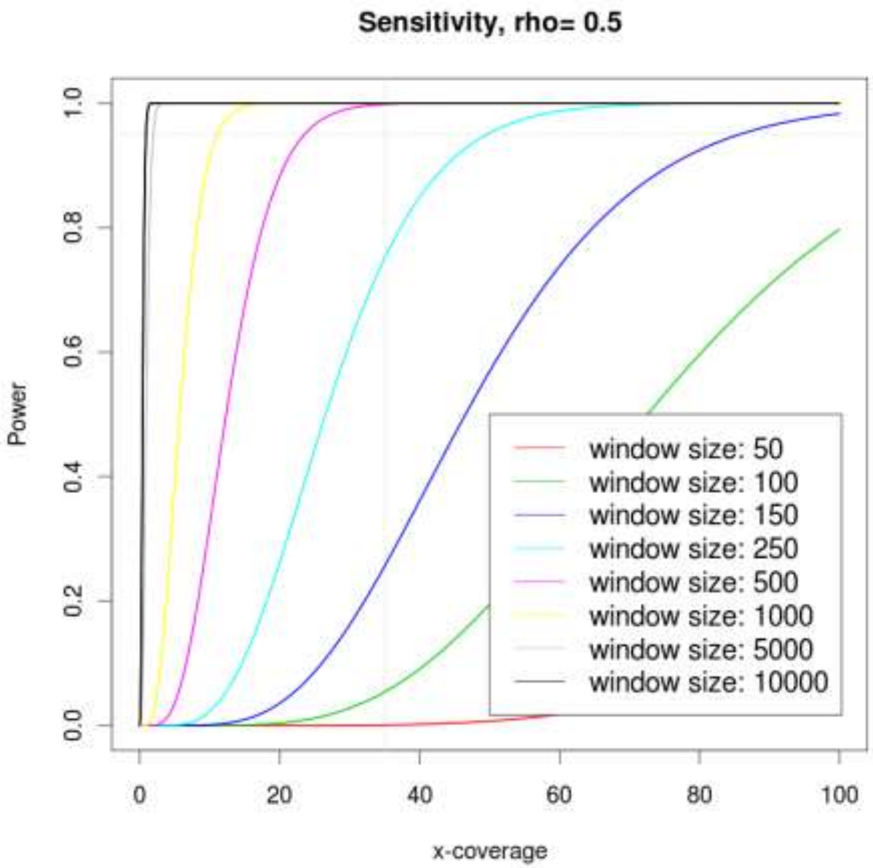
False Negative

Null: no CNV shift

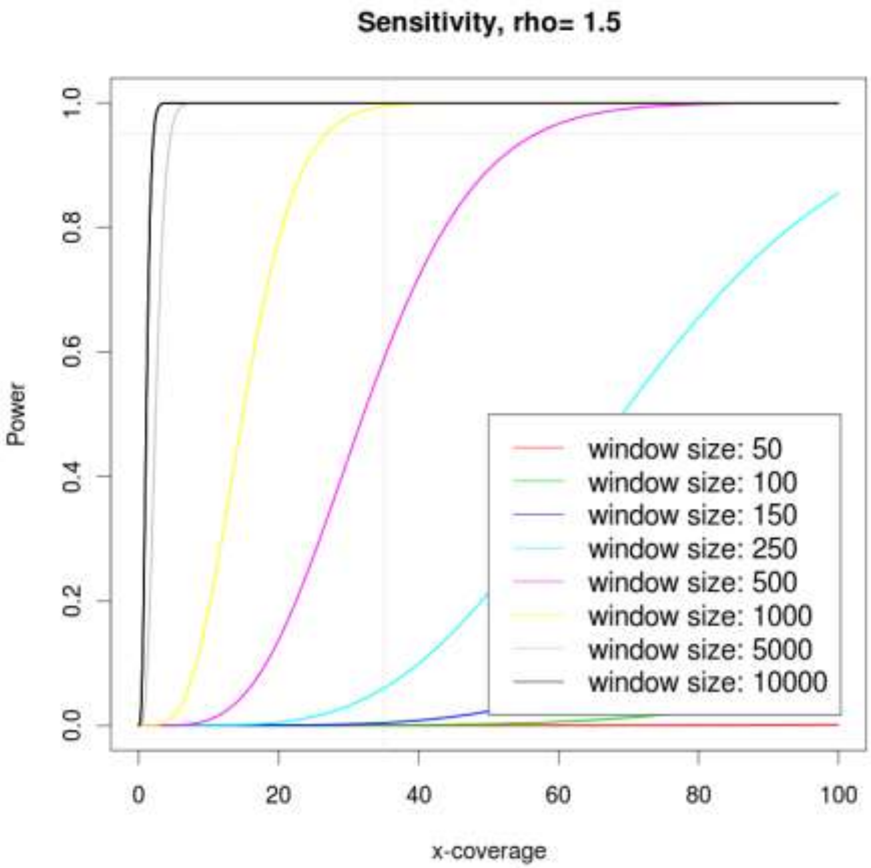
Alt: CNV shift

Power to detect CNV depends on depth-of-coverage

Deletion



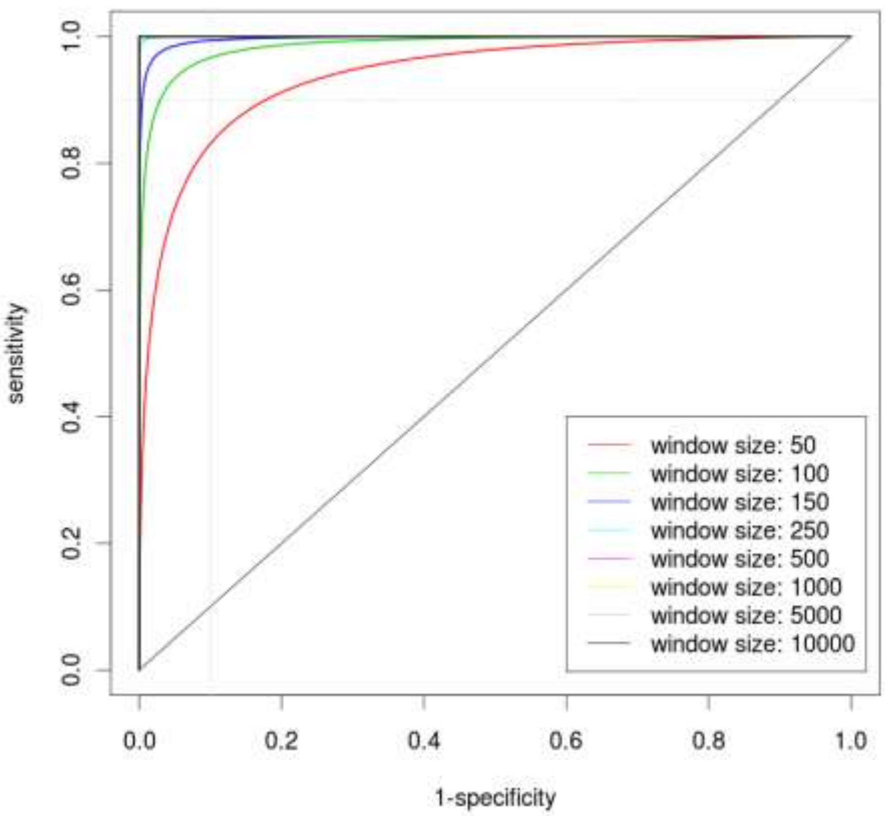
Duplication



It is generally harder to detect higher copy number as the variance increases linearly with the mean

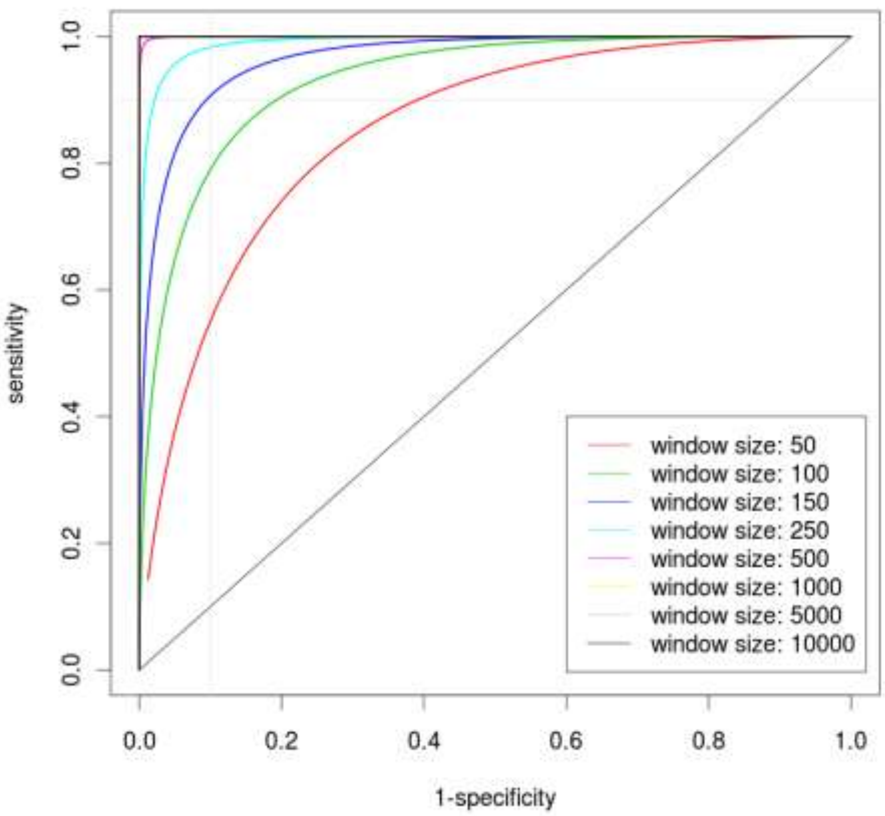
Deletion

rho=0.5, coverage 35x



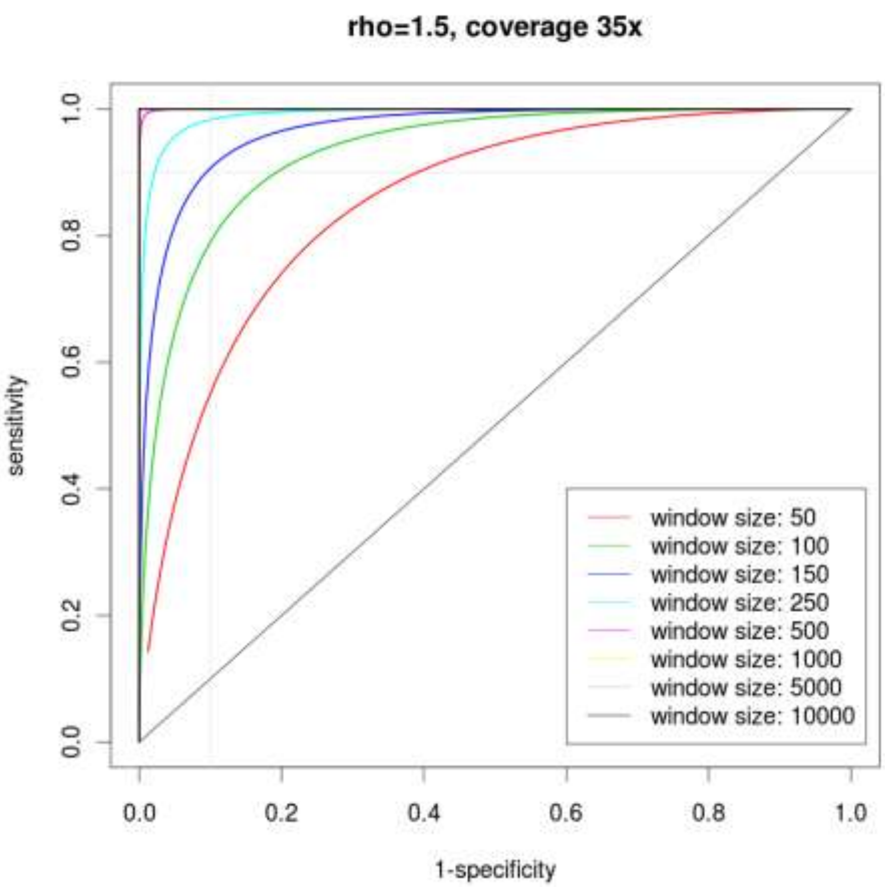
Duplication

rho=1.5, coverage 35x

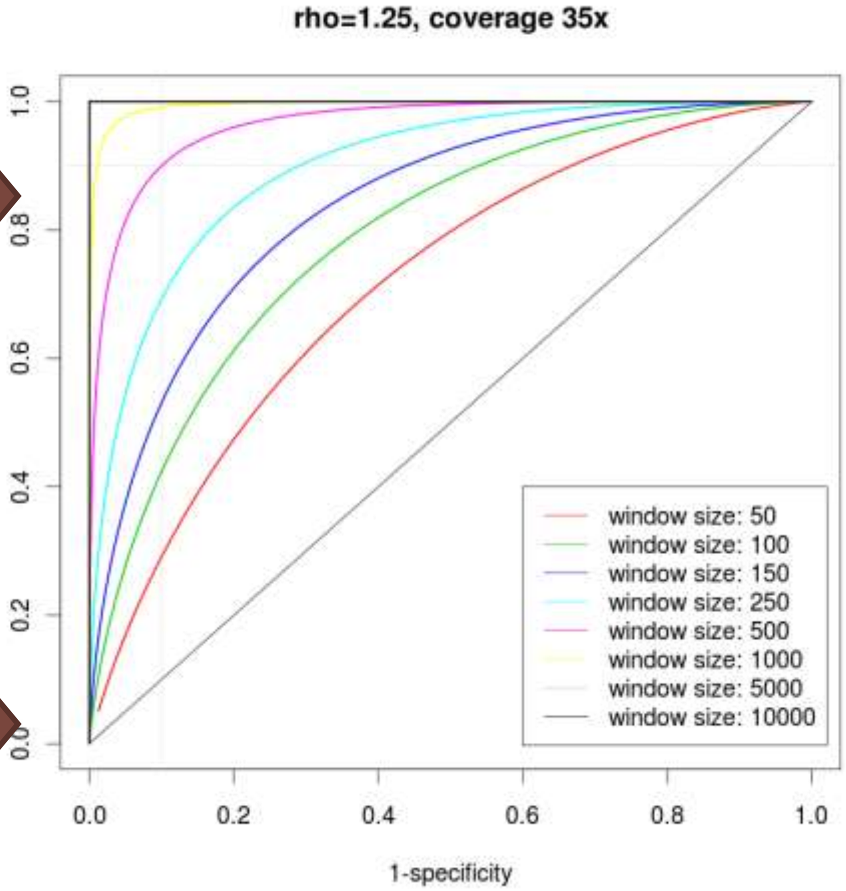


Issue: Admixture

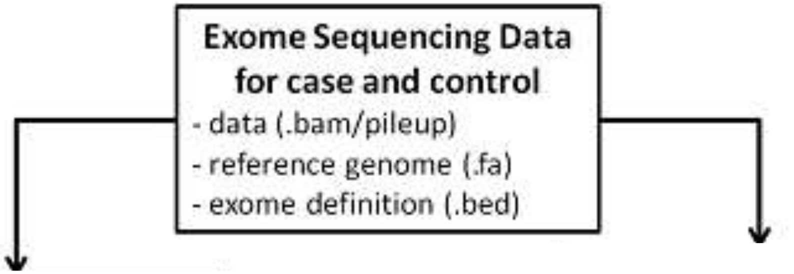
- Tumor sample is usually contaminated with normal cells
- Ratio will tend to 1, making it more difficult to detect CNV
- Have to estimate admixture rate prior to calling CNV otherwise power may be over/underestimated.



50% admixture



ExomeCNV Overview



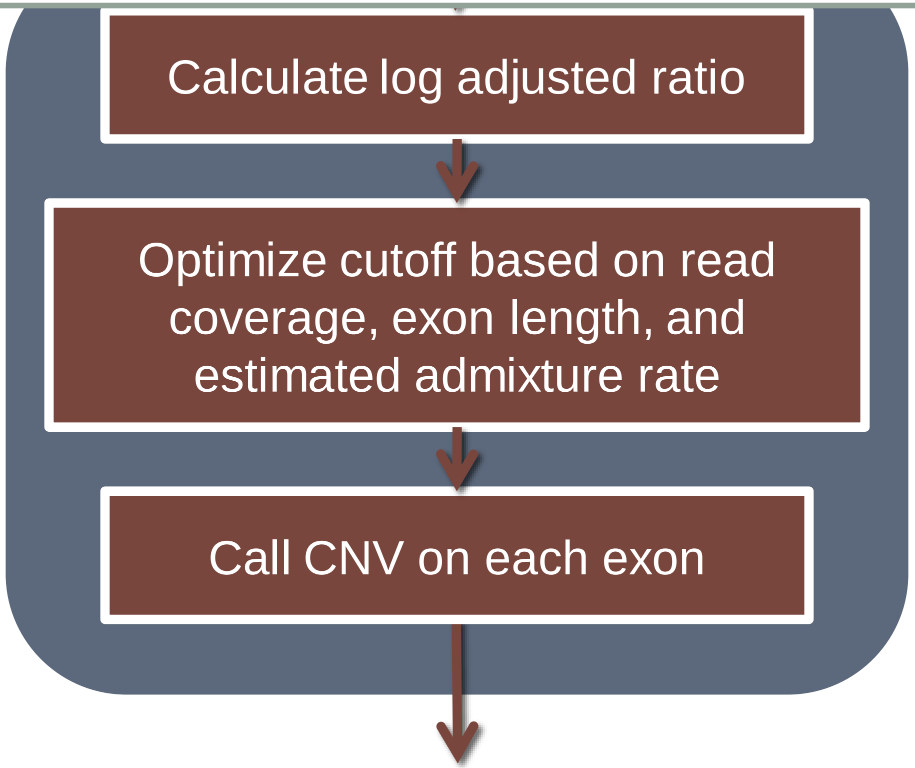
Exome Sequencing Data for case and control

- data (.bam/pileup)
- reference genome (.fa)
- exome definition (.bed)

```
source("http://bioconductor.org/biocLite.R")
biocLite("DNAcopy")
install.packages("ExomeCNV")
```

Exome CNV Calling Method

```
demo.eCNV = c()
for (i in 1:length(chr.list)) {
     idx = (normal$chr == chr.list[i])
     ecnv = classify.eCNV(normal=normal[idx,], tumor=tumor[idx,],
               logR=demo.logR[idx], min.spec=0.9999, min.sens=0.9999,
               option="spec", c=0.5, l=70)
     demo.eCNV = rbind(demo.eCNV, ecnv)
}
do.plot.eCNV(demo.eCNV, lim.quantile=0.99, style="idx", line.plot=F)
```



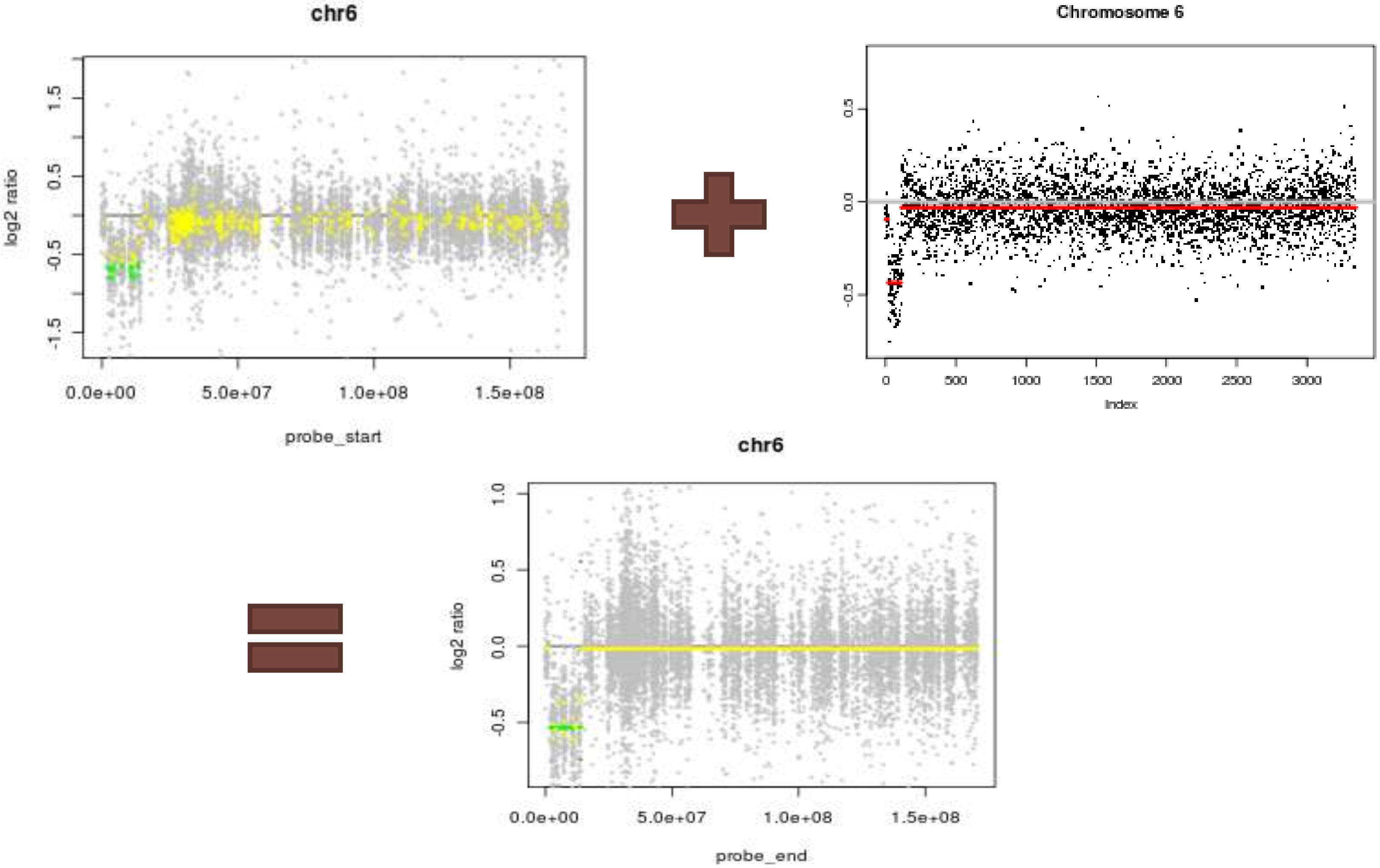
Calculate log adjusted ratio

Optimize cutoff based on read coverage, exon length, and estimated admixture rate

Call CNV on each exon

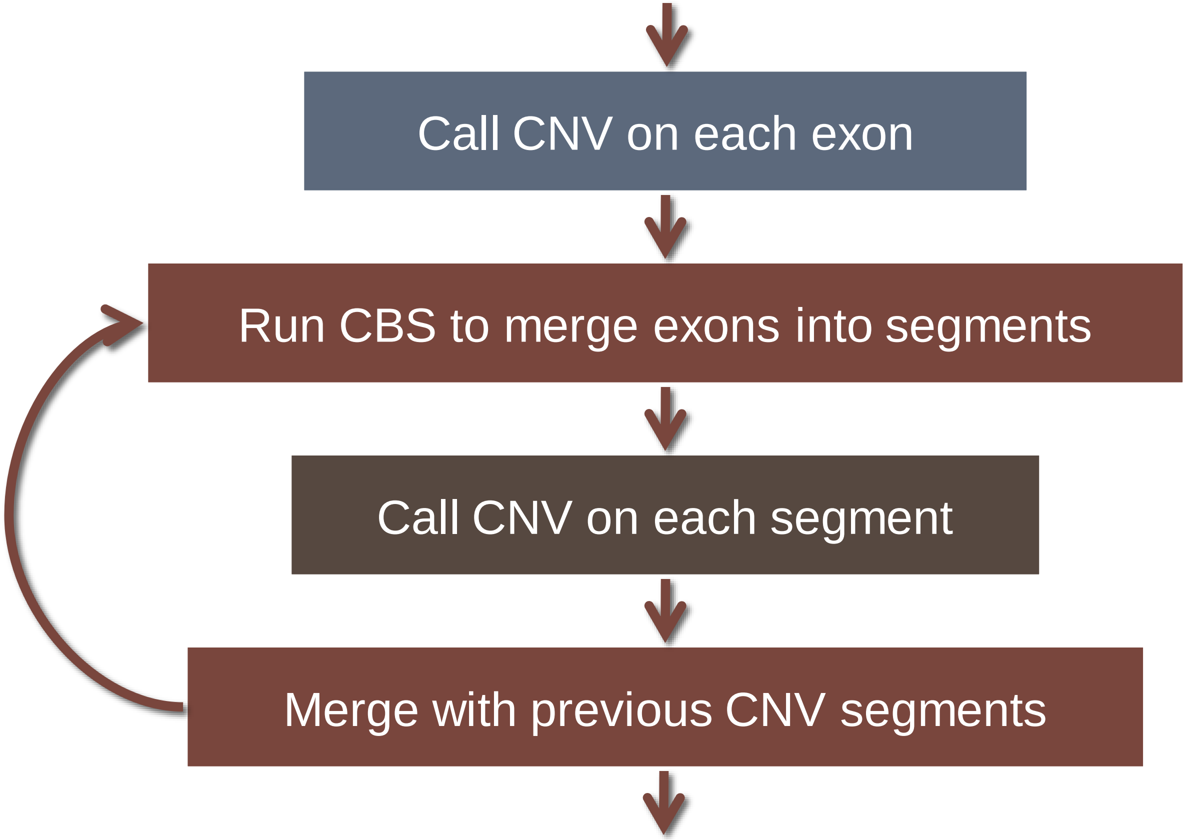
Merging exonic CNVs into segments

- Circular binary segmentation



Breakpoint Identification and Sequential Merging

```
demo.cnv = multi.CNV.analyze(normal, tumor, logR=demo.logR,
    all.cnv.ls=list(demo.eCNV), coverage.cutoff=5, min.spec=0.99,
    min.sens=0.99, option="auc", c=0.5)

do.plot.eCNV(demo.cnv, lim.quantile=0.99, style="bp", bg.cnv=demo.eCNV,
    line.plot=T)
```



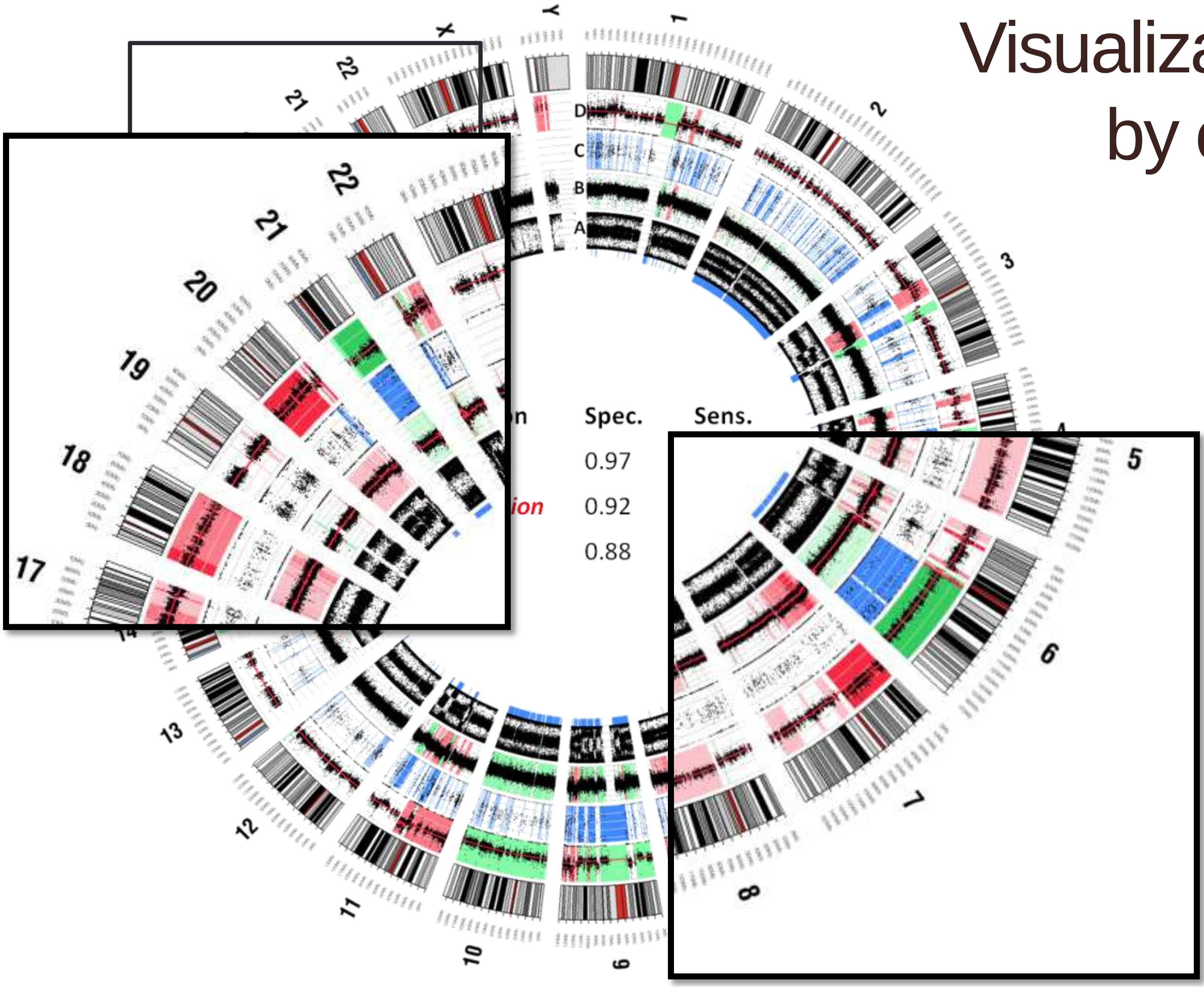
Call CNV on each exon

Run CBS to merge exons into segments

Call CNV on each segment

Merge with previous CNV segments

Visualization by circo



Resources

- https://secure.genome.ucla.edu/index.php/ExomeCNV_User_Guide

- JF Sathirapongsasuti, et al. (2011) Exome Sequencing-Based Copy-Number Variation and Loss of Heterozygosity Detection: ExomeCNV, *Bioinformatics*, 2011 Oct 1;27(19):2648-54. Epub 2011 Aug 9.

Thank you …