3rd Annual Practical GENOMICS August 26–28, 2013
From Biology to Biostatistics

# Mapping and Viewing Deep Sequencing Data

*bowtie2, samtools, igv*

Frederick J Tan
Bioinformatics Research Faculty
Carnegie Institution of Washington, Department of Embryology
tan@ciwemb.edu

27 August 2013

# What Is in Our Data Set?

## Illumina TruSight Autism Rapid Capture

*solution capture of exons from 101 genes*

## MiSeq 150 base, paired end reads

```
cd ~/autism
$ grep @M0 sample2_R1.fastq | head
$ grep @M0 sample2_R2.fastq | head
```

```
$ fastqc sample2_R1.fastq
```

216,362 reads

*How do we know if both reads the same?*

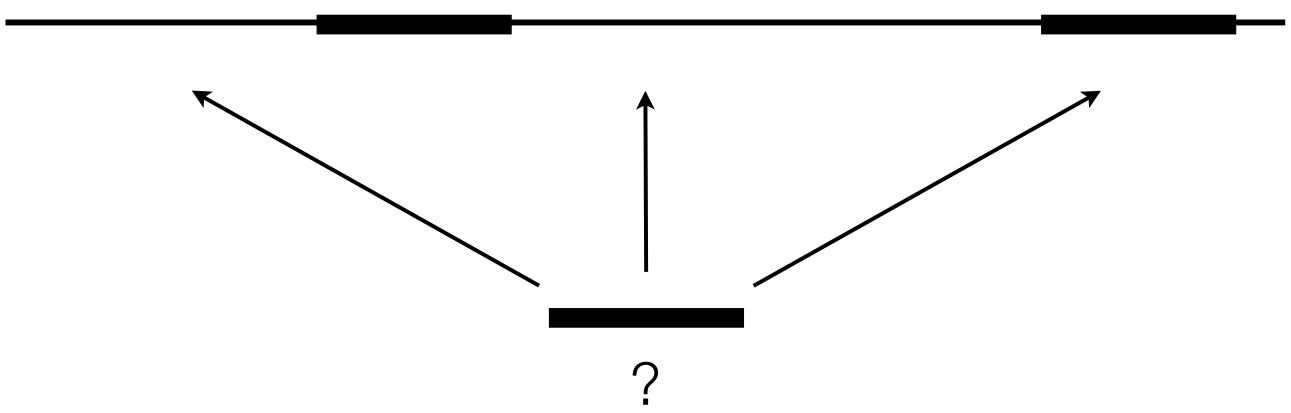
```
$ wc -l sample2_R2.fastq
```

Sample1 1.39 million read pairs

Sample2 0.22 million read pairs

illumina.com/clinical/translational_genomics/product/trusight_autism.ilmn

# The Mapping Problem




**GOALS**

Find match quickly

Find true match

Avoid missing a match

**COMPLICATIONS**

Large genome, many reads

Paralogs, pseudo-genes, repeats

Sequencing errors, polymorphisms

# Indices Address the Computational Complexity of Mapping Reads

```
$ bowtie2

No index, query, or output file specified!
Bowtie 2 version 2.1.0 by Ben Langmead (langmea@cs.jhu.edu, www.cs.jh
Usage:
  bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>} [-S <sam>]
```

```
  <bt2-idx>  Index filename prefix (minus trailing .X.bt2).
             NOTE: Bowtie 1 and Bowtie 2 indexes are not compatible.
```

*Which genome?*

Langmead, B, *et al.*, 2009 Genome Biology; Langmead and Salzberg, 2012 Nature Methods

# Build a Custom Reference Database

```
$ cd ~/genomes
$ head autism101.fa
$ tail autism101.fa
```

*How do we verify that we have the correct number of genes?*

```
$ grep ">" autism101.fa | more
$ grep ">" autism101.fa | wc
```

```
$ bowtie2-build autism101.fa AUTISM101
$ ls
```

# Map Reads to Reference with Bowtie2

```
$ cd ~/autism
$ bowtie2 -x ~/genomes/AUTISM101 -1 sample2_R1.fastq
          -2 sample2_R2.fastq -S sample2.sam
```

```
216362 reads; of these:
  216362 (100.00%) were paired; of these:
    64877 (29.99%) aligned concordantly 0 times
    148486 (68.63%) aligned concordantly exactly 1 time
    2999 (1.39%) aligned concordantly >1 times
    ----
    64877 pairs aligned concordantly 0 times; of these:
      13338 (20.56%) aligned discordantly 1 time
    ----
    51539 pairs aligned 0 times concordantly or discordantly; of these:
      103078 mates make up the pairs; of these:
        96967 (94.07%) aligned 0 times
        5008 (4.86%) aligned exactly 1 time
        1103 (1.07%) aligned >1 times
77.59% overall alignment rate
```

# SAM Stores Alignments and Reads

```
$ wc -l sample2.sam
$ less -S sample2.sam
```

```
@HD     VN:1.0  SO:unsorted
@SQ     SN:NEGR1        LN:879653
@SQ     SN:NTNG1        LN:341847
...
@PG     ID:bowtie2      PN:bowtie2      VN:2.1.0
M01121 83  RAI1   116639 42 150M = 116501 -288 AAGGT... AS:i:-9  XN:i:0  XM:i...
M01121 163 RAI1   116501 42 151M = 116639 288  NTTTC... AS:i:-19 XN:i:0  XM:i...
M01121 83  SHANK3 47360  42 137M = 47360  -137 GGGAA... AS:i:0   XN:i:0  XM:i...
M01121 163 SHANK3 47360  42 137M = 47360  -137 NGGAA... AS:i:-1  XN:i:0  XM:i...
M01121 83  EHMT1  5792   42 123M = 5792   -123 ACTCA... AS:i:0   XN:i:0  XM:i...
M01121 163 EHMT1  5792   42 123M = 5792   -123 NCTCA... AS:i:-1  XN:i:0  XM:i...
```

The SAM Format Specification (v1.4-r985)

The SAM Format Specification Working Group

September 7, 2011

## 1 The SAM Format Specification

SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to

samtools.sourceforge.net/SAM1.pdf

# A Utility That "explains SAM flags in plain English"

This utility explains SAM flags in plain English.

Flag: 83 Explain

Explanation:

- [x] read paired
- [x] read mapped in proper pair
- [ ] read unmapped
- [ ] mate unmapped
- [x] read reverse strand
- [ ] mate reverse strand
- [x] first in pair
- [ ] second in pair
- [ ] not primary alignment
- [ ] read fails platform/vendor quality checks
- [ ] read is PCR or optical duplicate

Summary:

read paired
read mapped in proper pair
read reverse strand
first in pair

picard.sourceforge.net/explain-flags.html

# Bowtie2 Manual Details Fields

12. Optional fields. Fields are tab-separated. `bowtie2` outputs zero or more of these optional fields for each alignment, depending on the type of the alignment:

| | |
|---|---|
| `AS:i:<N>` | Alignment score. Can be negative. Can be greater than 0 in `--local` mode (but not in `--end-to-end` mode). Only present if SAM record is for an aligned read. |
| `XS:i:<N>` | Alignment score for the best-scoring alignment found other than the alignment reported. Can be negative. Can be greater than 0 in `--local` mode (but not in `--end-to-end` mode). Only present if the SAM record is for an aligned read and more than one alignment was found for the read. Note that, when the read is part of a concordantly-aligned pair, this score could be greater than `AS:i`. |
| `YS:i:<N>` | Alignment score for opposite mate in the paired-end alignment. Only present if the SAM record is for a read that aligned as part of a paired-end alignment. |
| `XN:i:<N>` | The number of ambiguous bases in the reference covering this alignment. Only present if SAM record is for an aligned read. |
| `XM:i:<N>` | The number of mismatches in the alignment. Only present if |

bowtie-bio.sourceforge.net/bowtie2/manual.shtml

# Manipulate SAM/BAM Files with SAMtools

```
$ samtools
```

```
Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.19-44428cd

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
         index        index alignment
```

Li, *et al.*, 2009 Bioinformatics

# Determine Why Reads "Unmapped"

| Bit | Description |
|---|---|
| 0x1 | template having multiple segments in sequencing |
| 0x2 | each segment properly aligned according to the aligner |
| 0x4 | segment unmapped |
| 0x8 | next segment in the template unmapped |
| 0x10 | SEQ being reverse complemented |
| 0x20 | SEQ of the next segment in the template being reversed |
| 0x40 | the first segment in the template |
| 0x80 | the last segment in the template |
| 0x100 | secondary alignment |
| 0x200 | not passing quality controls |
| 0x400 | PCR or optical duplicate |

```
$ samtools view -Sf 12 sample2.sam | less -S
```

```
M01121 77  * 0 0 * * 0 0 AGACAGTGTTTCACCATGCTGGCCAGACTGGTCTCGAACTCCTGATCTCAGGCAGTC
M01121 141 * 0 0 * * 0 0 NCGGTAGCTCAAGCCTGTAATCCCAACACTTTGGGAGGCCGAGGCGGGGGGGGCGCC
M01121 77  * 0 0 * * 0 0 CCACAAAGATGTTCATCATGAAGAAAGCTACAATGATGATGTAGATGATGAAGAAGA
M01121 141 * 0 0 * * 0 0 GGGAAGGCAGCGGGCTGGGCCGTGTGGGCTGGGGGGCTTGGCAGGTCCTCACTTGGT
M01121 77  * 0 0 * * 0 0 GTCTCCTTCCATGCTAGAAGGAGACTTCCAGGCTGGAGGAAGAGGAGGCTTCCTCCC
```


## Basic BLAST

Choose a BLAST program to run.

| | |
|---|---|
| nucleotide blast | Search a **nucleotide** database using a **nucleotide** query<br>*Algorithms:* blastn, megablast, discontiguous megablast |

# Quickly Extract Sorted, Indexed BAMs

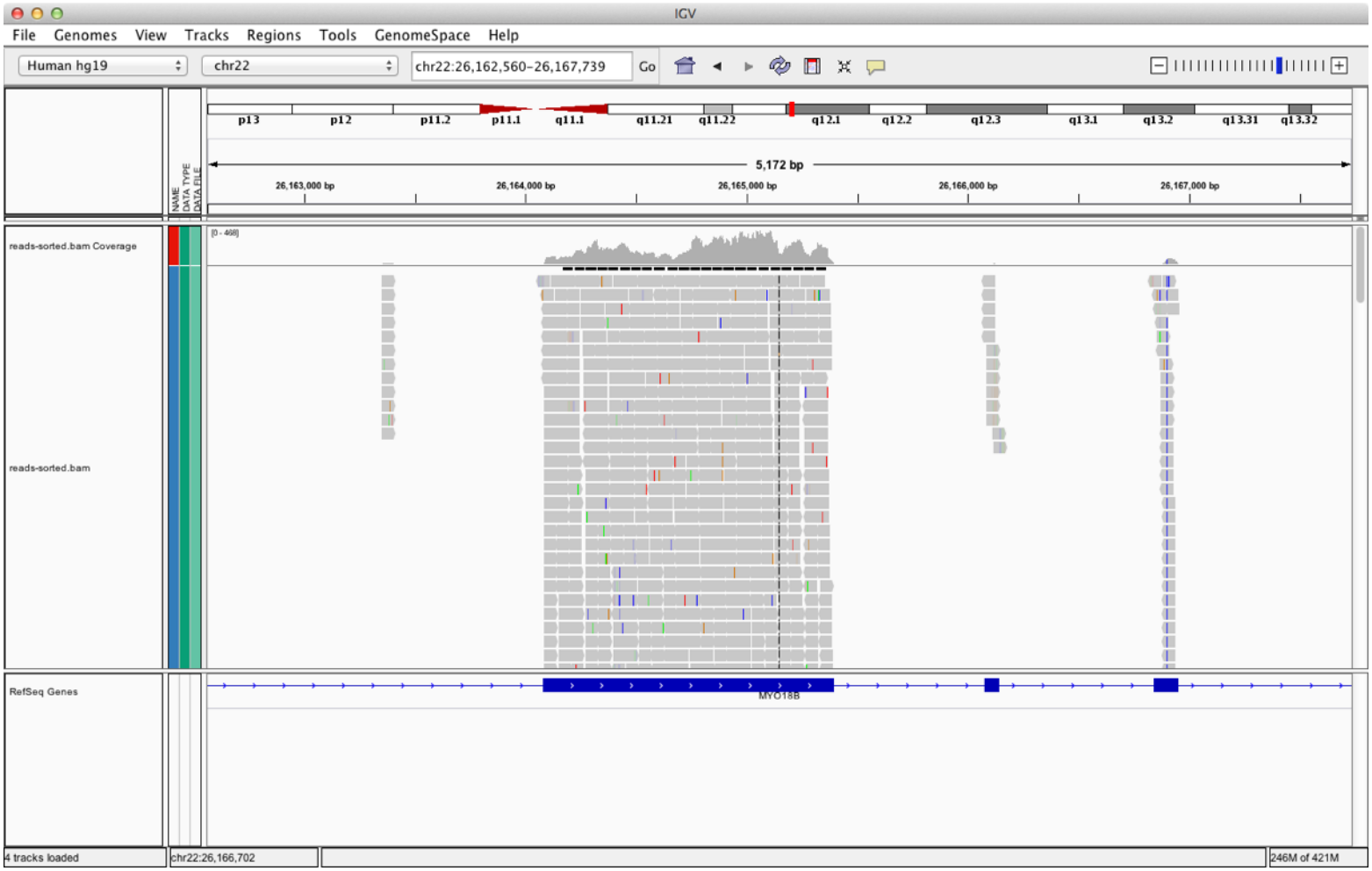
```
$ samtools view -bS sample2.sam > sample2.bam
$ samtools sort sample2.bam sample2_sorted
$ samtools index sample2_sorted.bam
```

```
$ samtools view sample2_sorted.bam PTEN:1000-2000 | less -S
```

```
M01121 99  PTEN 849 3  114M4D36M = 990  280 CCACCATCCAGCAGC
M01121 137 PTEN 862 23 101M4D49M = 862  0   GCTGCTGCCGCAGCC
M01121 163 PTEN 866 42 150M      = 971  256 CCGCAGCAGCCATTA
M01121 99  PTEN 874 42 149M      = 1047 324 GCCATTACCCGGCTG
M01121 73  PTEN 874 40 89M4D62M  = 874  0   GCCATTACCCGGCTG
M01121 137 PTEN 874 23 89M4D62M  = 874  0   GCCATTACCCGGGTG
M01121 73  PTEN 874 40 89M4D62M  = 874  0   GCCATTACCCGGCTG
M01121 163 PTEN 878 42 151M      = 969  242 TTACCCGGCTGCGGT
M01121 163 PTEN 878 42 151M      = 1151 424 TTACCCGGCTGCGGT
M01121 73  PTEN 888 24 75M4D76M  = 888  0   GCGGTCCAGAGCCAA
```

# “High-Performance Genomics Data Visualization and Exploration”

```
$ scp workshop@192.168.56.101:autism/sample2_sorted.bam* .
$ scp workshop@192.168.56.101:genomes/autism101.fa .
```




Robinson, *et al.*, 2011 Nature Biotechnology