Back-to-Basics: NGS Data Analysis 101

Presented By:
Jean Jasinski, Ph.D.
Field Applications Scientist
Agilent Technologies
Life Sciences & Diagnostics Group

Back-to-Basics: Agilent's Five Part 101 eSeminar Series

NGS 101

Tuesday,
October 8, 2013
1:00 pm ET

Alex Siebold, Ph.D.
Field Application Scientist

- NGS Background and Workflows
- Definition of NGS Terms
- Whole Genome Sequencing vs. Target Enrichment

RNA-Seq 101

Wednesday,
October 9, 2013
1:00 pm ET

Jean Jasinski, Ph.D.
Field Application Scientist

- How Does RNA-Seq Differ from DNA-Seq?
- What is Strand Specific RNA-Seq and How Does it Work?
- What is the Value of Targeted vs. Whole Transcriptome RNASeq?

Methyl-Seq 101

Wednesday,
October 9, 2013
4:00 pm ET

Alex Siebold, Ph.D.
Field Application Scientist

- Methylation Mechanisms and Significance
- Review of Comparative Technologies
- Introduction to Methyl-Seq

NGS Data Analysis 101

Thursday,
October 10, 2013
1:00 pm ET

Jean Jasinski, Ph.D.
Field Application Scientist

- Analysis Workflows, File Formats, and Data Filtering
- DNA-Seq vs. RNA-Seq Considerations
- Integrating Disparate Data Sets to Create a More Complete Story

NGS Panels 101

Friday,
October 11, 2013
1:00 pm ET

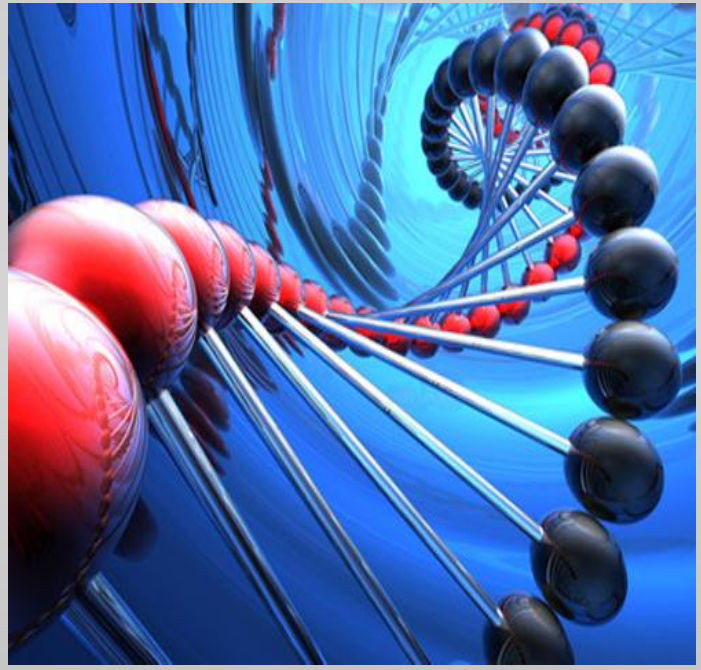
Adam Hauge
University of Minnesota

- Panel Design Process
- Quality at the Bench: Tips, Tricks, and Lessons Learned
- Considerations for Future Panels

Register at https://AgilentEseminar.Webex.com

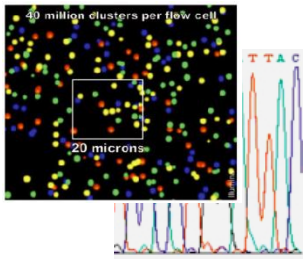
Topics for Today's Presentation

1. Overview/File Formats
2. Workflow
3. How RNA- and DNA-Seq Differ
4. Integrating Disparate Data Sets
5. Summary & Upcoming 101 eSeminars



Stages of NGS Analysis

Primary



Control Software

Data File (Reads + Quality)

Secondary

Data File (Reads + Quality)

FASTQ

ALIGNER

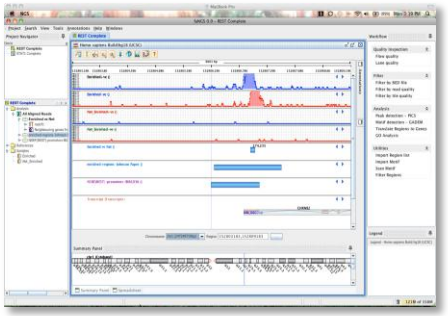
Reads aligned to genome

Tertiary

Reads aligned to genome

BAM, SAM

GeneSpring SureCall, Partek, …



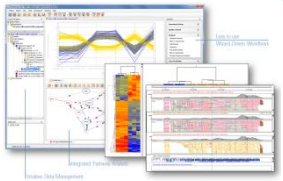
Typical Roles & Responsibilities
It Takes a Village

Core Facility
Service Bureau

Plan experiment
Obtain samples
Isolate DNA/RNA

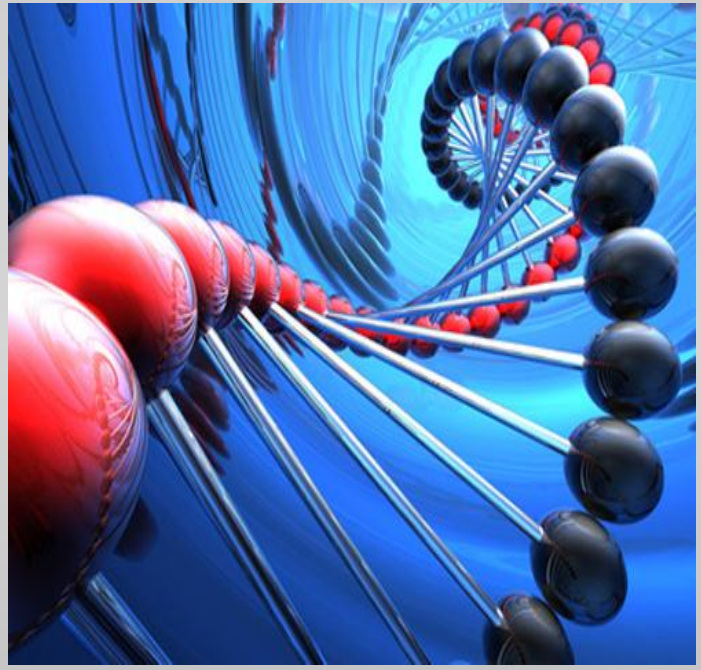
Alignment

Analysis



Topics for Today's Presentation

1. Overview/File Formats
2. Workflow
3. How RNA- and DNA-Seq Differ
4. Integrating Disparate Data Sets
5. Summary & Upcoming 101 eSeminars

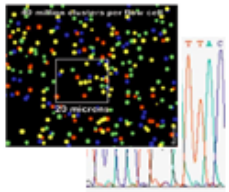


Primary Analysis Overview

- Sole responsibility of the sequencing platform vendor
- Converts physical signals to base calls as well as a quality score
- Quality score: measure of confidence in the base call

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

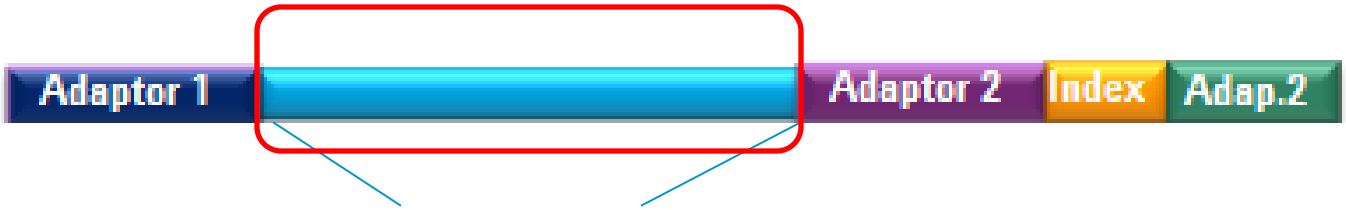


Data File (Reads + Quality)

FASTQ file

Primary Analysis
More Details

Sequencer Control Output



Sequence =
A,C,T,G,N + Qual
Score/base

@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

1) Demultiplex (separate reads based on barcode/index)
2) Trim Adapters
3) Filter bad reads (too short, too many N's)

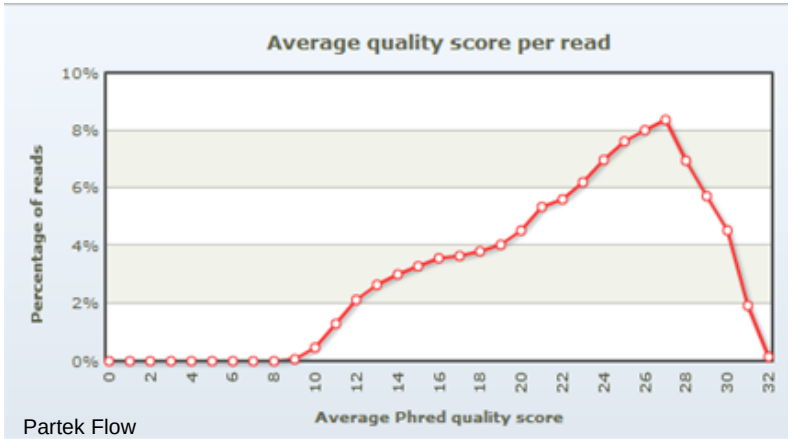
Primary Analysis
More Details

- Number of FASTQ files
  - 1 or 2 FASTQ files/lane (before demultiplex)
  - 1 file/sample for single-end or fragment reads
  - 2 files/sample for paired-end or mate-pair reads
- FASTQ processing steps all generate FASTQ (demux, filter, trim) so be sure you ask what processing steps have already been done
- Adapter removal

If sequence is too short, “other” adapter might need to be trimmed

Primary Analysis
Additional Filtering May be Required

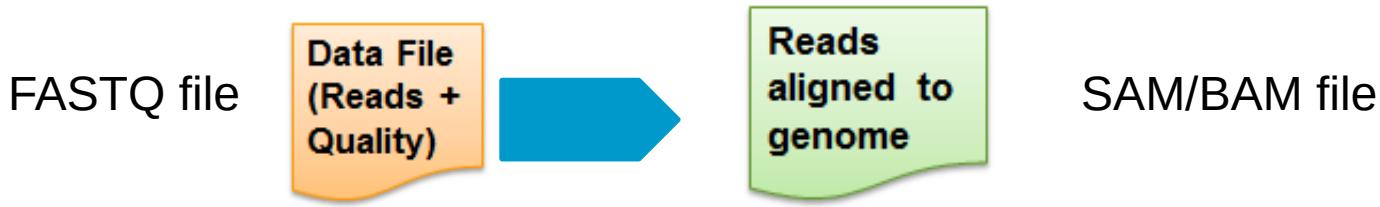
- For some sequencing chemistries, quality score drops off at the 3’ end or is lower for first few bases at 5’ end



- Trimming these bases may improve ability to align data
- Quality score of read is the average quality of all bases; can also reduce ambiguity by removing low quality reads

Secondary Analysis
Alignment

- Takes the sequences and quality scores from primary analysis and matches the reference sequence to return location of read(s)
- Most resource-intensive step of NGS analysis—requiring RAM, CPU, and disk
- *De facto* standard output file format now SAM/BAM (BAM is binary)
- Freeware or commercial products



Secondary Analysis
Alignment

- Compromise between speed and sensitivity
- Different aligners make different compromises; read length, read type, gaps, etc.
- Most aligners written by computer scientists; most run on Linux (command line interface)
- Tuned by a variety of parameters (command line options)

```
> bwa samse /CommonStorage\/enomeReferences/hg19/hg19/hg19.fasta $HOME/DataFiles/test/test.sai $HOME/
DataFiles\.est/test.fastq > $HOME/DataFiles/test/testBWA.sam
```

OPTIONS:

-n *NUM*
Maximum edit distance if the value is INT, or the fraction of missing alignments given 2% uniform base error rate if FLOAT. In the latte
automatically chosen for different read lengths. [0.04]

-o *INT* Maximum number of gap opens [1]

-e *INT* Maximum number of gap extensions, -1 for k-difference mode (disallowing long gaps) [-1]

-d *INT* Disallow a long deletion within INT bp towards the 3'-end [16]

-i *INT* Disallow an indel within INT bp towards the ends [5]

-l *INT* Take the first INT subsequence as seed. If INT is larger than the query sequence, seeding will be disabled. For long reads, this optior

-k *INT* Maximum edit distance in the seed [2]

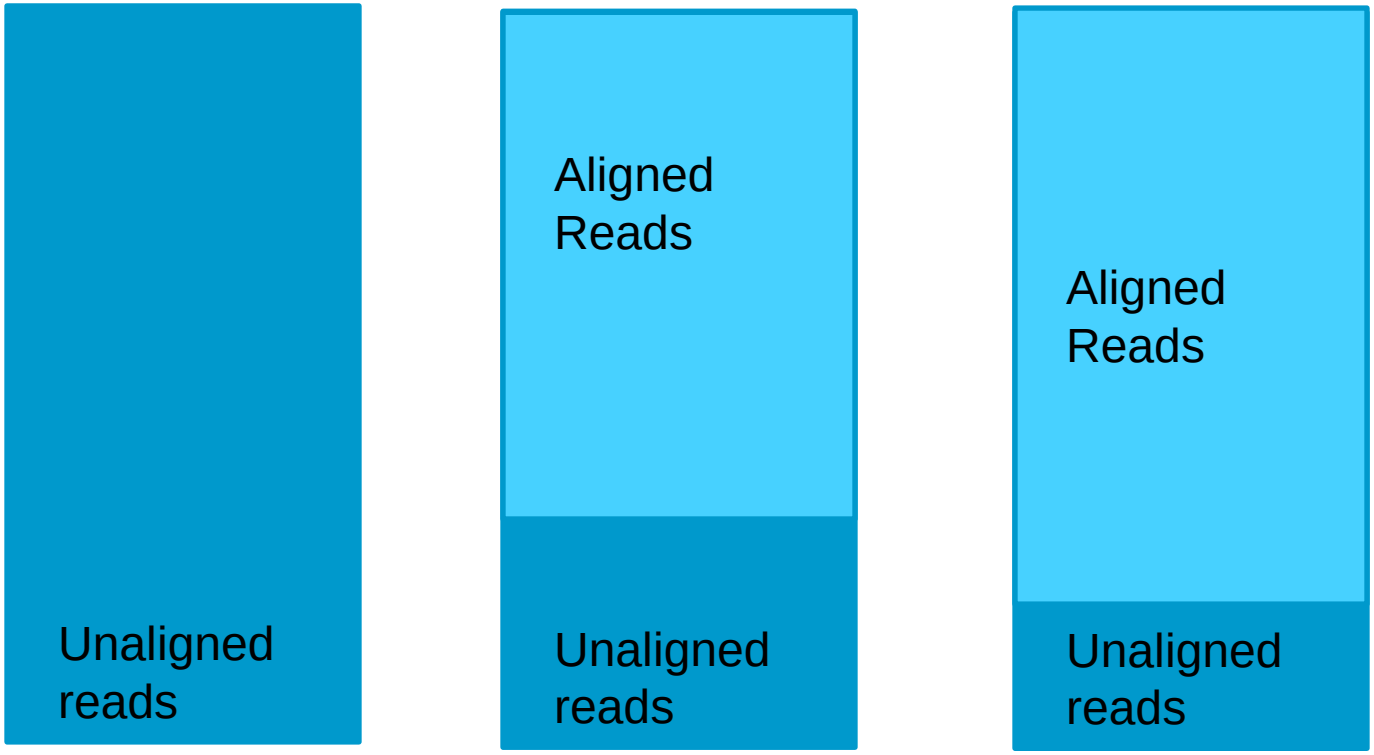
-t *INT* Number of threads (multi-threading mode) [1]

-M *INT* Mismatch penalty. BWA will not search for suboptimal hits with a score lower than (bestScore-misMsc). [3]

-O *INT* Gap open penalty [11]

Secondary Analysis
Alignment Details

- Alignment may be an interative process



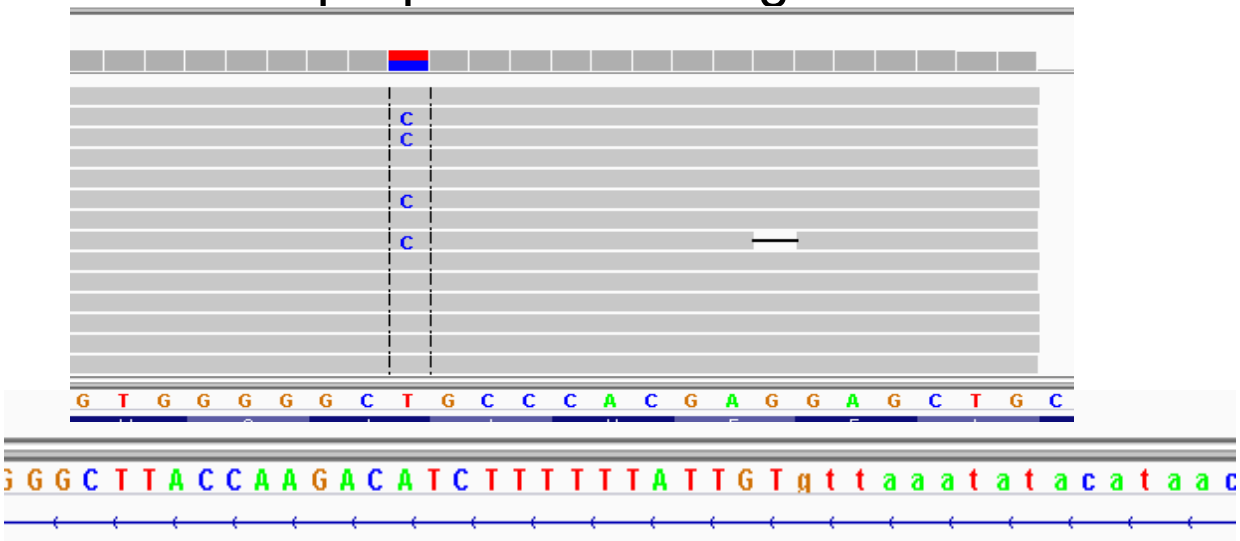
FASTQ file

SAM/BAM file after 1st run

SAM/BAM after second run with different aligner or less stringent parameters

Secondary Analysis
More details

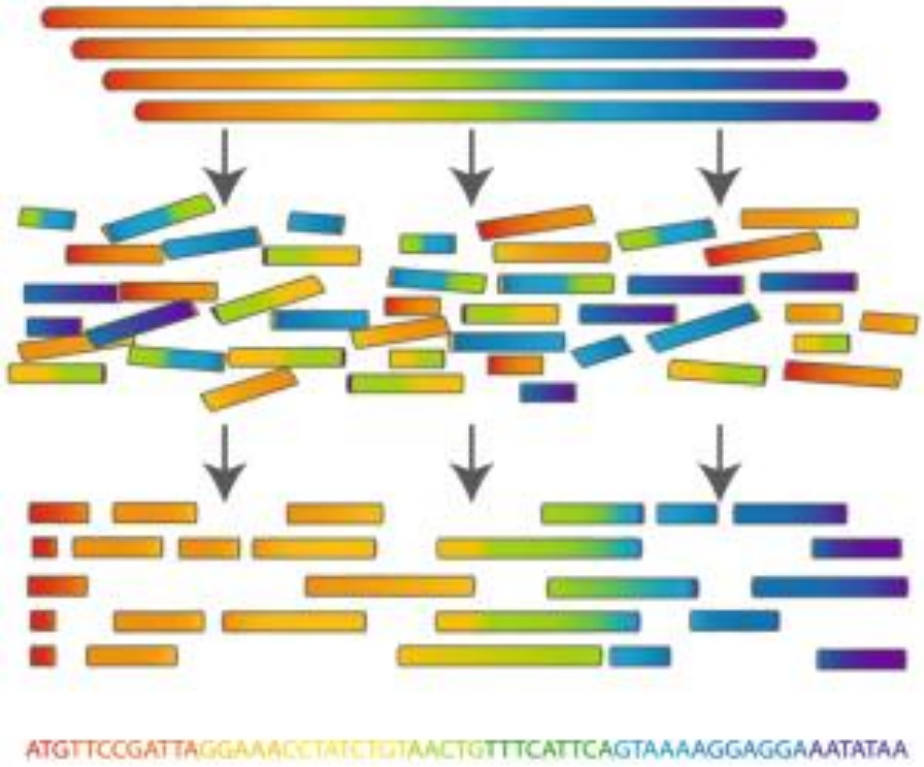
- Analogous to base quality, aligners may assign a mapping quality score
- Mapping Quality takes into account the average of the base quality of the reads as well as properties of the genomic location



- Reads may align to more than one location in genome; aligner options control whether to run all alignments, the first found, the best match

Secondary Analysis
Assembly

- Alignment requires a reference genome
- Assembly* creates genome from shotgun sequencing



Sample DNA

Fragmented
and Sequenced

Assembled into
contigs/scaffolds

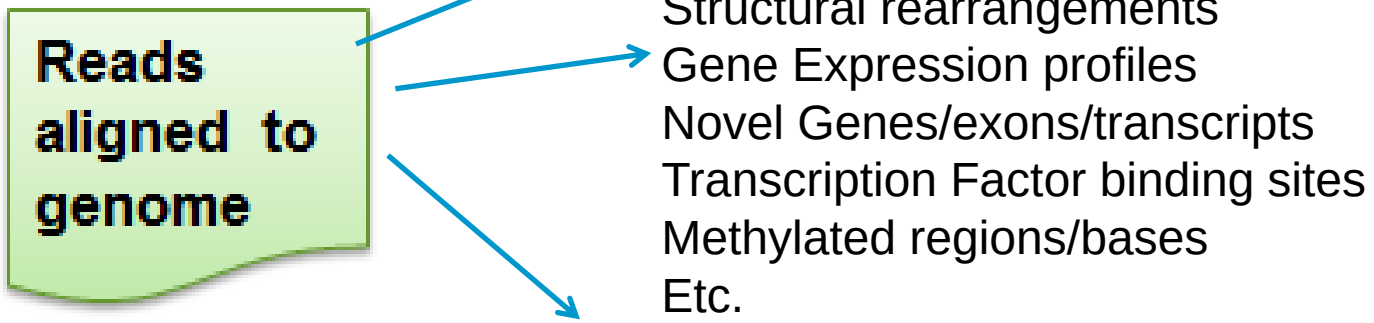
Rosalind.info

*DNAStar users term assembly for both assembly and alignment

Tertiary Analysis

- Tertiary analysis starts with aligned data (SAM or BAM format)
- Analysis diverges depending on NGS data analysis type: ChIP-Seq, Methyl-Seq, whole Genome sequencing, amplicon sequencing, RNA-Seq, small RNA-Seq, etc.
- Freeware and commercial software

SAM/BAM file



Variant Call File (one tertiary output file format)

•Tab-delimited format with each data line consists of the following fields:

Chromosome, position, variant id, reference/alternative alleles, quality, information (read depth), event, sample Id (optional), format (optional)

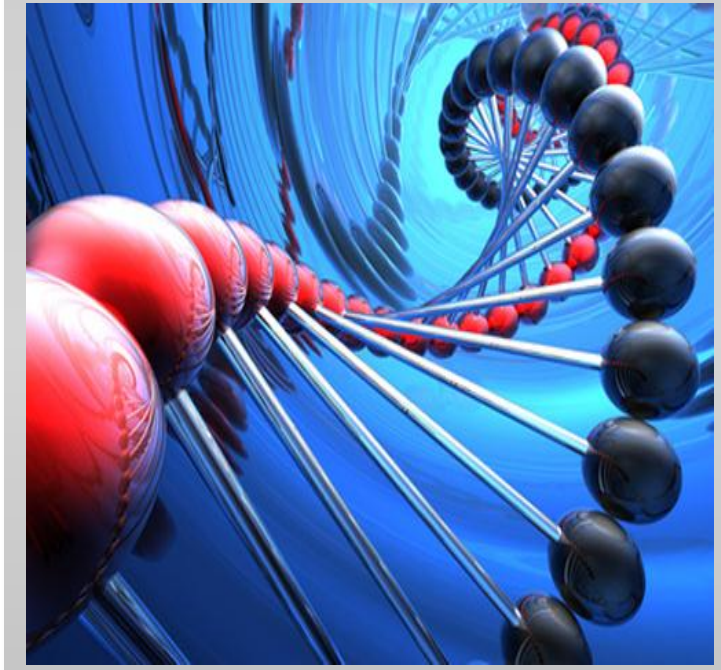
•Binary version is bcf

•May be shown as a Vcard (electronic business card) file in Windows

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA18507_DemoPane
chr4	1807894	rs7688609	G	A	9.52	.	DP=22;VD	GT:PL:GQ	0/1:39,0,152:41
chr4	55152040	rs2228230	C	T	101	.	DP=652;VI	GT:PL:GQ	0/1:131,0,255:99
chr7	55249063	rs1050171	G	A	208	.	DP=392;VI	GT:PL:GQ	0/1:238,0,255:99
chr10	43613843	rs1800861	G	T	222	.	DP=1224;\	GT:PL:GQ	1/1:255,255,0:99
chr11	32421533	rs5941766	T	C	195	.	DP=506;VI	GT:PL:GQ	0/1:225,0,255:99
chr11	32449661	rs5030167	G	A	222	.	DP=452;VI	GT:PL:GQ	1/1:255,255,0:99
chr13	28608241	rs6172913	A	G	69	.	DP=906;VI	GT:PL:GQ	0/1:99,0,255:99
chr16	89849480	rs6043402	C	T	174	.	DP=880;VI	GT:PL:GQ	0/1:204,0,255:99
chrX	66765627	rs6152	G	A	222	.	DP=440;VI	GT:PL:GQ	1/1:255,255,0:99

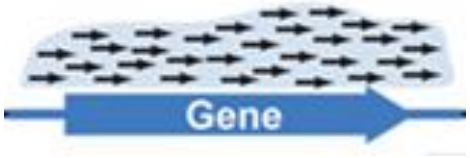
Topics for Today's Presentation

1. Overview/File Formats
2. Workflow
3. How RNA- and DNA-Seq Differ
4. Integrating Disparate Data Sets
5. Summary & Upcoming 101 eSeminars



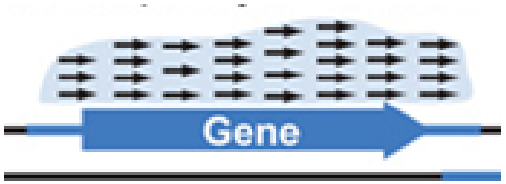
DNA-Seq Analysis
Library Analysis

- Three main types of sequencing approaches



Shotgun Sequencing

- Genome randomly fragmented
- High complexity
- Discovery
- Duplicates usually removed
- May be whole genome or targeted enrichment



Amplicon Sequencing

- Multiplex PCR-based targeted enrichment
- Low complexity
- Profiling
- Duplicates NOT removed

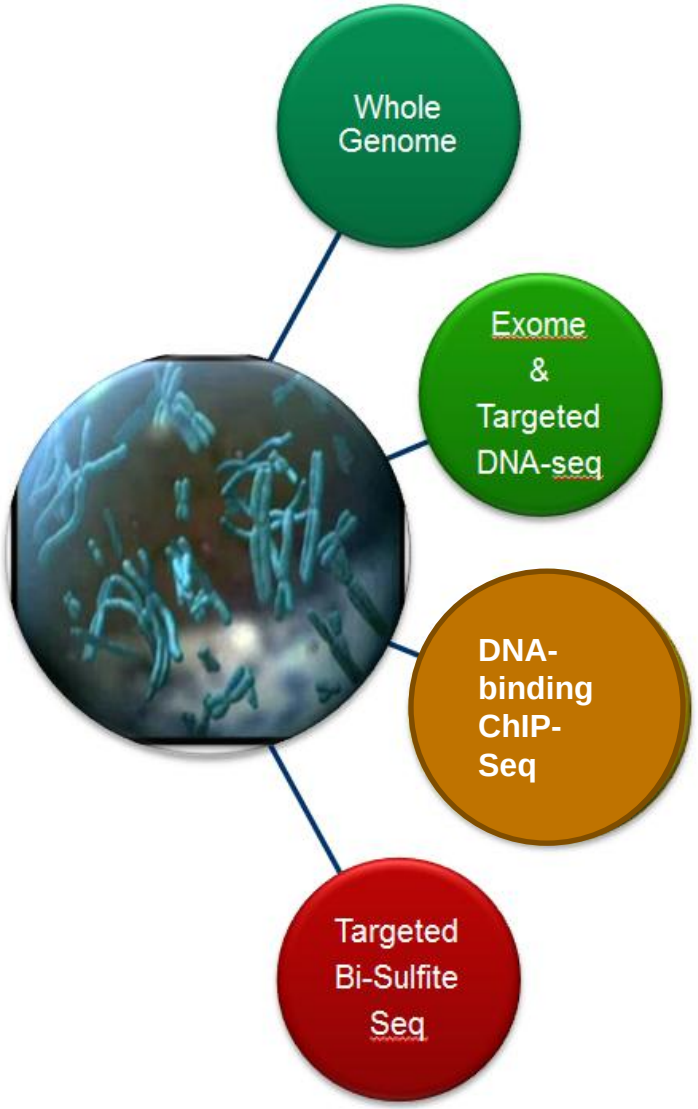
Hybrid approach (HaloPlex)

- Multiple, overlapping amplicons
- Medium complexity
- Discovery and profiling
- Duplicates NOT removed

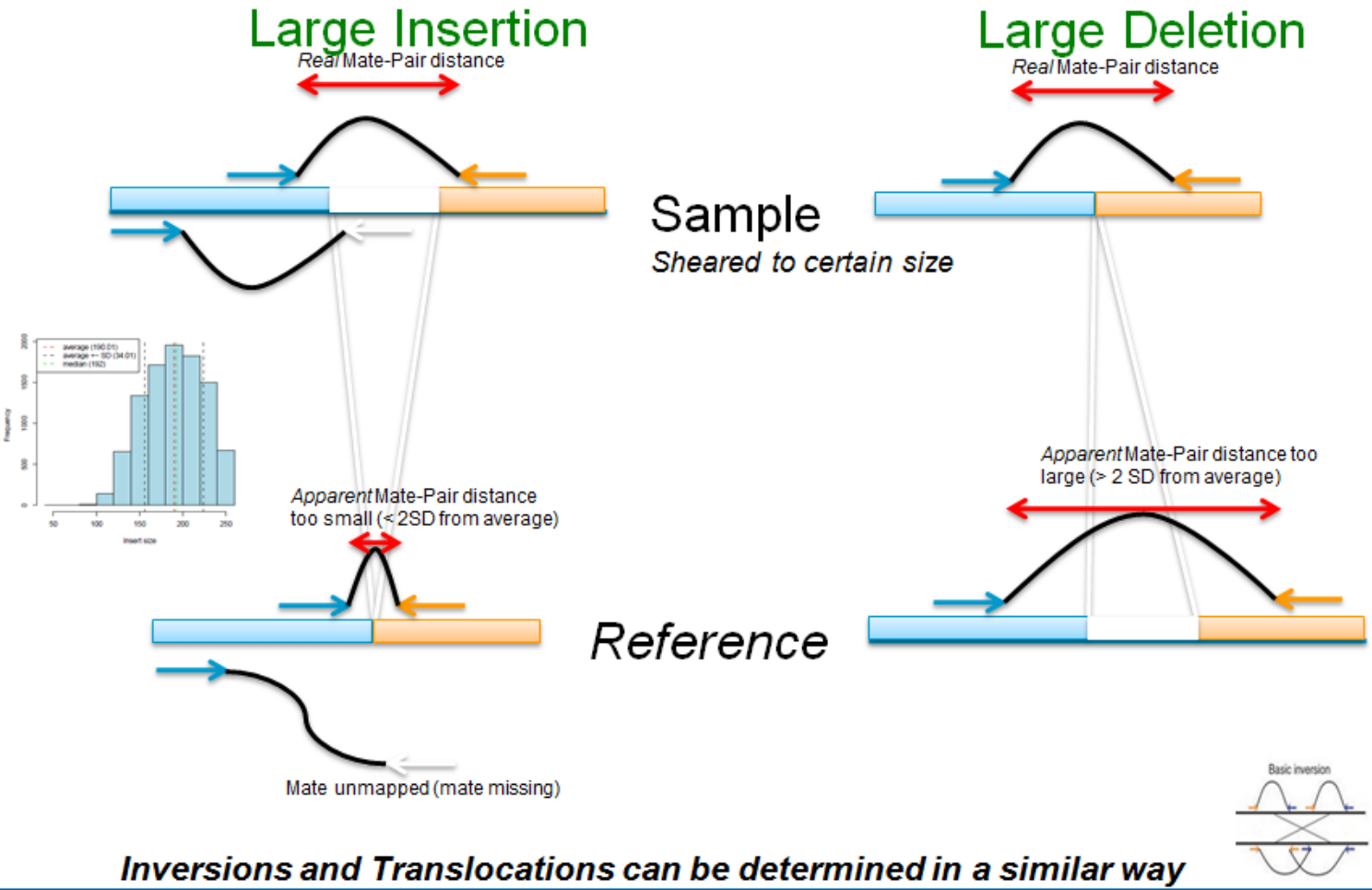
DNA-Seq
Tertiary Analysis

Types of DNA Variants Detectable using NGS

Large amplifications
Large deletions
Point mutations (SNP)
Inversions
Translocations
Copy number (CNV)
Fusions
Methylation status
DNA-binding profiles



DNA-Seq (Paired Ends)
Large Structural Rearrangements, Gene Fusions



Gene fusions can be detected from paired end reads by checking the orientation and alignment locations of paired ends

DNA-Seq
DNA-Binding Sites/Motifs

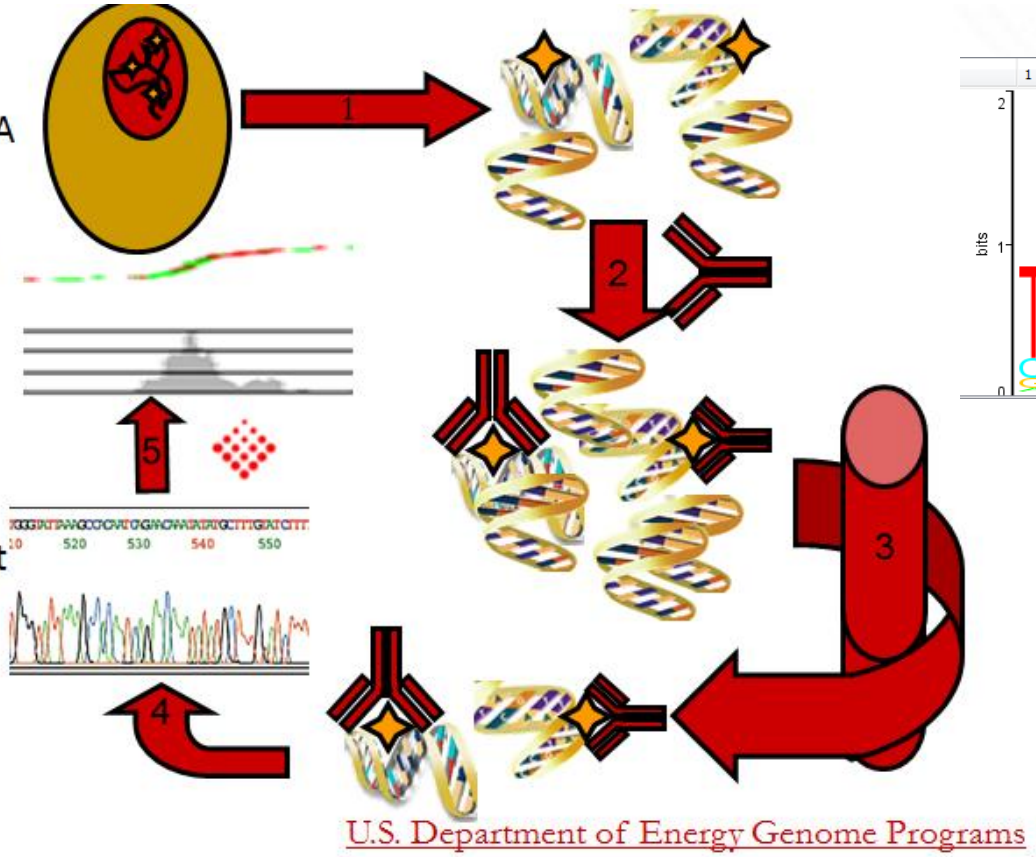
1. Collect and fractionate DNA

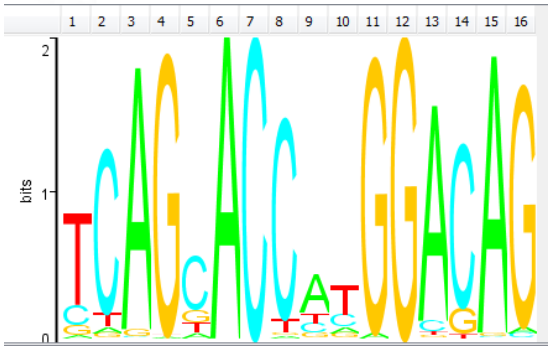
2-3. Enrich binding sites using IP

3b. PCR (Not Shown)

4. Sequence short reads

5. Align, detect Peaks and motifs





DNA-binding motifs

IP Antibodies may be for transcription factors, methylated Cs, etc.

DNA-Seq
Copy Number Variation Detection

- Most algorithms use the number of reads mapped to a certain location (depth of coverage) to estimate copy number; areas with more reads likely have higher copy number
- Use a sliding window to count the number of reads/window
- Less sophisticated algorithms (immature) compared to array CGH; often an “add-on” analysis rather than primary reason for DNA-Seq
- Requires complex library with duplicates removed
- Excellent review article:

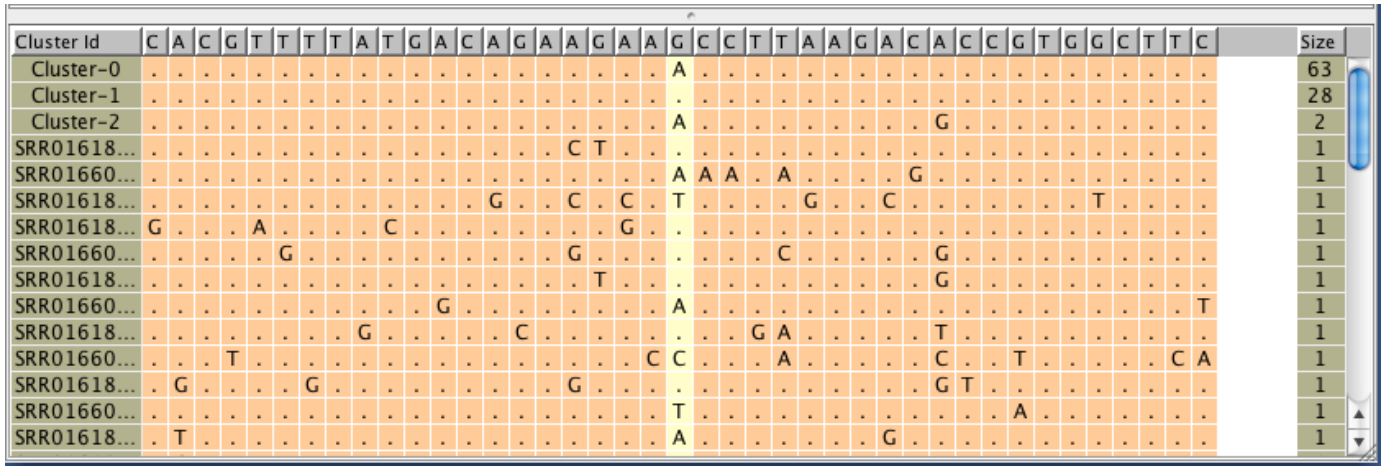
  Duan J, Zhang J-G, Deng H-W, Wang Y-P (2013) Comparative Studies of Copy Number Variation Detection Methods for Next-Generation Sequencing Technologies. PLoS ONE 8(3): e59128. doi:10.1371/journal.pone.0059128

DNA-Seq and RNA-Seq
SNP Detection

- Two main approaches:
  - SNPs detected against a reference genome sample
  - SNPs detected between samples (tumor vs. normal)
- Small insertion/deletions (indels) often detected in same step
- Can be detected in RNA-Seq as well as DNA-Seq
- Loss of Heterozygosity (LOH) analysis built upon SNP detection in DNA-Seq

DNA-Seq and RNA-Seq SNP Detection

- Many different algorithms available; differ in how base quality of read/reference are used, false positive/false negative rates, strand-bias, bi-allelic vs. all 4 bases,
- Different approaches for common/rare SNP detection
- Will require filtering and tuning of parameters (options)

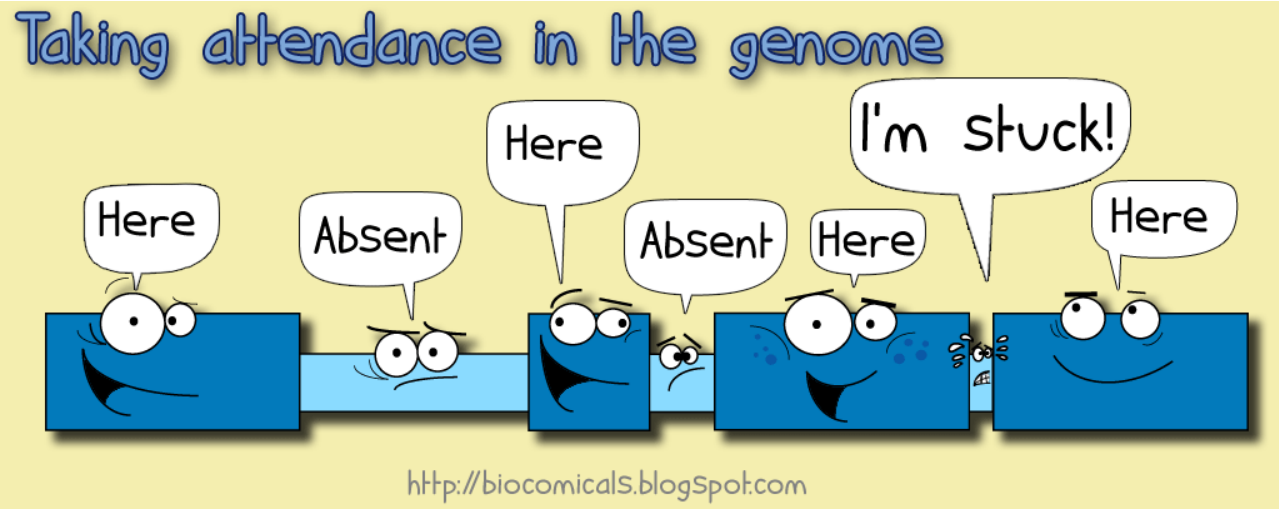


Cluster Id	C A C G T T T T A T G A C A G A A G A A G C C T T A A G A C A C C G T G G C T T C	Size
Cluster-0	. . . . . A	63
Cluster-1	. . . .	28
Cluster-2	. . . . . A . . . . . G	2
SRR01618...	. . . . . C T	1
SRR01660...	. . . . . A A A . A . . . G	1
SRR01618...	. . . . . G . . C . C . T . . . . G . . C . . . . . T	1
SRR01618...	G . . . A . . . . C . . . . . G	1
SRR01660...	. . . . . G . . . . . G . . . . C . . . . G	1
SRR01618...	. . . . . T . . . . . G	1
SRR01660...	. . . . . G . . . . . A . . . . . T	1
SRR01618...	. . . . . G . . . . C . . . . . T	1
SRR01660...	. . . T . . . . . C C . . . A . . . . . C . . T . . . . . C A	1
SRR01618...	. G . . . . G . . . . . G . . . . . G T	1
SRR01660...	. . . . . T . . . . . A	1
SRR01618...	. T . . . . . A . . . . . G	1

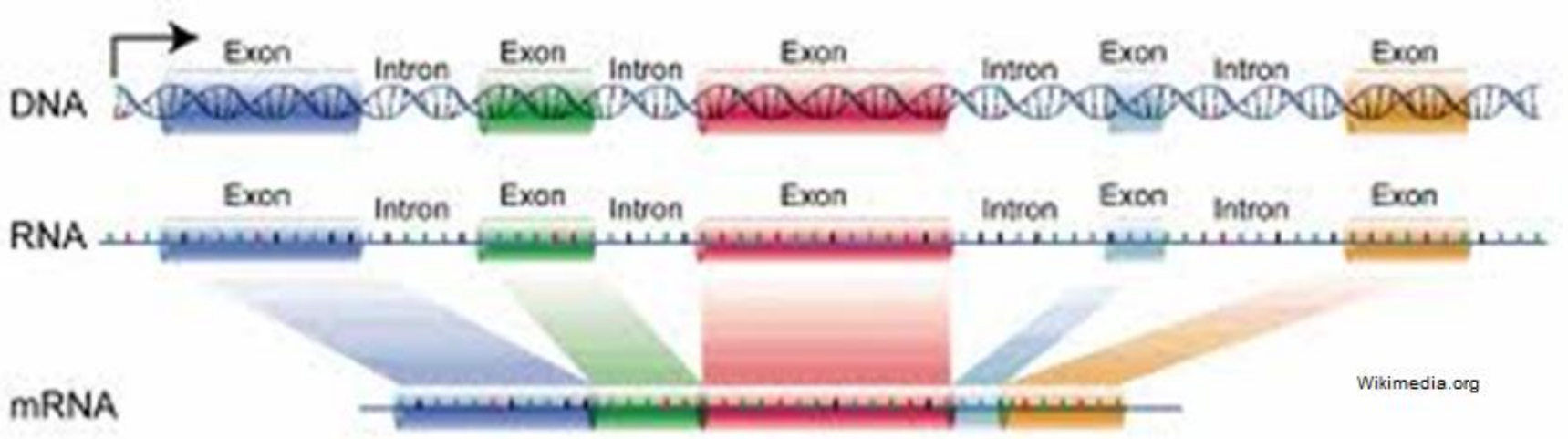
Variant Support View (GeneSpring)

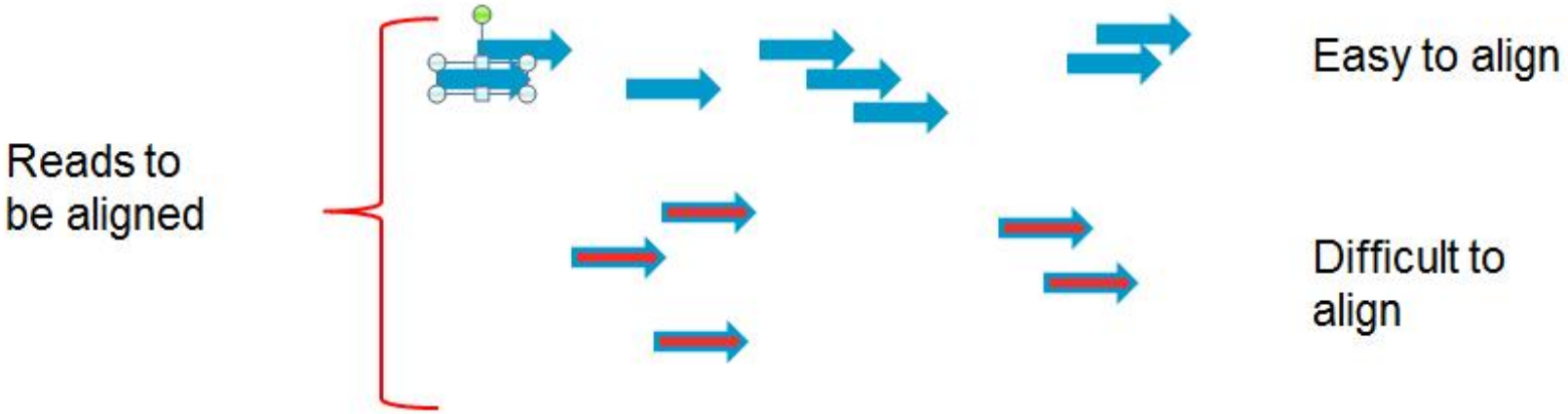
RNA-Seq
Alignment Three Approaches

- Align to reference genome but have to deal with gaps because mRNA is not contiguous with genomic sequence. Requires second step to map to transcriptome
- Align to transcriptome directly but novel sequences won't align
- Assemble sequences to create transcriptome (no known genome or transcriptome)



RNA-Seq
Junction Reads Require Different Aligner

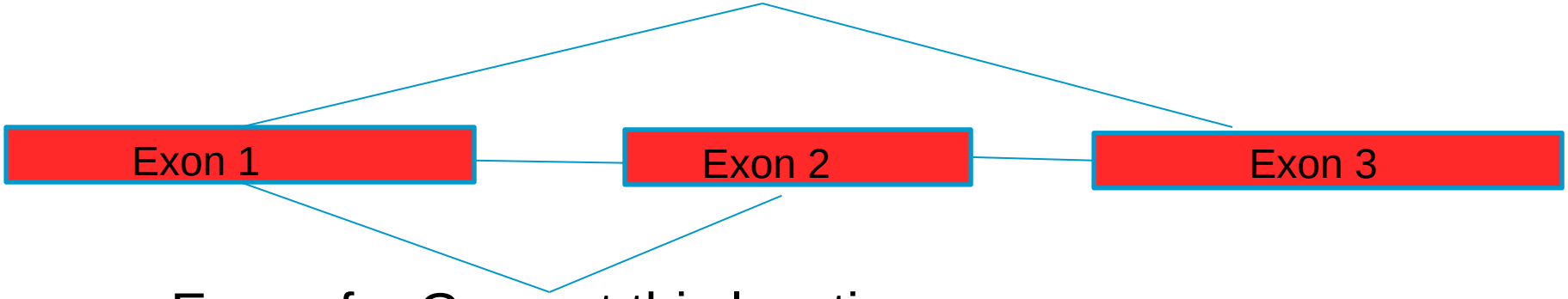




RNA-Seq
Align to Genome/Map to Transcriptome

Alignment aligns reads to genome locations

chr,start,stop



Exons for Gene at this location

Use an E/M (expectation maximization) algorithm to determine which reads go with which transcript

RNA-Seq
Align to Transcriptome

Genomic Coordinates

Reads: chr, start, stop

e.g., BRCA1

chr17: 41196312 -41277500

SNP would be:

Chr17 position 41196318

Use RefSeq start site to determine implication of SNP

Amplicon/Transcriptome Coordinates

Each targeted amplicon treated like a chromosome

e.g., Amplicon_name,start,stop or

NM_999999,25-125

SNP would be:

AMPL1 position 5

Tools use “modified” reference to determine consequence of SNP

Look at your aligned data to see which coordinate system was used; annotation of variants requires annotations to use the same system

RNA-Seq
Differential Gene Expression Normalization

Normalization of number of reads per gene/transcript needed due to:

1) Length of gene/transcripts (longer transcripts will have more reads mapped to them)

2) Number of reads per sample (samples with more reads will look like higher expressors)

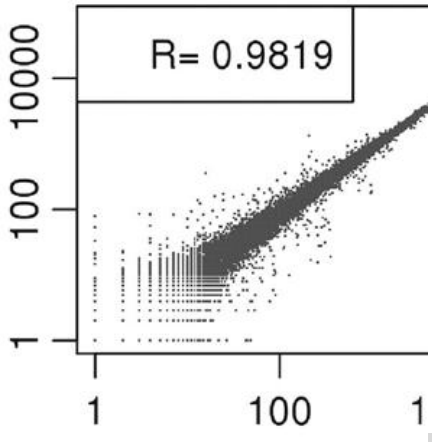
RPKM: Reads per Kilobase of exon transcript length per Million Mapped reads

=(Reads * 1,000,000)/(total exon length in Kb * mapped reads in sample)

DNA-Seq: most analyses involve analysis of single samples; paired analyses normalized for number of reads/sample or number reads/chr

RNA-Seq
Differential Gene Expression

- Requires normalization so genes can be compared to each other and samples can be compared to each other
- Count data (follows Poisson distribution); use right test
- There may be transcripts/genes with 0 reads detected; how to calculate fold change?
- Replicates provide more statistical power but simple Chi-Square test can be done (null hypothesis: each sample has the same distribution of reads; p-value provides statistical significance)



DNA-Seq and RNA-Seq Differences
Gene Fusions

DNA-Seq

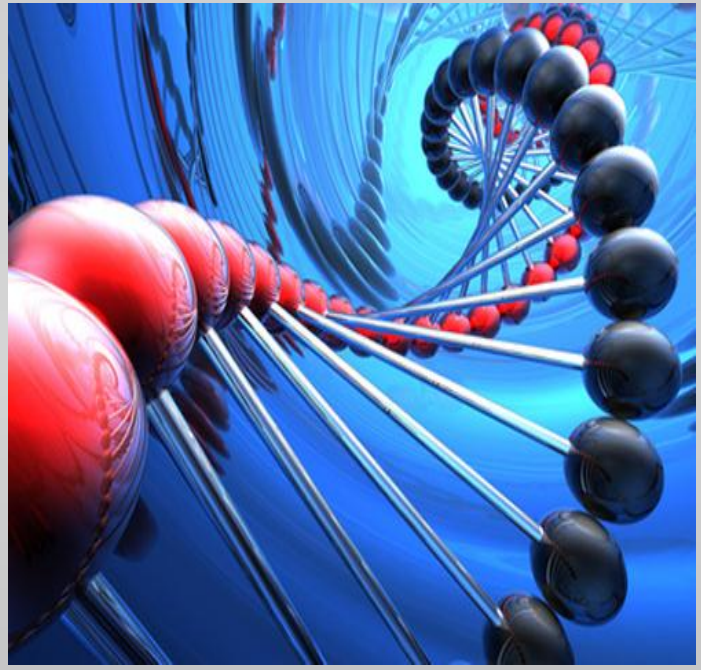
- Requires paired-end reads
- Fusion suspected if each paired end aligns to locations further away than expected or on different chromosomes
- P-value determined by how many of these reads are found vs. how many might be expected due to alignment errors
- No guarantee that fusion is actually expressed (functional)

RNA-Seq

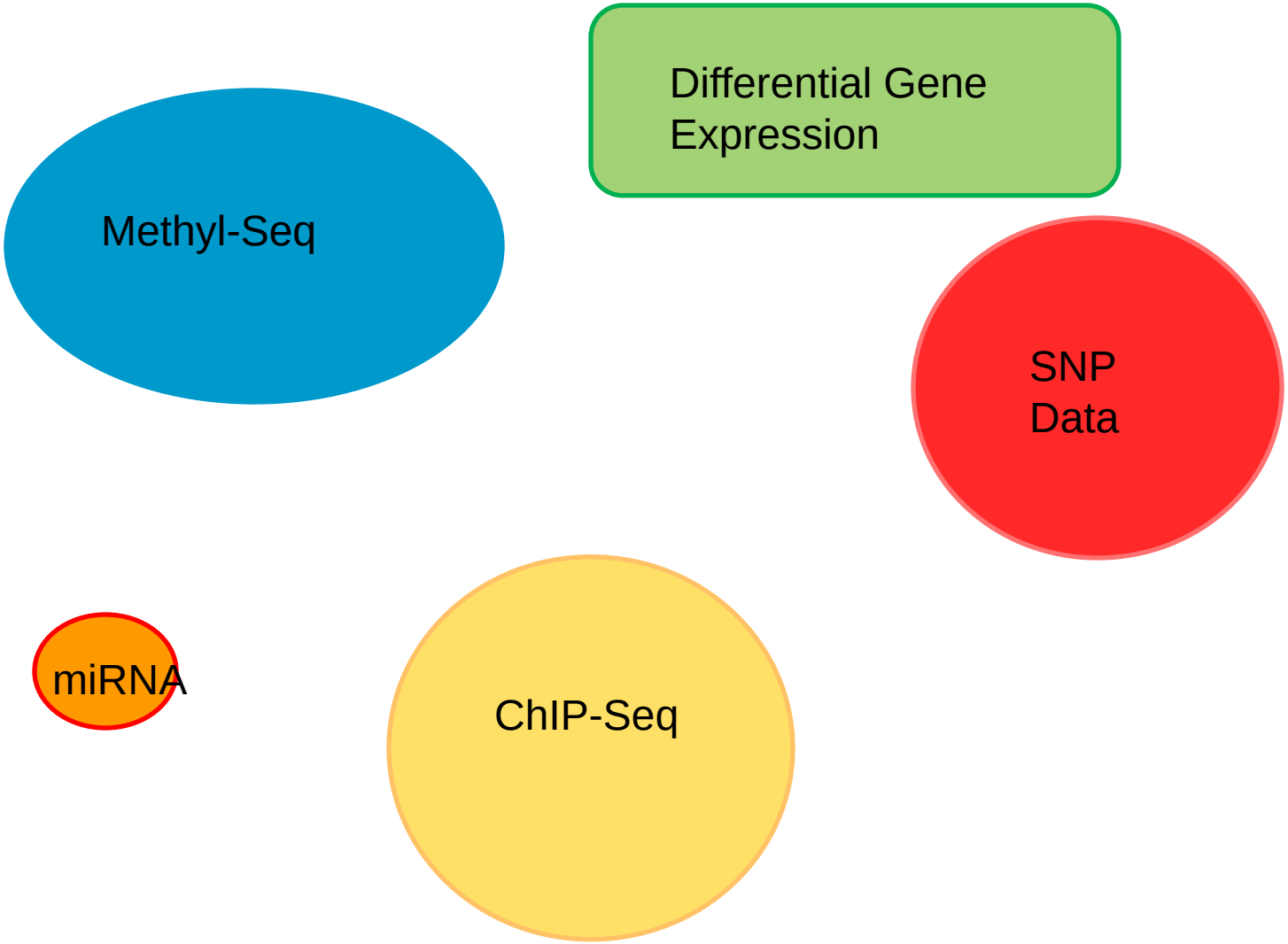
- Requires paired-end reads
- Fusion suspected if each paired end aligns to locations further away than expected or on different chromosomes
- Transcriptome smaller than genome; achieve deeper coverage for same number of reads
- Spurious chimeric transcripts may be generated during library prep
- Read-through transcripts and trans-spliced mRNAs will be detected
- Need to verify that sequence exists in genome

Topics for Today's Presentation

1. Overview/File Formats
2. Workflow
3. How RNA- and DNA-Seq Differ
4. Integrating Disparate Data Sets
5. Summary & Upcoming 101 eSeminars



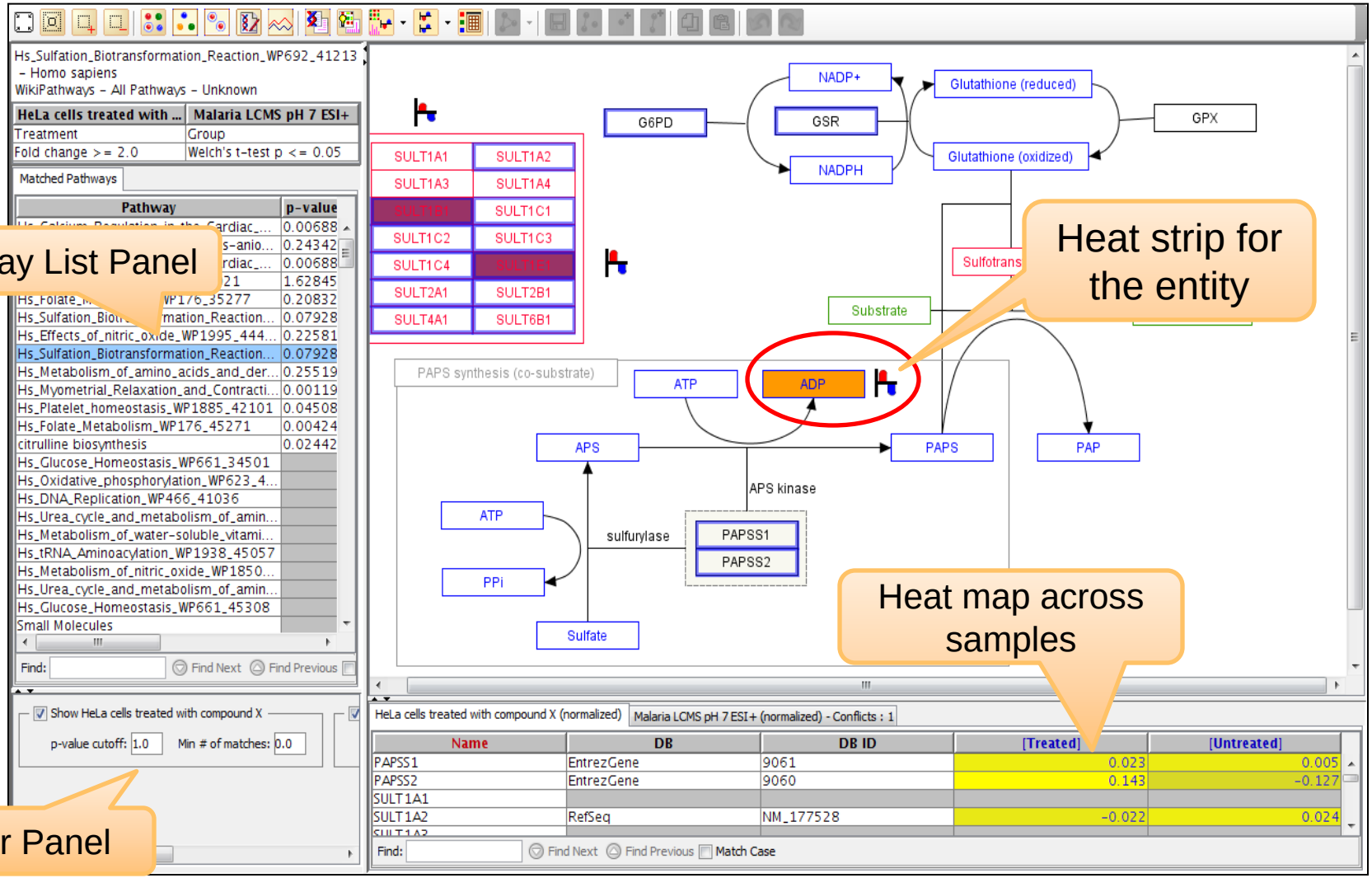
Integrating Disparate Data Sets



Integration can be done at many levels

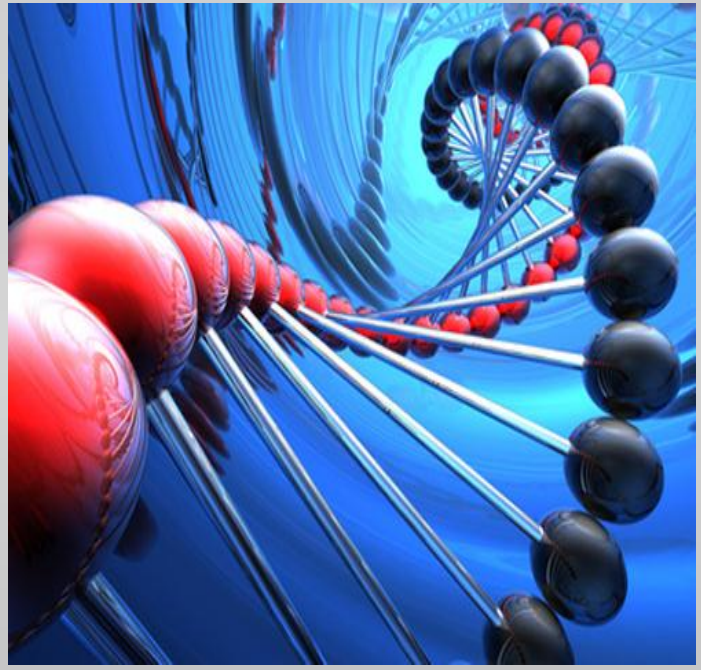
- List comparison via Venn Diagram
- Region List Comparisons
- Correlation of results
- Pathway/MultiOmic Analysis (integration at protein level with networks or canonical pathways)

Pathway View for a Single and Multi-Omic Analysis GeneSpring



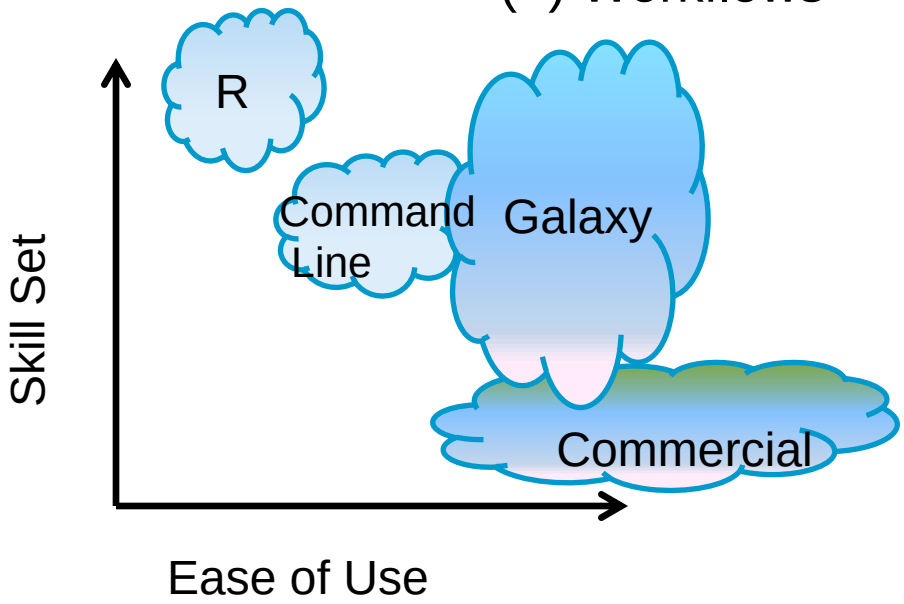
Topics for Today's Presentation

1. Overview/File Formats
2. Workflow
3. How RNA- and DNA-Seq Differ
4. Integrating Disparate Data Sets
5. Summary & Upcoming 101 eSeminars



Freeware vs. Commercial Software

Freeware	Commercial
(+) Price ($0)	(-) Price
(-) Support	(+) Support
(-) List-based (no graphics)	(+) Interactive
(+) Pipeline	(+) Workflows



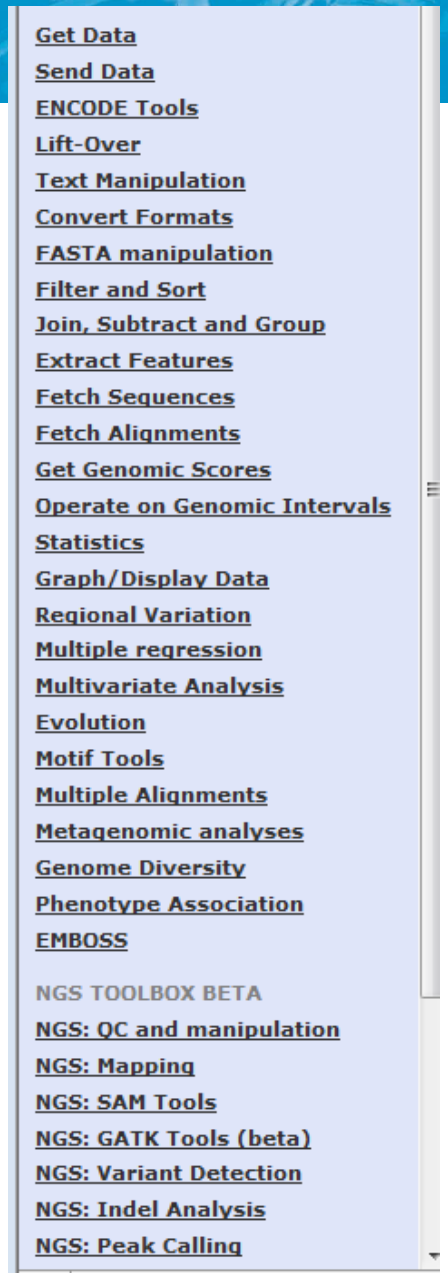
Using R

- Learn how to program and think like a programmer
- Know what the next step is; create pipeline from scratch
- Find or write a routine to handle the next step

```
> mydat <- data.frame(M) # Creates a dataframe
> names(mydat) # Give the names of each variable
[1] "obs"    "Height" "Weight" "BMI"
> str(mydat)   # give the structure of your data
'data.frame':   10 obs. of  4 variables:
 $ obs    : num  1 2 3 4 5 6 7 8 9 10
 $ Height: num  168 177 177 177 178 172 165 171 178 170
 $ Weight: num  88 72 85 52 71 69 61 61 51 75
 $ BMI    : num  31.2 23 27.1 16.6 22.4 ...
>
> View(mydat)  # Look at your data
>
> summary(mydat)  # Descriptive Statistics
      obs             Height          Weight          BMI
 Min.   : 1.00   Min.   :165.0   Min.   :51.00   Min.   :16.10
 1st Qu.: 3.25   1st Qu.:170.2   1st Qu.:61.00   1st Qu.:21.25
 Median : 5.50   Median :174.5   Median :70.00   Median :22.70
 Mean   : 5.50   Mean   :173.3   Mean   :68.50   Mean   :22.89
 3rd Qu.: 7.75   3rd Qu.:177.0   3rd Qu.:74.25   3rd Qu.:25.29
 Max.   :10.00   Max.   :178.0   Max.   :88.00   Max.   :31.18
>
```

What is Galaxy?

- Developed at Penn State
- Web interface/wrapper for many freeware NGS tools like alignment, genome browsing, filtering/file manipulation
- BLAST-like model (can use their computing resources or install locally)
- Fewer tools available in web than in toolshed (local install)
- Free



Tuxedo Suite: Free NGS tools

TopHat, TopHat2:
junction alignment
(calls Bowtie/2)

Bowtie, Bowtie2:
alignment

Cufflinks:
RNA-Seq tools
like cuffdiff,
cuffcompare,
cuffmerge

cummeRbund:
R package for
visualization of
cufflinks output

Glossary

Assembly: Process of creating a reference genome or transcriptome from shotgun sequenced data

Alignment: Assign genomic coordinates to sequences by comparing to a reference genome

Quantification/Mapping: Assign aligned reads to a particular transcript that overlaps the genomic coordinates

Normalization: Process of equalizing data between samples and genes so that read counts are comparable

Why we know NGS 101 & more…
Agilent Technologies Solutions for the NGS workflow

The Gold Standard for Sample QC

2100 Bioanalyzer Instrument & Kits

2200 TapeStation Instrument & Kits

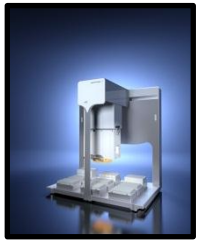
The Leader in NGS Target Enrichment

SureDesign

SureSelect

HaloPlex

Bravo Automation



NGS Analysis Software

GeneSpring NGS

SureCall

Validation Technologies

qPCR- Mx system & Brilliant reagents

Microarrays- CGH, CGH+SNP,
Gene Expression & miRNA

The NGS 101 eSeminar Series continues…

Event	Date & Time	Speaker	Topics
NGS Panels 101	Fri, Oct 11 1 pm ET	Adam Hauge, University of Minnesota	• Panel Design Process • Quality at the Bench: Tips, Tricks, and Lessons Learned • Considerations for Future Panels

Contact Us

Agilent Technologies

800.227.9770

Agilent_inquiries@agilent.com

www.agilent.com/genomics