Back to the Basics: Next-Generation Sequencing 101

Presented By:
Alex Siebold, Ph.D.

October 8, 2013

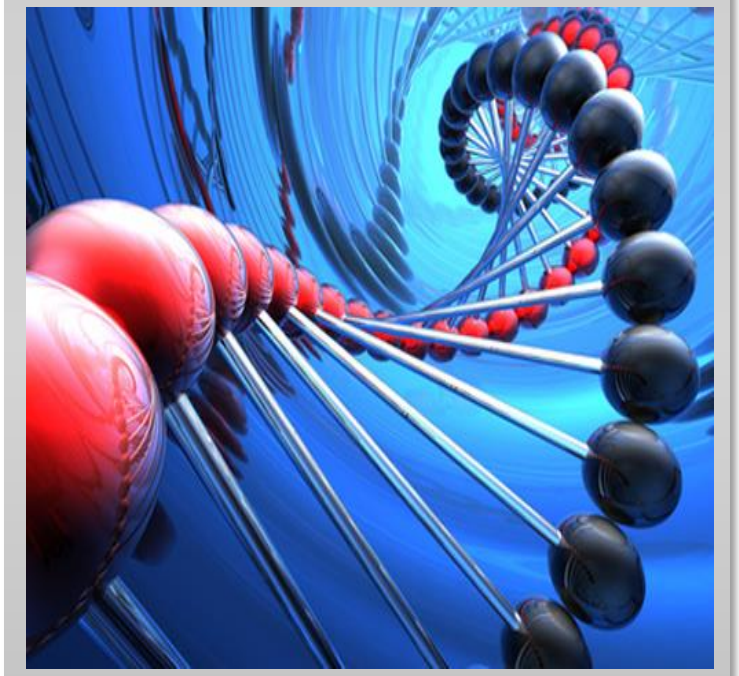
Field Applications Scientist
Agilent Technologies
Life Sciences & Diagnostics Group

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

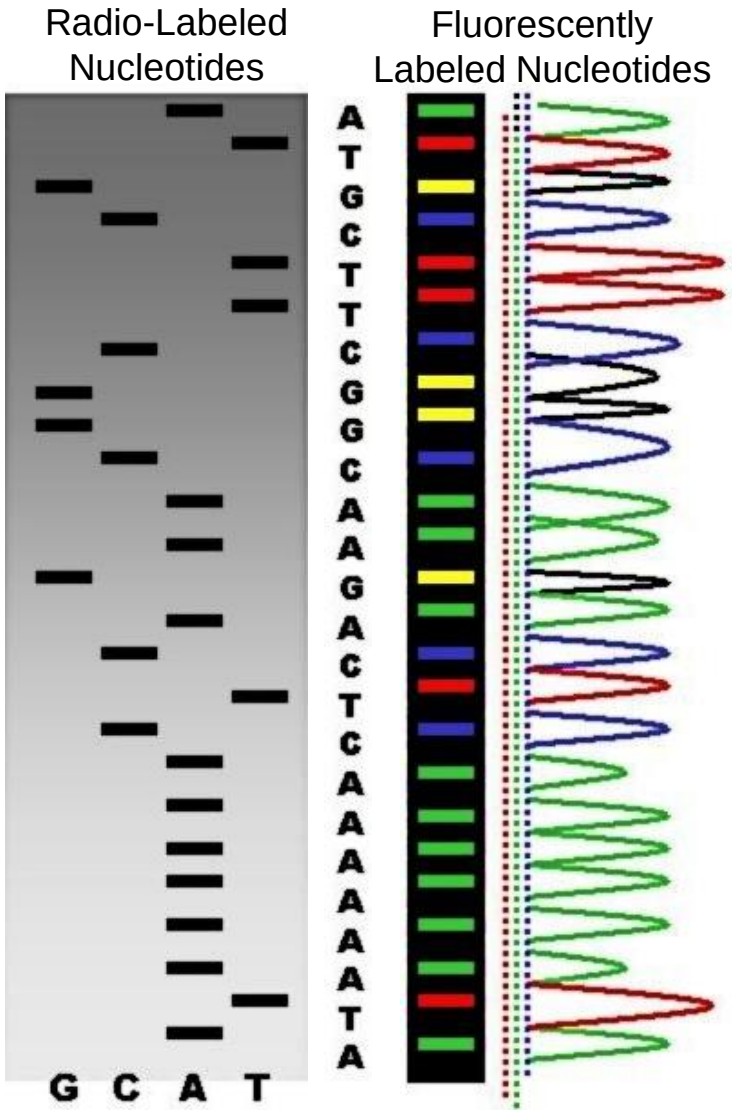
Event	Date & Time	Speaker	Topics
RNA-Seq 101	Wed, Oct 9 1 pm ET	Jean Jasinski, PhD Field Application Scientist	• How Does RNA-Seq Differ from DNA-Seq? • What is Strand Specific RNA-Seq and How Does it Work? • What is the Value of Targeted vs. Whole Transcriptome RNASeq?
Methyl-Seq 101	Wed, Oct 9 4 pm ET	Alex Siebold, PhD Field Application Scientist	• Methylation Mechanisms and Significance • Review of Comparative Technologies • Introduction to Methyl-Seq
NGS Data Analysis 101	Thu, Oct 10 1 pm ET	Jean Jasinski, PhD 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	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

Topics for Today's Presentation

1. What is Next-Gen Sequencing?
2. The NGS Library Prep Workflow
3. Whole Genome vs Targeted NGS
4. Reviewing NGS Terminology
5. Summary & Upcoming 101 eSeminars



What is Next-Gen Sequencing?
A Brief History



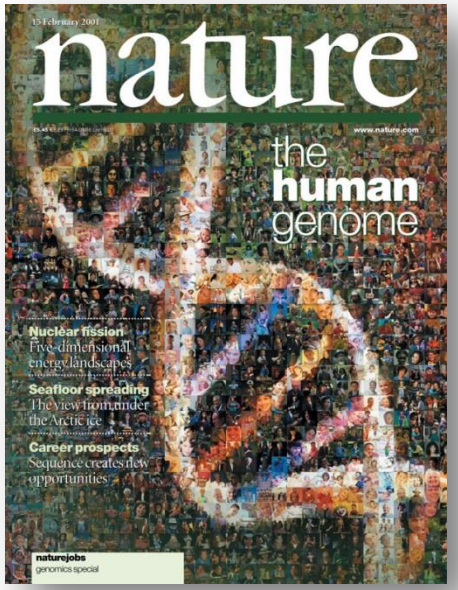
- Frederick Sanger (Sanger Sequencing)
  - "First Generation" (circa 1977)
    - Radiolabeled Nucleotides
    - Sequencing Gels
- Automated Capillary Electrophoresis
  - "Second Generation"
    - ABI 370 generate 500 Kilobases/day
      - Thousands of bases (Kb)
    - ABI 3730 generate 2.8 Megabases/day
      - Millions of bases (Mb)
    - Fluorescence based vs radiolabeling
    - Helped drive the Human Genome Project

The Cornerstone Driving Next-Gen Sequencing Technology

Research: 10 years

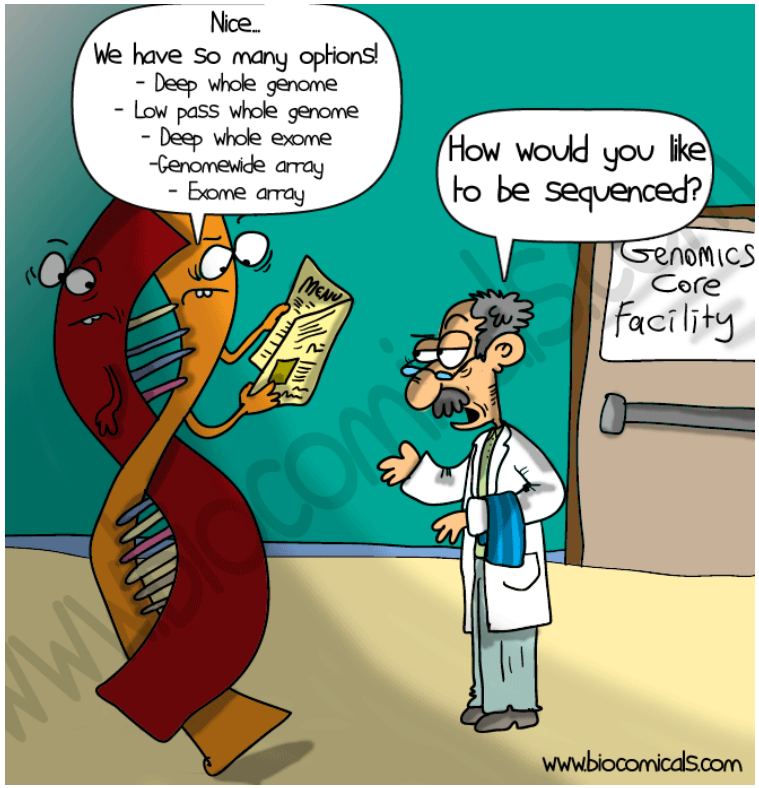
Cost: ~ $3 Billion

Completing The Human Genome…



…Priceless

What is Next-Gen Sequencing?
A Brief History



- Massively Parallel Sequencing
  - "Next-Generation Sequencing" (NGS)
    - Does not use Sanger method
    - Different Platforms = Different Chemistries
    - Very High throughput instruments
      - >100 gigabases of DNA sequence/day
- Desktop Sized Sequencing Instruments & Beyond!
  - "Next-Next Generation Sequencing"
    - Scaled down
    - Medium throughput
    - Individual Labs vs Core Facilities
- Some food for thought:
  - What will sequencing be like 5, 10, 15 years from now?

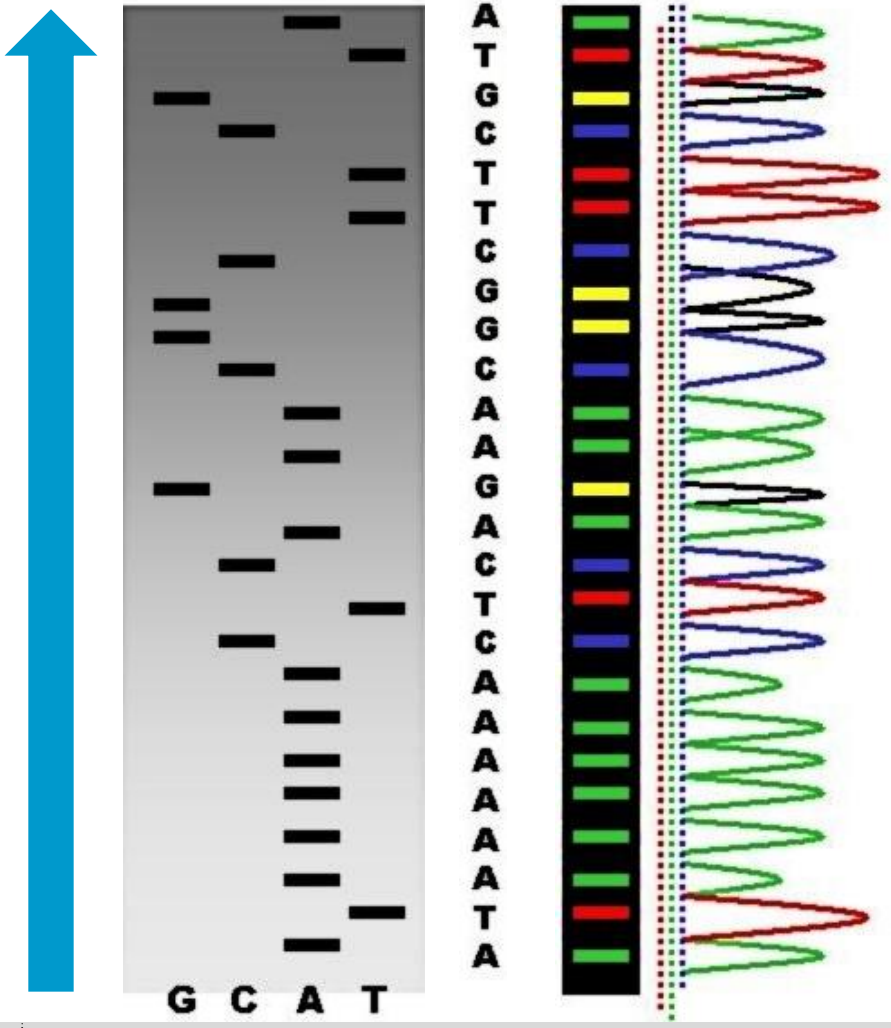
What is Next-Gen Sequencing:
Sanger Sequencing vs Next-Gen Sequencing

"Single" Read System/Run (i.e. 1 DNA Fragment)

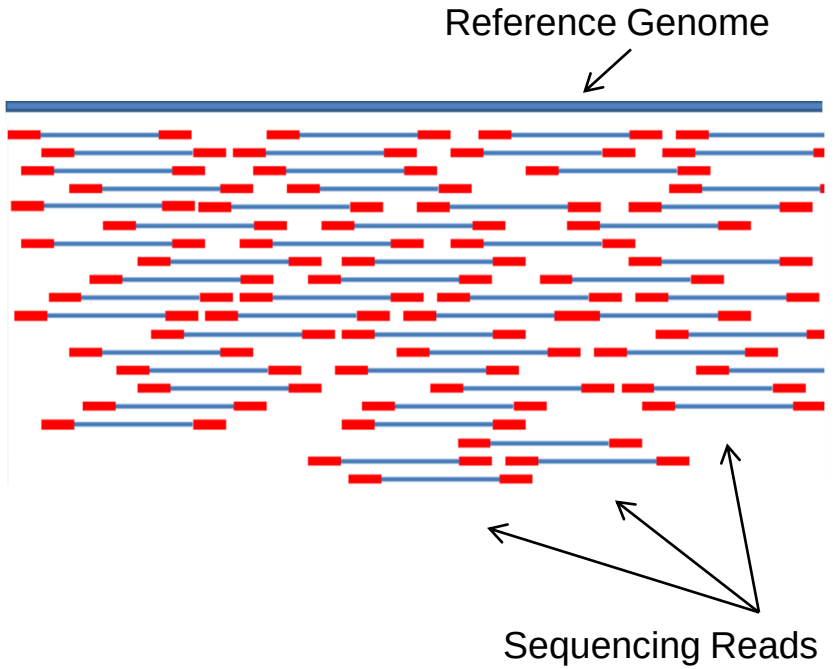
"Multi" Read System/Run (i.e. Thousands of Fragments)

Radio-Labeled Nucleotides

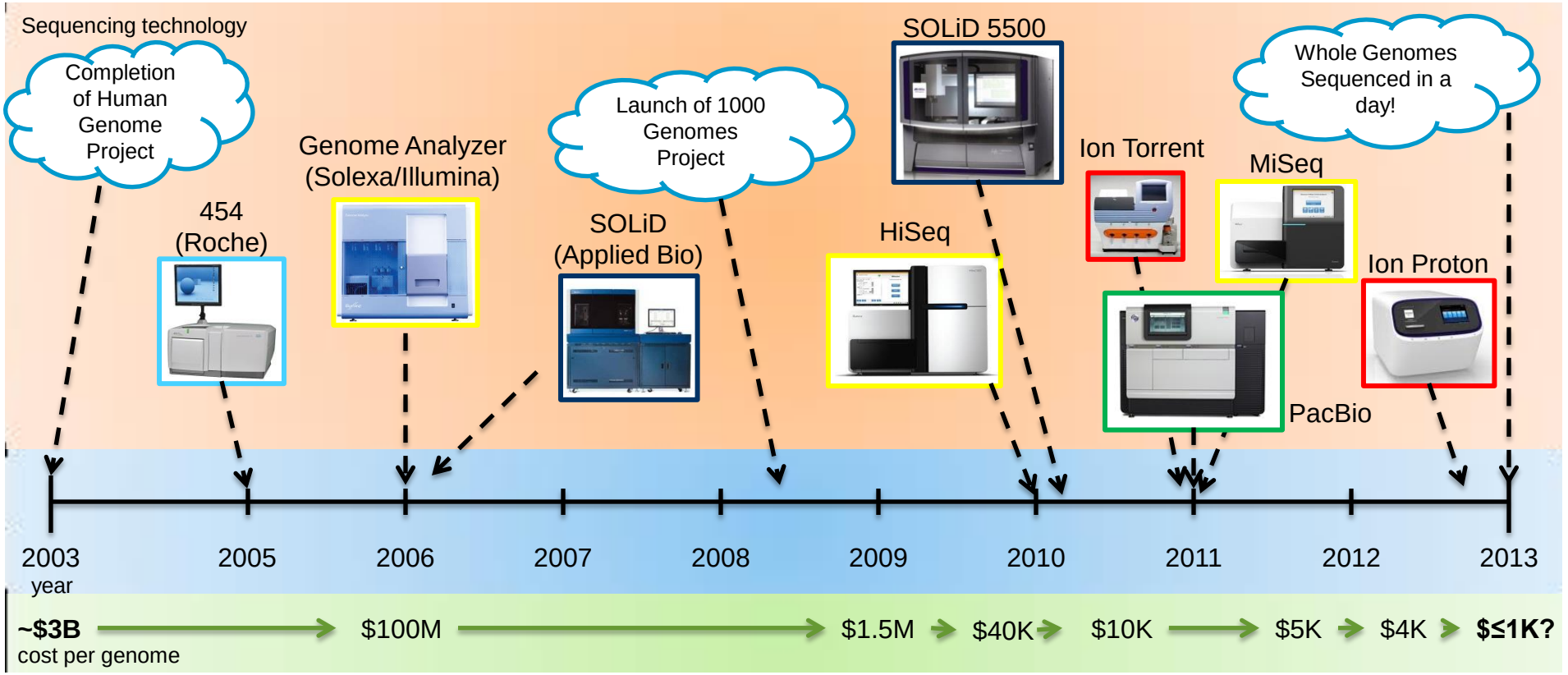
Fluorescently Labeled Nucleotides



Fluorescently labeled nucleotides of many different DNA fragments being sequenced in parallel

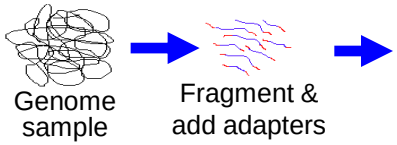


Next-Gen Sequencing Cost & Technology Timeline…



Next-Gen Sequencing Platforms: System Overview

Target Preparation

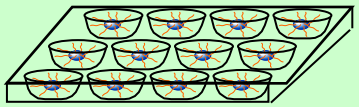
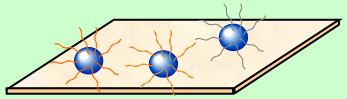
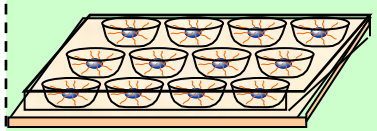


454/Roche

Illumina HiSeq/MiSeq

SOLiD (Life Tech)

Ion Torrent Ion Proton (Life Tech)

Target Amplification	Sequencing Format	Sequencing Chemistry & Imaging
On-bead emulsion PCR *(Clonal)*	 **Bead in defined microwell**	▪**A,G,C,T cycle controlled fluidics** ▪Sequential nucleotide extension - "pyrosequencing" ▪1-color bioluminescent imaging
Array-based "Bridge-PCR" *(Clonal)*	 **Random cluster on surface**	▪**A,G,C,T cycle controlled fluidics** ▪Sequential nucleotide extension ▪4-color fluorescence imaging ▪Flow Cells
On-bead emulsion PCR *(Clonal)*	 **Random bead on surface**	▪**A,G,C,T, cycle controlled fluidics** ▪Sequential 8mer ligation ▪4-color fluorescence imaging
On-bead emulsion PCR *(Clonal)*	 **Beads in defined microwell**	▪**A,G,C,T, cycle controlled fluidics** ▪Sequential nucleotide extension ▪Detects H^+ Ions ▪Semiconductor Chip Sequencing

Next-generation platforms have common elements and workflow

Great Places to Learn More About Sequencing Platforms!

1. PubMed – There are TONS of papers out there that review/compare NGS sequencing platforms

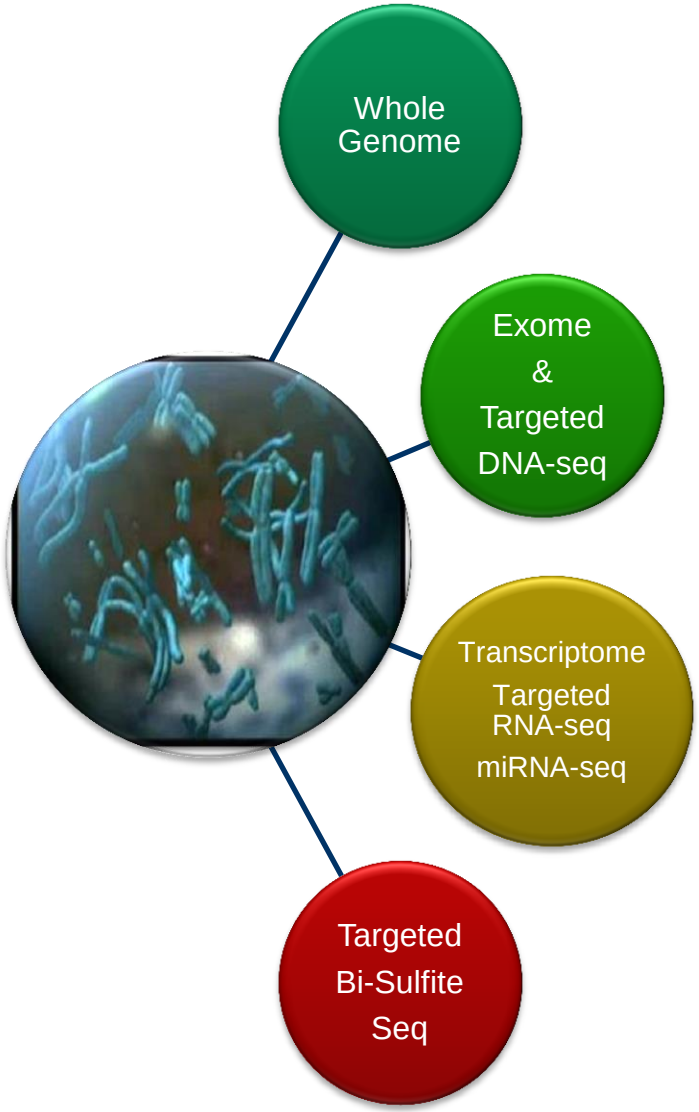
2. Some Favorite Websites:

   - www.youtube.com/watch?v=PMIF6zUeKko : Fantastic seminar of NGS technologies presented by Elaine Mardis, Ph.D., Genome Institute at Washington University in St. Louis.

   - www.SeqAnswers.com : Great message board for learning about & troubleshooting sequencing related topics.

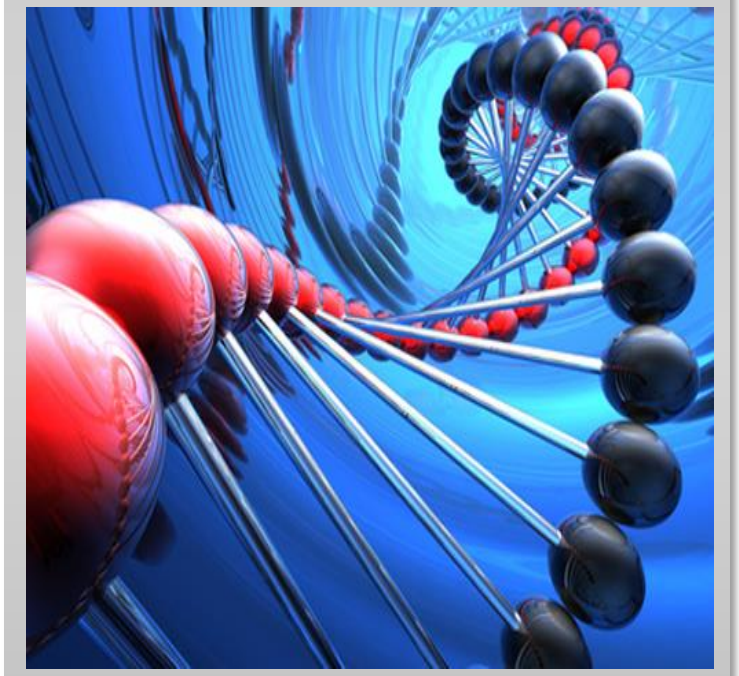
What can you do using NGS Technology: Applications for Basic and Clinical Research

Types of Variants Detectable using NGS
Large amplifications
Large deletions
Point mutations (SNP)
Insertions/Deletions
Inversions
Translocations
Copy number (CNV)
Fusions/splice variants
Gene expression data
Methylation status



Topics for Today's Presentation

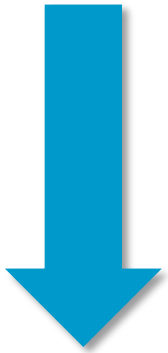
1. What is Next-Gen Sequencing?
2. The NGS Library Prep Workflow
3. Whole Genome vs Targeted NGS
4. Reviewing NGS Terminology
5. Summary & Upcoming 101 eSeminars



Overview of the NGS Workflow

Can be a 2 step or 3 step process…

Library Prep



Sequencing

Whole Genome
Whole Transcriptome

Library Prep

Target Enrichment
(subset of the initial library)

Sequencing

All Exons
Smaller # of Genes/Regions
(i.e. Sequencing Panels)

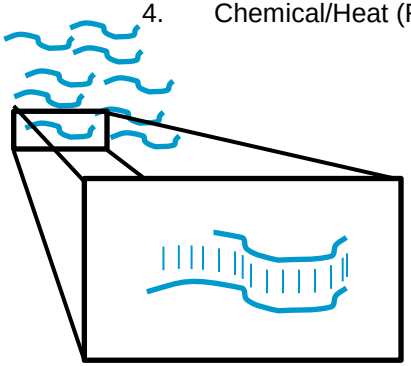
Learning the NGS Workflow: Generating a Sequencing Library

1. **Library** - A collection of DNA or cDNA fragments prepared for sequencing by a performing a series of enzymatic steps. These steps are commonly referred to as the **Library Prep.**

Isolate gDNA

Shearing/Fragmentation

1. Sonication **(**Covaris**)**
2. Restriction Enzymes
3. Transposons
4. Chemical/Heat (RNA-seq)

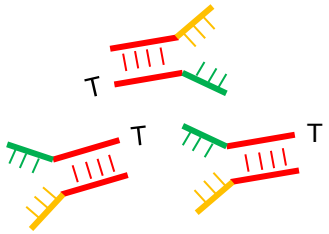


End Repair

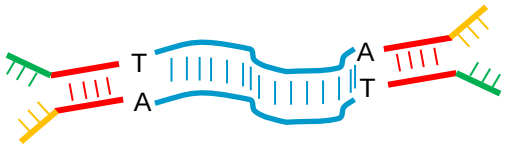
- Generate blunt end fragments

A-Tailing

- Add an “A” base to 3’ end of each strand

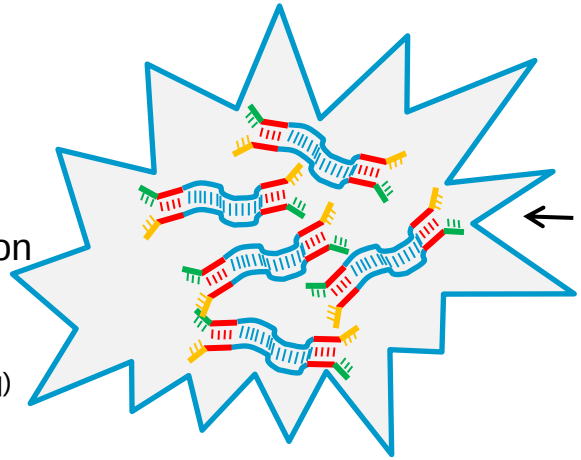


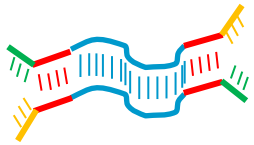
Y-Shaped Sequencing Adapters



Adapter Ligation

- **Adapters** are short DNA oligos that contain the primer sites used by the sequencer to generate the sequencing read
- Adapters can also contain short 6-8bp sequences called **indexes** or **barcodes**
- Incorporating barcodes allows different samples to be combined in the same sequencing run (**multiplexing**)

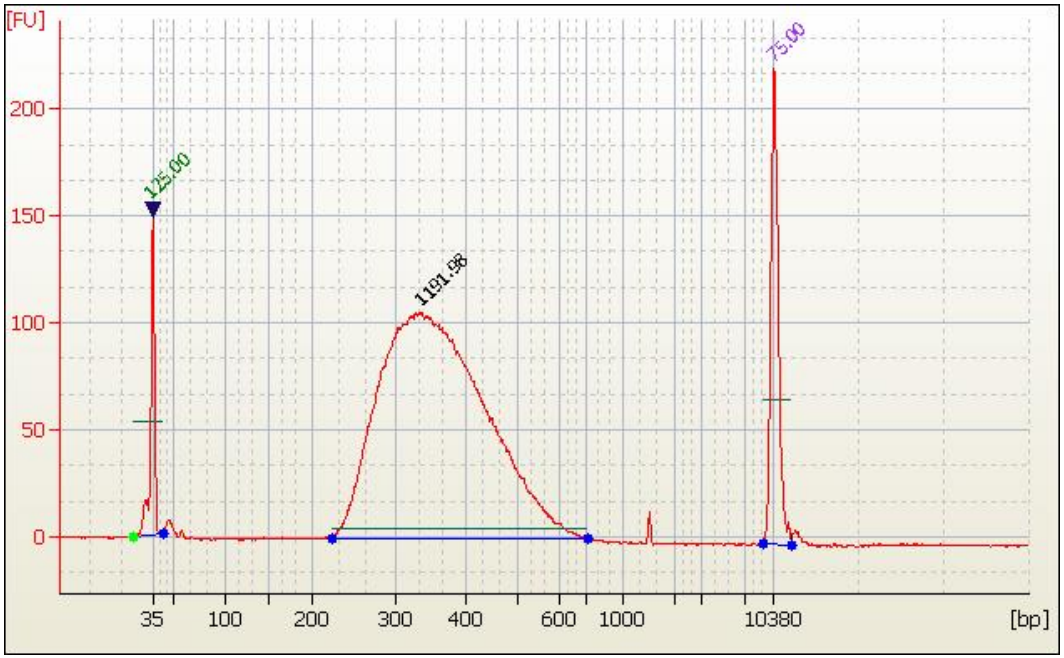




PCR

- Using PCR primers complementary to the adapters, DNA fragments with properly ligated adapters are selected for and amplified

Agilent's BioAnalyzer/Tapestation are frequently used for Quality Control of Sequencing Libraries



Electropherogram (i.e. trace) for a Standard Library Before or After Undergoing Target Enrichment

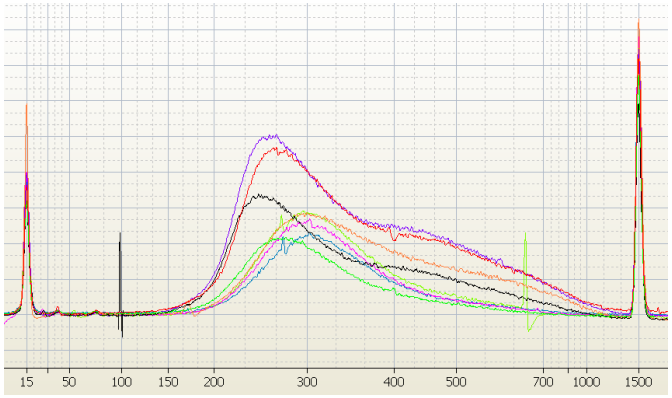
Agilent SureSelect

Illumina TruSeq

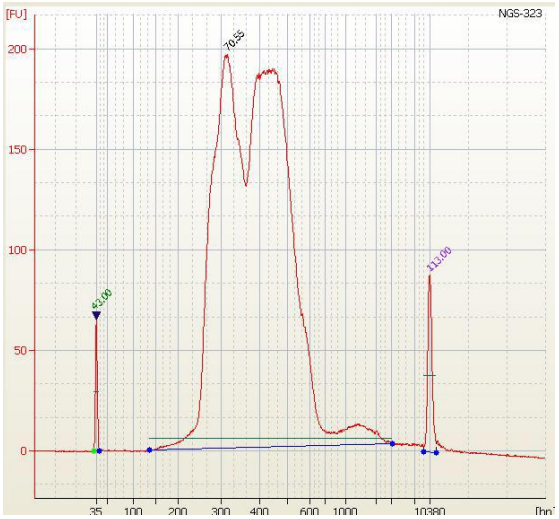
KAPA

NEB

NuGen etc…

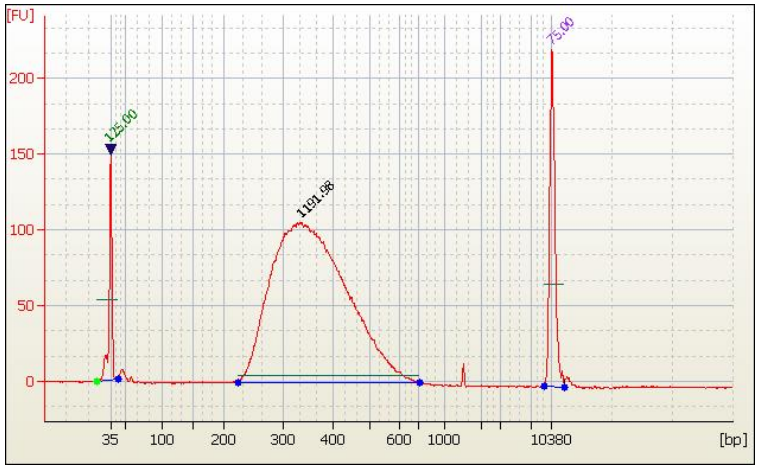


Over-Amplified: Reduce PCR

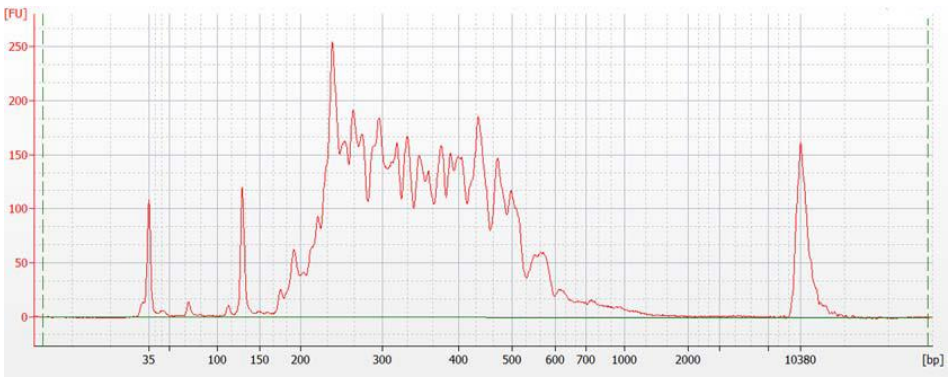


Over-loaded: Dilute and re-run

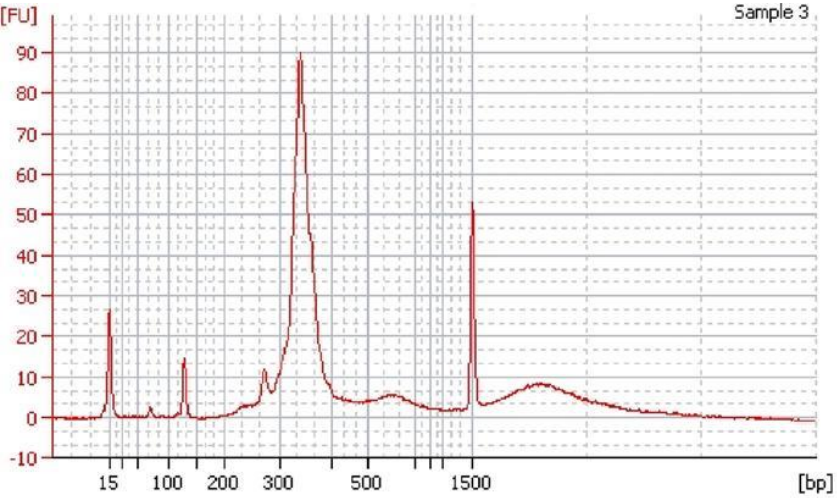
Different Library Preps Generate Different BioAnalyzer Traces



Agilent SureSelect Library Prep

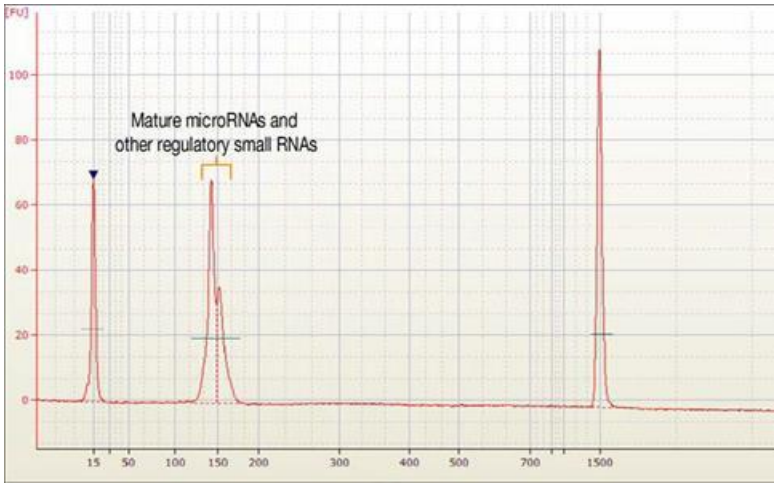


Agilent Haloplex Library Prep



TruSeq Custom Amplicon Library

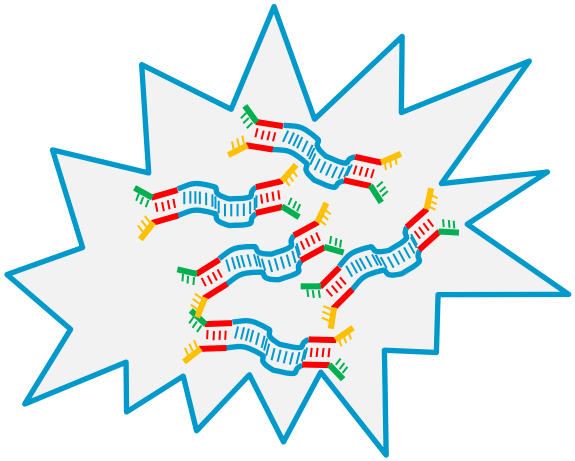
(adapted from Illumina protocol)



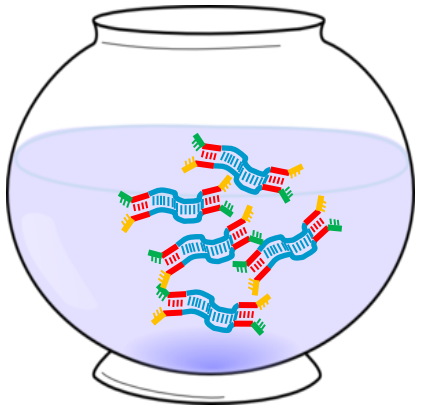
TruSeq Small RNA Library Prep

(adapted from Illumina Protocol)

So you've made a library….now what?



Sequence It!



Perform Target Enrichment

Target Enrichment: It's just like fishing…

Why perform target enrichment?

1. Sequence only your desired regions of interest (Exons, gene panels, intergenic regions etc...)!

2. Sequence more samples per lane/run (i.e. **Multiplex**)

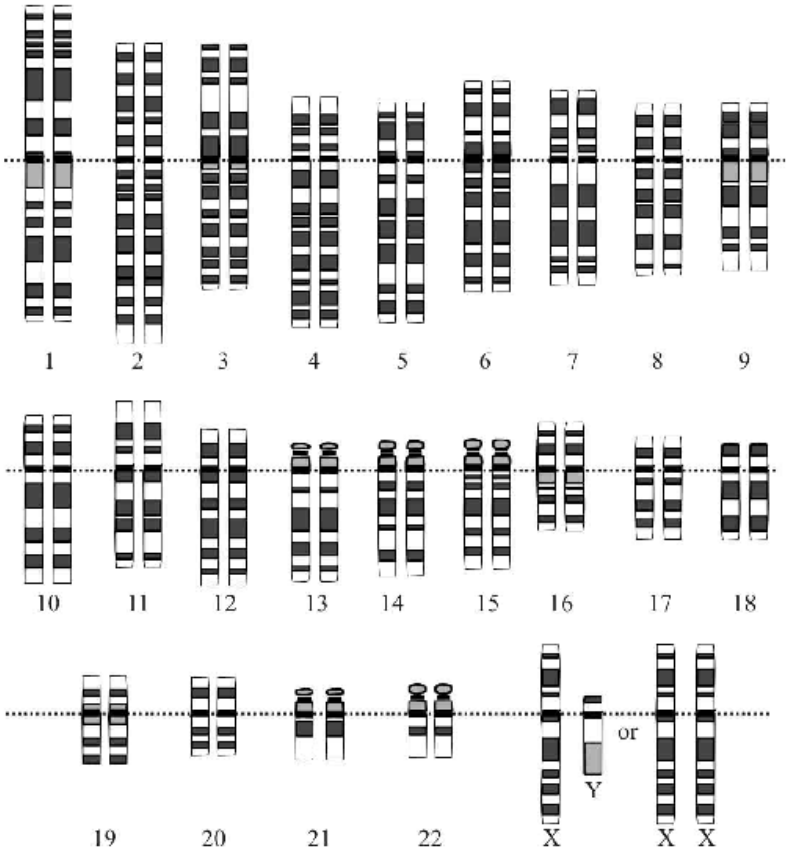
3. Save time and money

4. Faster time to results = Smaller datasets

5. Identify variants in samples with increased reliability and accuracy:
   More **Reads** in regions of interest =
   Higher **Depth of Coverage**

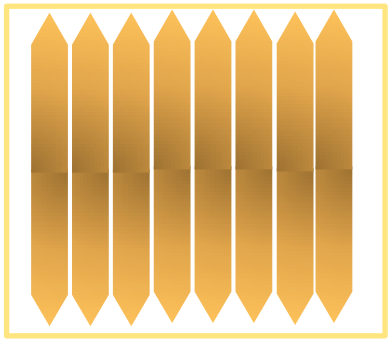
Target Enrichment Maximizes Your Sequencing Efficiency

Genome Size x Desired Depth of Coverage = Required Seq Depth/Sample



Human Genome
3Gb x 30 = 90Gb

Illumina HiSeq FlowCell

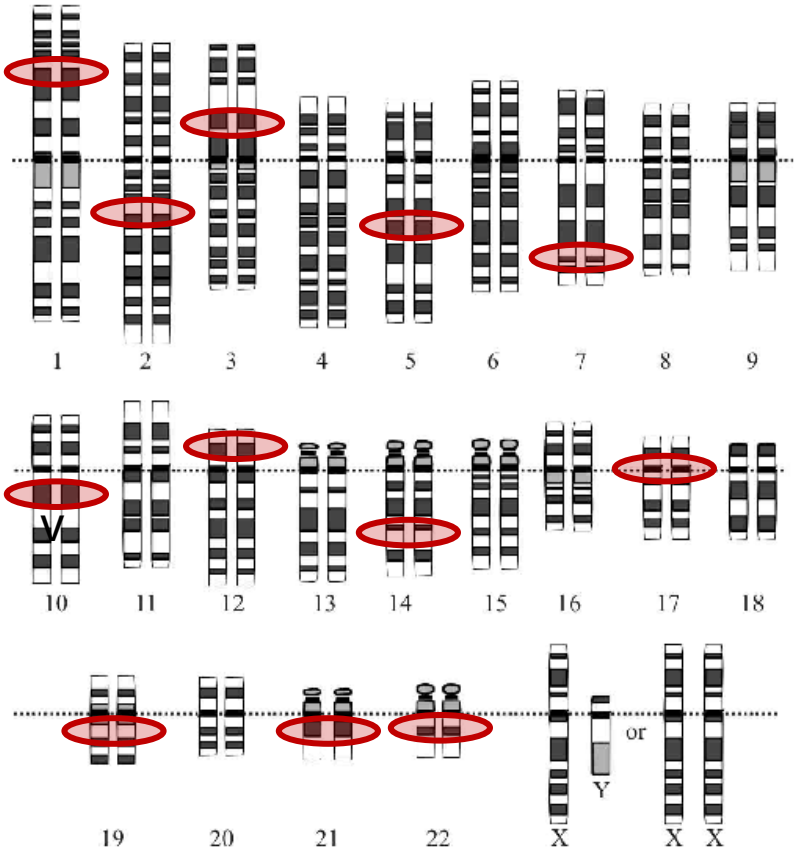


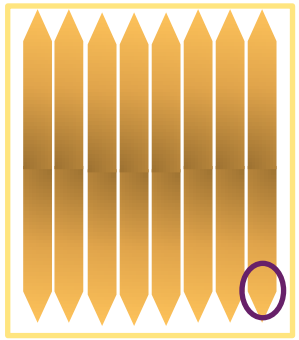
Illumina HiSeq 2000 37Gb/lane

~ 3 lanes per sample!
$$$$$

Target Enrichment Maximizes Your Sequencing Efficiency

Target Size x Desired Depth of Coverage = Required Seq Depth/Sample





Target = 50Mb x 100 = 5Gb

Target = 5Mb x 100 = 500Mb

Target = 500Kb x 100 = 50Mb

Target = 50Kb x 100 = 5Mb

Develop designs/panels for any sequencing capacity:

- High Throughput or Desktop

General Methods of Target Enrichment:

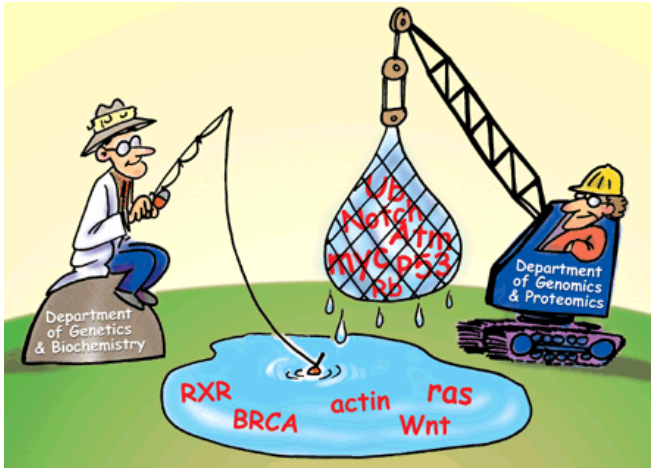
What is the basic concept?

1. Pull out the genes/regions of interest that you care about sequencing
   A. Capture the regions using biotinylated **baits**:
      - **In-solution hybrid capture**

   B. Use primers to selectively amplify the genes/regions you want to sequence:
      - **Amplicon sequencing**

2. Regions that are captured/amplified from initial library (i.e. **pre-capture library**) undergo additional amplification and processing creating a **post-capture library**

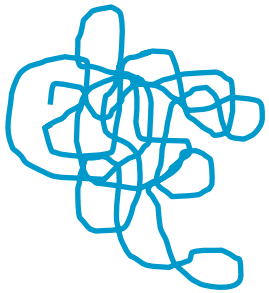
3. Off to sequencing!



(Adapted from www.sciencemag.org/cgi/content/full/291/5507/1221/F1)

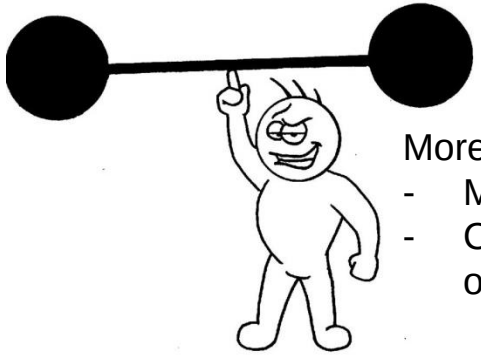
Learning the NGS Workflow: General Comparisons of Target Enrichment Methods

In-Solution Hybridization Capture



gDNA
- Micrograms
- Hundreds of nanograms
- Tens of nanograms ?

Typically Slower hyb time range: 3-72hrs



More Robust Data:
- Many unique reads
- Can find large variety of DNA aberrations

Amplicon Sequencing

gDNA
- Tens of nanograms
- And less…

Typically Faster (no hyb required)

Good but Limited Data:
- Few/No unique reads
- Best for small/point mutations

Six Things to consider beforehand…. Reviewing the NGS Library Prep Workflow

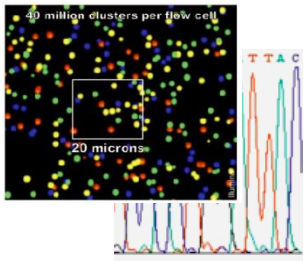
1. **What kind of sample am I using and how much do I have?**
   - High quality gDNA from cells or fresh/frozen tissue?
   - Degraded gDNA from Formalin Fixed Parafin Embedded Blocks (**FFPE**)?
   - Do you have micrograms, nanograms, picograms?
2. **What do I want to learn from the samples I prepare?**
   - Identify single nucleotide polymorphisms/variants (**SNPs/SNVs**)
   - Insertions and/or deletions (**InDels**)
   - More complex rearrangements: Translocations, Inversions, Copy # Variations (**CNVs**)
3. **One-size doesn't fit all**
   - Library preps & Target Enrichment technologies come in several flavors with their own sets of advantages & disadvantages
   - Know what to look for during the QC steps of a given library prep workflow
   - Do your homework & make a decision that meets *your* needs

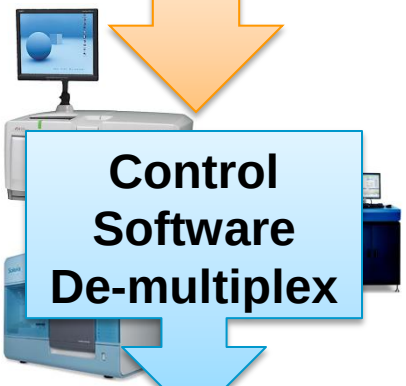
Six Things to consider beforehand Reviewing the NGS Library Prep Workflow

4. **Higher yields aren't necessarily better**
   - Especially when performing non-Amplicon based library preps…
   - The less PCR, the better the results (higher **Library Complexity**)
   - Over-amplification can lead to poor performance
5. **Set your expectations accordingly**
   - Poor quality and very low input starting materials may require special handling
   - More input required, Whole Genome Amplification
   - Results from high quality gDNA ≠ Results from FFPE gDNA
6. **Don't be afraid to ask for help!**
   - While sequencing costs have come down, it's still not cheap!
   - Reach out to your sequencing cores, other labs, or vendors for guidance

NGS 101 Analysis Teaser: What happens after the library is sequenced?

Primary





Control Software De-multiplex

FASTQs

"Raw" Data Files

Secondary

FASTQs

(Reads + Quality)

Trimming & Aligning Tools

BAM/SAM Files

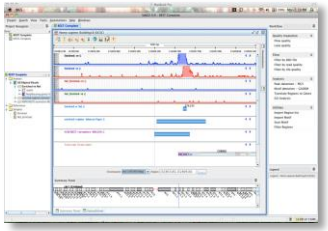
Reads aligned to genome

Tertiary

BAM/SAM Files

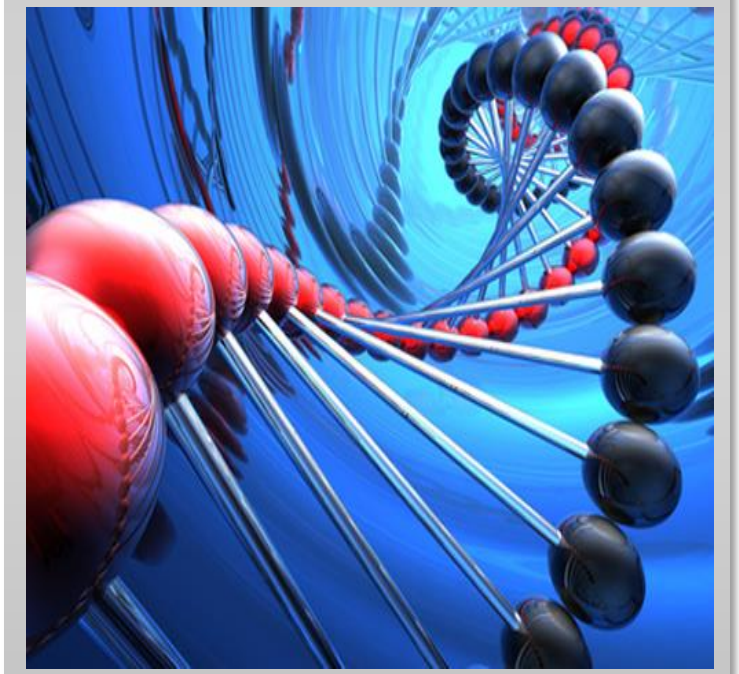
GeneSpring SureCall, Partek, Open source tools

Mutations/Variants Identified



Topics for Today's Presentation

1. What is Next-Gen Sequencing?
2. The NGS Library Prep Workflow
3. Whole Genome vs Targeted NGS
4. Reviewing NGS Terminology
5. Summary & Upcoming 101 eSeminars



Understanding Reads…
Types of Reads, Lengths of Reads, Depths of Reads

- **Single-End Reads**: Provide sequence from one end of a DNA insert
- **Paired-End Reads**: Provide sequence from both ends of a DNA insert.
  - Provides improved alignment of sequencing data
  - Better detection of chromosomal rearrangements: insertions/deletions/translocations and fusions.

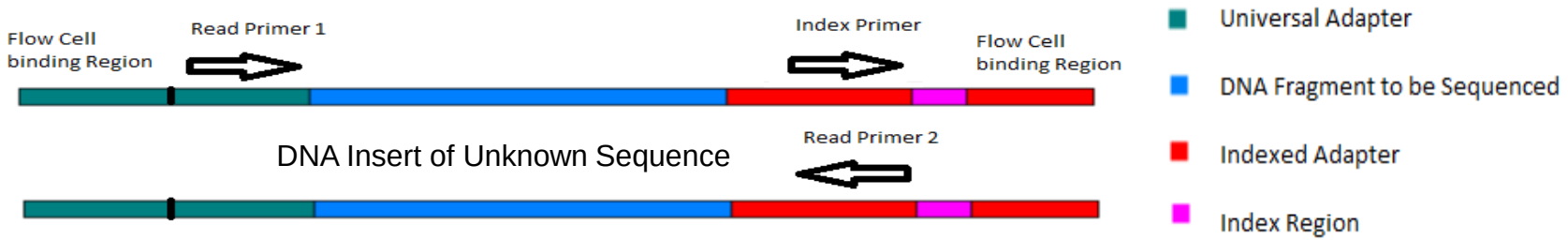


Figure Adapted from: tucf-genomics.tufts.edu

- **Mate Pair Reads**: Similar to paired-end but both reads come from a single strand of the DNA insert and the distance between the reads is often much greater.

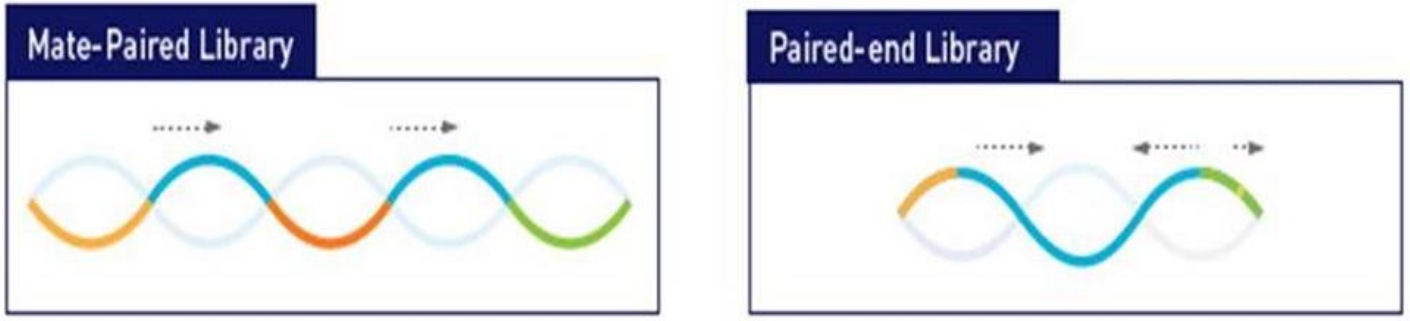


Figure Adapted from: GeneSpring NGS User Manual

Understanding Reads…
Types of Reads, Lengths of Reads, Depths of Reads

Read lengths vary across sequencing platforms:

- Short reads – Illumina, Ion Torrent/Proton, SOLiD
  - <100bp (ex. 1 x 36bp, 2 x 50bp, 1 x 75bp)
- Medium reads – Illumina, Torrent/Proton, Roche 454
  - >100bp but <1000bp (ex. 2 x 100bp, 2 x 150bp, 1 x 400bp, 1x 600bp)
- Long Reads – Roche 454, Pacific Biosciences (PacBio)
  - >1000bp (ex. 1x1000bp, >10,000bp (Avg. 3000-5000bp))

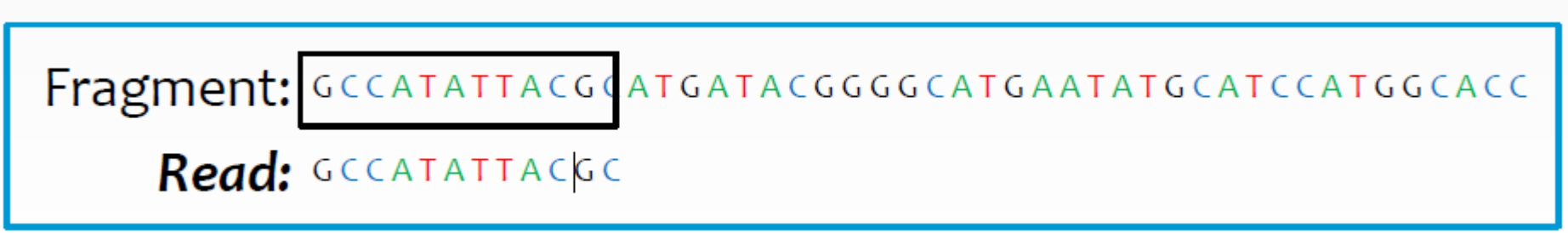


Figure Adapted from Ambry Genetics

Understanding Reads…
Types of Reads, Lengths of Reads, Depths of Reads

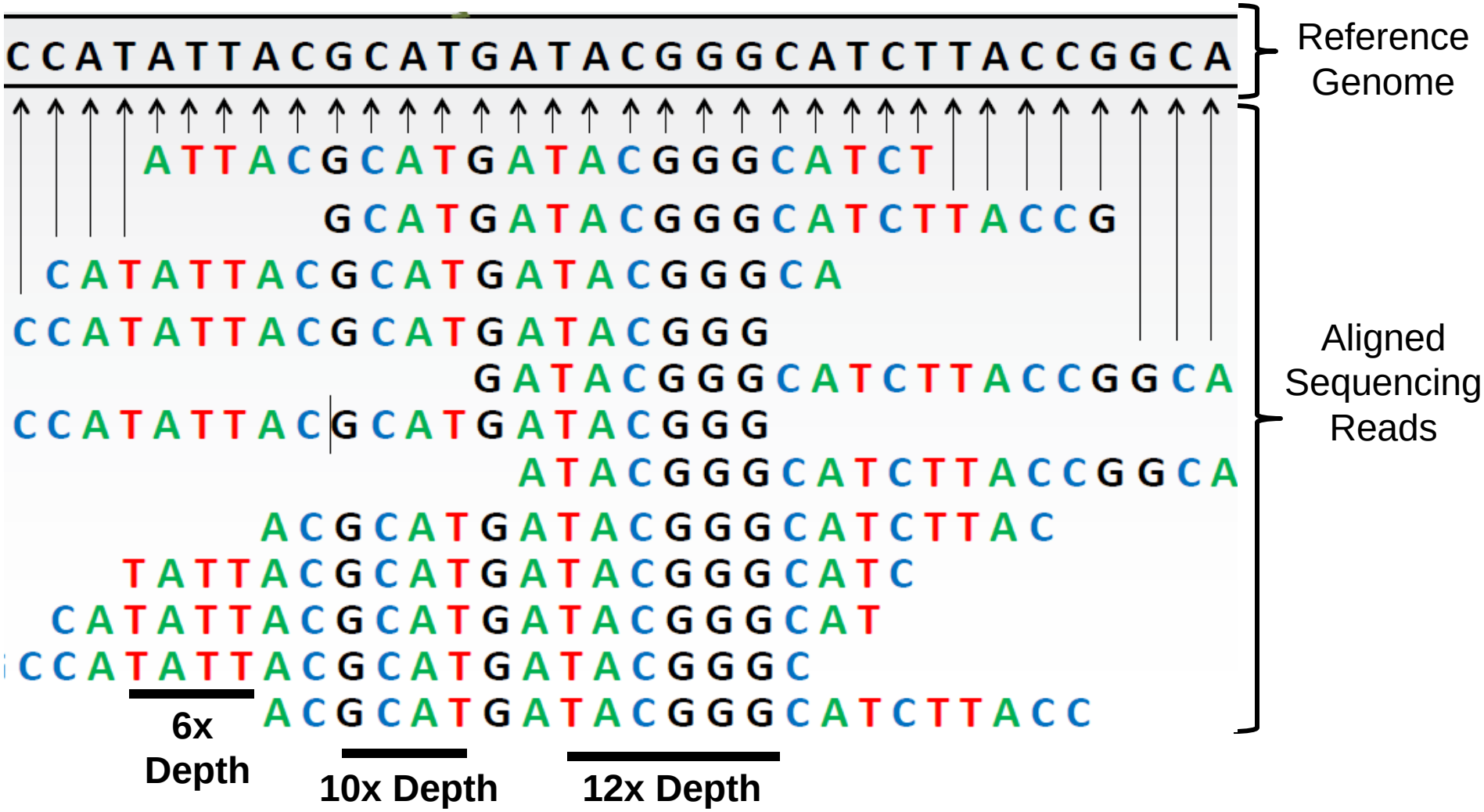


Figure Adapted from Ambry Genetics

A Beginner's Next-Gen Glossary: Walk the walk, talk the talk

1. **Library Preparation (Library Prep)** – The method(s) used to prepare DNA or RNA for next-generation sequencing.

2. **Sequencing Library (Library)** – A collection of DNA or cDNA fragments of a given size range with adapters ligated to each end that can be run through a sequencer. Libraries can be DNA or cDNA (cDNA libraries prepared when performing RNA-seq).

3. **Adapters** – Oligonucleotides of a known sequence that are ligated to each end of a DNA/cDNA fragment (i.e. insert). They provide the primer sites used for sequencing the insert.

4. **Index/Barcode** - Short sequences of typically 6 or more nucleotides that serve as a way to identify/label individual samples when they are sequenced together in a single sequencing lane/chip. Barcodes are typically located within the sequencing adapters.

5. **Multiplexing** – Mixing two or more different samples together such that they can be sequenced in a single sequencing lane or chip. Samples that are to be combined, need to be barcoded/indexed prior to being mixed together.

6. **Target Enrichment (Capture)** – Methods to allow one to isolate and/or increase the frequency of specific genes or other regions of interest from a DNA or cDNA library prior to being sequenced. The regions of interest are retained for sequencing and the remaining material is washed away.

7. **Baits** – Common name given to the oligonucelotide sequences (i.e. probes) that are responsible for identifying and binding to a given region of interest for performing target-enrichment.

8. **In-Solution Capture** – A method of performing target enrichment that requires samples to be hybridized to baits to select and enrich the sample for the desired regions of interest.

9. **Amplicon Sequencing** – A method of performing target enrichment that utilizes one or more pairs of PCR primers to increase the number of copies of the genes or other regions of interest that will ultimately be sequenced.

10. **Gene Panels** – Name frequently given to the selected regions of interest (this can genes or intergenic regions) that will be captured using some form of target-enrichment technology.

A Beginner's Next-Gen Glossary: Walk the walk, talk the talk

11. **Pre-Capture Library** – Common name given to the sequencing library that is created before that library undergoes some form of target-enrichment.

12. **Post-Capture Library** – Common name given to the sequencing library after it has completed some form of target-enrichment.

13. **Read** – Base pair information of a given length from a DNA or cDNA fragment contained in a sequencing library. Different sequencing platforms are capable of generating different read lengths.

14. **Single End Read** – The sequence of the DNA is obtained from the 5' end of only one strand of the insert. These reads are typically expressed as 1x "y", where "y" is the length of the read in base pairs (ex. 1x50bp, 1x75bp).

15. **Paired End Read** – The sequence of the DNA is obtained from the 5' ends of both strand of the insert. These reads are typically expressed as 2x "y", where "y" is the length of the read in base pairs (ex. 2x100bp, 2x150bp).

16. **Mate Pair Read** – The sequence of the DNA is obtained similar to paired-end reads, however the size of the DNA insert is often much greater in size (2-10kb in length) and the paired reads originate from a single strand of the DNA insert.

17. **Depth of Coverage** – The number of reads that spans a given DNA sequence of interest. This is commonly expressed in terms of "Yx" where "Y" is the number of reads and "x" is the unit reflecting the depth of coverage metric (i.e. 5x, 10x, 20x, 100x)

18. **Sequencing Depth** – The amount of sequencing a given sample requires to achieve a certain depth of coverage. This is frequently expressed as the number of reads a sample requires (ex. 40 million reads, 80 million reads) or the number of bases of sequencing a sample requires (ex. 4 gigabases, 100 megabases).

19. **Library Complexity** – The number of unique DNA fragments contained in a sequencing library.

20. **Electropherogram** – A graphical representation of the size and quantity of a DNA or RNA sample run through a BioAnalyzer, TapeStation or other instrument used for performing quality control.

21. **FFPE DNA/RNA** – Formalin Fixed Parafin Embedded DNA or RNA. When attempting to prepare sequencing libraries from these sample types, modifications are often required to standard library preparation protocols to accommodate the level of DNA/RNA degradation commonly found from samples stored using this technique.

A Beginner's Next-Gen Glossary: Walk the walk, talk the talk

22. **Call -** Referring to the identification of a given aberration detected in the sequenced sample when compared to the reference/normal genome.

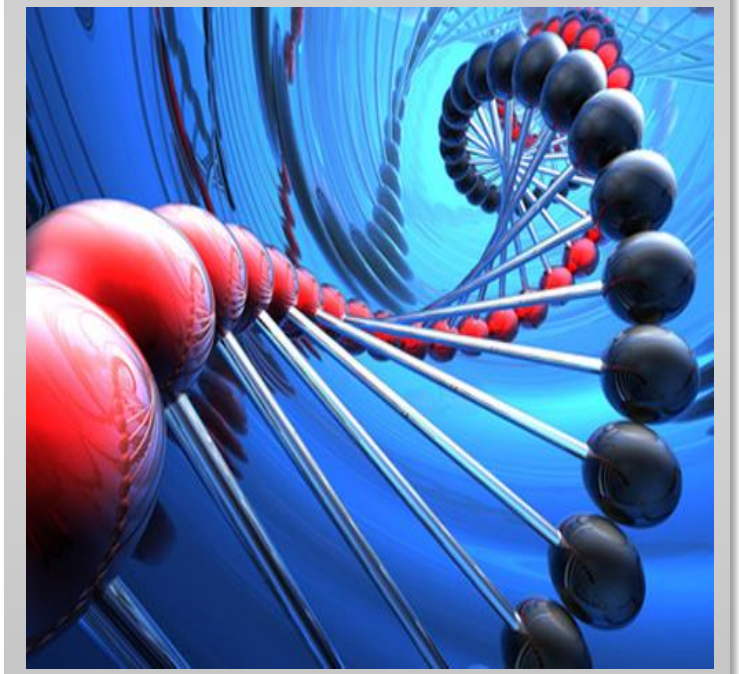
23. **SNP/SNV** – Referring to a Single Nucleotide Polymorphism or Single Nucleotide Variant detected in a sample.

24. **CNVs** – Referring to Copy Number Variation that is detected in sample.

25. **InDels** – One or more Insertion or Deletion event that is detected in a sample.

Topics for Today's Presentation

1. What is Next-Gen Sequencing?
2. The NGS Library Prep Workflow
3. Whole Genome vs Targeted NGS
4. Reviewing NGS Terminology
5. Summary & Upcoming 101 eSeminars



Agilent Technologies Knows NGS 101 & More… Offering Complete Solutions for NGS Workflows

The Gold Standard for Sample QC

2100 Bioanalyzer Instrument & Kits

2200 TapeStation Instrument & Kits

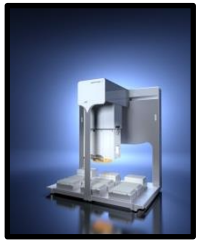
The Leader in NGS Target Enrichment

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

Back to the Basics: Agilent's Five Part 101 eSeminar Series Continues…

Event	Date & Time	Speaker	Topics
RNA-Seq 101	Wed, Oct 9 1 pm ET	Jean Jasinski, PhD 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	Wed, Oct 9 4 pm ET	Alex Siebold, PhD Field Application Scientist	• Methylation Mechanisms and Significance • Review of Comparative Technologies • Introduction to Methyl-Seq
NGS Data Analysis 101	Thu, Oct 10 1 pm ET	Jean Jasinski, PhD 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	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