What Nucleic Acids Are Sequenced???

Phillip McClean
September 2011

DNA is the most typical nucleic acid that is sequenced. Researcher will isolate total DNA, create some sequencing library, and those libraries will be sequenced using the Sanger approach or any of the modern next generation sequence technologies. But that is not the only type of nucleic acid that is sequenced. Often a researcher would like to study just a fraction of the genome rather than all of the DNA in the genome. Two alternate nucleic acid pools are DNA from exons and mRNA from specific tissues.

Exome sequencing is becoming popular. The term exome is refers the study of all of the exons of a genome at the same time. Why would a researcher focus just on the exons when a complete genome sequence project will capture the exons? There are two answers to this question. First, most mutant phenotypes are the result of mutations in exons. Therefore, by focusing preferentially on the exons, important mutations can be studied. Secondly, exome sequencing is more efficient if you goal is to only look for mutations in the coding region. For example, only 3% of the human genome is composed of genes. And further, on average only 1/3 of the gene consists of exons. Thus, just 1% of human DNA is exons. So, if this is your direct nucleic acid of interest, how can a project focus just on the exon fraction of the DNA?

The approach of choice to collect exon DNA is called ***exon capture***. The approach requires a reference genome sequence from which gene models have been built. Gene modeling defines the exon and intron boundaries for each predicted gene in the genome. With this information, the complete sequence of exon is available. The information is use to create a collection of long oligonucleotides whose sequence are highly similar to some part of a specific human exon. NimbleGen is a company that has developed the technology. They produce solution and microarray based systems that are used to capture exons. The SeqCap EZ Exome V2.0 solution system consists of 2.1 million exons that target ~300,000 exons from ~30,000 genes. This averages to about 10 oligonucelotides (=oligos) per exon. It is estimated that this collection of oligos will capture 36.5 Mb of human DNA. The arrary-based system (Sequence Capture 2.1M Human Exome Array) uses ~2.1 million long oligos that target 180,000 exons and 551 micro RNA exons. (NimbleGen also makes custom oligos for solution or array methods.)

To use either system, DNA is first fractionated into small fragments. Those fragments are denatured (made single-stranded) and either added to the solution or the array. Fragments complementary to the oligos are bound to the DNA. The DNA bound to the oligos is recovered, processed further and submitted for next generations sequencing. That sequence data is then used for specific analysis to address whatever question the researcher is trying to address.

RNA-seq (or RNA sequencing) is another approach gaining popularity. Researchers have focused on the RNA fraction since the early days of molecular genetics because it represents all of the genes that are expressed during a specific developmental stage or those genes that are expressed in response to external stimulus (such as biotic or abiotic stresses). The first method used to consider the genes expressed at a specific stages was EST sequencing. This

method required that development of a cDNA library (DNA copies of mRNA sequences) and many clones from the library were sequenced using the Sanger sequencing technique. The challenge was always to capture all of the mRNA in a specific tissue. This was almost never achieved because abundantly expressed mRNAs were predominant in the library while rare mRNAs were underrepresented in the library.

This limitation though has been overcome by using next generation sequencing techniques to collect sequence information for mRNA sequences collected from a specific (or multiple) tissues. The ability to collect such massive amounts of sequence data ensures that all of the mRNA transcripts will be sequenced. Furthermore, it has been shown that the number of copies of a sequence found in sequence collection is proportional to the number of copies that specific gene. Therefore RNA-seq techniques can be used to determine which specific genes are being expressed in a tissue and at what level those genes are being expressed.

Whatever type of nucleic that will be sequenced, it is important to first determine depth sequence depth that is required to address the research question. Often, a researcher needs just a low coverage (limited amount of sequence data). Yet the next generation sequencing procedures produce massive amounts of data. To leverage the large output, **bar coding and pooled sequencing** methods have been introduced. With this approach, during library preparation, different sequence tags are added to the different nucleic acid that will be sequenced. Then multiple libraries are pooled and sequenced. Since each library has its own bar code sequence, any sequence that contains that bar code sequence must come from a specific library. For example, four different libraries could be loaded onto a single lane of the sequencer. Therefore the sequence data from the lane will consist of one-fourth of each of the library. Software has been developed to keep this data separate as it comes off of the sequencer.

What Nucleic Acids Are Sequenced???

DNA - the most common nucleic acid sequenced.

- Isolate total genomic DNA
- Create some sequencing library
- Sequenced using the Sanger approach or a NexGen

Other subclasses nucleic acids are also sequenced

- Studies just a fraction of the genome
    - Two alternate nucleic acid pools
        - Exons
        - mRNA from specific tissues

Exome sequencing

- Ex**ome** - all of the exons of a genome
- Ex**omics** – the study of all of the exons of a genome at the same time

Why focus on the exons ?

- Most mutant phenotypes are the result of mutations in exons.
    - Important mutations can be discovered and studied
    - More efficient if your goal is to look for mutations in the coding region
    - Human genome
        - Only 3% of the genome is composed of genes
        - On average only 1/3 of the gene consists of exons
            - 1% of human genomic DNA is exons

Exon capture - an approach to collect exon DNA

- Requires a reference genome sequence with gene models
  - Gene modeling defines the exon and intron boundaries for each gene
- Long oligonucleotides highly similar to a part of a specific human exon are utilized to capture just exon sequences

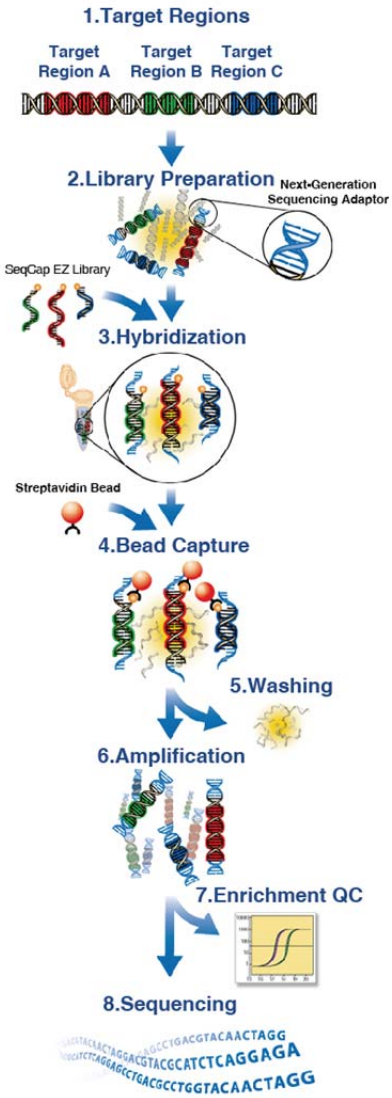
NimbleGen – developer of exon capture technology.

- NimbleGen SeqCap EZ Exome V2.0 solution system
  - 2.1 million oligonucleotides
  - Target ~300,000 exons
  - ~30,000 genes
  - Average - 10 oligonucelotides (=oligos) per exon
  - Captures ~36.5 Mb of human genomic DNA
- NimbleGen Sequence Capture 2.1M Human Exome Array
  - ~2.1 million long oligos
  - Targets 180,000 exons and 551 micro RNA exons
  - Captures ~50 Mb of human genomic DNA

How does exome capture work?

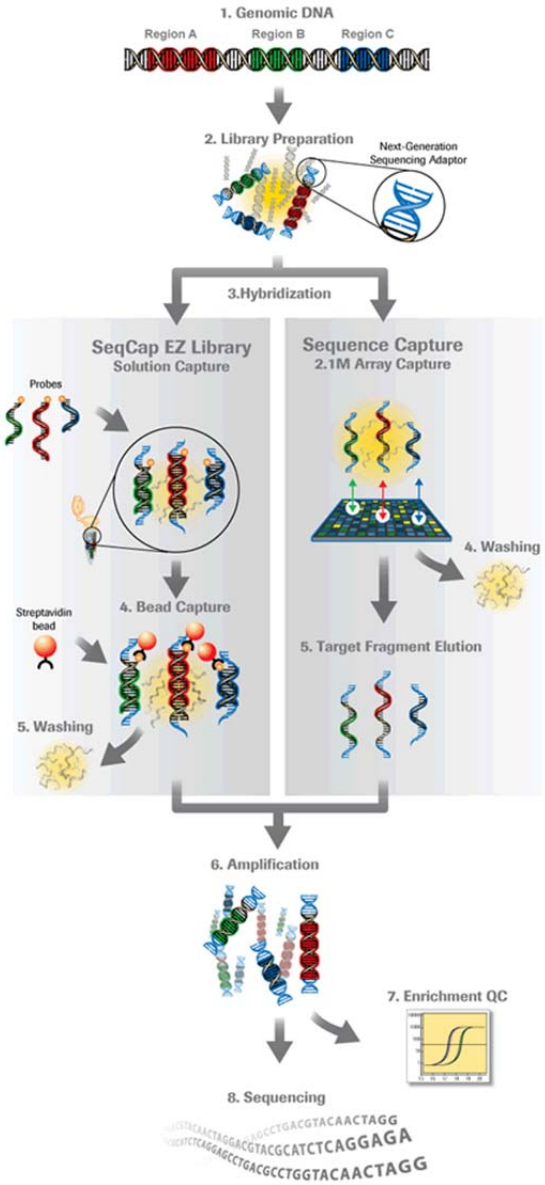
- Fractionated into small fragments
- Denature fragments (made them single-stranded)
- Hyrbridize fragments to oligos in solution or on array
  - Fragments complementary to the oligos are bound to the DNA.
- DNA bound to the oligos is recovered
- DNA sequenced
  - Data analyzed for mutant discovery

Solution Exome Capture



(www. NimbleGen.com)

Array Exome Capture



Sequencing RNA

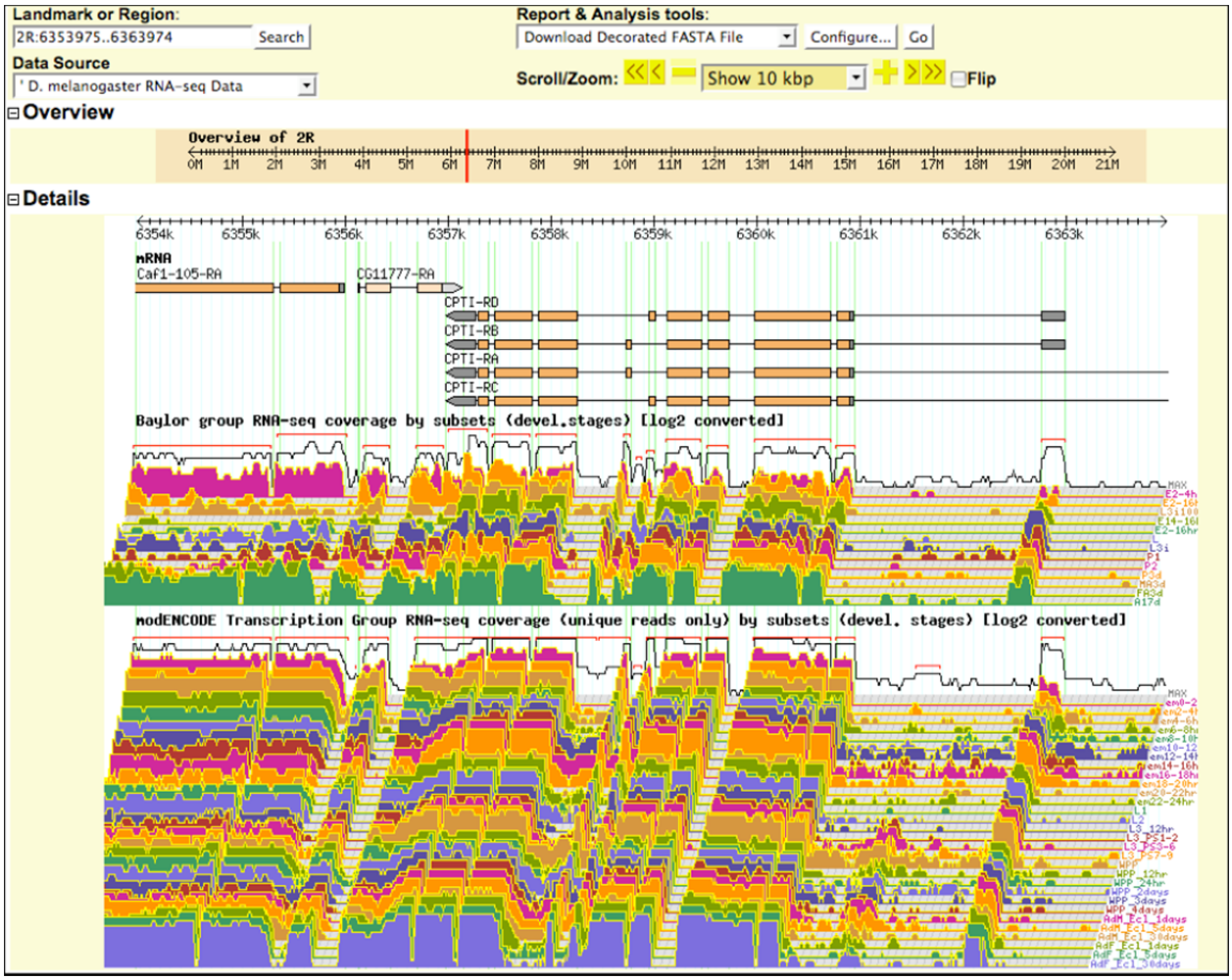
- RNA fraction studied since early days of molecular genetics
    - Why?
        - All expressed genes found in the RNA fraction.
- The first method used to consider the genes expressed at a specific stages was EST sequencing
    - Required that development of a cDNA library (DNA copies of mRNA sequences)
    - Many cDNA clones sequenced using the Sanger sequencing technique.
        - Data = EST, ***expressed sequence tags***
- Challenge - capture all of the mRNA in a specific tissue.
    - Never achieved
        - Abundantly expressed mRNAs were predominant
        - Rare mRNAs underrepresented

RNA-seq (or RNA sequencing)

- Limitation overcome by using next generation sequencing
- Massive amounts of sequence data ensures that all of the mRNA transcripts will be sequenced
    - Copy number of a sequence found in sequence collection is proportional to the number of expressed copies of that specific gene.
    - The expression pattern of specific genes can now be evaluated in detail

RNA-seq results

Notice only exon sequences are represented in the RNA-seq output



How much sequence do you need for a research project?

- Important question to address this research question
  - Low coverage (limited amount of sequence data) may be enough
    - But NexGen sequencing produces massive amounts of data

Bar coding and pooled sequencing

- A method to leverage the large output of NexGen sequencing
  - Cost of NexGen sequencing is spread over many samples
- How is it done?
  - Unique sequence tags added to fragments during library preparation
  - Multiple libraries are pooled and sequenced in a single lane
  - Sequences containing the same tag are evaluated together