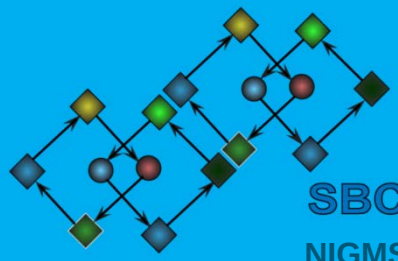


NIGMS funded Center

Experimental Methods in Systems Biology

Part of the Coursera Certificate in Systems Biology

Marc Birtwistle, PhD
Department of Pharmacology & Systems Therapeutics
Fall 2014, Week 1, Types of Measurements Part 1
Nucleic Acids

Icahn School of Medicine at Mount Sinai

Important Features of Any Experiment

- Given a specific question, one can then come up with answers to the following three key properties of the experiment:
  1. What biological system?
     - E.g. Do I look at human cell lines, a mouse, or yeast?
  2. What perturbation/treatment conditions?
     - E.g. What compounds should I apply to the system to elicit a relevant response?
  3. What measurements?
     - E.g. What transcripts do I need to look at, and/or do I need to look at protein levels instead?
- Often (but not always), if you can't design an experiment that only has a handful of conditions and measurements, results may be difficult to interpret
  - Usually the question is too complex or not significant
  - Exceptions are screening based studies, but those also typically have a specific question of interest

Outline

- We attempt to cover many important technologies but cannot possibly cover them all.
- This Lecture—Nucleic Acids
  - Population Average
    - qPCR
    - Microarrays
    - Deep Sequencing Technologies
      - Whole genome seq
      - Exome seq
      - RNA seq
      - Bisulfate seq
      - ChIP seq
  - Single Cell
    - qPCR
    - FISH
    - RNA seq
  - Some techniques can be applied to both population average and single cell
- Next Lecture—Protein and Protein States

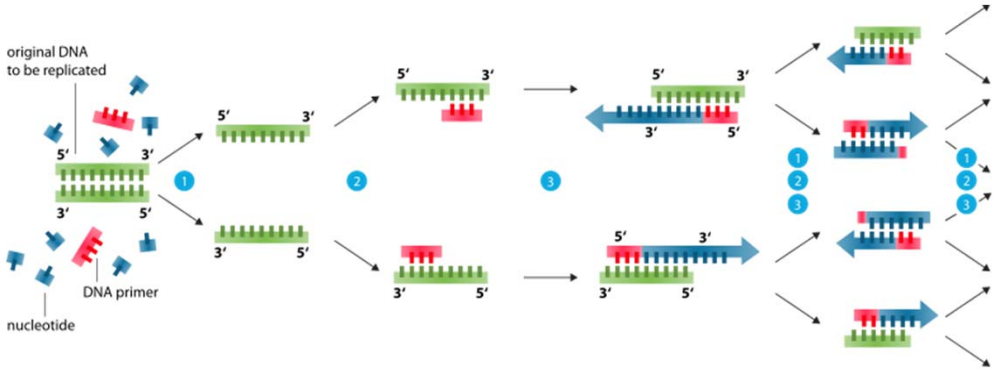
Population Average vs. Single Cell

- Population average techniques require a large amount of starting material.
  - Therefore their output reflects the average over typically ~1 million cells or more

- Single cell techniques are sensitive enough to detect analytes from individual cells.
  - A main challenge here is separating measurement noise from true cell-to-cell variability
  - It is becoming increasingly appreciated that many important phenomena are uncovered by observing at the single cell level

Population Average

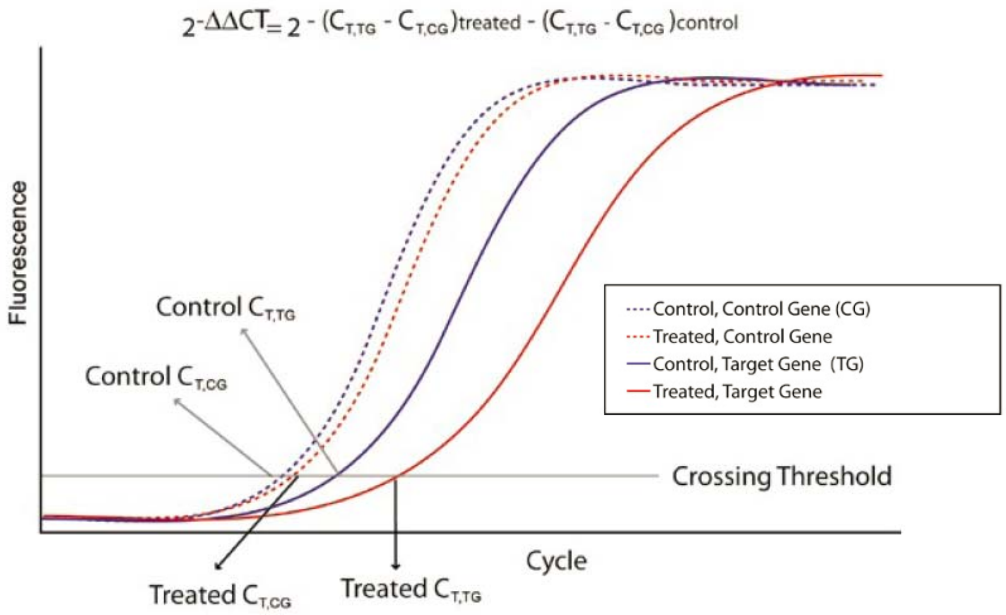
qPCR



http://en.wikipedia.org/wiki/Polymerase_chain_reaction#mediaviewer/File:Polymerase_chain_reaction.svg

- Quantitative polymerase chain reaction
- It measures the level of a particular nucleic acid (usually mRNA) which is defined by a primer pair sequence that amplifies it.
- Typically one needs to convert single stranded RNA to DNA with an enzyme called reverse transcriptase (RT)
  - Viral protein (e.g. HIV)
  - The copy is called complementary DNA (cDNA)
- First a quick review of PCR
- Works due to special polymerases that function at high temperatures (original one found in geyser microorganisms—Taq polymerase)
- Three steps (temperature changes):
  1. Melting—turn double stranded DNA into single stranded DNA (~95 C)
  2. Annealing—Allow primers to hybridize to the single stranded DNA (~55 C or so)
  3. Extension—Allow the polymerase to copy the primed single stranded DNA (~68 C for Taq—depends on polymerase)
- Repeat many times over (~30 usually).

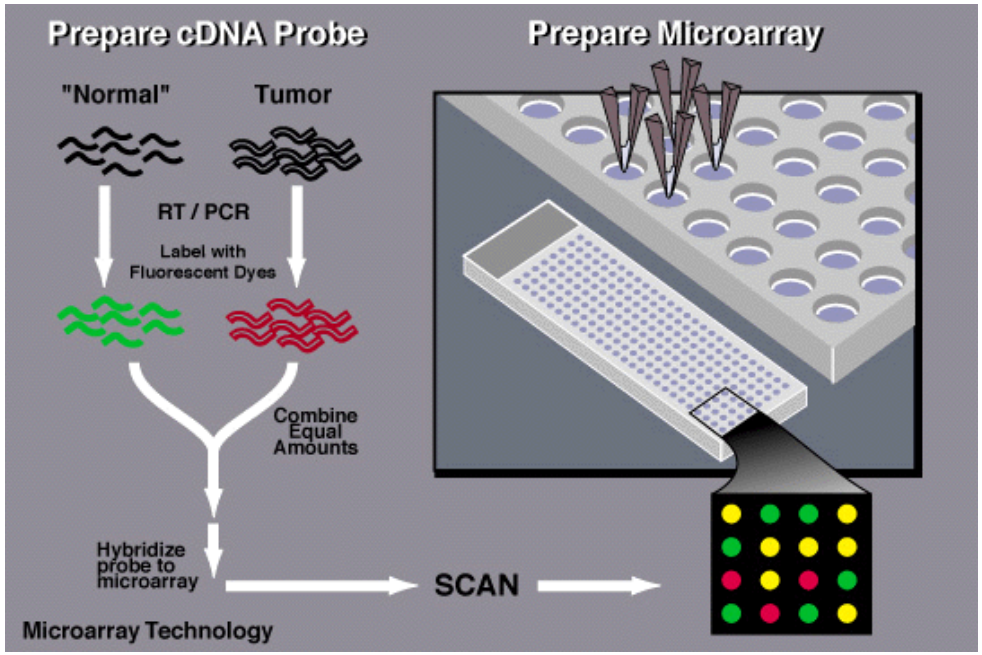
qPCR



VanGuilder et al., Twenty-five years of quantitative PCR for gene expression analysis, BioTechniques 44:619-626, 2008

- PCR is NOT a linear amplification!
  - Exponential growth means that the starting amount is distorted each PCR cycle.
    - 1→2→4→8→16→32 etc. (ideal scenario)
  - Beware, this is an issue with many quantitative nucleic acid assays
- PCR can be made more quantitative by measuring the number of cycles needed to reach a fluorescence intensity threshold
- Commonly called "real-time" because changes are monitored after every cycle in real time.
  - Not to be confused with RT-qPCR, where the RT stands for reverse transcriptase (turns RNA of interest into double stranded DNA which can be analyzed by PCR).
- The larger the number of cycles needed, the lower the original amount of the target.
- Medium throughput—10s to 100s, used more as a validation approach

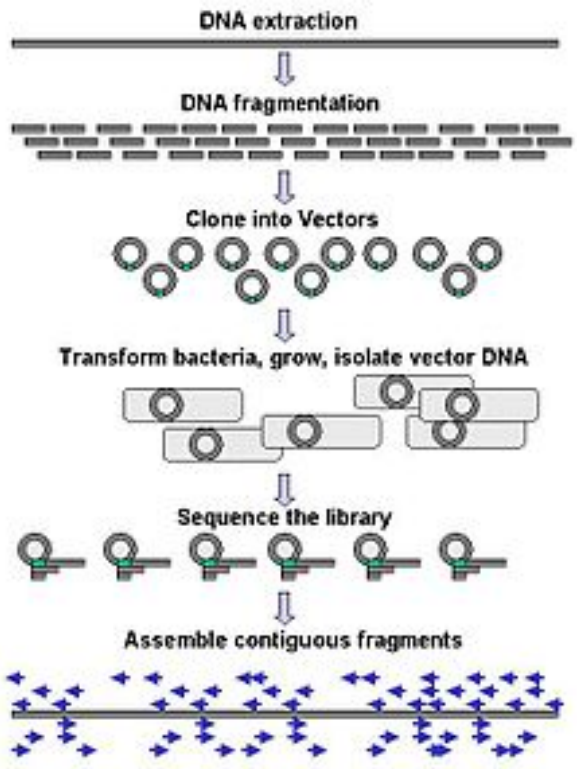
Microarrays



http://www.genome.gov/10000533

- Predominantly used for measuring transcript levels
  - There are other uses including genotyping, DNA mapping (DNA copy number, methylation) and many emerging technologies (Hoheisel, Nat. Genetics, 2006)
- Each spot has a different oligonucleotide
- Complementary oligos bind to each spot by base pairing, which is observable as increased fluorescence
- There are also bead-based methods (as opposed to slides or chips—Illumina)
- Workhorse in the 2000s, but next generation sequencing is emerging as superior (although still more expensive but probably not for long)

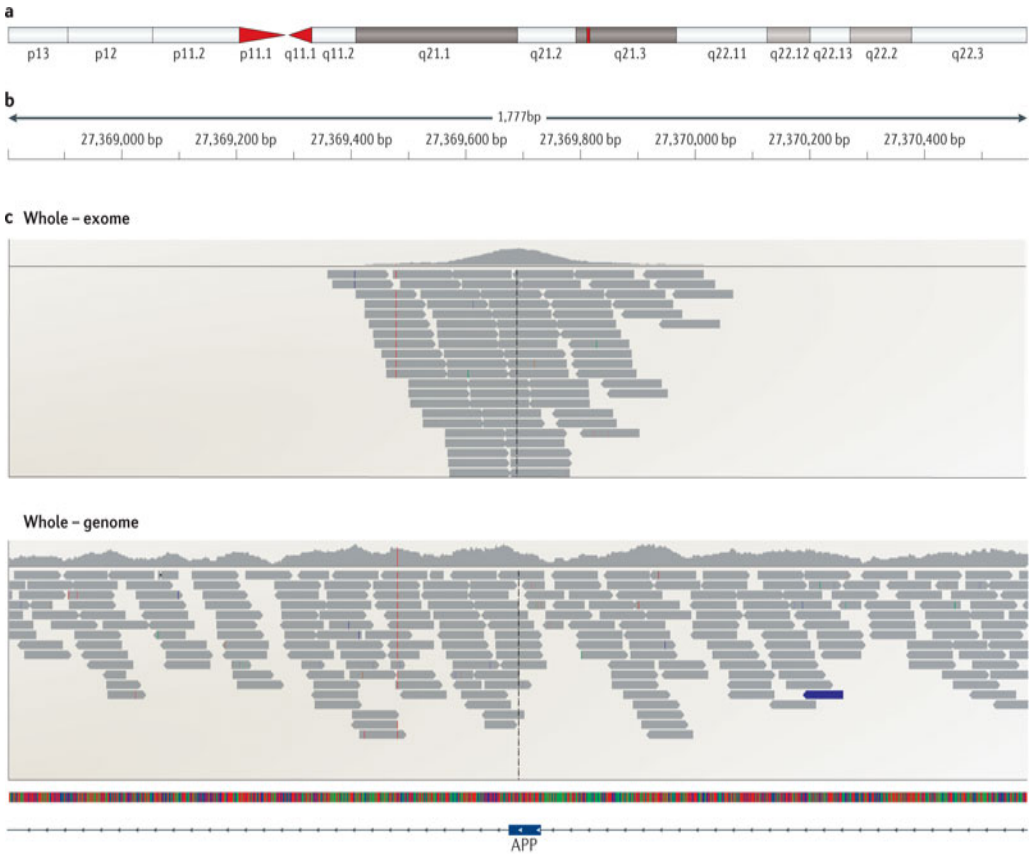
Sequencing-Based Technologies



http://en.wikipedia.org/wiki/DNA_sequencing

- High-throughput sequencing (to be discussed in greater detail later) has many applications
- Basic ability is to sequence MANY short pieces of DNA in parallel

Whole-exome vs. whole-genome

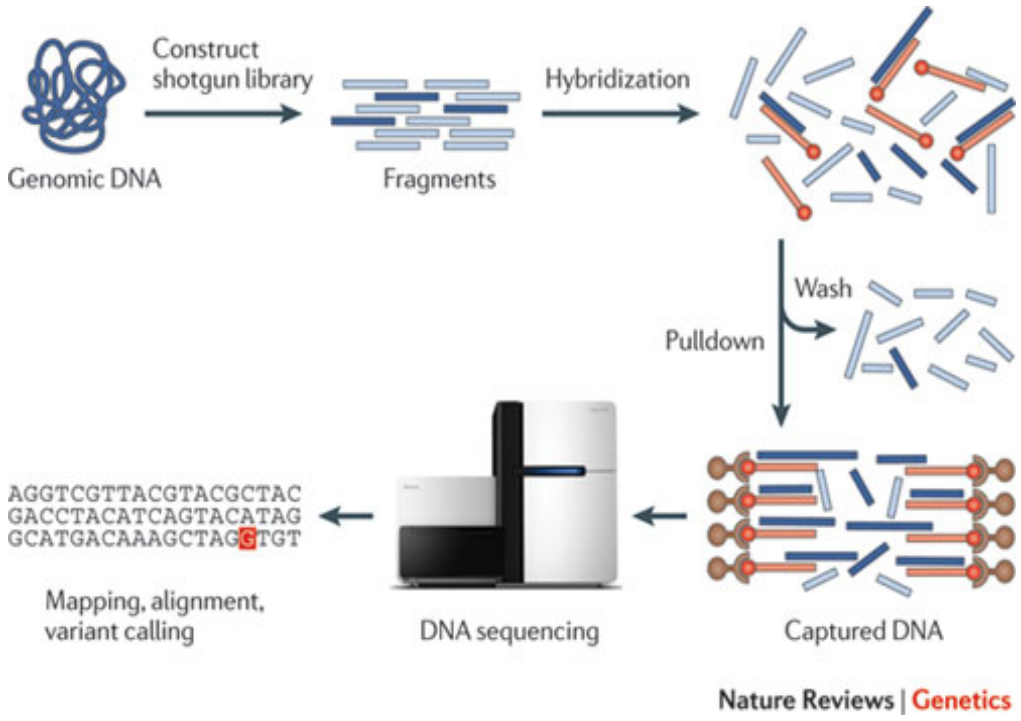


Bras et al., Nat. Reviews Neuroscience, 2012

Nature Reviews | Neuroscience

- A key parameter is “depth” of sequencing
  - Related to average number of short reads per base pair
- Exome sequencing gives more information for less short reads, because it is confined to exons
  - Exons are parts of the genome which end up in transcripts and thus are thought to have greater influence on average than most parts of the genome

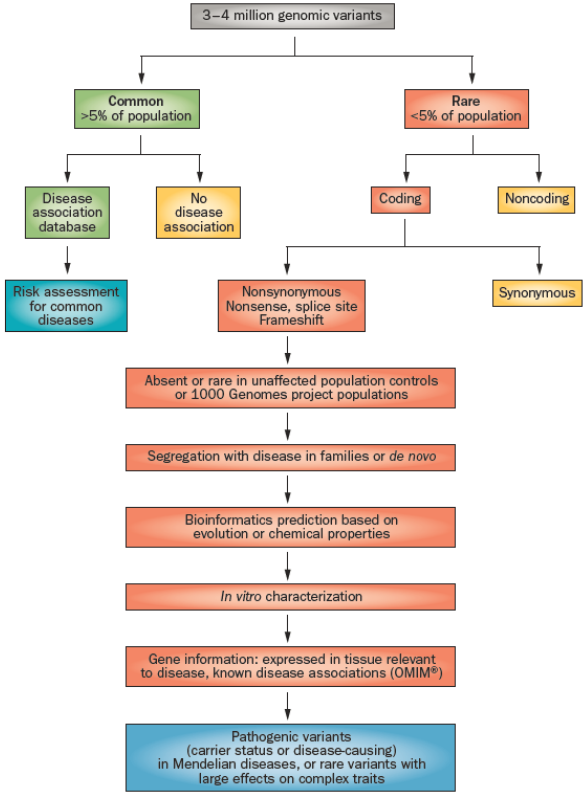
Workflow for exome sequencing



Bamshad, MJ et al, *Nature Reviews Genetics* **12**, 745-755 (2011)

- Key difference in exome sequencing is a separation step to isolate fragments that are exonic
  - Can be done by a variety of means including with microarrays

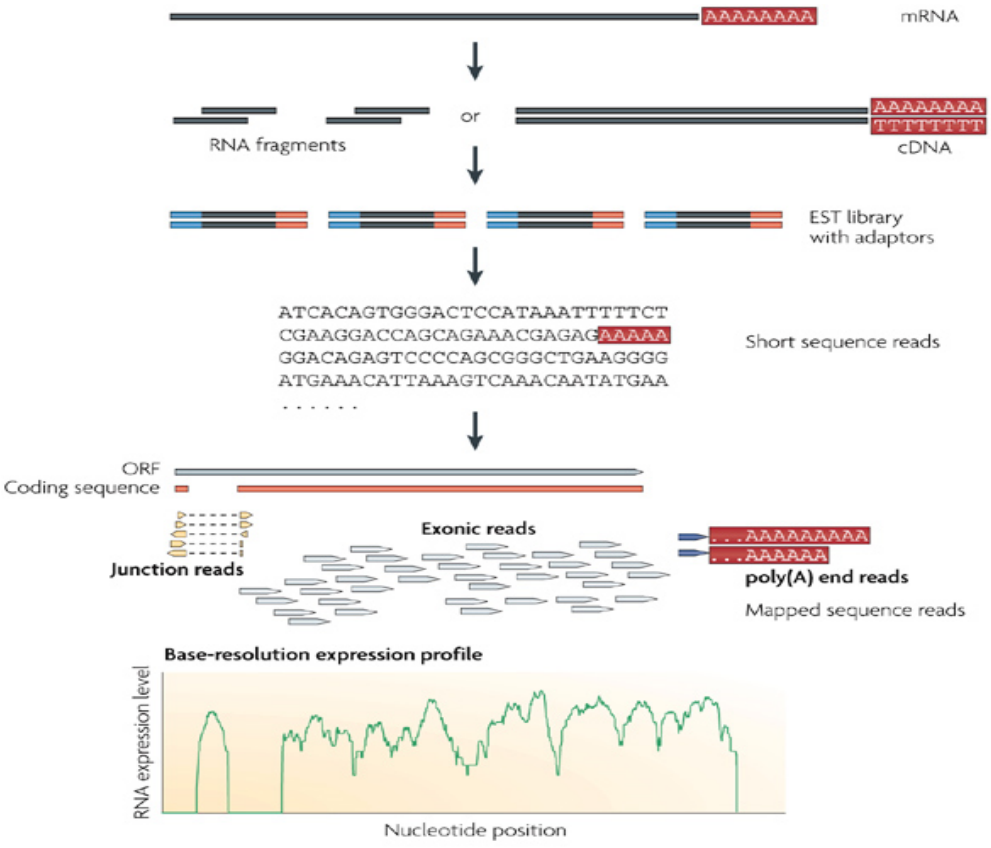
What can we learn from genome or exome sequencing data?



Foo, J.-N. *et al.Nat. Rev. Neurol (2012).*

- Many things! Examples:
  - Are certain small nucleotide polymorphism variants associated with disease?
  - Are there critical mutations (e.g. cancer)?
- Anything that is encoded in DNA sequence information can be learned from such data.
- However these data are categorical and usually not quantitative (unless done over a large population to estimate e.g. mutation frequencies)

RNA-Seq



Wang, Z et al Nature Reviews Genetics 10, 57-63 (January 2009)

- Apply massively parallel sequencing to RNA (mRNA or other types)
- One can
  - Quantify global gene expression
    - Related to number of sequences that align to a transcript (discussed in more detail later)
  - Quantify allele-specific expression
  - Quantify alternative splicing
  - Identify new transcript variants (e.g. Trapnell et al., Nat Biotech, 2010)

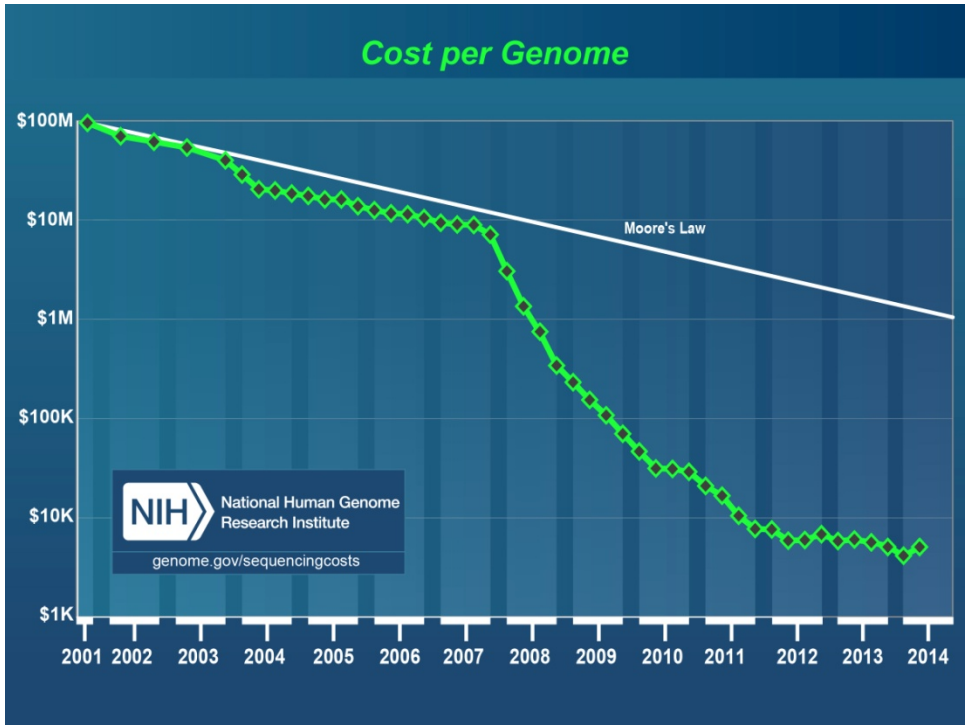
RNA-seq vs. microarray

Technology	Tiling microarray	RNA-Seq
Technology specifications		
Principle	Hybridization	High-throughput sequencing
Resolution	From several to 100 bp	Single base
Throughput	High	High
Reliance on genomic sequence	Yes	In some cases
Background noise	High	Low
Application		
Simultaneously map transcribed regions and gene expression	Yes	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes
Ability to distinguish allelic expression	Limited	Yes
Practical issues		
Required amount of RNA	High	Low
Cost for mapping transcriptomes of large genomes	High	Relatively low

Wang, Z et al Nature Reviews Genetics 10, 57-63 (January 2009)

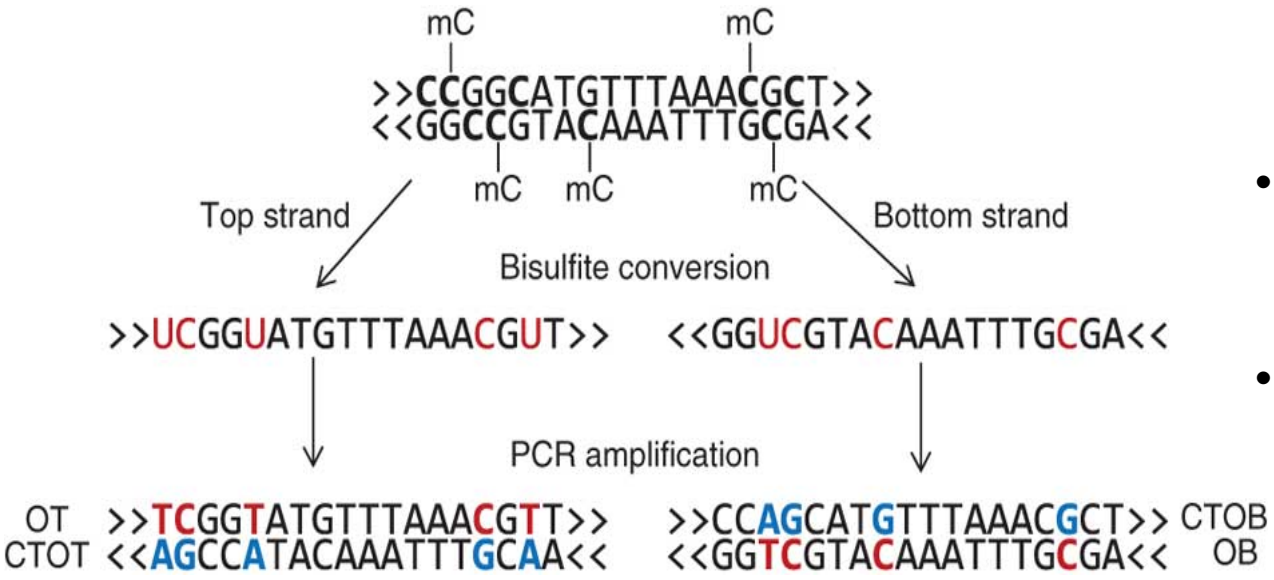
- Microarrays are currently cheaper, but not by a whole lot, and that may not last very long
- Microarray data can be hard to compare across different technologies
- RNA-seq has better potential for quantitation

Cost of High-throughput Sequencing Over Time



http://www.genome.gov/images/content/cost_per_genome2.jpg

Bisulfite sequencing for DNA Methylation

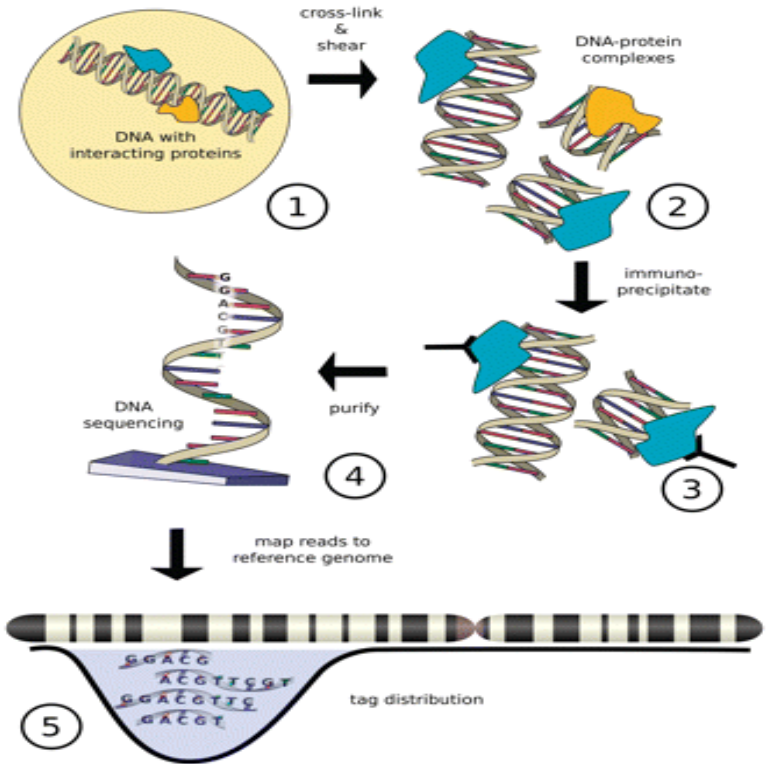


Krueger F *Nature Methods* **9**,145–151 (2012)

mC, 5-methylcytosine; OT, original top strand; CTOT, strand complementary to the original top strand; OB, original bottom strand; and CTOB, strand complementary to the original bottom strand.

- DNA methylation is associated with gene silencing and is a heritable epigenetic state.
- Bisulfite converts cytosine (C) to Uracil (U) which base pairs with adenine (A) (similar to thymidine (T)).
- Methylated C is not affected by bisulfite.
  - Thus treating DNA with bisulfite followed by whole genome sequencing can quantify DNA methylation levels.
- Bisulfite conversion followed by PCR yields four potentially different DNA fragments (OT, CTOT, OB, CTOB)

Chromatin Immunoprecipitation (ChIP) Sequencing

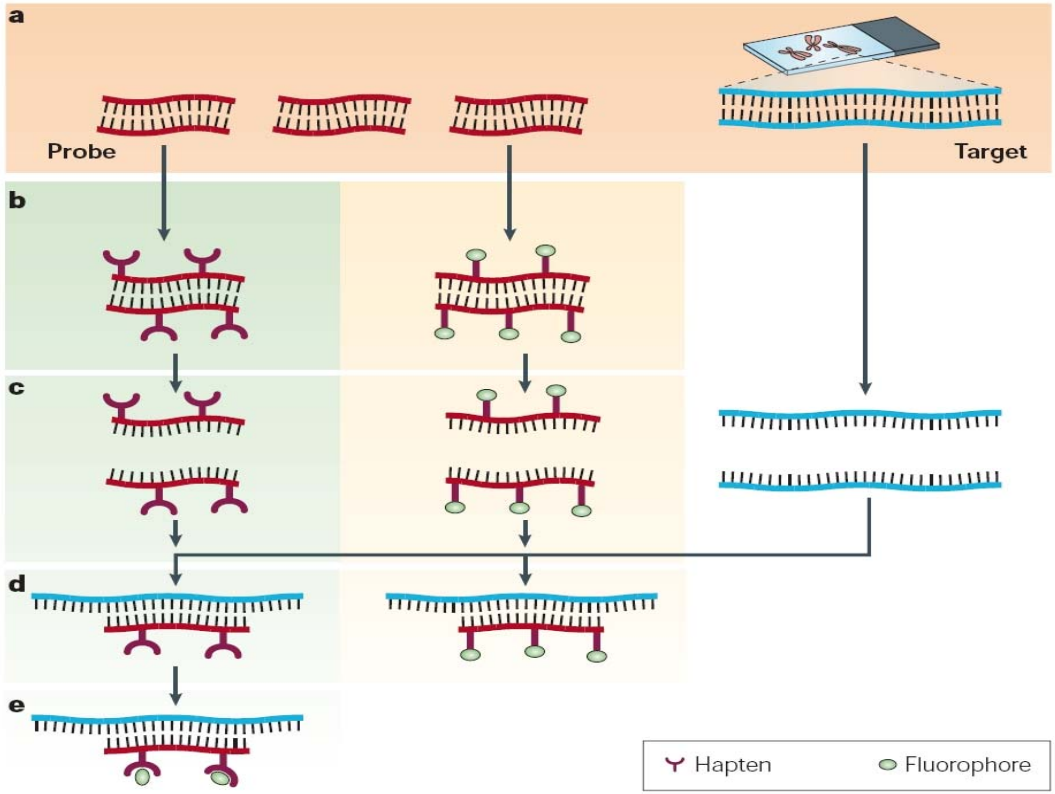


Szalkowski, A.M, and Schmid, C.D.(2010). *Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing banchmarking efforts.* Briefings in Bioinfomatics.

- Combine deep sequencing with chromatin immunoprecipitation (ChIP)
  - Immunoprecipitation is a general technique that allows one to isolate antibody epitopes from a mixture
  - Antibodies against particular transcription factors, against particular histone modifications, RNA polymerase, etc.
- When many reads align to a particular region of the genome, it is called a "peak"
  - A "peak" indicates that the antibody's target binds that region of the genome

Single Cells

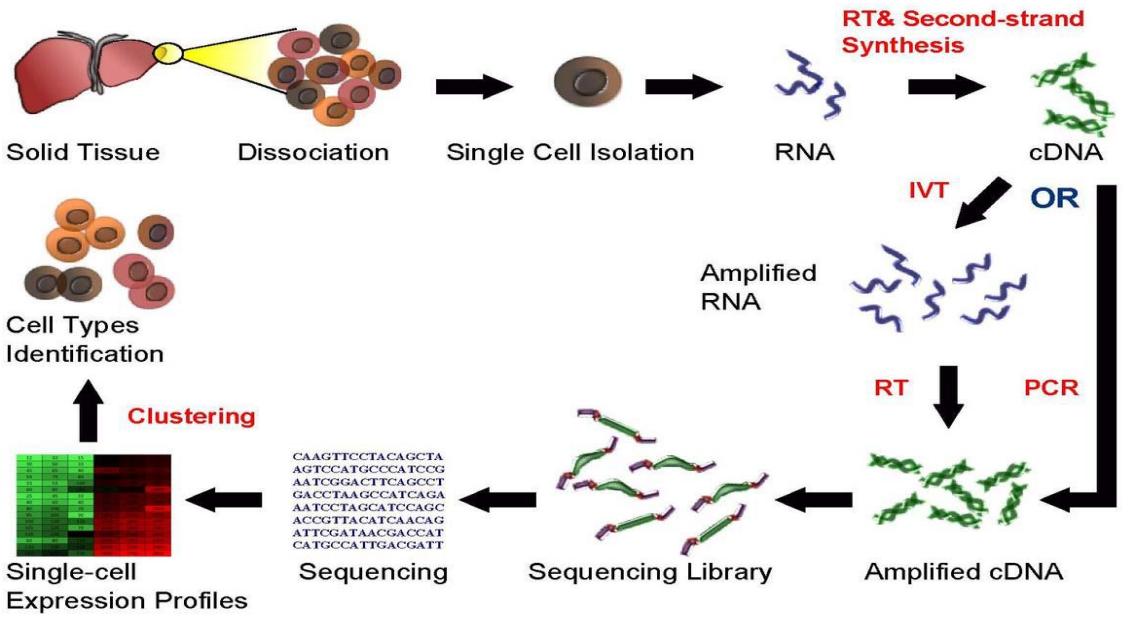
Fluorescence *in situ* hybridization (FISH)



Speicher, M. R. *et al.* *Nature Reviews Genetics* 6, 784 (2005).

- Uses labeled oligos to hybridize to target nucleic acid
  - mRNAs, genomic DNA, etc.
- Can be done for single molecules (Raj et al. Nat Methods, 5(10) 2008
- Can be highly multiplexed with super-resolution and fluorophore barcoding (Lubeck and Cai, Nat. Methods, 9(7) 2012)

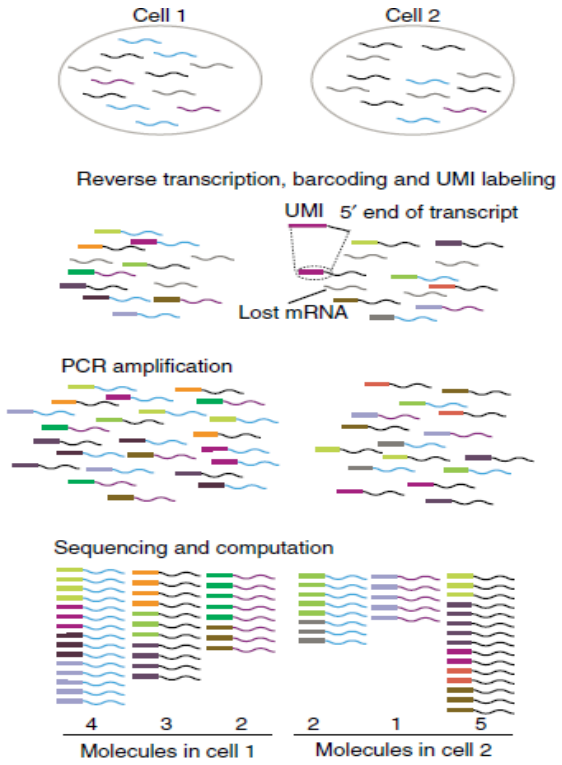
Deep sequencing is also applicable to single cells



http://en.wikipedia.org/wiki/Single_cell_sequencing

- Simply requires isolation of single cells and adequate amplification
- Is of course also applicable to exome and whole genome sequencing
- How can one be quantitative for mRNA with the reliance on amplification?

Unique Molecular Identifiers Allow Accurate Quantitation with Amplification



Islam et al., Nat Methods, 2014; Kivioja et al., Nat Methods, 2012

- Attach a set of 5 nucleotide barcodes (1,024 possible) to the 5’ end of (nearly) every mRNA.
- Sequence the 5’ ends* and count the number of unique barcodes that appear.
  - *enough to ensure unique alignment
- This is the number of mRNA molecules in the original sample.

Next Time—Measuring Proteins and Protein States