Introduction to Experimental Design of Sequencing-Based Studies

Michael C. Zody, Ph.D.
Harvard University CFAR
Workshop on Metagenomics
September 14, 2015

Logistics

- Introduction
- Please feel free to ask questions at any point
- Slides will be posted on workshop website
- One break at about 90 minutes
- Thanks to CFAR!

Course Outline

- Considerations before starting a sequencing experiment
- Steps of generating sequencing data
- Considerations before starting specific types of sequencing experiments

Course Outline

- Considerations before starting a sequencing experiment
- Steps of generating sequencing data
- Considerations before starting specific types of sequencing experiments

Considerations before starting a sequencing experiment

- What is the question you want to answer?
- How do you decide how much data to generate to answer your question?
  - Sensitivity (e.g. number of False Negatives)
  - Specificity (e.g. number of False Positives)
  - Cost
- Which factors influence the amount of data you generate?

What is the question you want to answer?

- What scientific result do you want?
- Is there an hypothesis you want to test?
  - Early sequencing was “hypothesis free” (i.e. the genome was the goal)
  - Now, it is affordable to sequence for a specific aim (i.e. What sequence do you need for that aim?)
- Understanding this shapes many decisions in designing the experiment

How do I decide how much data to generate?

- Is your sequencing result the final answer, or just a survey to generate preliminary data for follow up studies?
- What are the costs of false positives and false negatives, relative to the cost of the sequence?
- Four case studies highlight how projects involving SNP discovery might require different amounts of data based on these factors

Case 1: Tumor/normal sequencing

- Difficult problem, requires very low false positives and false negatives
- Trying to find somatic events (~1-2 / Mbp)
- FP rate approaching 1 / Mbp swamps signals
- FN runs the risk of missing real tumor variants
- Every sample is unique, so the cost of following up (orthogonal resequencing, custom genotyping) is high

High coverage, high variant calling stringency

Case 2: Microbial evolution

- Example: Sequencing a drug resistant microbe to find functional changes
- Low tolerance for false negatives, because you want to find a variant in a small genome
- Relatively high tolerance for false positives because the functional mutation is most likely a coding change, so triage of calls for follow up is effective

High coverage, low variant calling stringency

Case 3: Vertebrate evolution

- Example: Sequencing to find signatures of selection
- Relatively high tolerance for false negatives, because specific sites of variation are not important
- Low tolerance for false positives because background noise from sequencing errors can obscure the signature of selective sweeps

Low coverage, high variant calling stringency

Case 4: Population SNP discovery

- Example: Sequencing multiple strains or individuals from one species to design a SNP array
- High tolerance for both FN and FP because the experiment is just a first pass
- Only need sufficient SNPs to design the array
- Array design and testing will identify FPs (Rate of SNPs failing to work on the array will likely exceed the false positives from discovery)

Low coverage, low variant calling stringency

Which factors influence the amount of data I generate?

- Number of samples
- Type of read
- Type of library
- Number of reads
- Read length
- Complexity of library
- Which sequencing machine to use

Consideration: Number of samples

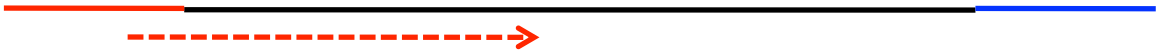
- How many different samples do you need for your experiment?
- Do you need biological replicates?
- Do you need technical replicates?
- Do you need controls, such as:
  - Resequencing your reference genome to control for alignability
  - Generating unenriched controls for ChIP-Seq

What is a read? What is a library?

- Definition of “read”: A single sequence from one fragment in the sequencing library (one cluster, bead, *etc*.)
- If generating paired reads, then 2 reads derived from each fragment in the library
- Definition of “library”: A collection of DNA fragments that have been prepared to be sequenced
- Definition of “coverage”: The number of reads spanning a particular base in the genome

Consideration: Type of read

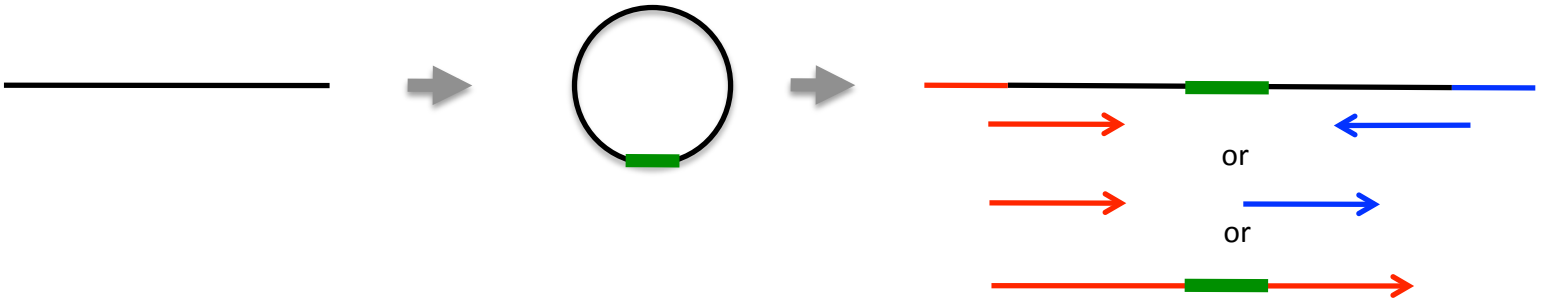
- Fragment reads (come from fragment libraries)
  - Single read in one direction from a fragment



- Paired end reads (come from fragment libraries)
  - Two reads from opposite ends of the same fragment
  - Reads point towards each other

Consideration: Type of read

- Mate Pair Reads (come from Jumping Libraries)
  - Long fragment of DNA is circularized
  - Junction is captured (e.g., by biotinylated adapter)
  - Remainder is cleaved (many methods)
  - Ends are sequenced
  - Read orientations depend on the exact method



Consideration: Why choose one type of read?

- Fragments
  - Fastest runs (one read per fragment), least cost
  - Some technologies only make one read
- Paired reads
  - More data per fragment
  - Help with assembly and alignment
  - Same library steps as fragments, but yields more data

Consideration: Why choose one type of read?

- Mate Pairs (Jumping Libraries)
  - Advantages over paired ends:
    - Paired end separation limited by fragment size
    - Some platforms can't read second strand of fragment
  - Only way to make long links, which are very useful for:
    - Assembly and alignment across repeats and duplication
    - Identification of large structural variants
    - Phasing of small variants
  - Drawback: Requires much more input DNA than paired ends

Consideration: Number of reads

- How much data do you need to generate to answer your question?
- This depends on the level of completeness & accuracy you want
- You have to decide before beginning the experiment what level of completeness & accuracy you want, and this determines how much data to generate
- Analogy: Trying a protocol in the lab that requires 1ug of DNA with 0.1ug may end up working, but it may not

Consideration: Read length

- For most experiments, the longer the reads are the better
- Exception: longer poor-quality reads are not as useful as shorter high-quality reads
- Some experiment types have more stringent requirements for minimum read length

Consideration: Complexity of library

- Definition of “complexity”: the number of distinct fragments in the library
- After amplification, you may have many copies of the same initial fragment (which does not increase complexity)
- For most experiments, sequencing the same fragment multiple times is not useful and may be detrimental to your analysis

Consideration: Which sequencing machine to use

- Type of read/library:
  - Illumina & Ion: all
  - 454: fragment, mate pair
  - PacBio: fragment
- Read length:
  - Illumina: short (≤150 bp) on HiSeq, medium (≤400 bp) on MiSeq
  - 454: long (450-750 bp)
  - Ion: medium (200-400 bp, 100-200 for paired end)
  - PacBio: very long (thousands of bp)

Course Outline

- Considerations before starting a sequencing experiment
- Steps of generating sequencing data
- Considerations before starting specific types of sequencing experiments

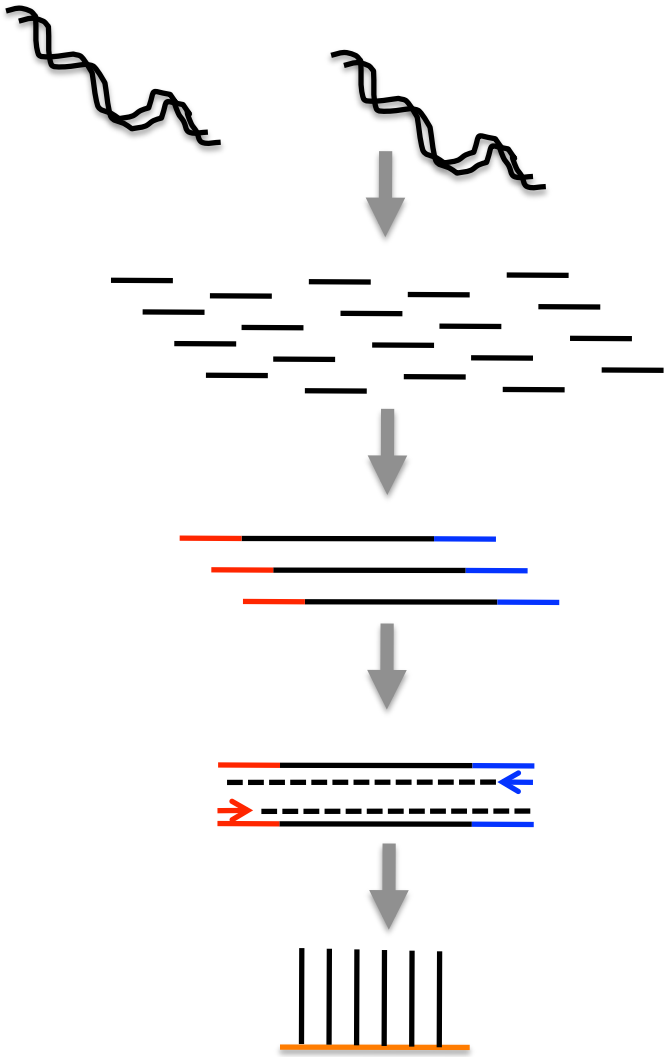
Steps of generating sequencing data

- Steps of library construction and sequencing
- Making Fragment libraries (to generate fragment or paired end reads)
- Making Jumping libraries (to generate mate pair reads)
- Pooling with or without barcoding
- Possible artifacts of library construction
  - PCR-based artifacts
  - Sequencing of primers, adapters, and tags

Steps of Library Construction

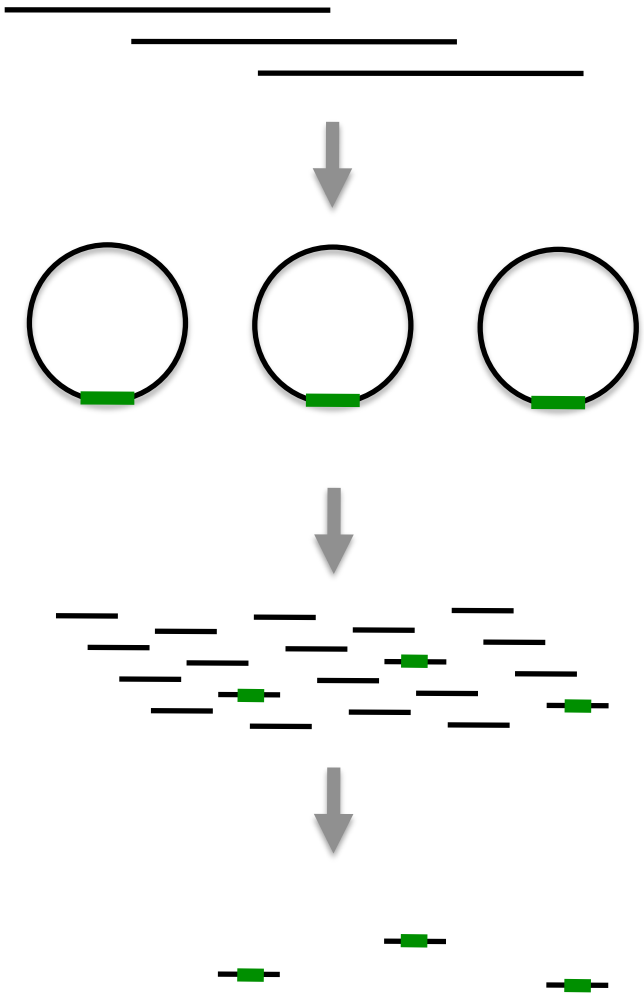
- Add adapters containing:
  - Barcodes (for multiplexing)
  - Sequencing primers
  - Amplification primers
  - Sequence for substrate attachment
- Amplify fragments by universal PCR
- Optionally pool barcoded libraries

Steps of Fragment Library Construction



- Extract DNA
- Fragment and possibly size select (300-600 bp)
- Add adapters
- Amplify
- Select single molecules
- Amplify in clusters/ beads

Steps of Jumping Library Construction



- Extract DNA, fragment and size select (2-40 kb)
- Circularize with labeled adapters
- Fragment and size select (300-600 bp)
- Select fragments containing labeled adapters
- Proceed as for fragment library

Pooling with barcoding

- Unique DNA tags identify samples
- Allows multiple distinct samples on one run
- Advantages:
  - Reduced cost of sequencing for small samples
  - Analysis is identical to unpooled data
- Disadvantages:
  - Some small throughput loss due to barcode fails
  - Data mis-assignment from bad barcode reads
  - Increased per sample cost for library construction

Pooling without barcoding

- Mix input DNA without identification
- No way to definitively separate data from different samples afterwards
- Advantages:
  - Single library prep for a number of samples
  - No yield lost to barcodes
- Disadvantages:
  - Loss of all individual associations
    - Loss of ability to use replicates!
  - No check on accuracy of pooling

PCR-based artifacts

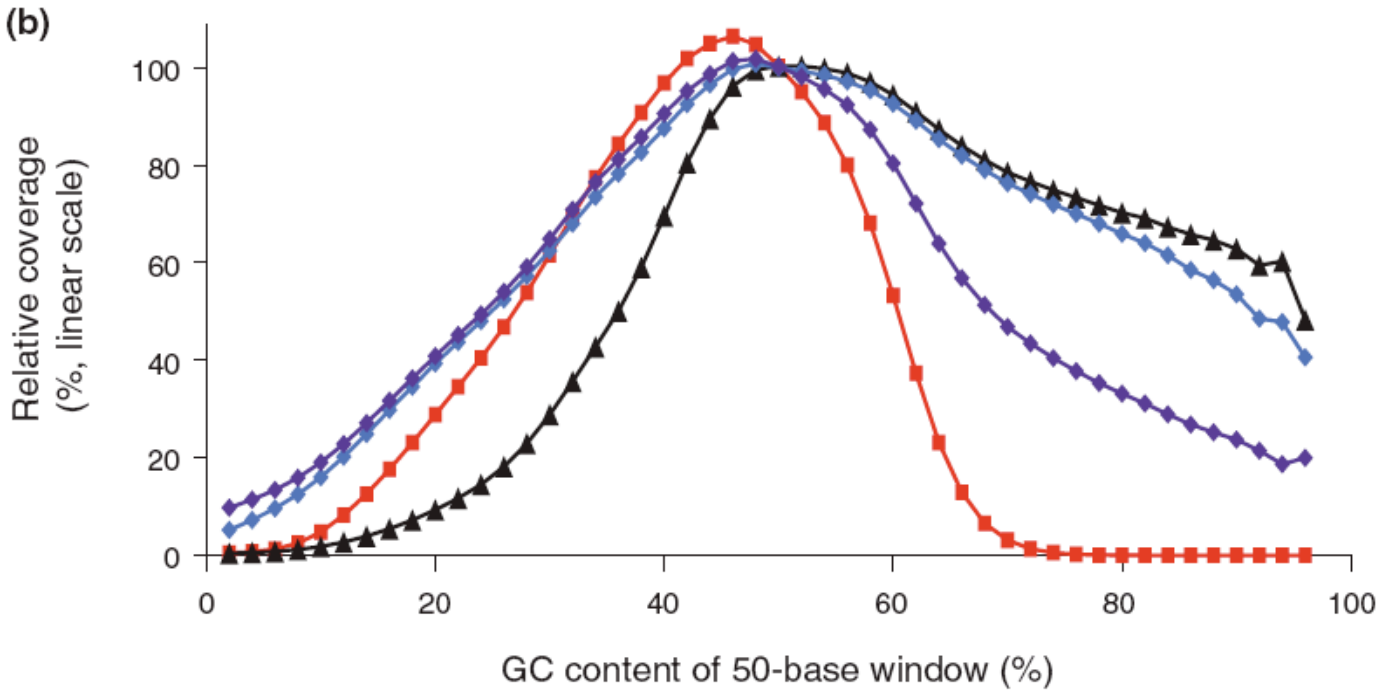
- Most libraries are PCR amplified during construction
- After library construction, single molecules are isolated and then amplified again for sequencing
- Errors from library construction PCR will not be detectable as sequencing errors
- Regions with secondary structure or extreme GC content:
  - Will amplify poorly and be underrepresented
  - May form small or weak clusters with poor sequence quality
- PCR may form chimeric sequences (especially in targeted designs)
- PCR amplification may result in duplicated sequences

PCR Errors: How Much PCR?

- You may be doing more PCR than you think
- Initial amplification of sample
- Targeting PCR
- Library amplification
- 100 rounds of PCR is equivalent to a 2 order of magnitude drop in polymerase accuracy

PCR-based artifacts: PCR bias

- Most PCR protocols work best for ~50% GC
- Extreme GC sequences are underrepresented



Red = standard PCR protocol
Other colors = modified PCR protocols

From Aird et al., Genome Biology *(2011)*

Sequencing of primers, adapters & tags

- Not every base you sequence is useful
- Primers will be present if you used PCR to target your input DNA
  - Sequence from primers does not represent target
  - Variation seen (or not) under primers is not real
  - Overlapping products will allow analysis of the primer-covered regions
- Short fragments may read through to adapter
- Custom barcodes or other tags may get sequenced too, though most vendor tags will be removed automatically

Course Outline

- Considerations before starting a sequencing experiment
- Steps of generating sequencing data
- Considerations before starting specific types of sequencing experiments

Types of sequencing experiments

- Resequencing
- Genome assembly
- RNA-Seq
- Metagenomics

Types of sequencing experiments

- Resequencing
- Genome assembly
- RNA-Seq
- Metagenomics

Example uses of resequencing

- SNP discovery and genotyping
- Population sequencing
- Structural variant discovery and genotyping
- Comparative genomics of closely related species

Considerations before a resequencing experiment

- Considerations for all resequencing experiments
  - Working with a reference genome
  - Aligning reads to a reference
  - Alignability
  - Read length and type
- Considerations for specific types of resequencing experiments
- Targeted resequencing

Working with a reference genome

- How good is the reference?
  - Completeness
  - Accuracy
- How representative is it of your genome(s)?
- Sequence won't align if
  - Absent from the reference
  - Too diverged from the reference

Aligning to a reference genome

- Aligning long sequences is relatively easy
  - Abundant information to predict true alignments
  - Can trim sequences based on alignment
- Short reads are harder
  - Less information per read
  - Often need full length alignments
  - For diverged sequences, may not match at all
  - Many more sequences, so speed of the aligner matters

Alignability

- Not all of the reference will be useful for alignment because some parts are too similar for unique alignments (duplications, recent repeats, gene families)
- Longer reads and pairing increase alignability
- Example from human genome resequencing:

	No pairing	400 bp pair	6000 bp pair
36 bp read	85%	96%	-
100 bp read	93%	97%	98%

Adapted from The 1000 Genomes Project Consortium, Nature *(2010)*

Read length and type

- Read length matters for alignability
- Paired end reads also help with alignment
  - Aligning one end uniquely localizes other end
  - Aligners may use this to run more sensitive alignments
  - Allows finding highly variant regions and small indels if the other read from that pair aligns cleanly
- Paired end reads are necessary for structural variant discovery and genotyping
- Mate pairs (from jumping libraries) are very useful for structural variant analyses but of relatively little use for SNPs and small indels

Considerations for specific types of resequencing experiments

- SNP discovery and genotyping
- Population sequencing
- Comparative genomics of closely related species

Considerations: Sequencing depth for SNP discovery

Type of Experiment	Coverage Required
Haploid SNPs/divergence	≥ 10 x
Diploid SNPs/divergence	≥ 30 x
Aneuploid/somatic mutations	≥ 50 x
Population sequencing	≥ 200 x

Example: Haploid SNP discovery

- You know there is only one base-pair at each locus, so you should make the majority call
- Assuming a uniform 1% error rate, what is the probability that the majority call from your sequencing is actually right?

Depth of coverage at the locus	% of time that majority call is correct	% of time there was no majority call	% of time that majority call is an error
1	99.000	0.00	1.00
2	98.010	1.98	0.01
3	99.970	0.00	0.03
4	99.941	0.06	<0.001
5	99.999	0.00	<<0.001

SNP discovery: Adjusting for random sampling

- Previous graph assumed uniform coverage
- What are the probabilities if the reads are theoretically randomly distributed?

Average depth of coverage across genome	% of time that majority call is correct	% of time there was no majority call	% of time that majority call is an error
1	62.475	37.153	0.372
2	85.646	14.075	0.279
3	94.409	5.432	0.158
4	97.786	2.134	0.081
5	99.110	0.851	0.039
8	99.938	0.059	0.004
10	99.987	0.012	<0.001

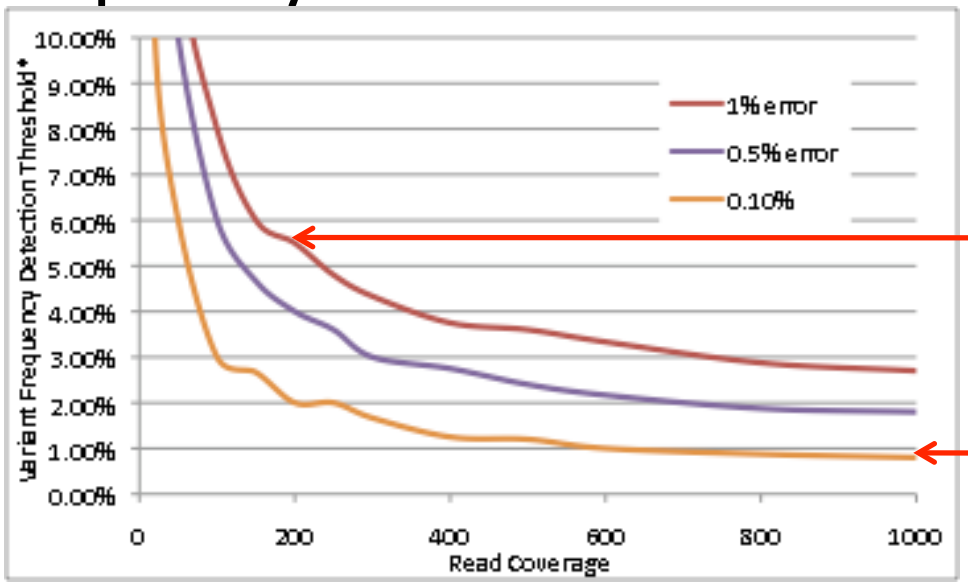
- In reality, distribution will be worse because reads are non-randomly distributed

SNP discovery: Diploid or aneuploid samples

- Diploid samples require twice as much coverage
  - Want to be able to call heterozygotes
  - Need to see each allele as often as you would for a haploid organism
- Aneuploid or somatic mutation samples
  - Cannot rely on expected 1:0 or 1:1 allele ratios
  - Often unique variants, and thus are harder to confirm

Considerations: Population Sequencing

- Example: Want to find all real variants in pooled or host/environmental samples
- What coverage do we need to find a variant at a given frequency?



1% error + 200x
= 5.5% variant

0.1% error + 1000x
= 1% variant

* Lowest frequency of call which exceeds Poisson error probability after Bonferroni correction for 10kb genome

Considerations: Population Sequencing

- Where is the sampling bottleneck?
- Generating more reads than input molecules doesn't improve calling
  - The accuracy and sensitivity of calling is limited by the sampling of the population, not the reads
- With limiting amounts of input, consider using a barcoding scheme that tags input molecules

Considerations: Comparative genomics of closely related species

- Comparative genomic analysis is most effective when species are less than a few % diverged
- Using a more diverged reference:
  - Requires more sensitive (time consuming) algorithms
  - Results in loss of alignability (reads are not placed)
  - Is worse if the divergence is due to insertion/ deletion

Targeted sequencing

- Mostly similar to whole genome resequencing
- Targets specific regions (e.g., exome) by:
  - PCR amplification
  - Hybrid selection
  - Targeted genome amplification
- Involves some special analysis considerations

Pros & Cons: Targeted sequencing

- Pros:
  - Significant cost savings if target <<< genome
  - Can achieve higher coverage on target
- Cons:
  - Cost of targeting reagents can be high
  - Some sequenceable regions very hard to target
  - Variability of coverage is higher
  - Targeting may introduce bias
  - Challenging to identify duplicates in targeted sequence

Considerations: Targeted sequencing

- Targeting introduces additional bias
- More coverage required to overcome this (want 3 times or more as much average depth)
- Many off-target reads are generated
  - Not all reads will come from targeted regions
  - Need to bulk up coverage to overcome this
  - Amount will depend on specificity of the targeting

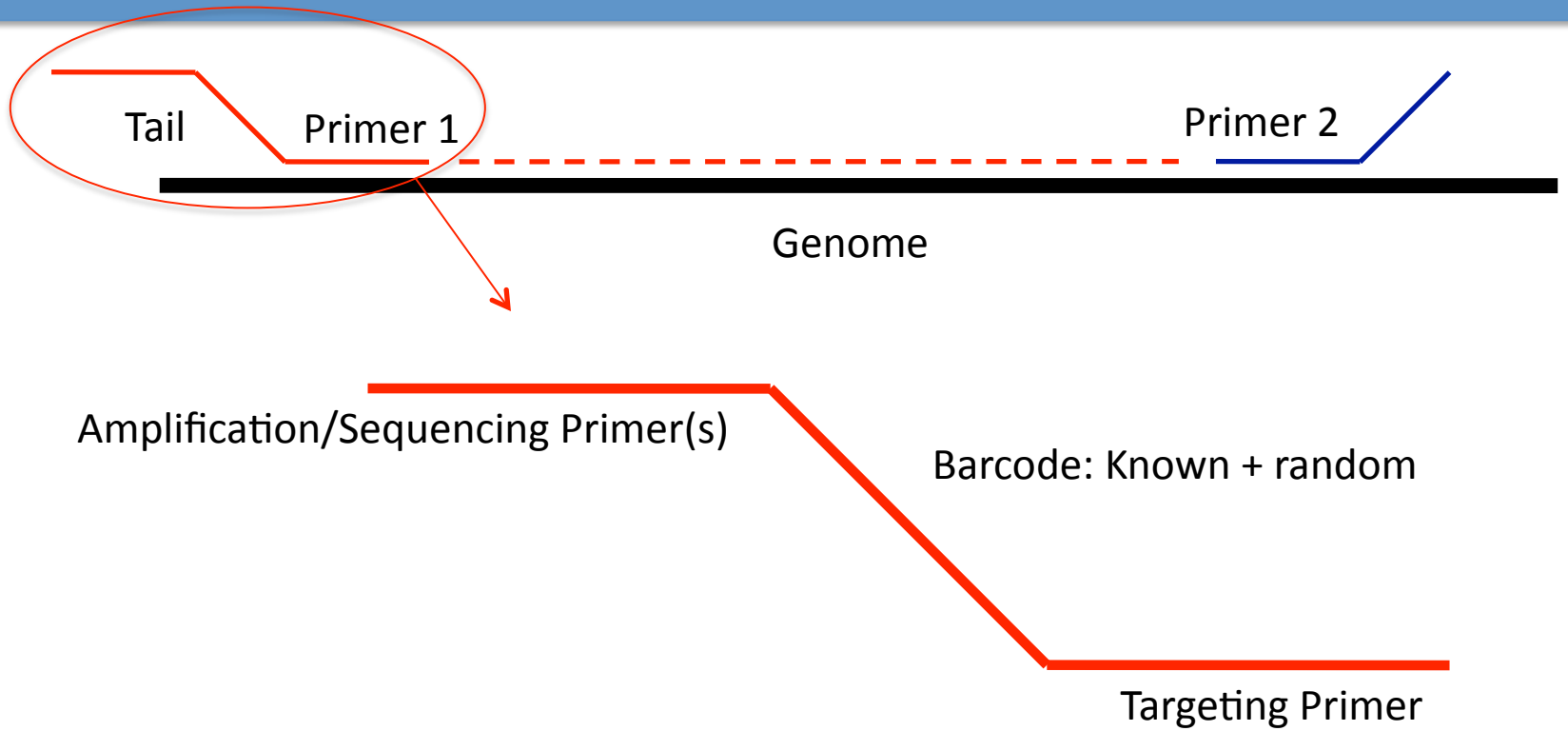
Considerations: Targeted sequencing

- Targeted sequences often include repeats and duplications, and thus some untargeted regions may be sequenced as well
- Need to align to whole genome (not just to the part you targeted) to ensure that unique hits to targeted regions are the best hits for that read in the genome

Considerations: Targeted sequencing

- Some targeting generates identical fragments
  - Hard to find PCR duplicate reads
  - Many or all starts and ends are the same
- Can use a random barcoding scheme in the amplification to tag fragment of origin

Tagged Primers



Types of sequencing experiments

- Resequencing
- Genome assembly
- RNA-Seq
- Metagenomics

Example uses of genome assembly

- Generate a reference genome
- Alternative method of SNP discovery (even if you have a reference)
  - Mostly for small, haploid genomes
  - Provides better diversity calling for small indels and particularly difficult-to-align regions
- Discover structural variants
  - *De novo* assembly is the only way to get the sequence of a novel insertion
  - Complex structural variants can be more easily discovered through de novo assembly than read alignment to a pre-existing reference

Steps of a genome assembly experiment

- Choose your sample(s)
- Extract DNA from samples
- Fragment the DNA (may need to do this into multiple sizes)
- Library construction (probably need to make multiple libraries)
- Sequencing

Genome assembly considerations: Depth of coverage

- Very deep coverage needed
  - For short reads (Illumina, Ion, SOLiD): 50x – 100x
  - For longer reads (454, PacBio): 20x
- Common issue is not having sufficient coverage for *de novo* assembly

Genome assembly considerations: Type of reads

- Long reads help greatly
  - Provide connectivity through low coverage
  - Resolve repetitive/duplicated regions
- Paired reads necessary
- Jumping libraries (& mate pair reads) are not always necessary, but yield much better connectivity

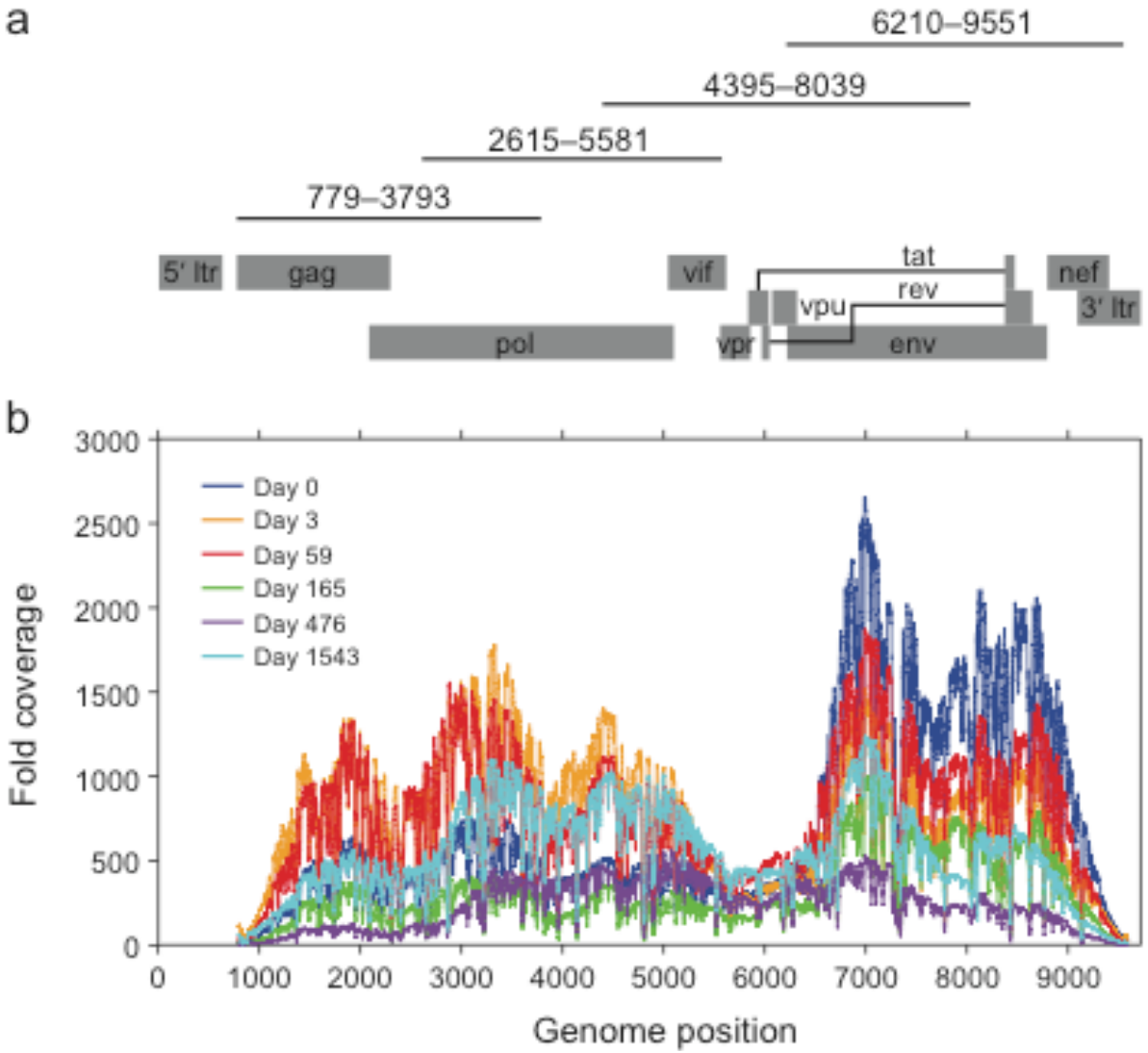
Genome assembly considerations: Genome complexity and composition

- Repeat content of genome
  - More repetitive genomes require more coverage
  - Paired end reads and jumping libraries more important
- GC content of genome
  - Genomes with extremes of GC content will have more bias in representation
  - Greater average coverage will be required to assemble through extreme GC regions

Genome assembly considerations: Viral Genomes

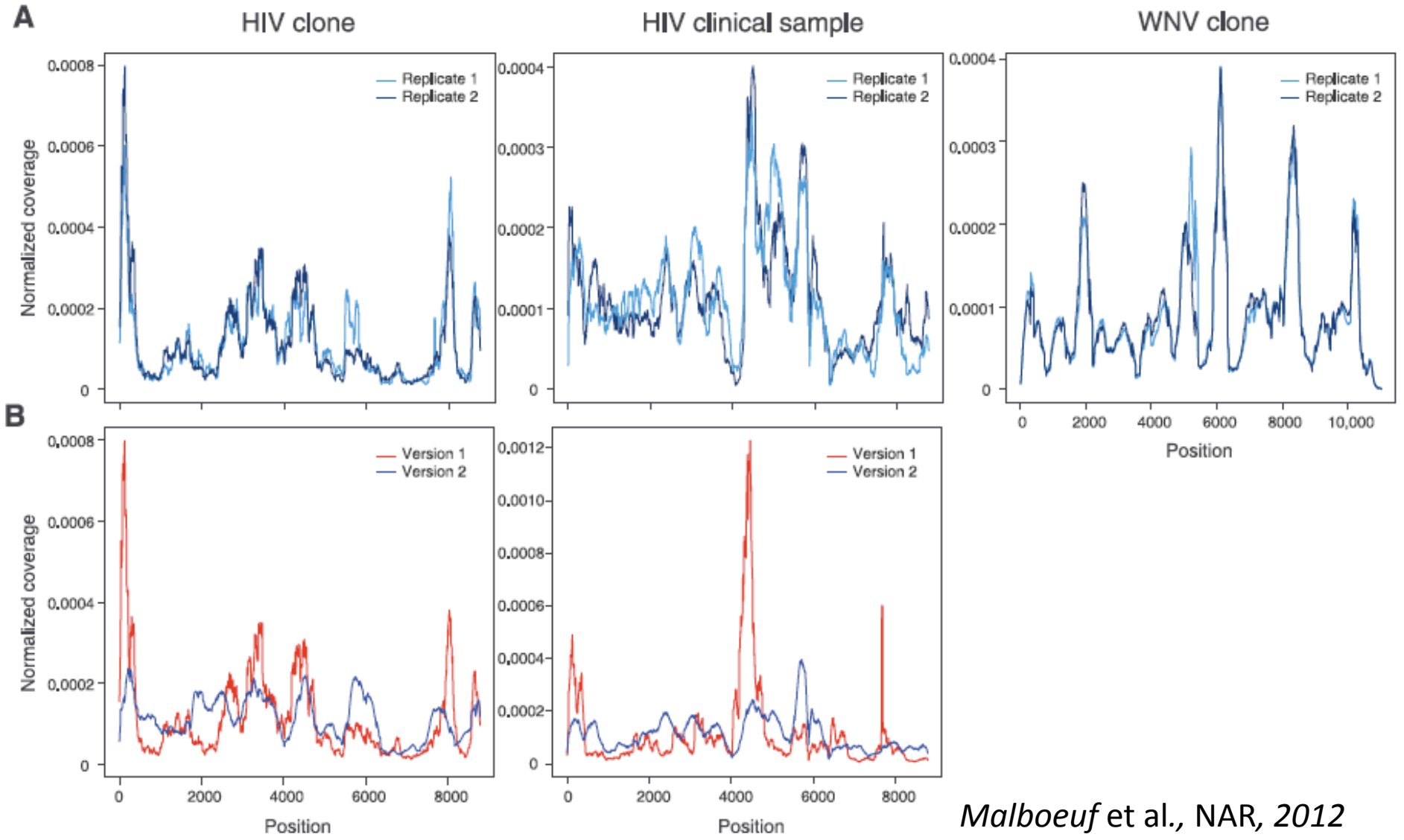
- Viral genomes can be difficult to assemble despite their small size
- High internal variability
- More variability of coverage due to amplification techniques required

Genome assembly considerations: Highly variable viral coverage



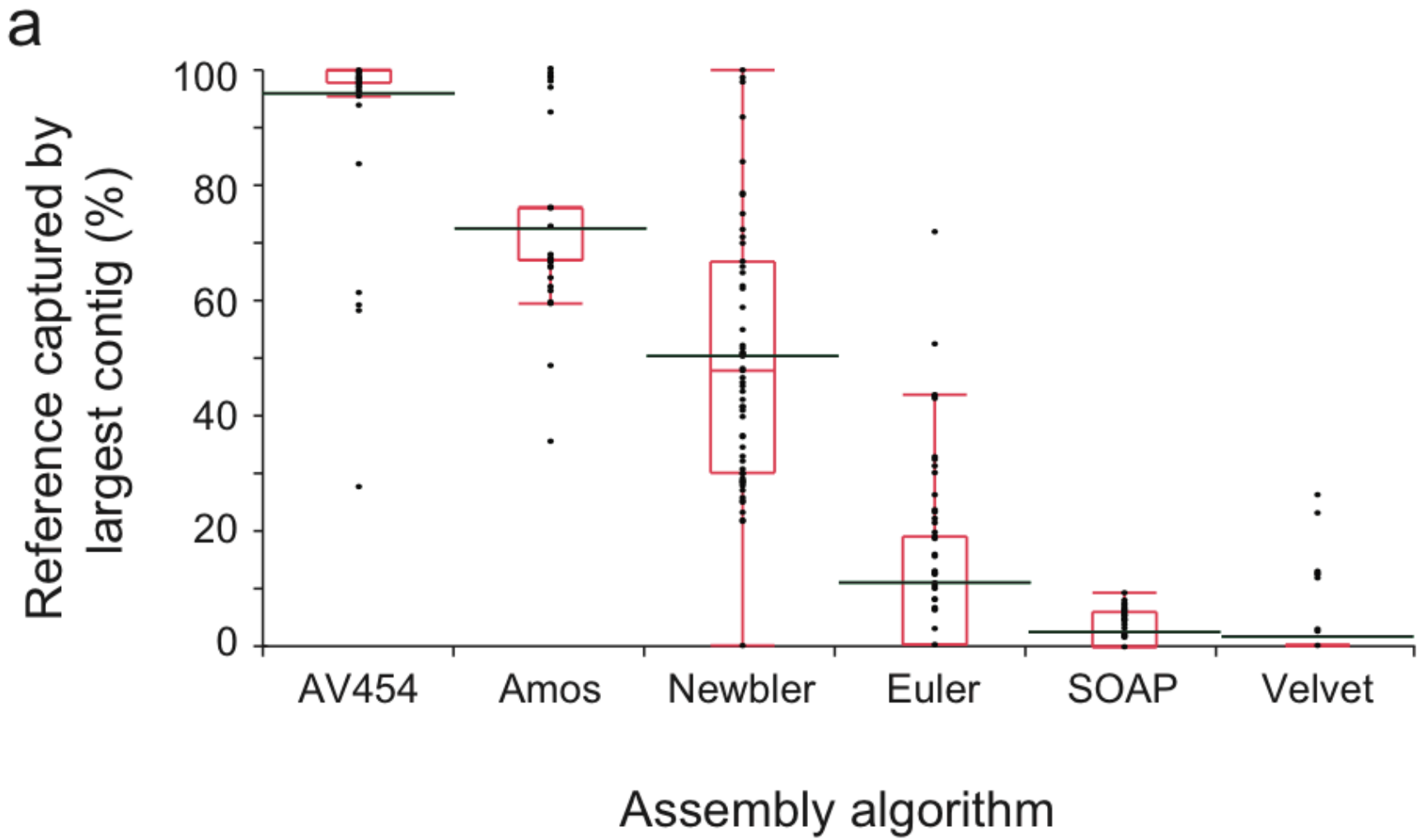
Henn *et al.*, *PLoS Pathog.*, 2012

Genome assembly considerations: Highly variable viral coverage



Malboeuf et al., NAR, *2012*

Genome assembly considerations: Assembler performance on viruses



Henn *et al.*, *PLoS Pathog.*, 2012

Types of sequencing experiments

- Resequencing
- Genome assembly
- RNA-Seq
- Metagenomics

Example uses of RNA-Seq

- Global expression differences
- Annotating genes from a newly sequenced genome
- Discovery of novel genes or transcripts
- Discovery of antisense or other regulatory transcripts
- Variability of isoform expression across conditions

Steps of an RNA-Seq experiment

- Extract RNA from samples
- Enrich for mRNAs
- Make cDNA from RNA
- Fragment the cDNA
- Library construction
- Sequencing

Considerations before an RNA-Seq experiment

- Number of samples needed (conditions and replicates)
- Number of reads needed
- Optional specialized techniques
- Length of reads
- Single end or paired end sequencing
- Two methods of analysis:
  - Align then assemble
  - Assemble then align
- Measuring transcript levels by RNA-Seq

Number of samples needed

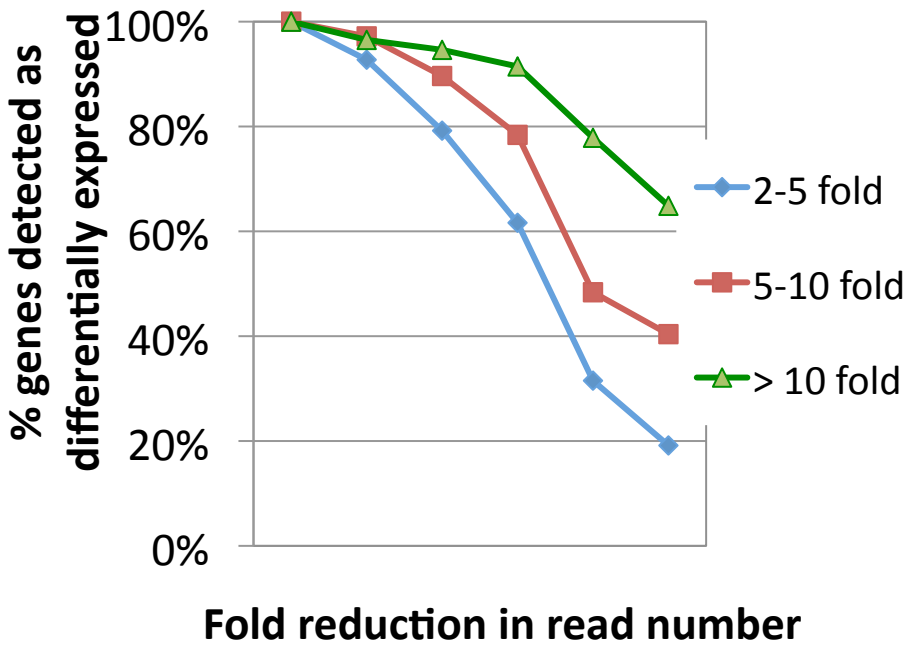
- Number of conditions or tissues determined by experiment:
  - For differential expression, what are you comparing
  - For novel discovery, what are the relevant tissues, conditions, or time points?
- Number of replicates determined by biological variability among replicates
- Website to help estimate optimal power: Scotty
  - http://euler.bc.edu/marthlab/scotty/scotty.php

Number of reads needed

- Need enough reads to identify (and quantify) all transcripts of interest
- How abundant are transcripts of interest?
- What fraction of all transcripts in the cell are in your transcripts of interest?

Number of reads needed

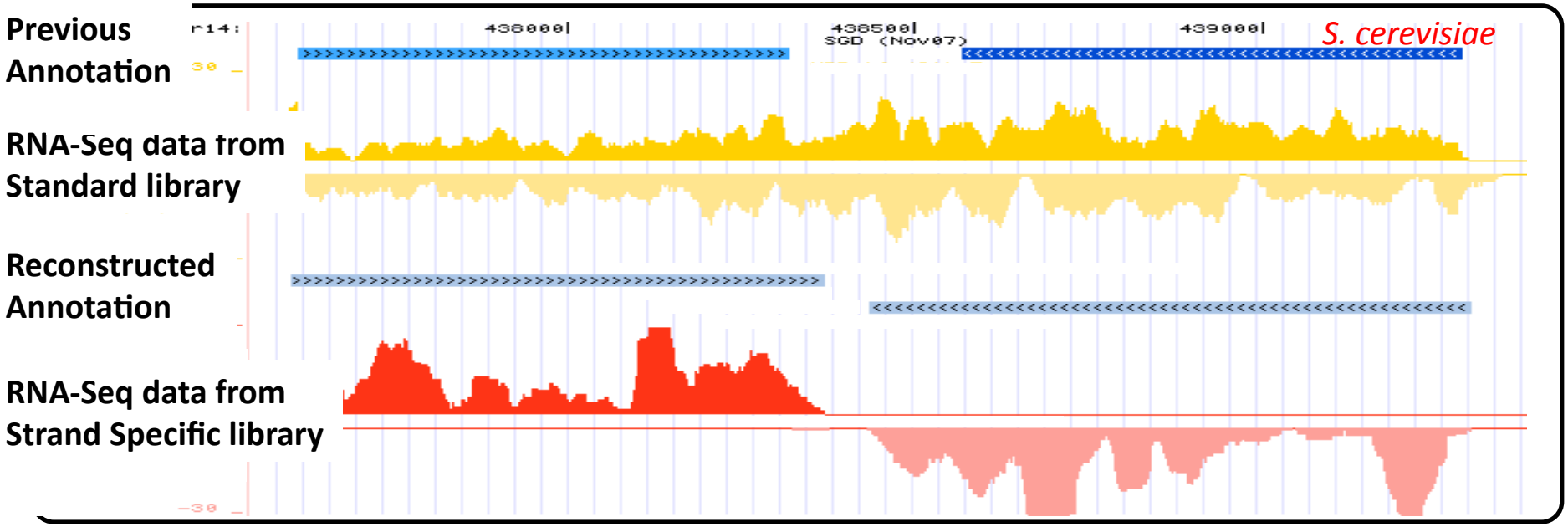
- How large are expression differences?
- Determines significance of the statistical difference



Optional: Strand-specific libraries

- Standard techniques make sequencing libraries that lose the strand of the transcripts
- Multiple special techniques exist to preserve strand information (Levin, *Nat. Methods*, 2010)
- Strand-specific libraries make it easier to annotate:
  - Starts and stop of overlapping genes on opposite strands
  - Low abundance transcripts
- Cons: extra steps, extra cost

Strand-specific libraries



Joshua Levin and Moran Yassour

- Better resolution of overlapping genes
- Can therefore improve annotations

Length of Reads

- Longest high quality reads you can get
- Reads should be at least 75 bp to take advantage of the best analysis tools
- Caveat: very long reads may create a problem if they span more than 2 exons

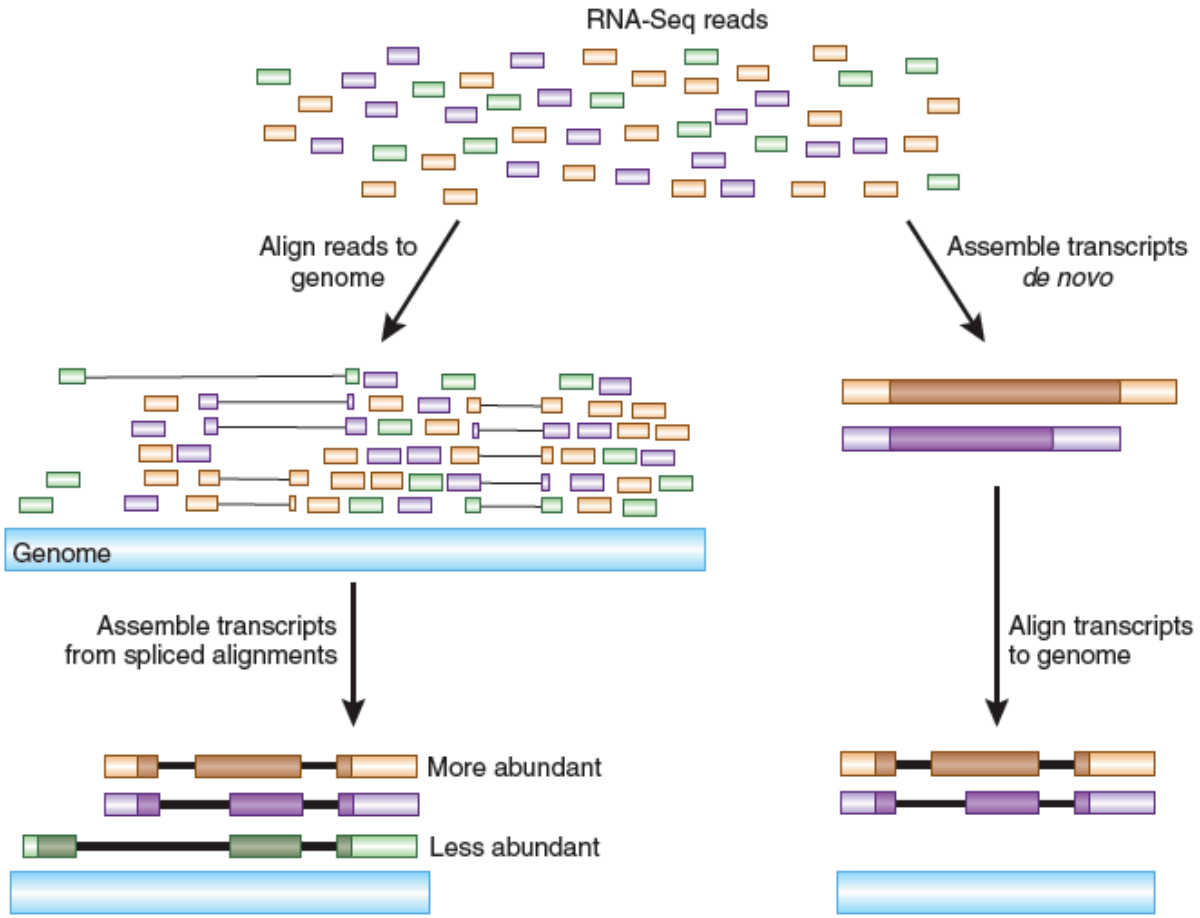
Single end or paired end sequencing

- Always generate paired reads if possible
- Read pairing is used to assemble transcripts
- Exception: Aligning to known transcripts for expression

Example RNA-Seq Runs

- Human expression (per condition):
  ¼ lane HiSeq, 76bp paired
- Vertebrate annotation (per tissue):
  ¼ lane HiSeq, 101 bp paired, strand-specific
- Bacterial and fungal annotation:
  1/12 lane HiSeq, 101 bp paired, strand-specific

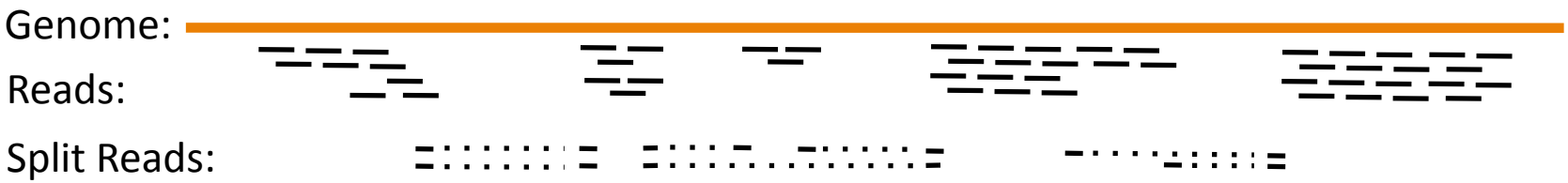
Two methods of RNA-Seq analysis



from Haas and Zody, *Nat. Biotechnol.*, 2010

Analysis method: Align first

- Leverages the genome to guide construction of transcripts
  - Allows complete reconstruction at lower coverage
  - More power to detect low abundance transcripts
- Dependent on having a good reference genome



Analysis method: Assemble first

- Works without a reference
- Must use if organism has no or poor quality reference genome
- Good for reconstructing full length models for moderate to high abundance transcripts
- Does poor job of reconstructing models for low abundance transcripts

Measuring transcript levels by RNA-Seq

- Read count from a transcript is proportional to transcript levels, with two considerations:
  - Transcripts differ in length
  - Experiments differ in total read count

Measuring transcript levels by RNA-Seq

- Read count from a transcript is proportional to transcript levels, with two considerations:
  - Transcripts differ in length

    Normalize: divide read count by length in kb
  - Experiments differ in total read count

    Normalize: divide read count by millions total reads
- Resulting value in **RPKM**
- For paired end sequencing, count each fragment once whether one or two read align = **FPKM**

Caveats exist when measuring expression by RNA-Seq

- RNA-Seq values can be compared across different experimental conditions
- Current programs that perform statistical tests on RNA-Seq data are of variable quality
- Programs like Cufflinks and Cuffdiff are reasonable for comparing genes or isoforms in different conditions, but not perfect
- Genuine differences between conditions are easiest to show with statistical significance if several replicates are used in analysis

Examples of caveats when measuring expression by RNA-Seq

- PCR duplicates don't represent actual counts of RNA fragments, so you need to remove them for quantitation
- Need to be careful about variance:
  - Biological Variance, e.g. Biological variability between replicates of the same conditions may be greater than what is needed to determine statistically significant gene expression changes between conditions
  - Statistical Variance, e.g. When you align reads, they may map to multiple isoforms or multiple paralogs, so you need to assign those reads fractionally to get total transcription levels

Types of sequencing experiments

- Resequencing
- Genome assembly
- RNA-Seq
- Metagenomics

Example uses of metagenomics

- Characterize species present in an environment
- Determine differences in an environmental population measured at different times or conditions
- Associate metagenomic results with environmental conditions (e.g., host health)

Steps of a metagenomic experiment

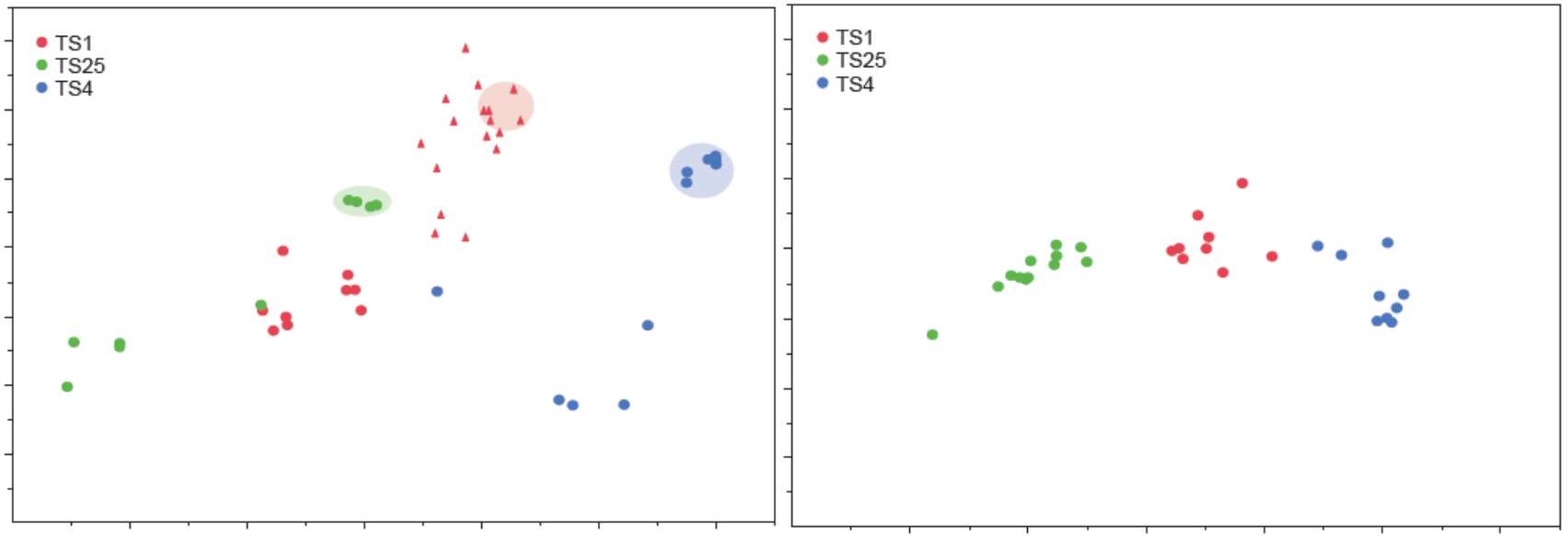
- Extract DNA from samples
- Fragment the DNA (or amplify 16S if not doing whole-genome shotgun sequencing)
- Library construction
- Sequencing

Considerations before a metagenomics experiment

- Reproducibility of metagenomic data depends on:
  - Sample Prep
  - Sequencing Technology
  - Analysis tools
  - Read length and read depth
- Results are not consistent across different experimental designs, but are comparable within identical designs

Metagenomics: Different sample preps

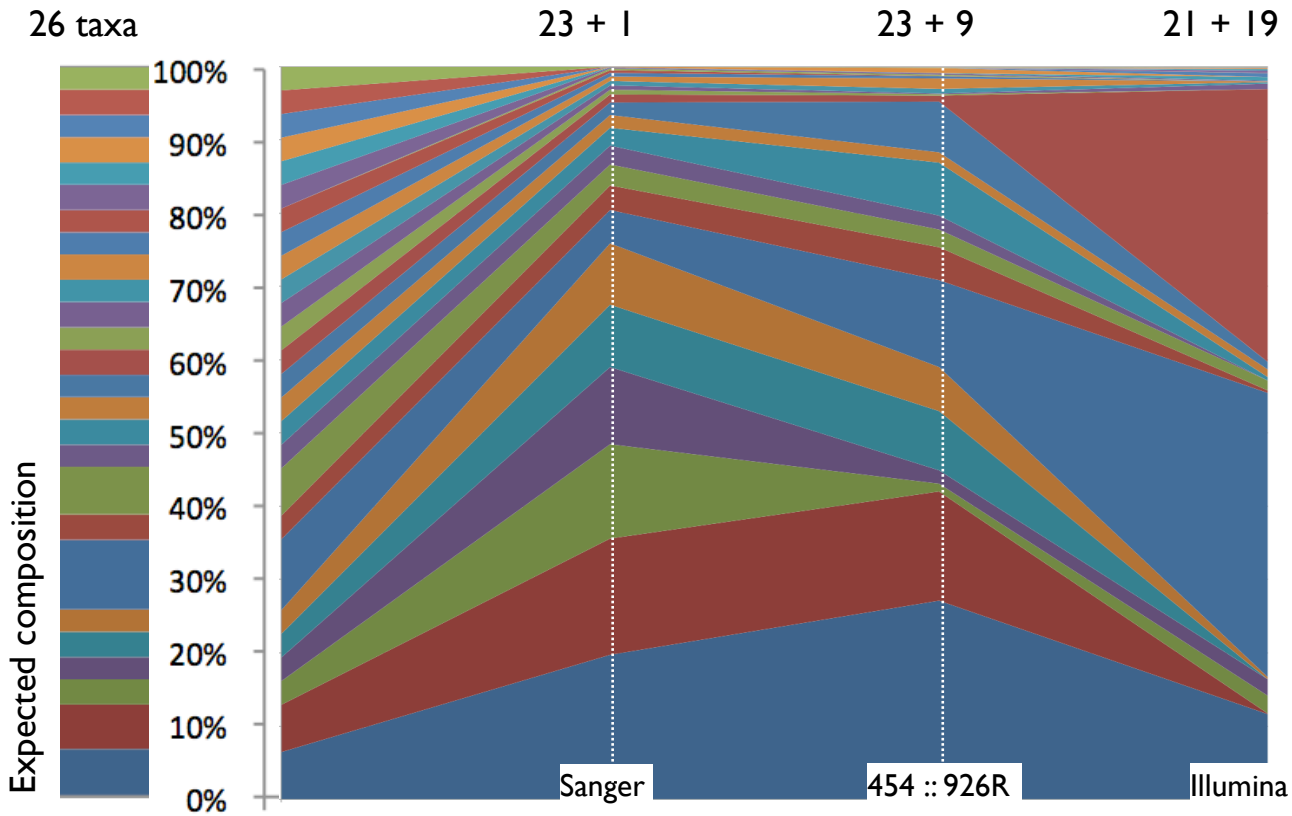
- PCA plots from three samples (colors) sequenced by three groups using different (left) versus identical (right) protocols for sample prep



Human Microbiome Project Data Generation Working Group, submitted

Metagenomics: Different sequencing technologies

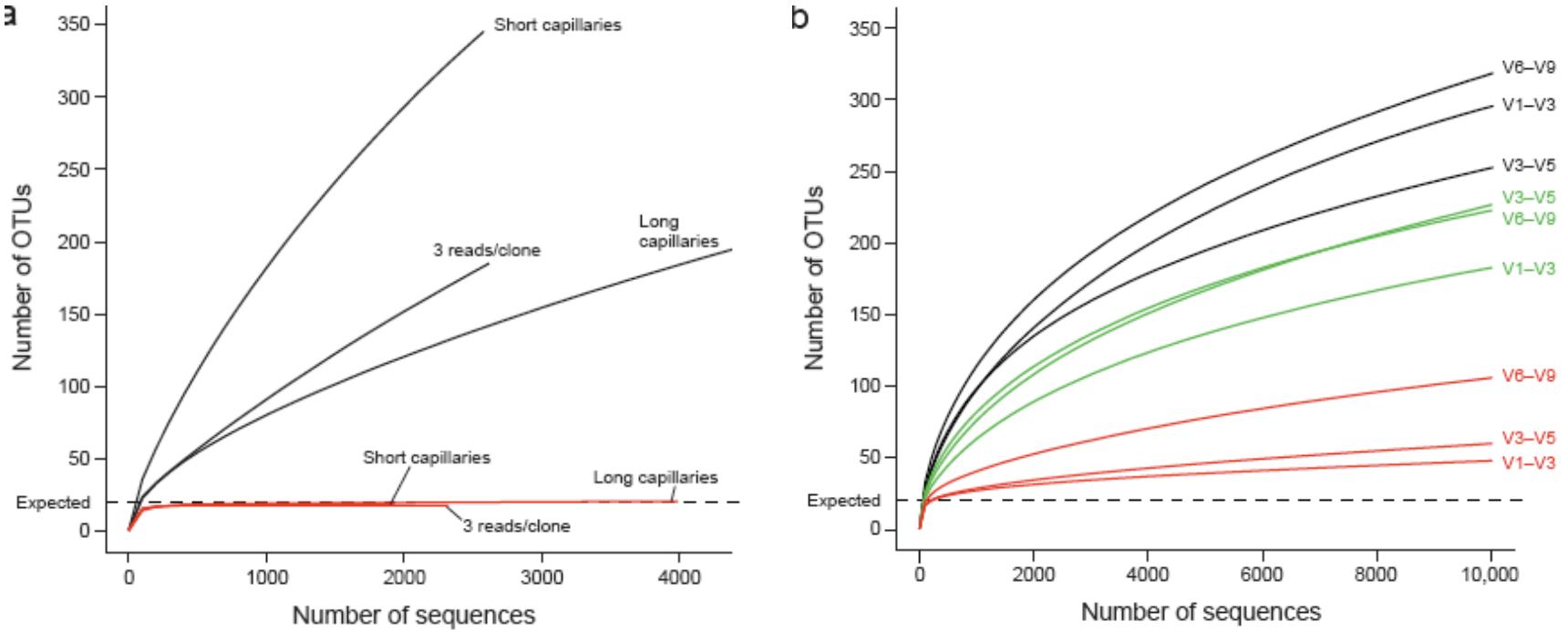
- Same (known) mock community sequenced on 3730, 454, and Illumina



From Dirk Gevers

Metagenomics: Different sequencing & analysis techniques

- Mock community of 21 samples sequenced by 3730 (left) and 454 (right) show greatly different numbers of taxa when filtered differently



Human Microbiome Project Data Generation Working Group

Metagenomics: Considerations for read type

- Length is very important for all strategies
  - For 16S, length provides more of target
  - For WGS, better assemblies
  - More chance of indentifying gene from single read
- Importance of pairs depends on strategy
  - For 16S, provides more length only
  - For assembly methods, very important
  - Will not help much with direct gene finding

Course Outline

- Considerations before starting a sequencing experiment
- Steps of generating sequencing data
- Considerations before starting specific types of sequencing experiments
  - Resequencing
  - Genome assembly
  - RNA-Seq
  - Metagenomics