Genomic Study Design

Michael Zody

Workshop on Genomics

October 2011

Overview

1. Introduction
2. Global considerations for sequencing studies
3. General technology issues
4. Design criteria for specific experiment types
5. Dealing with the data you have

Why Does Study Design Matter?

Question:

"What should I use to do this analysis?"

Answer:

"A data set with enough power."

What Do We Mean by Power?

- Technical: what is the chance that my experiment could detect a real event/effect/ difference in my samples? (at a given FP rate)
- Traditional: do I have enough samples to see this difference?
  - Typically the assay gives a result for most samples
- Sequencing: am I generating enough sequence for each sample to get a result?

Example: Genotyping

- Genotyping chip:
  - ~95% samples work
  - ~99% loci work (for well-designed chip)
  - ~99% yield of calls per sample
- Could also genotype by sequencing
  - All of those parameters would be dependent on coverage

What Really Happens on the Chip

- The chip is binding DNA to features and "counting" it by fluorescence intensity
- If you don't get enough intensity, the chip calling software won't make a call
- You never see those intermediate results, which are estimates of single molecular events
- If your chip protocol wants 1 µg of DNA and you use 0.1 µg, you'd probably get nothing

What Happens with Sequencing

- You get a pile of reads across your SNP
- This is a precise count of alleles observed
- You need enough counts to know what's there
- The counts will vary by site
- If your protocol says you need 50 reads per site, and you have 5, you could still make some calls, but they wouldn't be much better than the chip did on 10% of the needed DNA

Overview

1. Introduction
2. Global considerations for sequencing studies
3. General technology issues
4. Design criteria for specific experiment types
5. Dealing with the data you have

Global Considerations

- What is the question?
- What kind of sequencing experiment is this?
- What other resources are available?
- How good does the answer need to be?
- What technological factors are limiting?
- Are there other ways to do this?

What Is the Question?

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

What Kind of Sequencing?

- There are many different sequencing designs
  - You may need more than one!
- The kind of sequencing design or designs you use will influence or be influenced by:
  - Goal(s) of the experiment
  - Available genetic materials
  - Existing “omics” resources
  - Sequencing capacity/cost
  - Analytic methods

What Resources Exist?

- All sequencing analyses except *de novo* assembly require a reference genome
  - If a suitable reference doesn't already exist, *de novo* assembly will be required
- Other resources may be useful if they exist
  - Gene annotations
  - Variation calls (divergence data for inter-species)
  - Chip data for SNPs, expression, ChIP
  - Genetic or other maps

Using a Reference Genome

- How good is the reference?
  - Completeness
  - Accuracy
- How representative is it of your genome(s)?
  - Sequence absent from the reference won't align
  - Using a diverged reference (more than a few %)
    - 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

How Good an Answer?

- Is your sequencing result the final answer, or just a starting point for something else?
- What are the costs of false positives and false negatives relative to the cost of the sequence?
- For example, identifying single base variants might have very different needs depending on the project

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

Case 2: Microbial Evolution

- Sequence an evolved (e.g., drug resistant) microbe to find functional changes
- Low tolerance for false negatives
  - Should be able to find a variant in a small genome
- Relatively high tolerance for false positives
  - Functional mutation is most likely a coding change, so triage of calls for follow up is effective

Case 3: Vertebrate Evolution

- Sequencing to find signatures of selection
- Relatively high tolerance for false negatives
  - Specific sites of variation are not important
- Low tolerance for false positives
  - Background noise from sequencing error can obscure the signature of selective sweeps

Case 4: SNPs for Model System

- Sequencing multiple strains or individuals from a model organism to design a SNP array
- High tolerance for both FN and FP
  - 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

What Factors are Limiting?

- Material/biological
  - Sufficient samples of good quality and quantity
- Sample Prep
  - Can libraries be made from your material?
- Sequencing
  - Read length
  - Pairing
  - Number of reads
  - Complexity of library

Is There a Better Way?

- Many things *can* be done by sequencing
- Other options exist
  - Gel assays
  - Microarrays
  - Capillary sequencing
  - A different kind of sequencing experiment

Case 1: Association Study

- Could sequence every genome
- For organisms with existing chips/arrays
  - Could run more samples on array
  - Could follow up with custom local array
  - Use sequencing once you have a target
- Without existing resource, might be better off to generate that resource first
  - Cheaper for large samples even with start cost

Case 2: SNPs for Model System

- Could generate light (5-10x) coverage of several individuals/isolates to identify SNPs
  - Can only call good SNPs at deeply covered sites
- Could generate light (<0.1x) coverage of several individuals by capillary shotgun
  - SNPs can be accurately called from a single read
  - Likely to yield a variant every 2 reads
- Total cost may be less by capillary

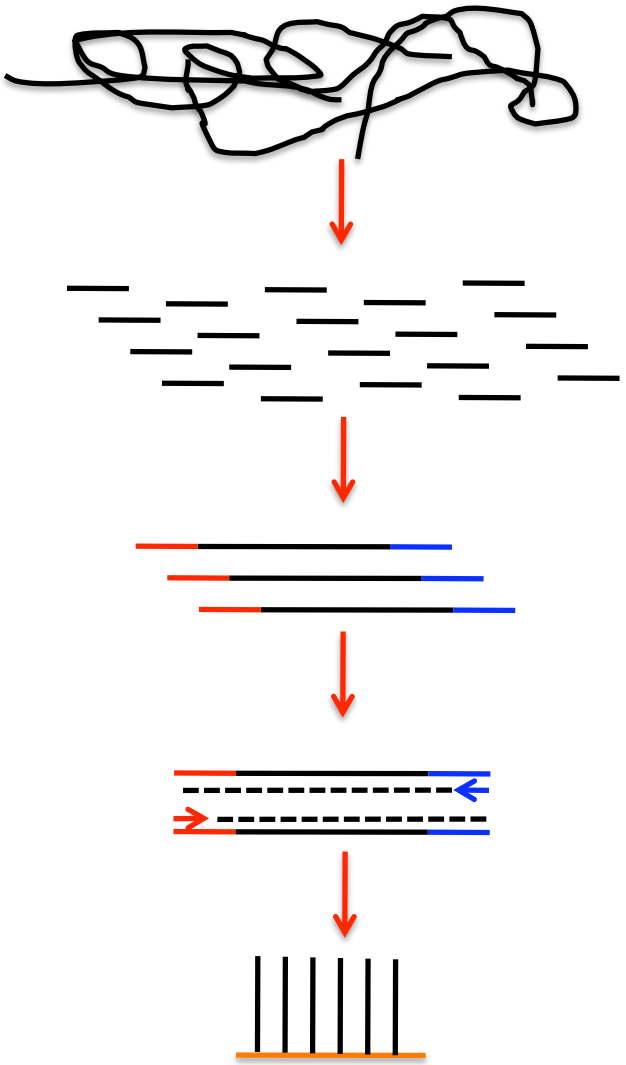
Overview

1. Introduction
2. Global considerations for sequencing studies
3. General technology issues
4. Design criteria for specific experiment types
5. Dealing with the data you have

General Technology Issues

- Sample prep
- PCR artifacts
- Pooling and barcoding
- Types of read data
- Primers, adapters, and tags
- True single molecule
- Aligning reads
- Controls and replicates

Generic Sample Prep



- DNA
  - Extracted/prepared
- Fragment and size (?)
  - Shear, size select
  - 300-600 (<1000) bp
- Add adapters
  - Generic ends
- Amplify (!)
  - Usually needed
- Select single molecules
  - Amplify in cluster/bead

Fragment Size

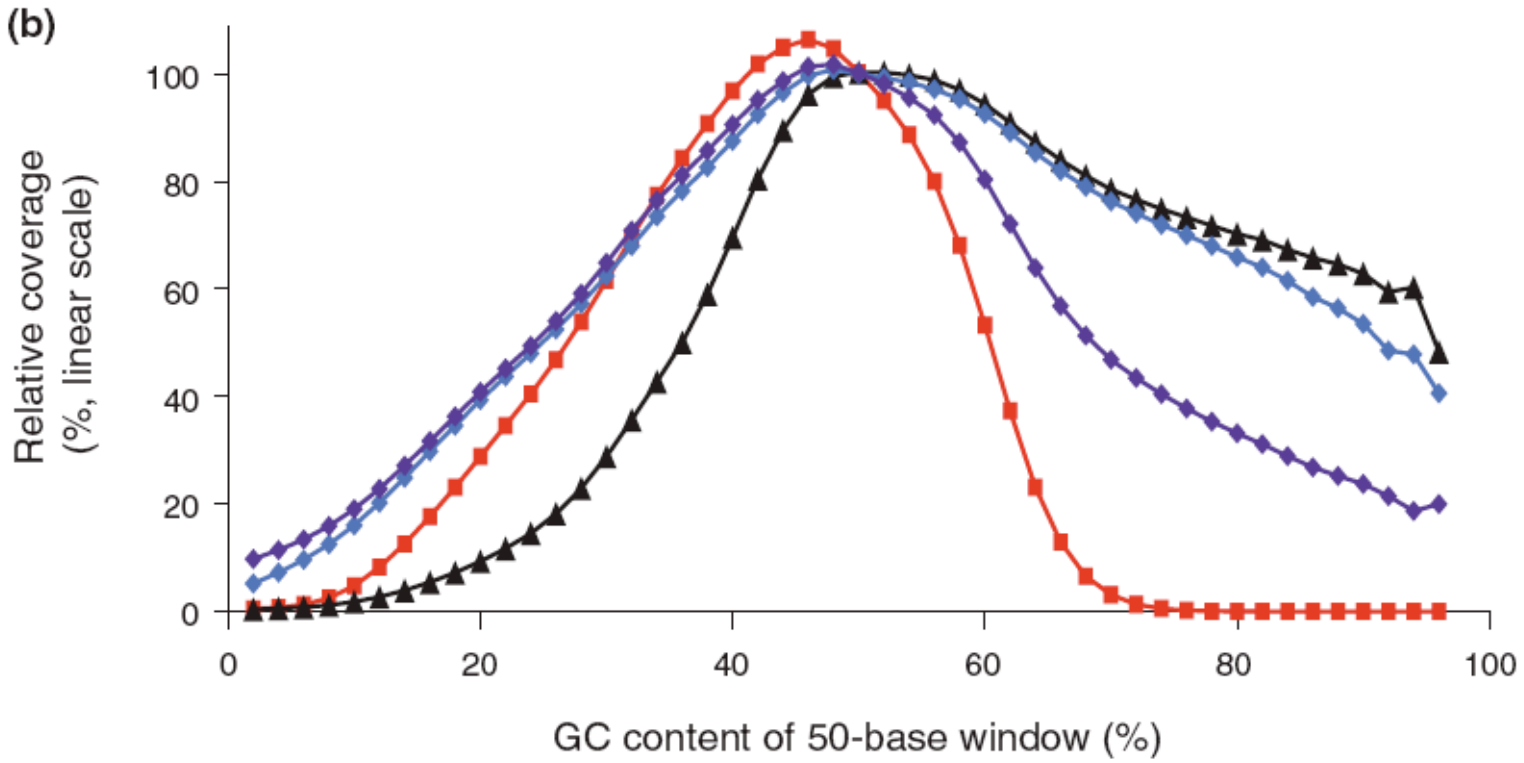
- Most sequencers want input 300-600 bp
  - Long links are a special process
- For longer fragments, shear first, then adapter
  - Genomic, cDNA
- Short, “wrong size”: concatenate then shear
  - SAGE-type DGE, some exon targeting
- Right size amplified products (PCR) can be tailed directly with adapters

PCR Artifacts

- Most libraries see PCR during prep
  - Targeting or amplification of adaptered fragments
- After PCR, there is a single molecule stage
  - Errors of PCR will be “true” bases at this step
    - Undetectable by quality metrics as errors
    - Rate can be quite high, e.g. 1/3000
      - PCR error 1/100,000 x 30 cycles ~ 1/3000
- Chimeric sequences (esp. in targeted designs)
- Duplicated sequences

PCR Bias

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



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

Pooling with Barcoding

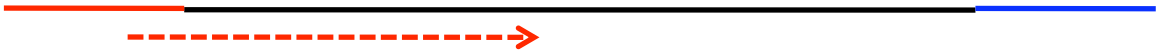
- Unique DNA tags identify samples
- Allows multiple distinct samples on one run
- Advantages
  - Reduced cost of sequencing for small samples
  - Analysis is identical to unpooled data
- Disadvantages
  - Some small throughput loss due to barcode fails
  - Increased per sample cost for library construction

Pooling without Barcoding

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

Types of Read Data

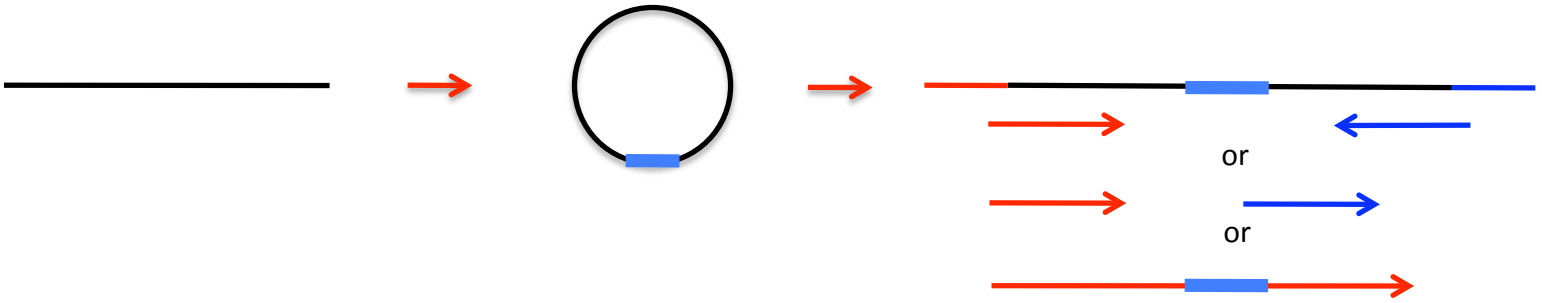
- Fragment reads
  - Single read in one direction from a fragment



- Paired end reads
  - Two reads from either of the same fragment
  - Pointing towards each other

Types of Read Data

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



Why Different Library Types?

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

Why Different Library Types?

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

Primers, Adapters, and Tags

- Not every base you sequence is useful
- Primers will be present if you PCR-targeted
  - Sequence from primers does not represent target
  - Variation seen (or not) under primers is not real
  - Overlapping products will allow analysis
- Short fragments may read through to adapter
- Custom barcodes or other tags
  - Most vendor tags will be removed automatically

True Single Molecule

- Most high throughput methods isolate a single molecule, but sequence amplified clusters
- A few technologies are true single molecule
- True single molecule techniques are less subject to amplification-related bias
- Single molecule techniques have no redundancy, so higher error rates
- Most still require abundant starting material

Aligning Reads

- 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, many more sequences, so speed matters

Controls and Replicates

- You can publish next gen analyses without!
- They can be useful
- Resequencing the reference
  - If DNA (or appropriate input) exists for reference individual, sequence will control for alignability
- Unenriched samples (for ChIP-Seq)
- Replicates (for RNA-Seq)
  - Variance of read depth is larger than Normal

Overview

1. Introduction
2. Global considerations for sequencing studies
3. General technology issues
4. Design criteria for specific experiment types
5. Dealing with the data you have

Sequencing Experiment Types

- Resequencing/variant discovery
- Targeted resequencing
- *De novo* assembly
- RNA-Seq
- ChIP-Seq
- Metagenomics
- Sequencing as an assay for something else

Resequencing Design

- Requires a genome!
  - Quality of the resequencing bounded by genome
- Considerations:
  - Alignability
  - Coverage
  - Read length
  - Read pairing
- What do you want to find?

Alignability

- Not all of the reference is accessible
- Parts are too similar for unique alignments
  - Duplications, recent repeats, gene families
- Longer reads and pairing increase alignability
  - Example, for 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)*

Coverage for SNP Finding

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

- Why do we need so much?

Example: Haploid SNPs

- Know there is only one base at each locus
- Make majority call
- How likely is a correct majority call?
  - Assume uniform 1% error rate

Depth at Locus	% Correct Majority	% No Majority	% Error Majority
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	

Adjusting for Random Sampling

- Probabilities if reads randomly distributed:

Average Cov	% Correct Maj	% No call	% Error Maj
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

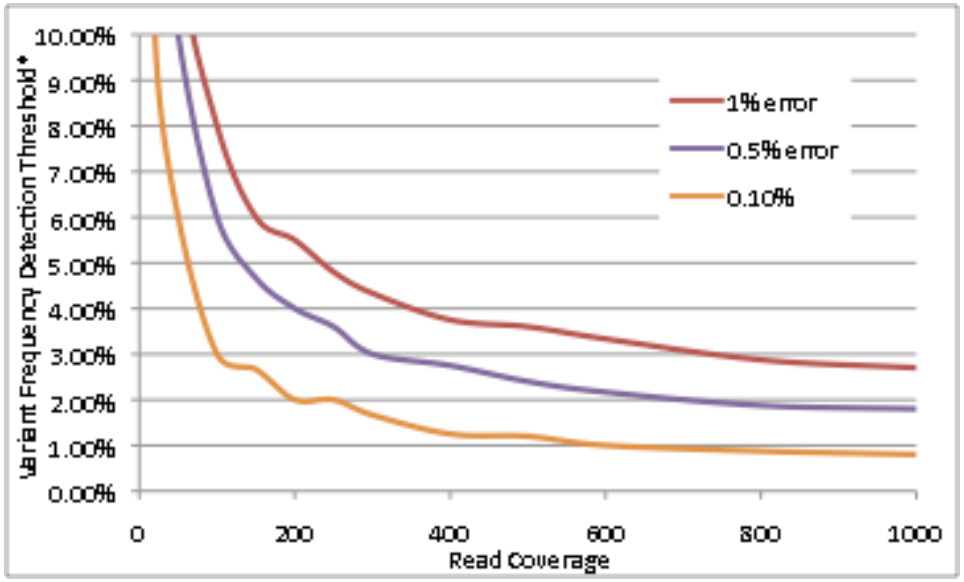
- Reality will be worse: reads are not random

Diploid or Aneuploid Samples

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

Continuous Variation

- Pooled or host/environmental samples
- Want to find all real variants
- What sensitivity do we have at X coverage?



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

Read Length and Pairing

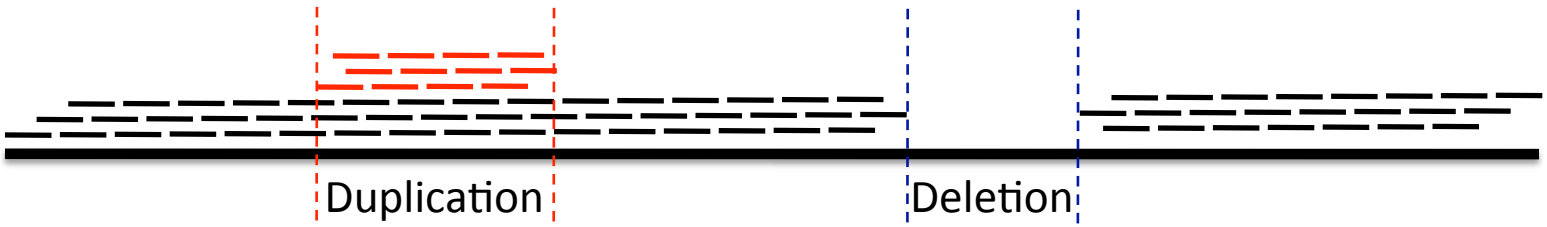
- Read length only matters for alignability
  - (For equal total coverage)
- Paired end reads also only help alignment
  - Aligning one end uniquely localizes other end
  - Aligners may use this to run more sensitive align
  - Allows finding highly variant regions and small indels if the other read aligns cleanly
- Long links are of relatively little use for SNPs

Discovery of Structural Variants

- Read depth can identify copy number changes
- Paired end spacing can mark regions of insertion, deletion, or rearrangement
- Long reads can be aligned at multiple places (split-read alignment) to find breakpoints
- *De novo* assembly (global or local) can find novel insertions and define breakpoints

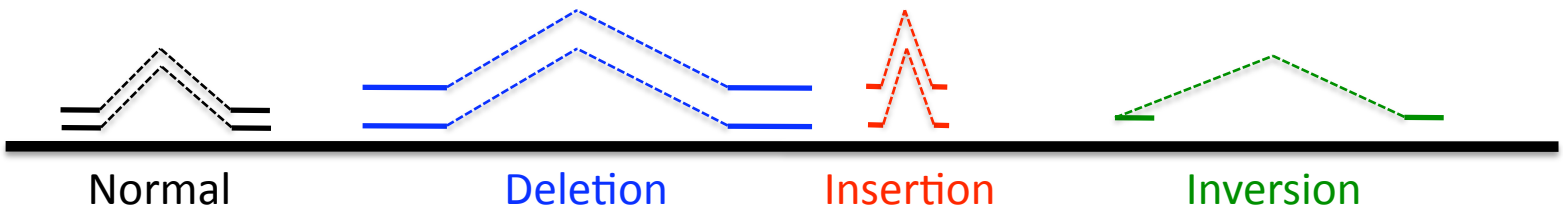
Read Depth Analysis

- Can use depth of coverage to estimate copy
- Caveats:
  - How many copies of the duplication in reference?
  - How similar are copies?
  - Are events homozygous or heterozygous?



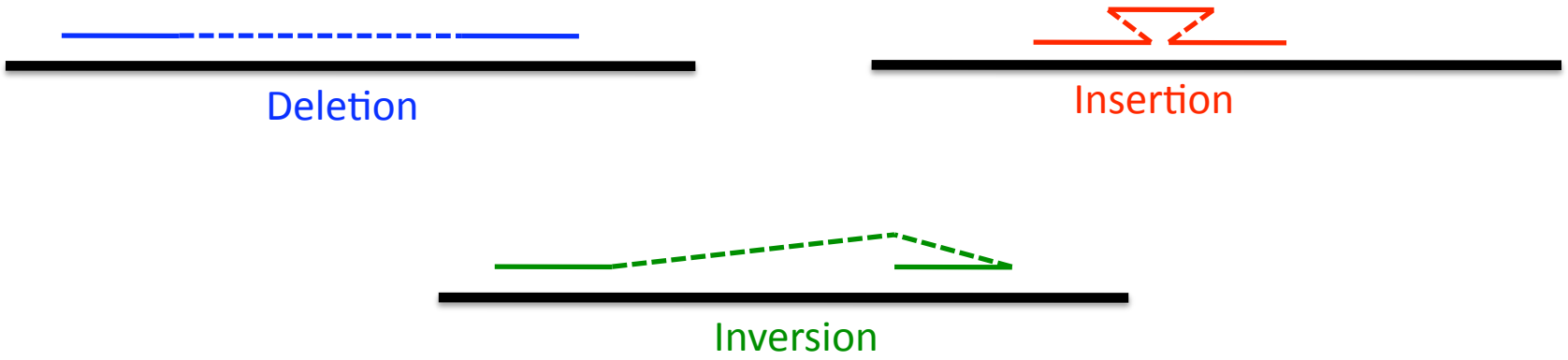
Paired End Analysis

- We expect a certain orientation and spacing
- If these vary, they signal rearrangement
  - Deletion: reads too far apart
  - Insertion: reads too close together
  - Inversion: too far apart and wrong orientation
- Works better with long pairs (jumps)



Split Read Alignments

- Gives base level breakpoint resolution
- Only works with long reads
  - Short reads have too many spurious splits
- Caveat: breakpoints are often duplicated
  - Reads won't split if single alignment is as good



De Novo Assembly

- Assemble whole genome
  - Align contigs to reference
  - Look for insertions, deletions, rearrangements
- Use paired ends to identify potential events
  - Assemble unaligned reads whose mates align near the event
  - Iterate this to build up an insertion or a deletion breakpoint

Targeted Resequencing

- Mostly similar to whole genome resequencing
- Targets specific region or regions (e.g., exome)
  - PCR amplification
  - Hybrid selection
  - Targeted genome amplification
- Some special analysis considerations

Targets Require More Coverage

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

Alignment and Targeting

- Targets including repeats and duplications
  - Pull other copies of those sequences
  - Need to align to whole genome to insure that unique hits in target are best in genome
- True of off-target reads even if targets unique
- Aligning to targets first and then only aligning hits to whole genome can save some compute

Should You Target at All?

- Significant cost savings if target <<< genome
- Can achieve higher coverage on target
- Drawbacks
  - Cost of targeting reagents can be high
  - Some sequenceable regions very hard to target
  - Variability of coverage is higher
  - Miss untargeted sequences (does it matter?)
  - Targeting may introduce bias

De Novo Assembly

- Why assemble?
  - No genome reference
  - Identify novel insertions, other structural variants
  - Alternative method of SNP finding
    - Mostly for small, haploid genomes
    - Provides better diversity calling for indels and particularly difficult to align regions

Requirements for Assembly

- Very deep coverage (at least 50x, 100x better)
- Long reads help greatly
  - Provide connectivity through low coverage
  - Resolve repetitive/duplicated regions
  - For 454 (450 bp reads), can assemble at 20x
- Paired reads necessary for complex genome
- Long links (jumps) are not always necessary, but yield much better connectivity

Coverage Requirements

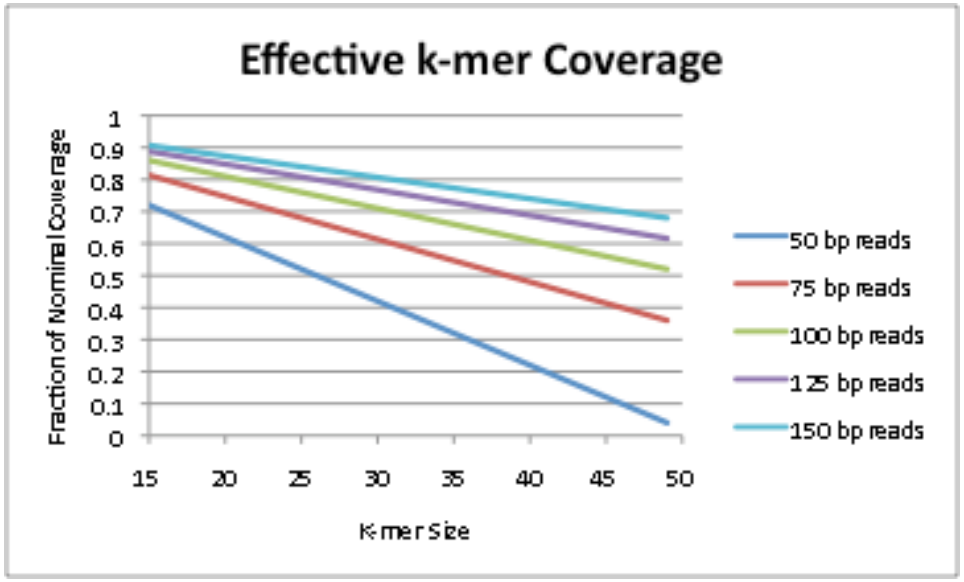
- Next Gen assembler use DeBruijn graphs
- Chains of k-mers

```
ATGTGTACGTACGTACGTA
 TGTGTACGTACGTACGTAT
  GTGTACGTACGTACGTATC
   TGTACGTACGTACGTATCC
    GTACGTACGTACGTATCC?
```

- To continue assembling, next k-mer must overlap last and extend one base
- What is the probability that a read exists that will extend the graph?

Read Coverage ≠ k-mer Coverage

- If the last k-mer starts at position i, there must exist a read that starts at or before i + 1 and extends to i + k
- Effective coverage is (L - k + 1) / L



How Large Does k Need to Be?

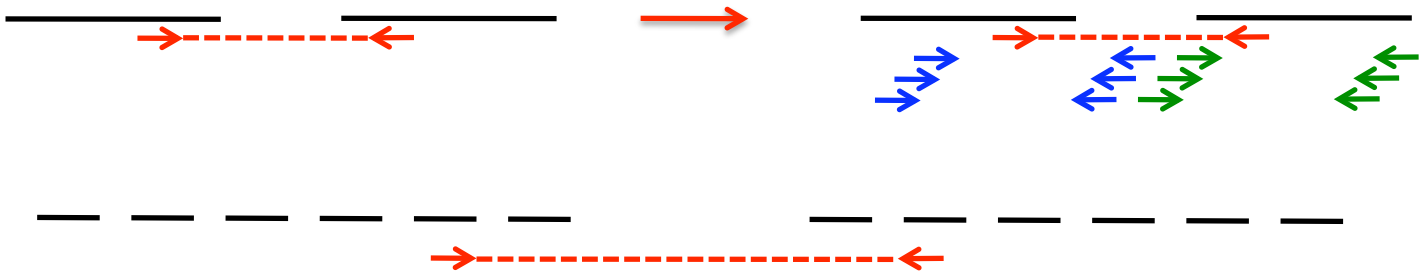
- Cannot directly resolve repeats longer than k
- Almost all genomes have some long repeats

K	*E. coli* (%)	*S. cerevisiae* (%)	*A. thaliana* (%)	*H. sapiens* (%)
200	98.5	95.9	97.4	97.6
160	98.3	95.6	97.1	97.2
120	98.2	95.2	96.6	96.6
80	98.0	94.7	95.4	95.2
60	97.8	94.4	94.4	93.1
50	97.7	94.2	93.4	91.2
40	97.6	93.9	92.2	88.3
30	97.4	93.5	90.4	83.4
20	97.0	92.9	86.5	71.8
10	0.0	0.0	0.0	0.0

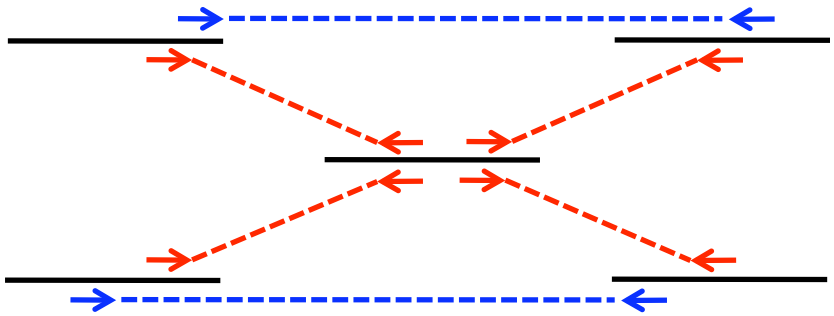
From Butler et al., Genome Res. *(2008)*

Paired Reads in Assembly

- Span and fill gaps



- Resolve repeats



RNA-Seq

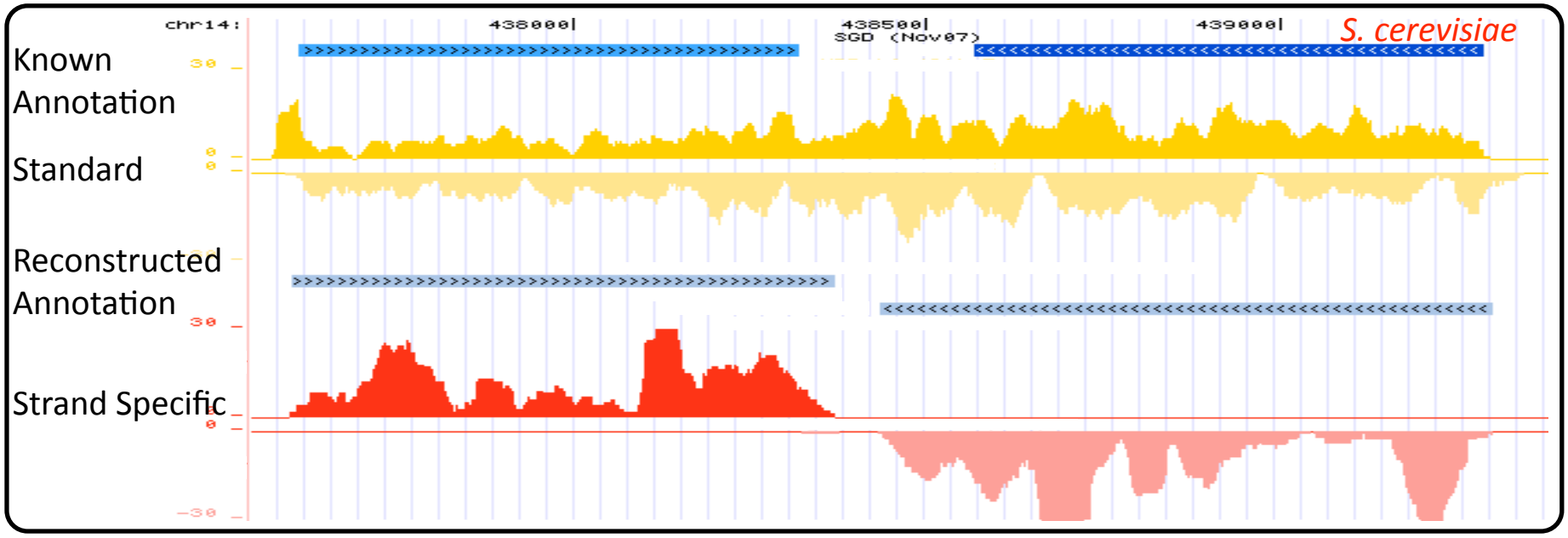
- Capture information about transcriptome
- Using a reference genome
  - Like resequencing, but additional challenges
  - Coverage is uneven
  - Reads may be spliced
- Without a reference genome
  - Align to existing ESTs or cDNAs
  - *De novo* assembly of transcripts

Making RNA-Seq Libraries

- Number of standard protocols for prep
- Strand-specific libraries
  - Strand of reads match strand of transcribed RNA
- Normalization of libraries
  - Reduce high abundance transcripts
- Hybrid selection for RNA-Seq
  - Enrich specific transcript targets

Strand-Specific Libraries

- Better resolution of overlapping genes
- Detection of anti-sense transcripts



Joshua Levin and Moran Yassour

Strand-Specific Methods

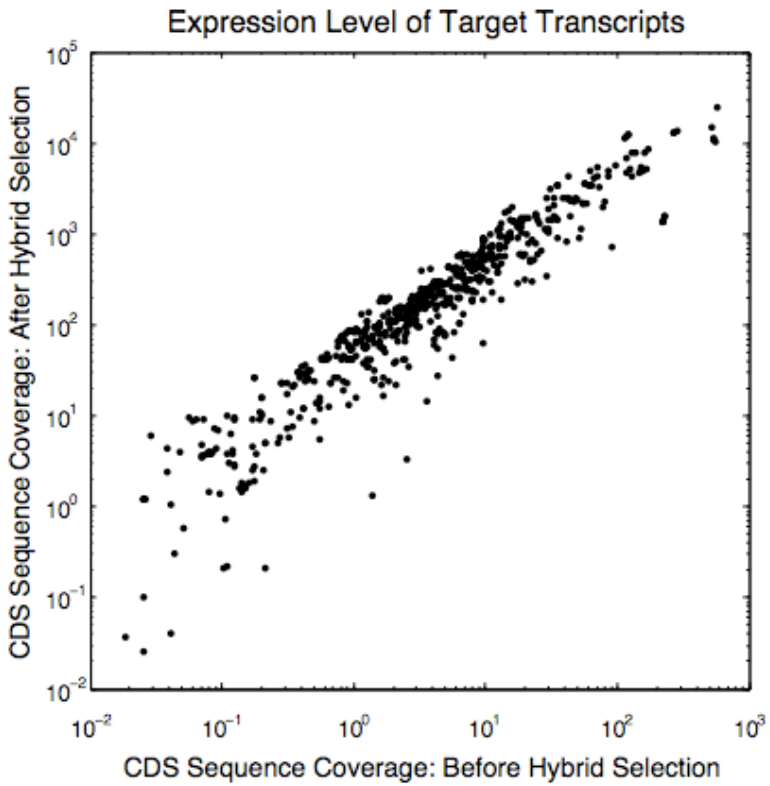
- Multiple methods available
  - Compared in Levin *et al., Nat. Methods* (2010)
- Factors
  - Ease of use
  - Input RNA required
  - Degree of strand specificity
  - Uniformity of coverage

Normalization

- Input material for RNA-Seq may span several orders of magnitude
- Complete and quantitative sequencing of low abundance transcripts requires many reads
- Or, reduce the abundance of most common
- Normalizing excludes expression quantitation
- Will not rescue very low abundance
  - Removing top 50% abundance only doubles low

Hybrid Selection

- Similar to targeted resequencing
- Capture from fully prepared library
- Increases coverage of target genes
- Maintains relative expression levels!
- Low input/mixed RNA



From Joshua Levin & Mike Berger

Reads for RNA-Seq

- Read length is very important!
  - Detecting spliced reads much easier if longer
  - Short reads align in exons or to mature transcripts
  - Spanning multiple exons confirms isoforms
- Pairs are also very important
  - Pairs landing in different exons confirm transcript structure, includes pairs in non-adjacent exons
  - Help with unique placements in reference aligns
  - Scaffold *de novo* assembly of complete transcripts

Digital Gene Expression

- Specifically target transcript tagging sequences (as in SAGE) instead of full transcripts
- Sequencing single tags
  - Requires very short reads, fragment only
  - Requires many reads
- Concatenate and shear (traditional SAGE)
  - Longer reads much better
- Not much advantage over full length RNA-Seq

ChIP-Seq

- Also applies to other methods of non-sequence-based enrichment (e.g., methyl-cap)
- Goal: identify regions of the genome with that feature and quantify their occupancy
- Fragment length is often set by capture
- Sequence is not important except as a means to identify genomic position

Reads for ChIP-Seq

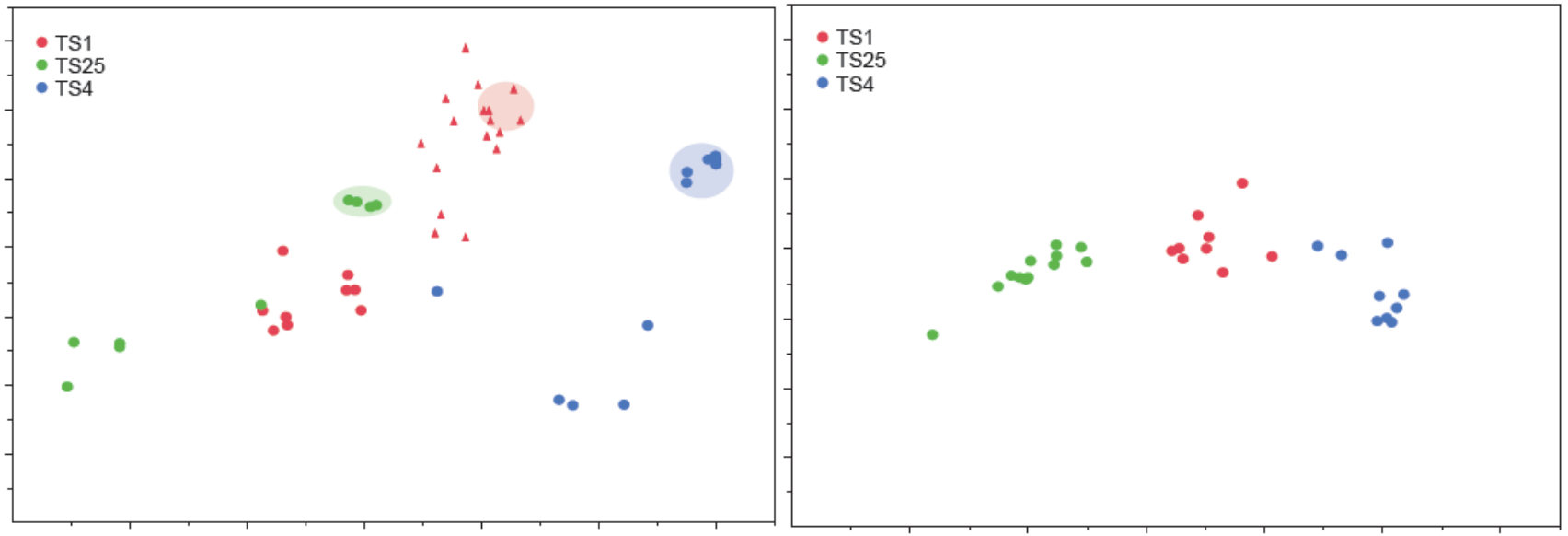
- Need only be long enough to place uniquely
- Paired ends are not needed
  - Other end of fragment can be inferred
  - Would be helpful only for unique placement
- Loci not uniquely alignable often excluded
- Smaller enrichment requires more reads
  - Non-specific capture
  - High background rate (e.g., nucleosomes)

Metagenomics

- Inherently pooled sequencing, but harder
  - May not have a reference (WGS metagenomics)
  - No known bound on number or breadth of taxa
- Many factors can affect results
  - Sample Prep
  - Sequencing Technology
  - Read length and read depth
  - Analysis tools

Different Sample Preps

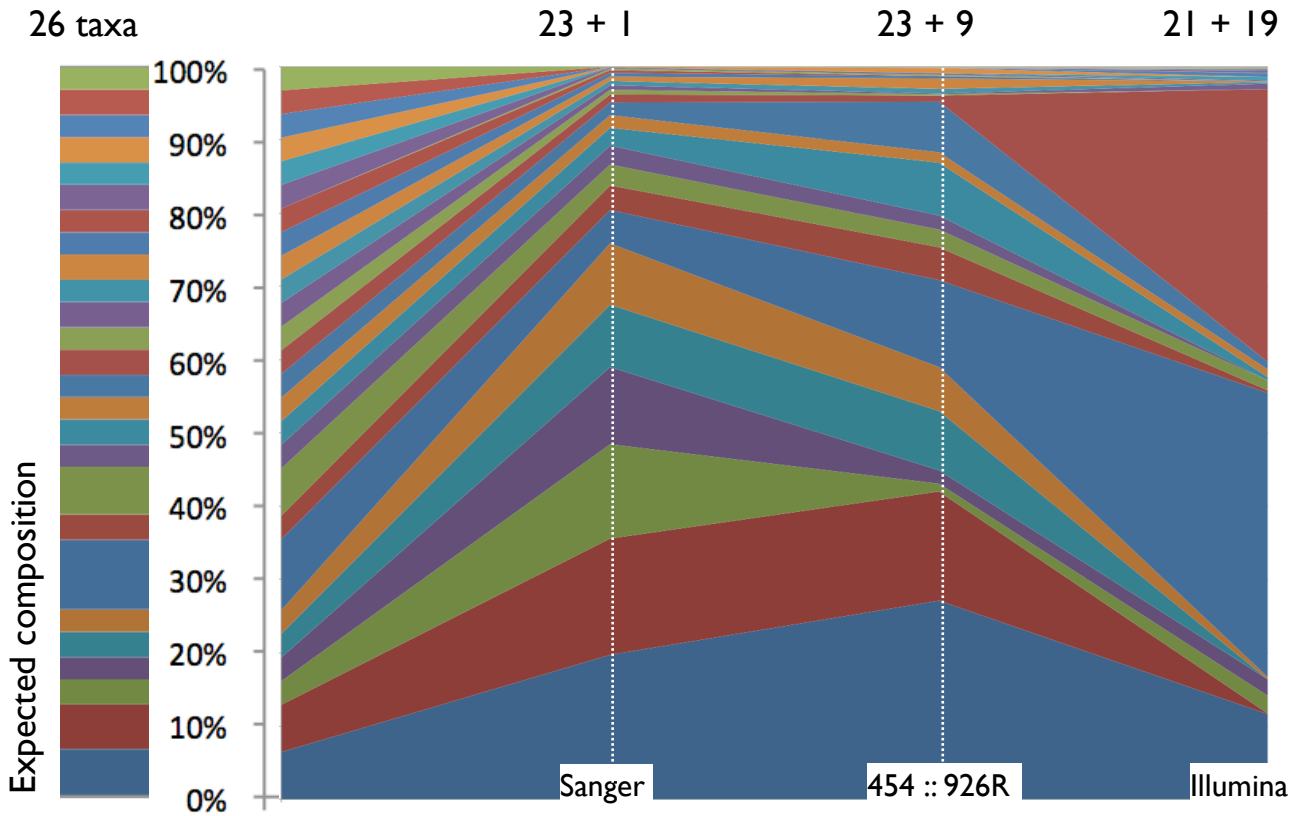
- PCA plots from three samples (colors) sequenced by three groups using different protocols (left) and identical protocols (right)



Human Microbiome Project Data Generation Working Group, submitted

Different Sequencing Technologies

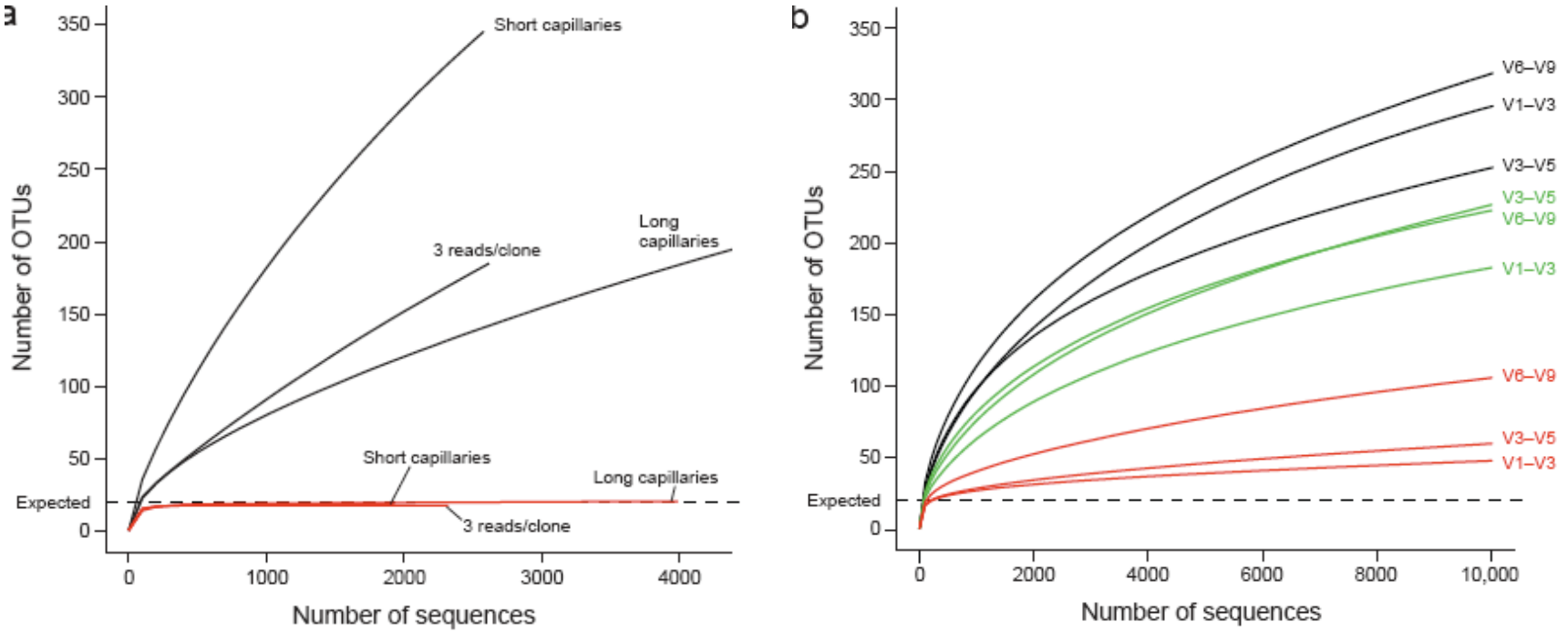
- Same mock community (known) sequenced on 3730, 454, and Illumina



From Dirk Gevers

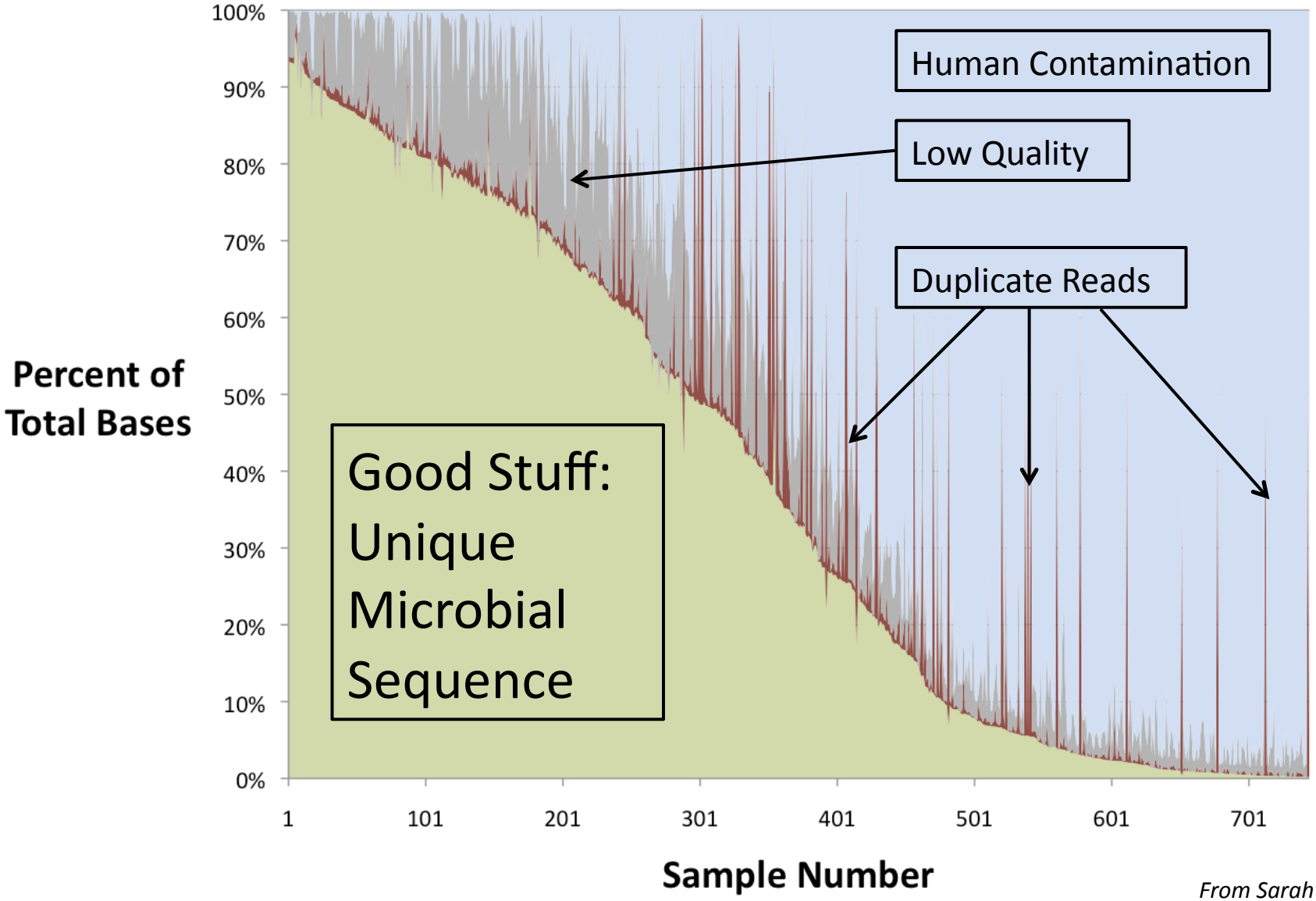
Different Sequencing and Analysis

- Mock community of 21 samples sequenced by 3730 and 454 and filtered differently show greatly different numbers of taxa



Human Microbiome Project Data Generation Working Group, submitted

Filtering Whole Genome Data



From Sarah Young

Reads for Metagenomics

- 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

Sequencing as an Assay

- Sequence DNA that is tagging something else as a readout
- Ligate two disparate pieces of DNA
  - Hi-C (chromosome conformation)
  - Protein-protein interaction with DNA tails
- Use DNA tails to label, count by sequencing
- Reads long, accurate enough to distinguish
- As many reads as possible

Overview

1. Introduction
2. Global considerations for sequencing studies
3. General technology issues
4. Design criteria for specific experiment types
5. Dealing with the data you have

Suboptimal Data

- Sometimes you can't design the experiment
  - Samples or resources are limited
  - Data were already collected
  - Piggybacking on another experiment
- How can you make the best of this?

- Understand what you really cannot do

Filter Your Analysis

- Most common failure is lack of sufficient data
- Identify regions where data are sufficient
  - Sites with enough coverage for SNP calls
  - Contigs consisting of multiple read lengths
  - Genes with reasonable transcript coverage
- Analyze these sites and try to extrapolate
- Check that the sites you are analyzing are reasonable tails and not outliers or artifacts

Combine Data

- You may have multiple samples or even multiple data types that could be used
- Merged data may still tell you something about your samples as a whole
- Applies to pooling before sequencing to save costs, or combining different data sets
- Check that results do not correlate with a particular subset of the data

Scale Back the Goal

- Is there a simpler question that could be answered with the available data?
- Even if the data are underpowered to prove something, they may be able to disprove it
- Can you use the existing data to suggest what or how much data would be needed to really address the question?

Off the Map

- In most cases, there is no pushbutton to answer your scientific question
- The tools taught in this (or any) sequencing analysis course are only the starting point
  - These are standard methods to take raw data and turn it into calls of valid genomic features
- Most of your analysis for a given project will occur after you have used these methods

Thanks!

- Thanks to the following people who provided slides or references to help with this talk:
  - Ashlee Earl, Dirk Gevers, Joshua Levin, Michael Ross, David Jaffe, Brian Weiner, Sarah Young

- Please ask questions!