Genomics Part 2: Common Questions in Genomics

Bridging the Bench-Machine Learning Gap

Dr. Emily A. Beck

Dr. Jake Searcy

Learning Objective: Be able to answer the following questions?

What type of sequencing do I need?
Do I need single-end or paired-end reads?
If using short-reads what length do I need?
How much coverage/depth do I need?

What kind of sequencing do I need

Last time we discussed:

- Targeted sequencing (Sanger)
- Short-read sequencing (Illumina)
- Long-read sequencing (PacBio/NanoPore)

There are other types of sequencing but we will stick with this for now.

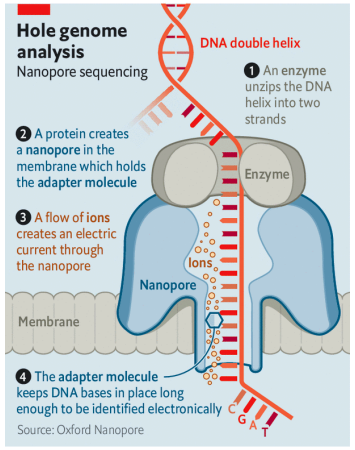
Most of what we are starting with will be in the context of genomics but we can talk more about applications to transcriptomics and metagenomics as well

What kind of sequencing do I need?

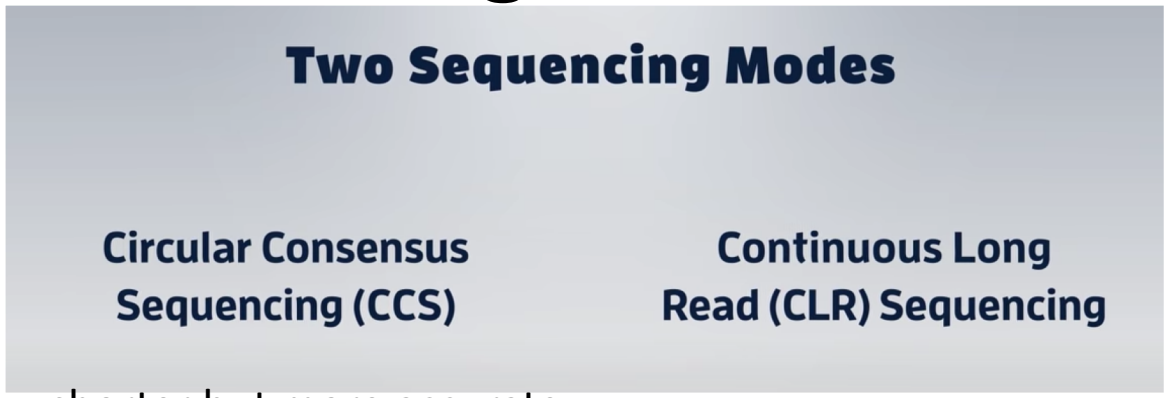
Short reads cannot sequencing through highly repetitive regions and can sometimes struggle to identify structural variation in comparative genomics

Long read sequencing can solve some of these issues but usually with a much higher error rate.

Quick aside about long read error rates



The Economist



shorter but more accurate (HiFi)

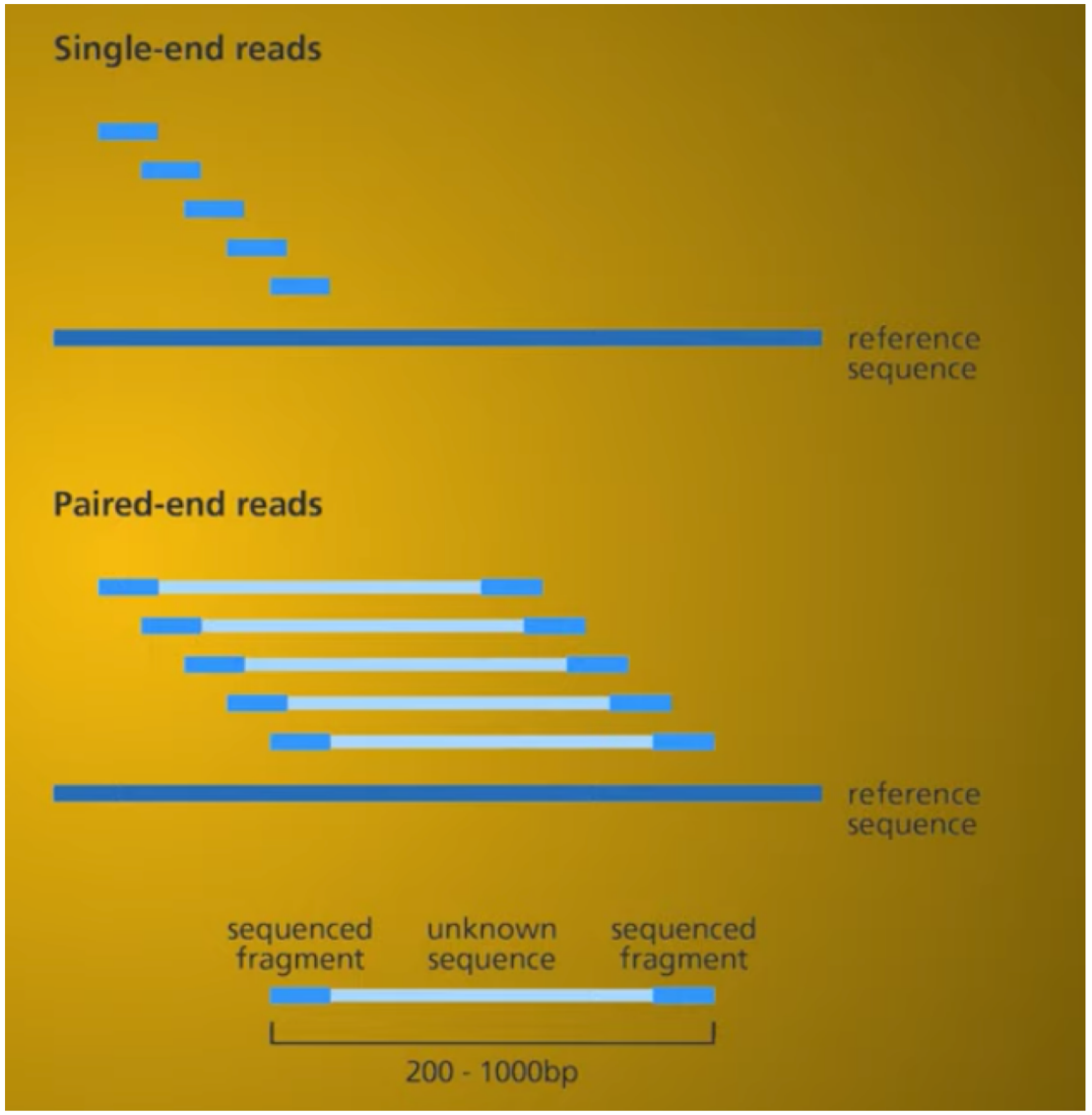
Longer but less accurate

Why are error rates high in long-read seq?

Nanopore: the speed cannot be controlled of the DNA moving through the pore- systemic problem with the tech when going too fast you get errors.

PacBio: High random error rate. CCS does not reduce error rate, but increases the depth allowing for error correction $$$$$$$$

Single-end vs Paired-end reads

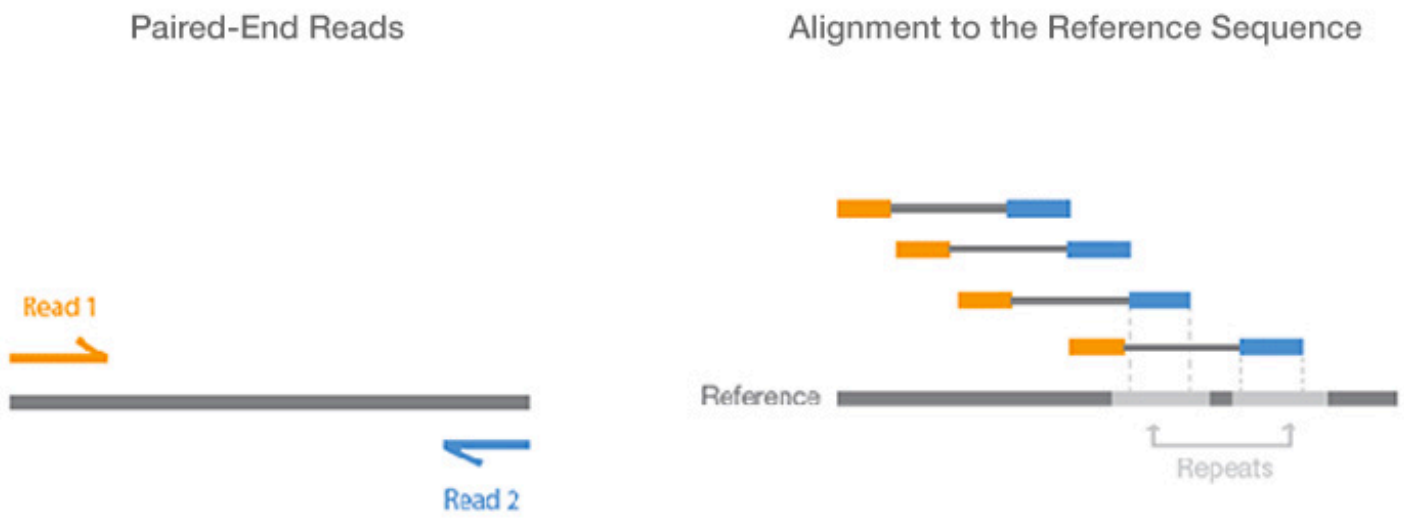


In single-end seq you sequence a specified fragment length from one direction

In paired-end seq you also sequence the fragment from the opposite side so you gain high quality information about the sequence itself but also the distance between the high quality sequences which helps with assembly

(MakeTheBrainHappy-Scientific Exploration)

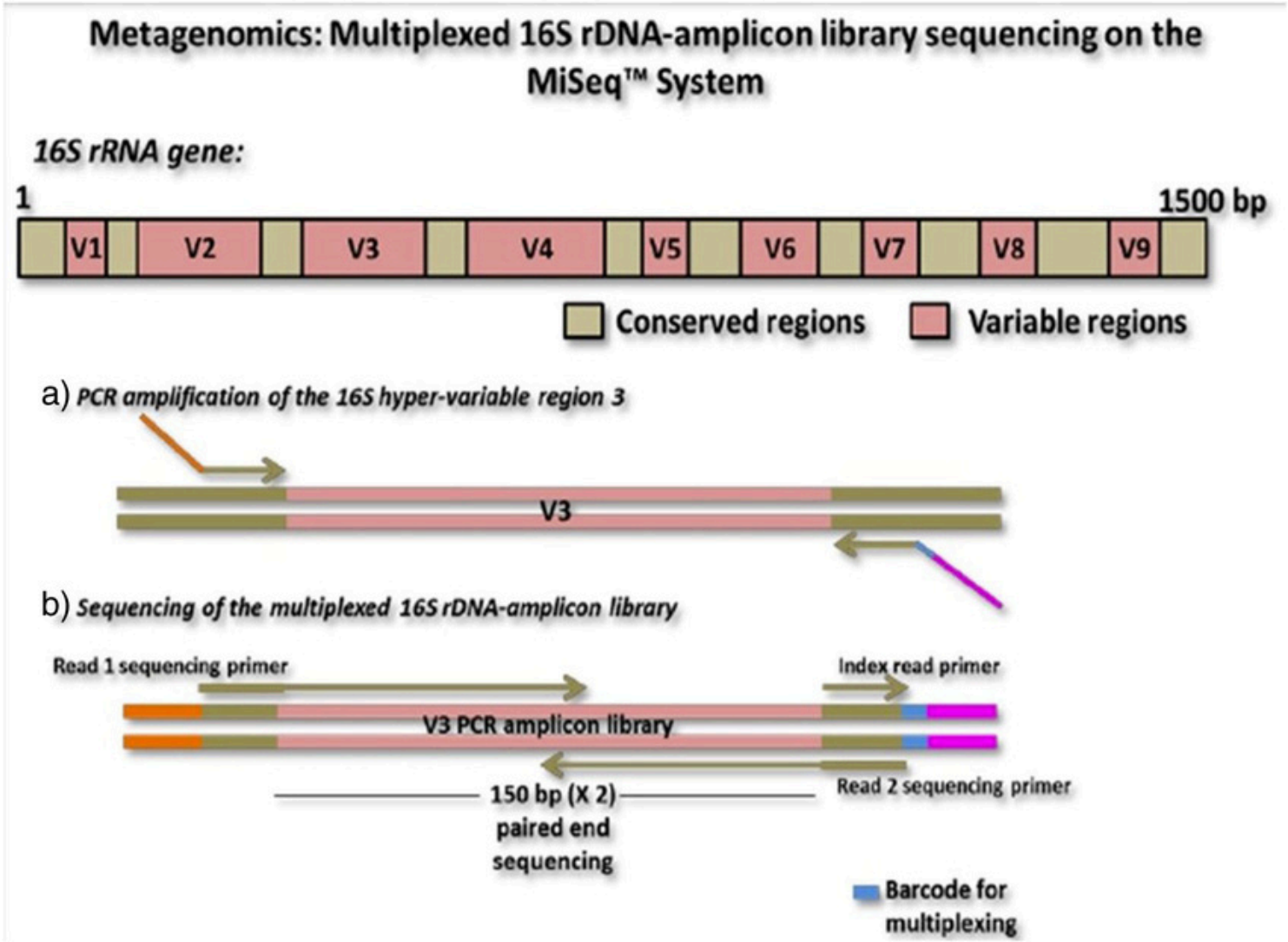
Paired-end reads can help power through repetitive sequences



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

illumina®

Depending on the type of library prep you may need to use paired-end data: Example 16S-seq



(Ranjith et al. 2016)

What length of reads do I need?

Historically, Illumina read lengths have included 50bp, 100bp, 150bp, and 250bp

New Novaseq (GC3F) has several options

Run Type (single lane)	Run Configurations	Read Pairs per lane**
S4 300 Cycle	PE 2x150	2-2.5 billion+
S4 35 Cycle	SR 53*, PE 2x32*	2-2.5 billion+
S1 300 Cycle	PE 2x150	650-800 million+
S1 100 Cycle	SR 118*, PE 2x59* 10X scRNA	650-800 million+
SP 500 Cycle	PE 2x250	325-400 million+
SP 300 Cycle	PE 2x150	325-400 million+
SP 100 Cycle	SR 118*, PE 2x59* 10X scRNA	325-400 million+

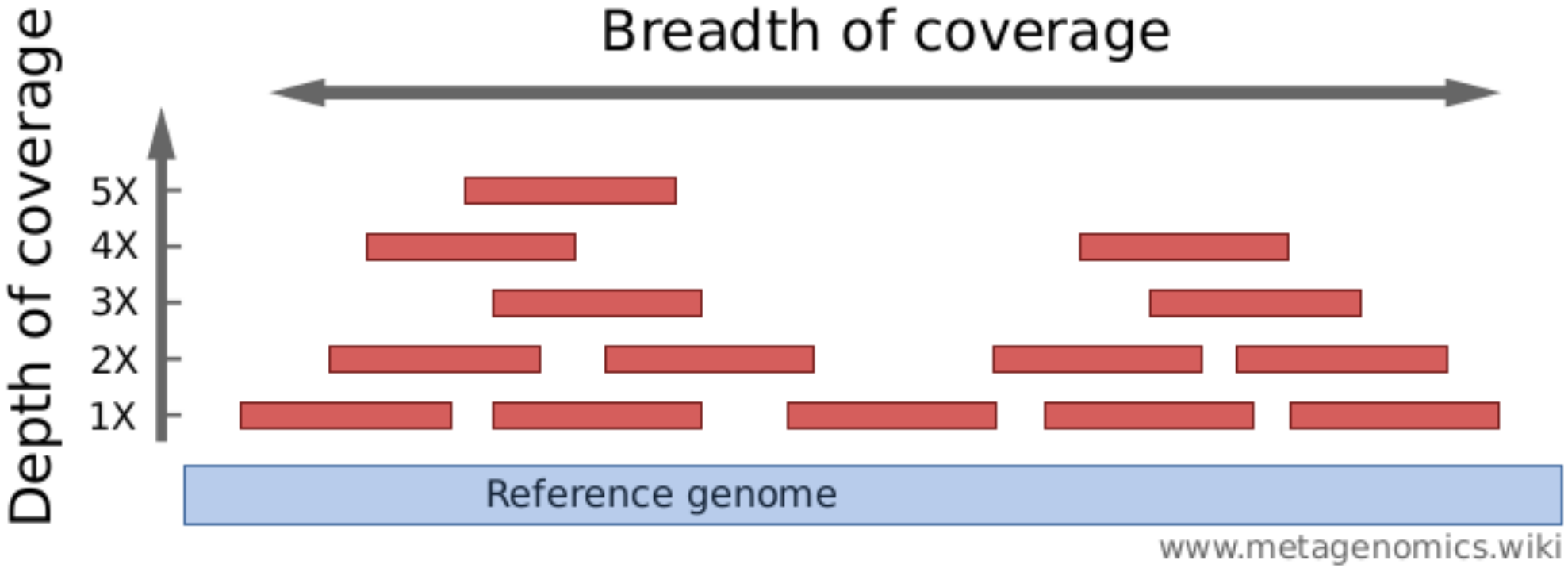
“SP” S Prime
“S1” NovaSeq6000 S1
“S4” NovaSeq 6000 S4

“SR” Single-Read (single-end)
“PE” Paired-End

The longer the reads the more info you will get, but you don’t always need all that info depending on your question.

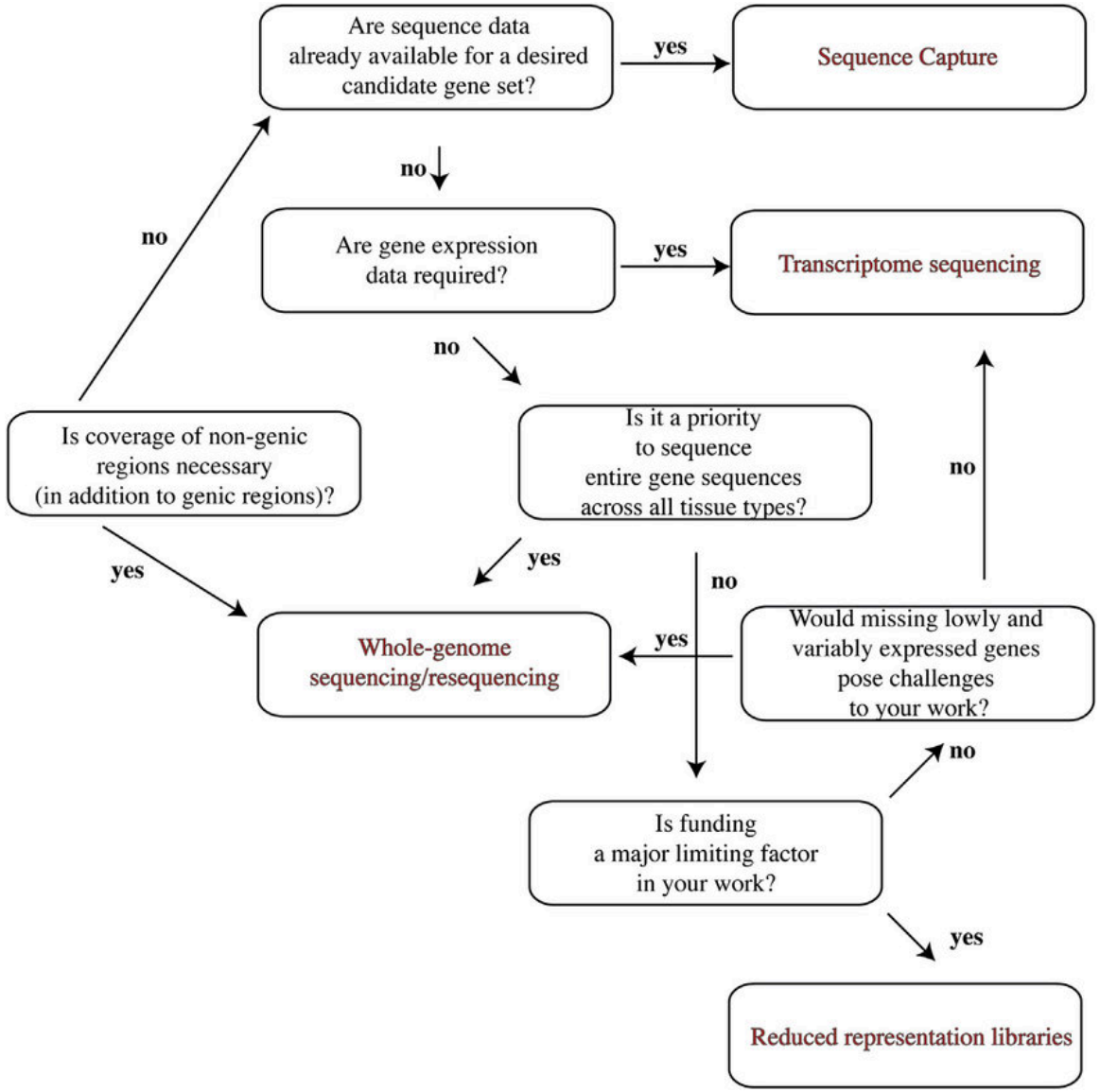
Let’s think through an example concerning differential expression

Most Common Question EVER!!!!
How much sequencing do I need?



(www.metagenomics.wiki)

Let's talk coverage first:
Could reduced-representation sequencing be the right choice?



RRS is a great choice for genetic mapping studies where you are sequencing 100s of individuals

Most Common Question EVER!!!!
How much sequencing do I need?

(I wish this answer were more straight forward)

For Genomics:

(1) What is the size of the genome?
(2) Do I have a reference genome?
(3) Am I combining sequencing technologies?

For Transcriptomics: (can be even trickier if you are the first person to do this in your organism)

(1) Do I have a reference genome?
(2) Do I care about splice variants?
(3) Am I interested in low abundant genes?

Most Common Question EVER!!!!
How much sequencing do I need?

(I wish this answer were more straight forward)

For Genomics:

(1) Wha

(2) Do I

(3) Am I

How many samples am I pooling into the lane???

For Tran o do this in your org

(1) Do I have a reference genome?

(2) Do I care about splice variants?

(3) Am I interested in low abundant genes?

Think about a project you have worked on or pick a problem to tackle

Take a few minutes to list pros and cons of each sequencing type and how you would attack the problem. What steps do you need to do? Where are your QC steps?