Experimental Design

4 Main Stages

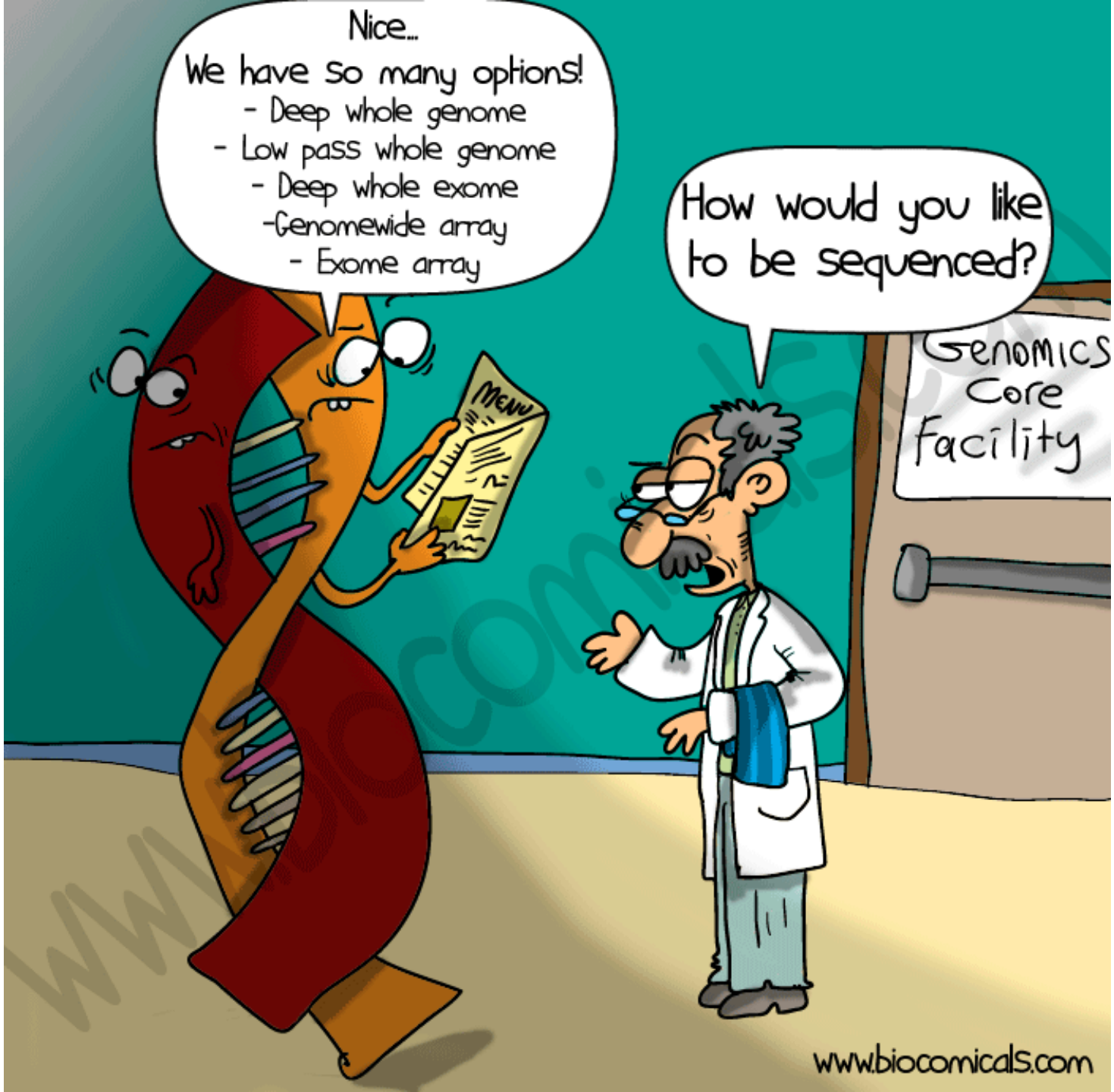
1	2	3	4
Biological Question	Design & Conduct Experiment	Prepare NGS Library & Sequence	Sequencing Analysis

Class time

2. DESIGN & CONDUCT EXPERIMENT

Types of sequencing

- Clonal sequencing
  - Expect a single genome to be present. Highly resilient to sequencing errors.
- Population sequencing
  - Multiple subpopulations and genomes expected to be present. Highly subjective to sequencing errors especially when making quantitative analysis.
- Amplicon/targeted sequencing
  - Know all/most reads correspond to specific genomic locations. Can be dominated by sequencing errors.


Nice...
We have so many options!
- Deep whole genome
- Low pass whole genome
- Deep whole exome
-Genomewide array
- Exome array
MENU
How would you like to be sequenced?
Genomics Core Facility
www.biocomicals.com

3. PREPARE NGS LIBRARY & SEQUENCE

Standard Library Prep

1. Fragment DNA
    1. Enzymatic, sonication, acoustic, nebulization
2. Blunt DNA
3. “A”-Tail DNA
4. Ligate adapters
5. PCR

Standard library prep sufficient

- Clonal samples
  - Each base 0 or 100%
  - 50% possible in diploids
- Low to moderate coverage depth populations (<100)
  - Standard Illumina error rate 1%
  - Much more on this later
- Good reference
  - Typical insert sizes 250 – 700 bp

Standard library prep lacking

- High coverage populations (>100)
  - Error rate of 1% sets limit of detection at 1% regardless of depth
    - Error rate reduction
- Non-model organism
  - Difficult to generate good reference assemblies using 200 - 700bp
    - Mate-pair libraries several kb inserts
    - Combine with other long read sequencing solutions
- Repeat-mediated rearrangements
  - Repeats often 1.5kb+ long, difficult to get reads on both sides
    - Mate-pair libraries again

3. Prepare NGS Library & Sequence

ERROR RATE REDUCTION

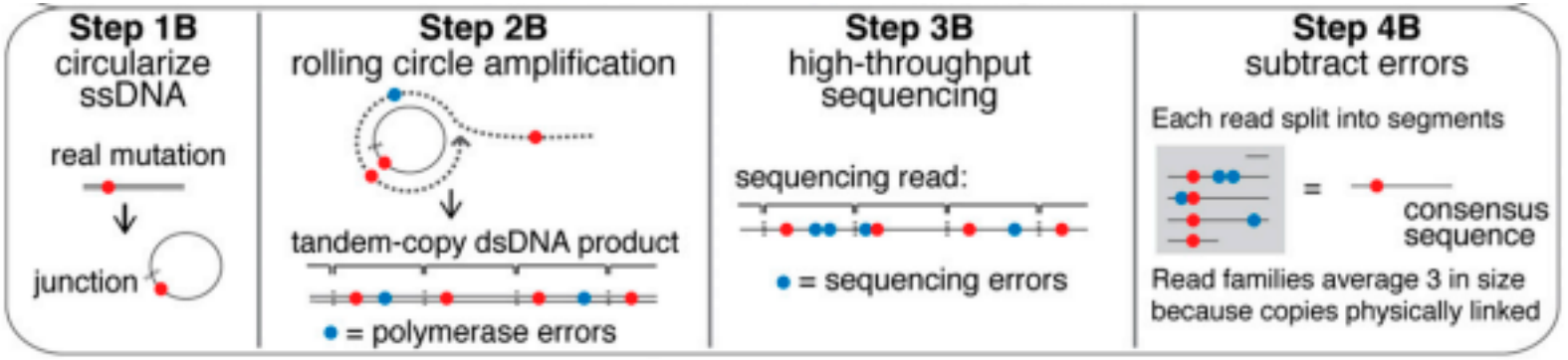
Basic Principle

- Because majority of errors are randomly distributed along reads, several alternative library preparation methods exist to read the same original fragment of DNA multiple times to reduce error rates.
- 3 main ways.
  1. Molecular indexing
  2. Circle sequncing
  3. Short insert size approximately read length

Alternative Library Preparation

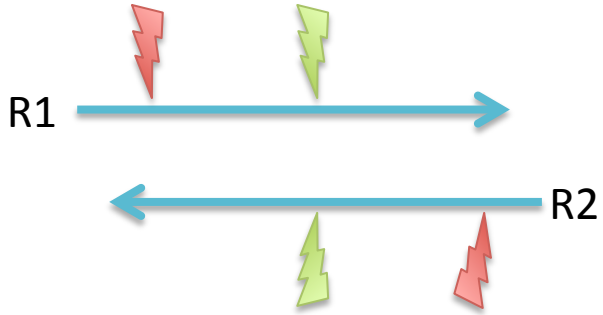
- Duplex sequencing (Schmidt et al 2012 PNAS)
  - Molecular index to identify original DNA fragments

- Circle sequencing (Lou et al 2013 PNAS)
  - Rolling circle amplification to reread same fragment



Alternative Library Preparation

- Double read
  - Fragment DNA size to ~read length.
  - Can be done with any paired end read.



3. Prepare NGS Library & Sequence

“LONGER” READS

Illumina Options

- Technology:
  - Illumina: 2 x 600 miSeq runs
- Mate pair library:
  - Generates 2 outwardly facing reads separated by up to 25kb.
  - Allows connection of otherwise distant locations on a single read.

Read Sequence Quality control

4. SEQUENCING ANALYSIS

Garbage In, Garbage Out

Data Preprocessing

- Massaging standard illumina data is likely to be less effective than better experimental planning and design using alternative library preparation methods.

Read Sequence Quality Control Questions

- Contaminated with other samples?
- Adapter dimers present?
  - Reads with no insert present.
- Adapters present on ends of reads?
  - Insert size smaller than read length.
- 3’ end of reads quality decline?

Onto the Tutorials.

- Some more basic bash/linux interrogation about reads, and working with TACC.
- FastQC can answer all questions about raw read quality and is pretty much the only game in town.
- FASTX toolkit nice set of tools for improving data when you identify a problem.