Genome Variant Analysis 2015

General Introduction

Background

- Key research interest identifying rare variants and determining how to leverage that information.
- TA two years
- Co-teach last year
- Self-taught computational biologist

A nod to the past …

- Scott Hunicke-Smith
- Director of Genome Sequencing Analysis Facility

- Jeffrey Barrick
- Assistant Professor MBS & ICMB

Disclaimers up front

- Royal “we” – Tutorials written without clean tense, we likely means me so don’t blame Sean.
- Spelling – Has never been a skill I possess, hopefully will only be noticeable if I write on board.
- Typos – Will likely be your biggest problem in using the commands we provide. Tutorials difficult to put together in format.
- Names – I usually use this opportunity to apologize for my inability to remember people’s names with a funny anecdote, with small class might not be an issue.

Where to start

- Many say “don’t know where to start” their data analysis once they have their data files.
- Typically should have “started” weeks-months ago in planning experiments.
- Not all libraries are created the same, and can drastically effect analysis.

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-moderate coverage depth populations (<100)
  - Standard Illumina error rate 1%
  - Much more on this later
- Good reference
  - Typical insert sizes 200 – 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
- Repeat-mediated rearrangements
  - Repeats often 1.5kb+ long, difficult to get reads on both sides
    - Mate-pair libraries again

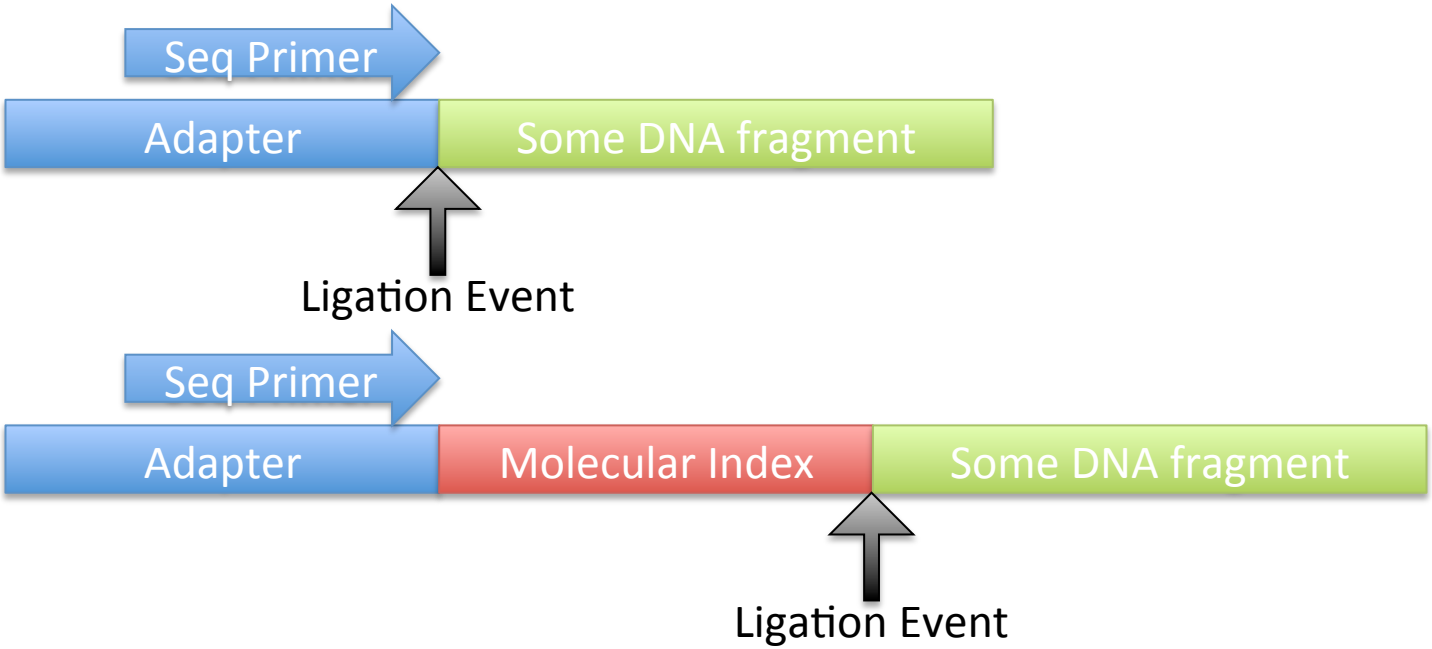
Error rate reduction

- Key = reading the same fragment of DNA multiple times independently. 2 main ways.

1. Molecular indexing
2. Circle sequncing

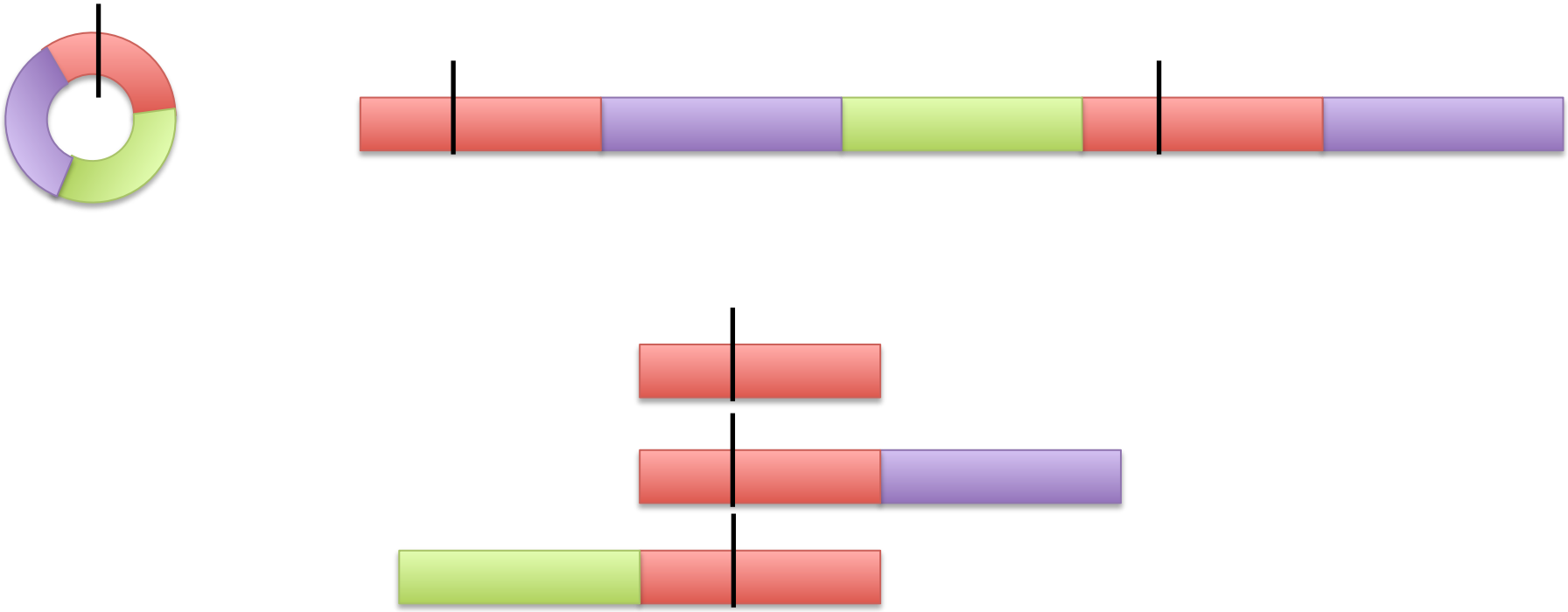
Molecular Indexing

- Ligating adapters with degenerate sequences in the sequence read to fragmented DNA.
  - Schmitt et al PNAS 2012



Circle Sequencing

- Circularize DNA fragments, rolling circle amplification, standard library prep.
  - Lou et al, PNAS 2013



Mate Pair Library

- Generates 2 outwardly facing reads separated by up to 25kb

Computers Computers Computers

- Millions of reads, 100s of bp long, mapping to millions-billions of base long references.
- Windows is your enemy, linux/Mac is your friend.
- TACC is a time machine that lets you get stuff done much faster
  - This is where we will start the class.