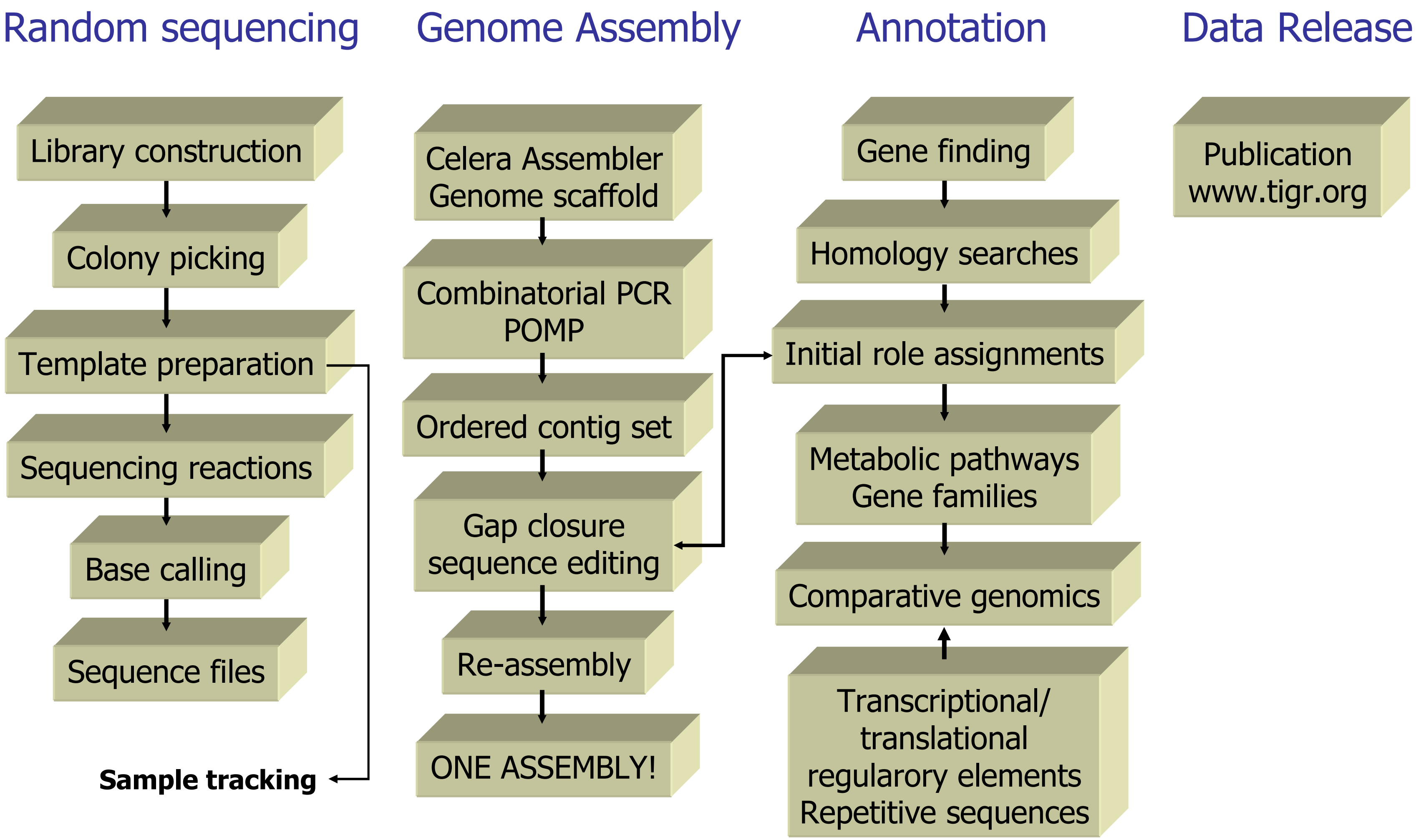
Assembly Checklist

Michael Schatz

August 17, 2006
University of Hawaii

A Genome Sequencing Project



Random sequencing

- Library construction
- Colony picking
- Template preparation
- Sequencing reactions
- Base calling
- Sequence files

Sample tracking

Genome Assembly

- Celera Assembler Genome scaffold
- Combinatorial PCR POMP
- Ordered contig set
- Gap closure sequence editing
- Re-assembly
- ONE ASSEMBLY!

Annotation

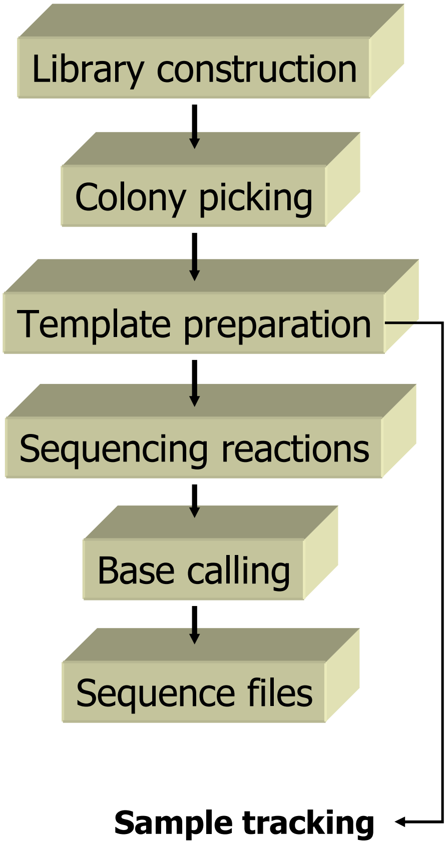
- Gene finding
- Homology searches
- Initial role assignments
- Metabolic pathways Gene families
- Comparative genomics
- Transcriptional/ translational regularory elements Repetitive sequences

Data Release

- Publication www.tigr.org

Library Issues

Random sequencing



Library construction

Colony picking

Template preparation

Sequencing reactions

Base calling

Sequence files

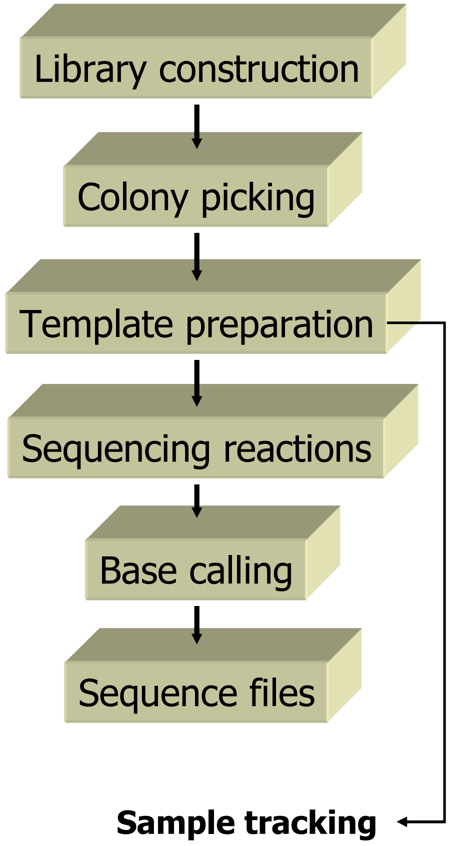
Sample tracking

- Uniform Random Sampling of Genome
  - Test: K-mer statistics to ensure uniform coverage
  - Action: Check early, check often
  - Number of reads to sequence is *dependent* variable
    - Num Reads = $\frac{(\text{Coverage} * \text{Genome Size})}{\text{Read Length}}$

- Size Selection of Libraries
  - Test: Histogram of Insert Sizes
  - Action: Resize libraries in frg file
  - Prefer Mixture of Small (4kb) and Large (10kb)
    - BAC Libraries if Possible

Sequencing Issues

Random sequencing

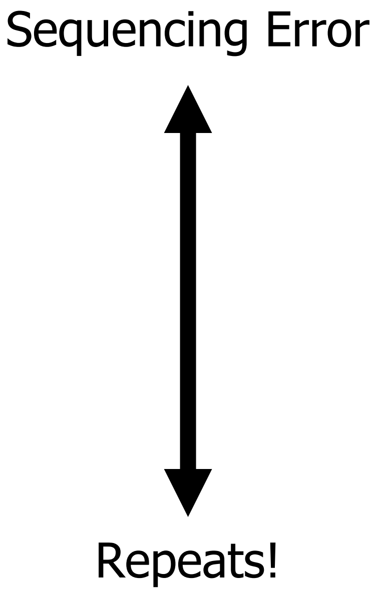


- Contamination / Multiple Replicons
  - Test: Histogram of GC Content of Reads
  - Action: Partition Replicons and assemble separately
  - Multiple Replicons may have widely varying coverage -> Inaccurate A-stat -> Poor Contigs & Scaffolds

- Tracking Issues
  - Test: Mis-oriented mates in non-repeats
  - Action: Make sure mates are complete and correct!

Trimming Issues

Genome Assembly



Sequencing Error

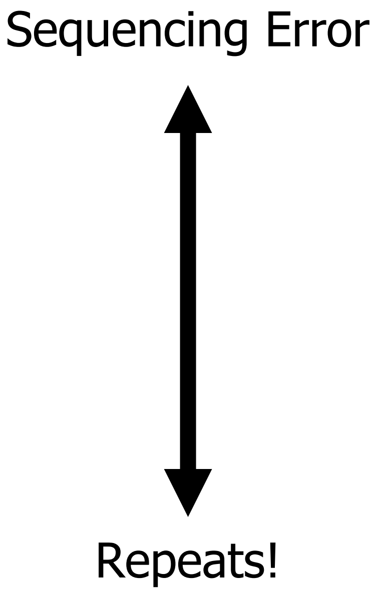
Repeats!

- Vector trimming
  - Test: Check for missing 5′ overlaps
  - Action: Retrim vector more aggressively with Lucy
- Quality Trimming
  - Test: High Singleton Reads Rate
  - Action: Retrim
  - Action: Raise Unitigger Error Rate
    - Unitigger –e
    - ERATE in runCA.euk
    - utgErrorRate in runCA-OBT
- Experimental Overlap-Based-Trimming
  - runCA-OBT.pl
- Note: Repeat Masking is NOT necessary for CA

Assembly Issues

Genome Assembly

- A-Stat Problems
  - Test: Large fraction of degenerate (> 15%)
  - Action: Set genome size estimate smaller
    - grep genome unitigger.err
    - Unitigger –l
    - utgGenomeSize in runCA-OBT
- Localized Mis-assemblies
  - Test: Use cavalidate
    - Especially SNP analysis!
  - Action: Try lowering unitigger error rate
  - Action: Try local re-assembly
    - nucmer local assembly to original assembly
    - stitchContigs to fix global assembly



Sequencing Error

Repeats!

Important URLs

- CBCB Homepage
  - http://www.cbcb.umd.edu/~mschatz/
- Celera Assembler
  - http://wgs-assembler.sourceforge.net
- AMOS
  - http://amos.sourceforge.net
- MUMmer
  - http://mummer.sourceforge.net
- AutoEditor
  - http://www.tigr.org/software

Check Frequently for Updates!

Conclusions

- Assembly is an inherently difficult problem
  - Blue sky with millions of pieces
  - Good coverage is key to success

- Repeats are forks in the road
  - Need mate-pair “map” to navigate

- Be aware of potential size/quality tradeoffs
  - Bigger is not always better

THANK YOU!