Screening Reads

- Prefer accurate reads: 98%+
  - Clip reads if predicted accuracy drops
- Remove vector and other contaminants
  - Alignment & identification
  - Especially other recent genome projects
- Screen for known repeats

Evaluate Overlap

- Compare each fragment with each other
  - This is why Illumina seqs so hard! N^2
- Example
  - >= 40 bp overlap, <= 6 mismatches
  - Either a true overlap, or a repeat
  - Figure this out asap!
  - Fragments with excessive numbers of overlaps are probably repeats

Unambiguous Contigs

- Combine fragments with only one possible assembly into longer sequences
  - Perfect matches
  - Match no other: no conflicting overlaps
- Drosophila
  - 3.158M reads => 54K unitigs
  - Still might be wrong
- Can extend unitigs up to one read length into repeat regions

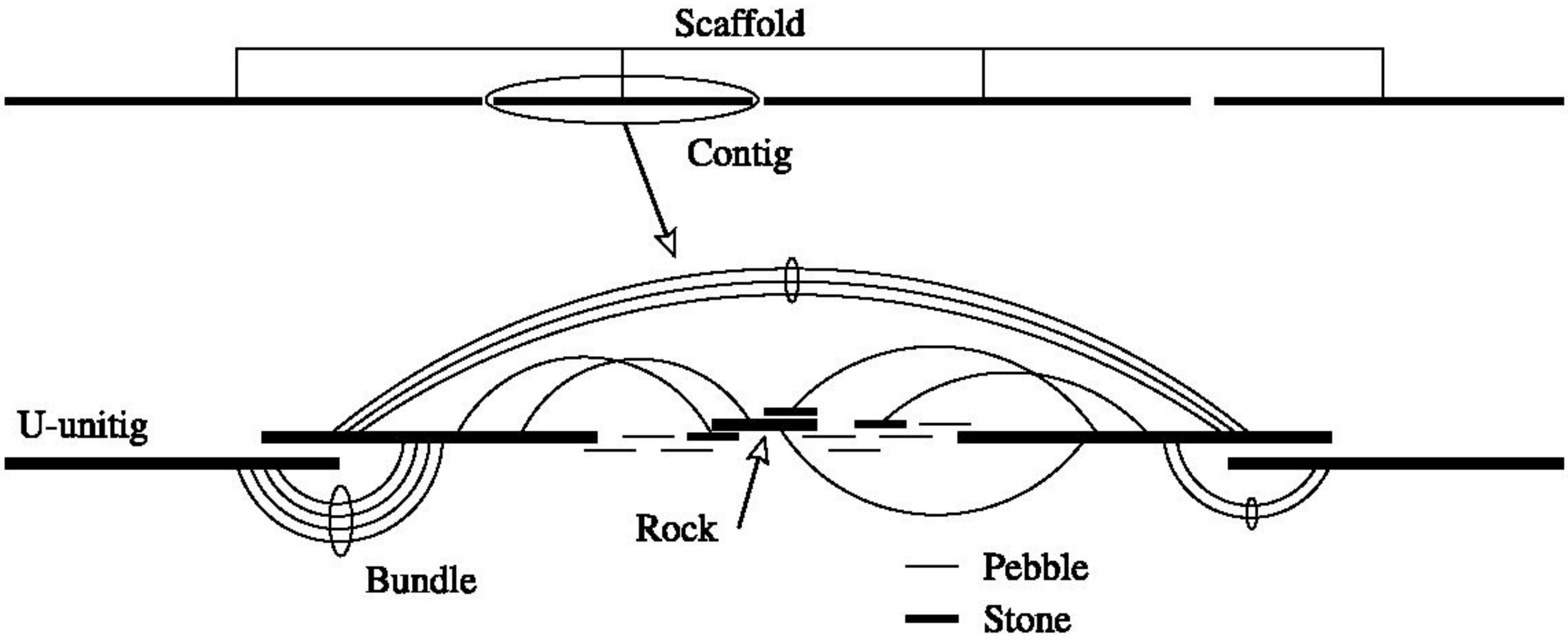
Scaffolds

- Set of ordered, oriented contigs
- Gaps of approximately known size
  - BAC ends in two different contigs
  - BAC library of tight known size range
  - Same concept for other paired end reads
- “Bundle” if more than one placement
  - The more mate pairs, the more reliable
- Map scaffolds with FISH, recombination

Place Repeats

- Placement evidence from mate pairs
  - Multiple = rocks
  - Single = stones
  - None = pebbles
    - Basically just guessing. Statistical

Scaffold Visual



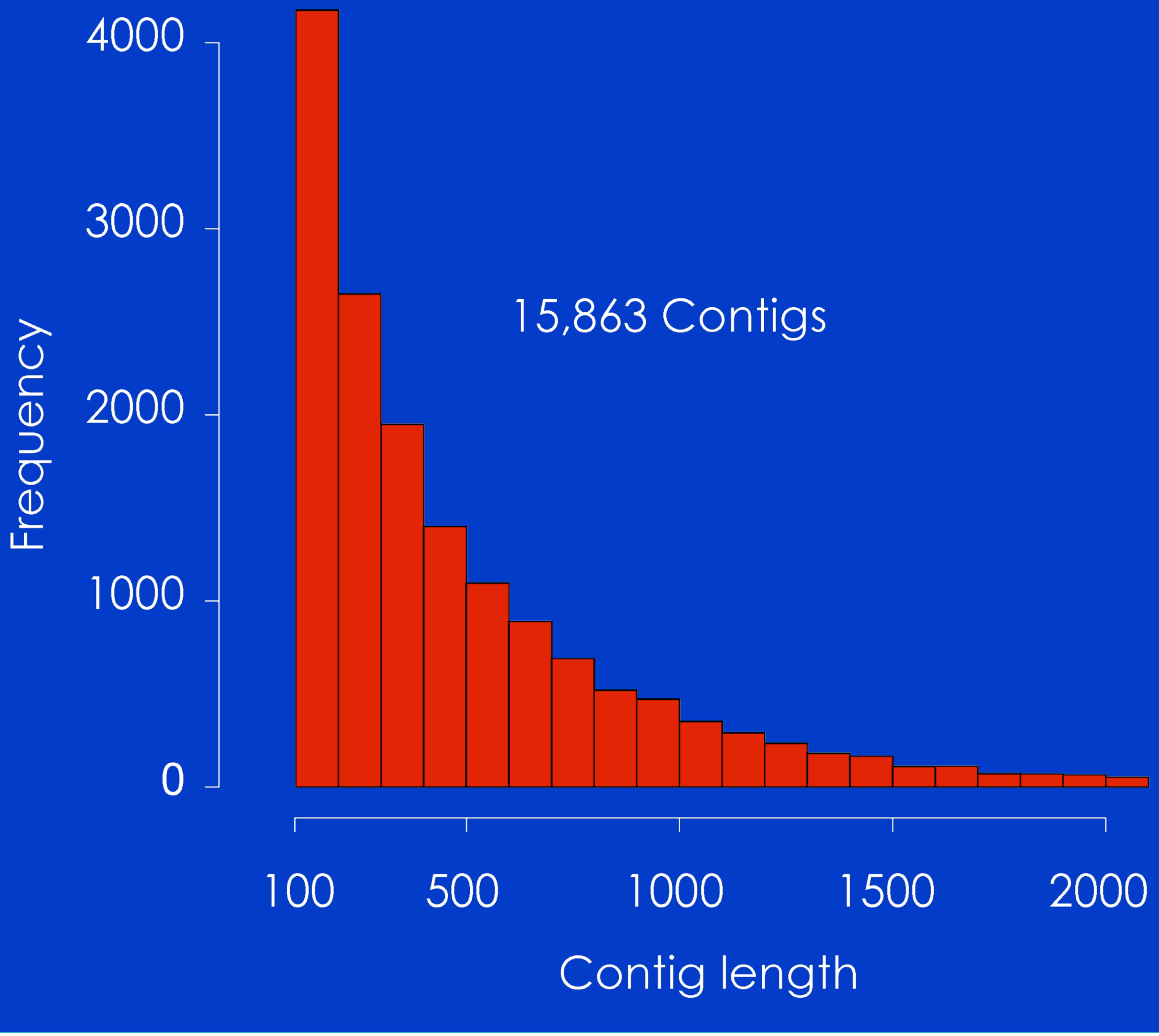
Finishing and Validating

- Manual review and adjustment
- Quality control
  - PCR
- Overlaps, mis-assembly, etc.
- Gaps can reflect unresolvable repeats or low coverage in a region
- More intense sequencing in a region
  - Compare to other sequencing efforts (Celera)

NextGen Assembly

- More faster cheaper shorter error-prone
- Bacterial example: Mycobacterium spp.
  - ~4 Mb genome
- Solexa/Illumina, de novo assembly with VELVET assembler
- 50x coverage = 2Gb, 36 bp reads

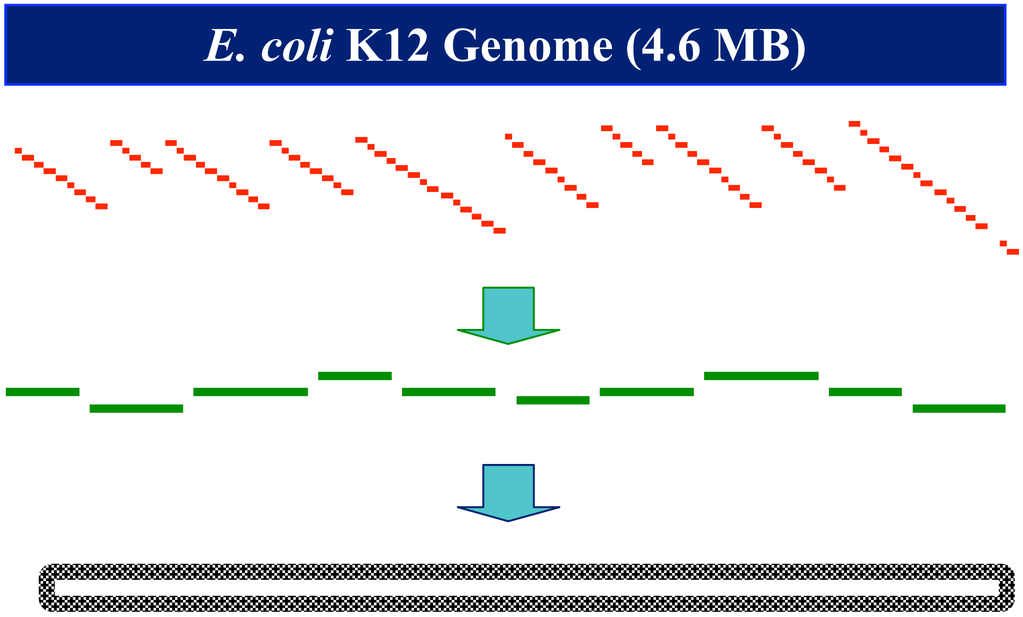
Assembly Result



454 E. coli Assembly

old 250 bp reads

454 Read Type	Genome Coverage	Number of Contigs/ Scaffolds
Shotgun	15×	98
	+	
3 Kb Jump	18×	7
	+	
20 Kb Jump	20×	1

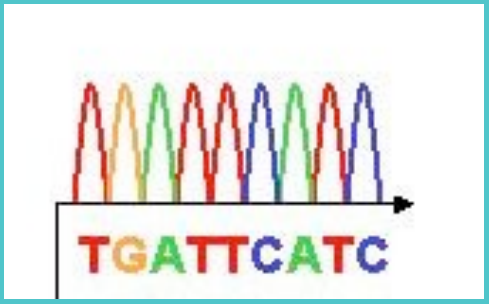


So, this is 20x total coverage

Consensus Accuracy: ~ 99.999%

Open Questions

- What is the most efficient way to combine various sequencing methods?
  - Solexa paired ends versus 454 single reads
    - SOLiD for its accuracy?
  - 454 paired end 3Kb, 8Kb, 20Kb mix
  - Sample repeat regions ahead of time
- Do you have to have BACs for a eukaryote genome?
- Any tricks to finish off gaps efficiently?
- Priors: low heterozygosity, few repeats


TGATTCATC