Towards the $1000 genome – the impact of new DNA technologies on Finishing

Niranjan Nagarajan

Post-doctoral Fellow

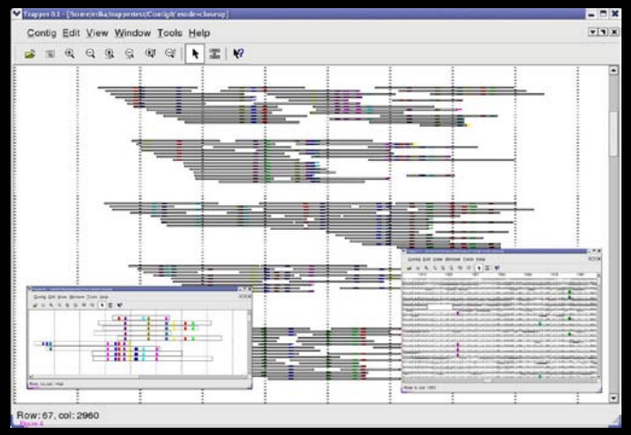
Center for Bioinformatics and Computational Biology

University of Maryland, College Park

Sequencing

- Fast & Cheap (~ 2-4 Days)
  - 454: 5 Bacterial Genomes in one run!
  - Illumina/SOLiD: 10-20 Genomes …

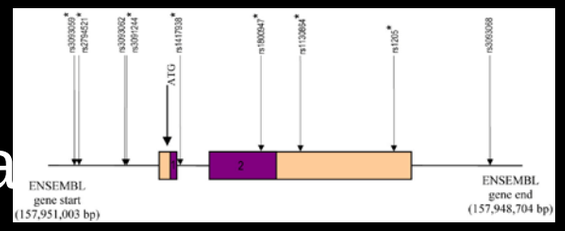
Vs Finishing



- Slow & Laborious (~ 2-4 Months)
  - 454: 100s of Contigs
  - 100-1000 PCRs and Finishing Sequences needed

Closing the Gap

- Advantages
  - Order-based Analysis (Rearrangements, Regula
    control, …)
  - No missing pieces – Short elements, functional
    repeat-rich regions (CRISPRs)
  - Assembly Validation

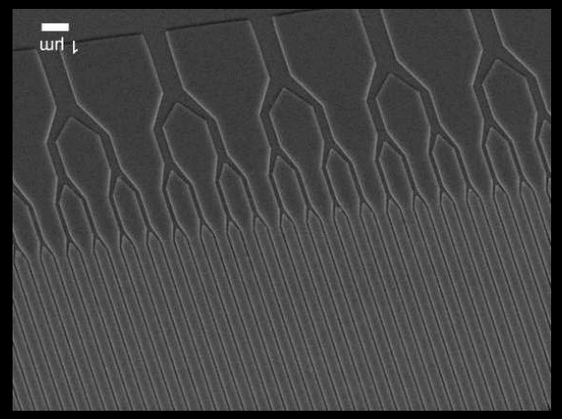


Papaya Genome – 48,408 contigs
Macaque Genome – 301,039 contigs

How can we reduce the time and effort?

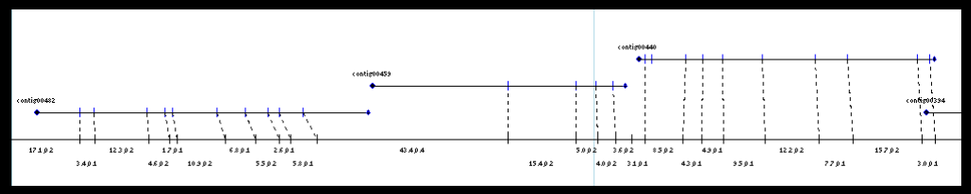
Wet-lab

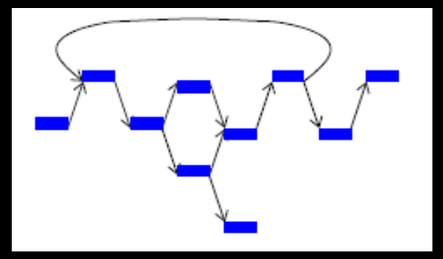
- Improved mate-pair protocols
  - Multi-fragment Reads?
- Genome-wide Maps
  - Optical Maps
  - Nanocode Maps
  - Nanofluidics



In-silico

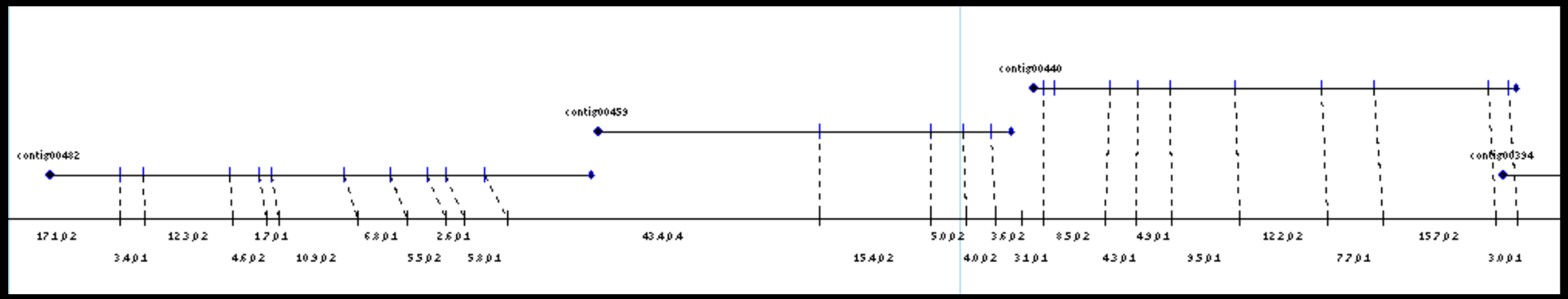
- Hybrid-assemblers
- Map-based Scaffolding
- Assembly Models





Map-based Scaffolding

Genomic Sequences



Genomewide Map

Challenges

- Limitations of optical mapping
  - Sizing errors
  - Missing fragments
- Errors in sequence data
  - Sequencing errors
  - Incorrect Assembly
- Repeat sequences

Related Work

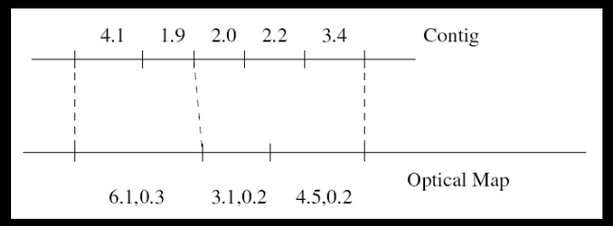
- Heuristics, Manual Analysis (< 10 Contigs, Sanger sequencing) (Zhou et al. 2002, Reslewic et al. 2005)
- Restriction fingerprints (Ben-Dor et al. 2003, Engler et al. 2003)
- Aligning Optical Maps (Valouev et al. 2005)
- Validating Optical Maps (Antoniotti et al. 2001)

SOMA (http://www.cbcb.umd.edu/soma/)

Matching

- Scoring function – two levels
- Dynamic Programming

$$\sum_{k=1}^{j} \left(\frac{c_k - o_k}{\sigma_k} \right)^2$$



$$\left| \sum_{i=s}^{t} c_i - \sum_{j=u}^{v} o_j \right| \leq C_\sigma \sqrt{\sum_{j=u}^{v} \sigma_j{}^2}$$

Match Significance

- Non-chromosomal sequence, Misassembled sequences
- Permutation Test
- Stopping criteria for C_σ search

Non-unique matches

- Repeats, Small sequences
- F-test

$$\frac{\chi^2{}_{\text{best}}/n_{\text{best}}}{\chi^2{}_{\text{other}}/n_{\text{other}}}$$

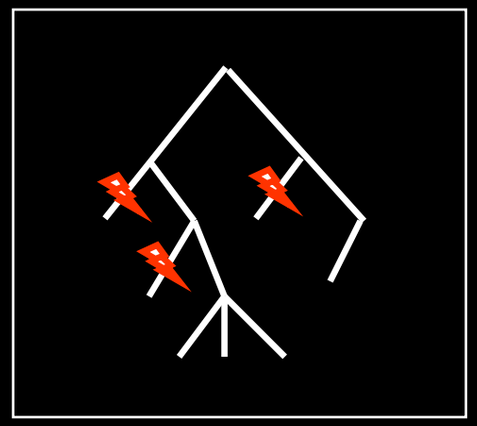
Niranjan Nagarajan, Timothy D. Read and Mihai Pop. **Scaffolding and validation of bacterial genome assemblies using optical restriction maps.** *Bioinformatics* (2008)

Global Placement

- Why? – Non-unique, overlapping matches, Multiple Maps?
- What is a good placement?
  - Most Sequences
  - Most confident placement
  - Map Coverage
- Placement ⇔ Interval Scheduling (NP-hard)

```
procedure SCHEDULE(< i_1, ..., i_n >)
 if n = 0
   then return (empty schedule)
   else { for each i ∈ {instances that overlap i_1 or belong to the same job as i_1}
            do reduce the weight of i by the weight of i_1
          Eliminate all instances with non-positive weight
          result ← SCHEDULE(< i_2, ..., i_n >)
          if i_1 ∪ result is feasible/non-overlapping
            then return (i_1 ∪ result)
            else return (result)
```

Greedy Algorithm
2-Approximation, O(nlogn) running time



Enumeration with pruning
Optimal, $O(k^n)$ worst case

1. Place unique matches
2. Remove conflicts with placed matches.
3. Repeat from 1.

Heuristic
O(n) running time

Results

- Synthetic datasets with *real* characteristics
  - 454 Sequencing: *Y. kristensenii* (YK, 488 contigs) & *Y. aldovae* (YA, 375 Contigs)
  - AflII Optical Maps (YK: 350 sites, YA: 360 sites)
  - 100 Artificial maps = Permutations of real optical map
  - Artificial in-silico maps = Artificial map **superimposed** on contigs with fragments inflated by factor **C** and sites omitted with probability **p**

Program	Experiment	Correctly placed (%)
Basic Matching	C = 2, p = 0	87
Basic Matching	C = 2, p = 0.1	80
Matching & Tests	C = 2, p = 0.1	95
Matching & Tests	C = 4, p = 0	99

Sequence Placement

Experiment	Greedy		Enumeration		Heuristic	
	Cov.	Acc.	Cov.	Acc.	Cov.	Acc.
C = 2, p = 0	0.77	0.87	0.83	0.86	0.70	0.99
C= 2, p = 0.05	0.70	0.76	0.74	0.76	0.49	0.94
C = 4, p = 0	0.72	0.78	0.77	0.78	0.64	0.96

Without C_σ search: 30% of matches are not significant.

The heuristic approach is the most accurate!

Assembly Results

Yersinia Genomes (relatives of the plague agent)

Strain	Coverage of AflII Map (%)	Coverage of NheI Map (%)	Merged Assembly (%)	Map + Assembly graph (%)
Y. kristensenii	84			94
Y. aldovae	83			86
Y. mollaretii	76	81	85	88
Y. fredriksenii	81	85	86	88
Y. bercovieri	73	77	79	85
Y. Intermedia	87	82	86	92
Y. rohdei	85	83	87	90
Y. ruckerii	80	80	82	86

Watson genome (454 Sequencing) vs. Nanocode maps (*Schwartz Lab, Univ. of Wisconsin*)

Chrom. Number	Sequence Size in reference genome (Mbp)	Map Size (Mbp)	No. of contigs	Proportion of map covered by contigs (%)
22	35.6	35.6	354	77
21	34.8	33.7	175	86
20	60.4	59.0	335	92
19	56.7	49.8	581	77
18	75.7	77.7	368	86
17	78.9	78.2	689	80
16	80.2	78.1	656	84
15	82.8	81.1	561	93

Other Projects

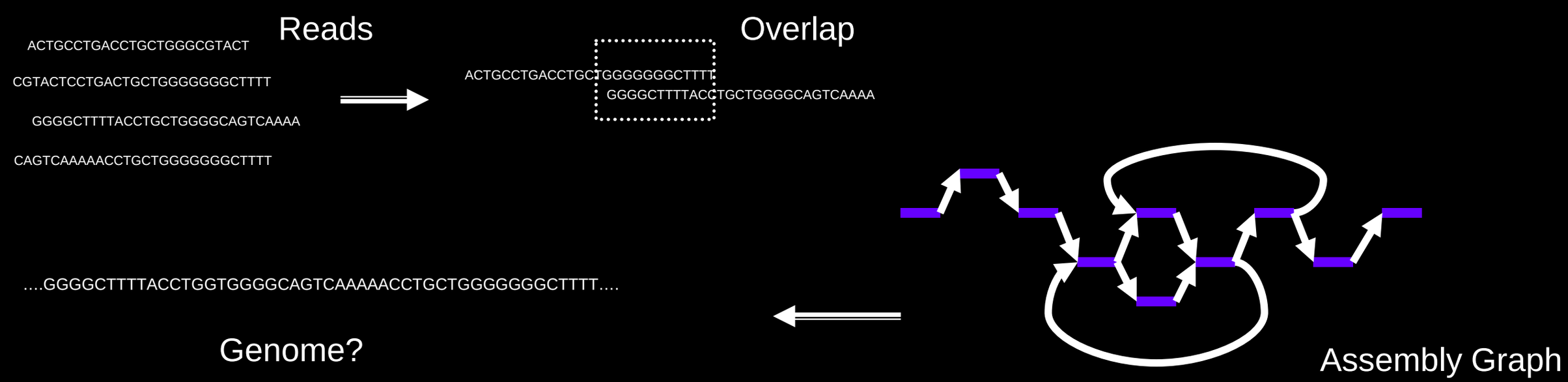
Vibrio Harveyi, 3 *Acinetobacter Baumannii* strains, …

Finishing Effort?

Strain	Draft Genome (Size in Mbp)	# of gaps (> 10 Kbp)
Y. kristensenii	4.36	37 (8)
Y. aldovae	3.64	39 (14)
Y. mollaretii	4.34	56 (24)
Y. fredriksenii	4.72	33 (20)
Y. bercovieri	3.87	57 (25)
Y. Intermedia	4.62	53 (15)
Y. rohdei	4.15	40 (16)
Y. ruckerii	3.34	39 (17)

- Case Study I (*Yersinia rohdei*)
  - 43 PCRs, 26 Finishing Reads
  - 7 more gaps to close

Assembly Models



- Parametric Complexity
  - Assembly as a function of Read Length, Repeat Complexity and Coverage
  - How best to exploit short-read sequences and mate-pairs

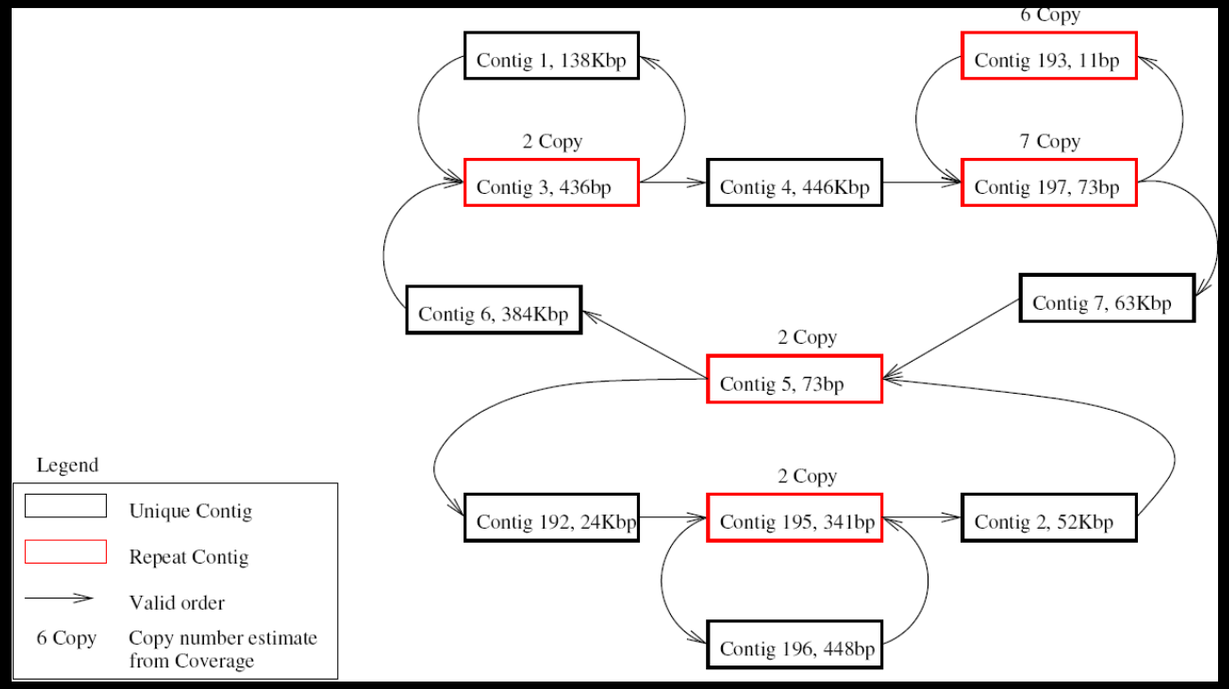
Niranjan Nagarajan and Mihai Pop. **Parametric complexity of sequence assembly: Theory and applications to next generation sequencing**. *Journal of Computational Biology* (2009)

When is assembly easy?

- Implications
  - Contigs can be made to span Repeats much longer than the Read Length!
  - Efficient, exact algorithms can be designed for Genome Assembly in the presence of High Coverage.
  - Mate-pairs can be used to efficiently constrain short-read assembly graphs
- Case Study II (*Aggregatibacter aphrophilus*)
  - Sequencing: 22X coverage, 100bp reads (454)
  - Assembly: 146 Contigs, N50 = 99 Kbp
  - Mate-pairs: 2 Kb and 4Kb library, 3X coverage (Sanger)
    - Very little new sequence (< 0.2%)
    - Constrained assembly graph: 18 unique and 6 repeat contigs (AMOS-Hybrid pipeline)
  - Closure with 42 PCRs and 17 Finishing Reads!

In-silico Closure

- Case Study III (*Rickettsia prowazekii*)
  - Assemble regions in Contig Graph that have a unique path through them



Rickettsia prowazekii (1.1 Mbp) contig graph

- Work in Progress
  - Automated Tool for Common Subpath Assembly
  - Merging Contig Graph and Optical Map information

Acknowledgements

Mihai Pop, U. Maryland

Timothy Read, Emory U.

Chris Cook (NMRC), MariaPia DiBonaventura (AMNH)

Opgen Inc (Travel Sponsorship)