Background

- De novo assembly.
- Based on 3 computational techniques :
- *Pairs of reads with potential overlap* : A fast comparison method (Pearson and Lipman, 1988; Altschul et al., 1990) is used to quickly find pairs of reads with a potential overlap.
- *Overlaps between reads and construct alignment of reads* : Dynamic programming methods (Needleman and Wunsch, 1970; Smith and Waterman, 1981) are used to compute overlaps between reads and to construct alignments of reads in contigs.
- *Construct contigs and supercontigs* : A maximum-weight spanning tree method (Kruskal, 1956) is used to construct contigs and supercontigs.
- Works in three phases.

Phase 1

- Reads partitioned into subsets of similar sizes.
- Compute overlaps between reads in the subset and reads in the whole set. The comparisons for the subsets are performed in parallel.
- The pairs of reads with two close word matches of 12 bp are quickly located.
- For each pair of reads, an overlap between the reads is computed by a banded dynamic programming algorithm.
- A region of a read is identified to be highly repetitive if it occurs in many overlaps.
- Overlaps involving only highly repetitive regions are removed.
- The remaining overlaps are called unique overlaps.

Phase 2

- Poor ends of each read are determined and removed based on unique overlaps.
- Unique overlaps are ranked in decreasing order of overlap strength. Overlap strength depends on the similarity level of the overlap, and the depths of coverage for the read positions in the overlap.
- The overlaps of strength greater than a cutoff are called good overlaps. (default : cutoff 4500)
- Reads assembled into contigs by processing the good overlaps in the decreasing order.
- Links of read pairs between contigs are ranked in decreasing order of link strength.
- Contigs are connected into supercontigs by processing the links in the decreasing order.

Phase 3

- Consensus sequences for the supercontigs in each group are computed.
- For a supercontig, attempts are made to close gaps between contigs in the supercontig, with repetitive reads that are linked by read pairs to the supercontig. The resulting contigs in the supercontig are considered one at a time.
- For the current contig, a multiple alignment of reads in the contig is constructed and a consensus sequence is generated from the alignment.
- Read base quality scores are used in the computation of multiple alignments and generation of consensus sequences.

PCAP - Major programs

- PCAP consists of several main programs for generating an assembly.
  - *Pcap* program computes pairwise overlaps between reads.
  - *bdocs* program uses these overlaps to calculate the coverage depths at each region of the genome.
  - *bclean* program removes overlaps between reads with extremely high coverage depths (typically repetitive regions of the genome).
  - *bcontig* program builds the assembly layout, placing each read into an ungapped region of contiguous sequence known as a “contig,” and then assembling the contigs into larger gapped structures known as “supercontigs.”
  - *bconsen* program generates the consensus sequences of the contigs.

PCAP – minor programs

- The PCAP package also contains a few minor programs for formatting an assembly and collecting statistics on it.
  - *bform* program combines a number of files of consensus sequences into a single file and compiles lists of all reads that were either used or omitted from the assembly.
  - *bpair* program reports the status of read pairs at the contig level and at the read level.
  - *n50* program collects the N50 lengths (a standard measure of the distribution of contig length) and counts of contigs and supercontigs.
  - *xstat* program reports the distribution of the distances of read pairs in supercontigs.

Autopcap

The autopcap script automatically runs the major and minor programs to produce a small-scale assembly.

- Usage: autopcap FileOfFileNames [options]
- FileOfFileNames is a file of file names
- Options (default value):
- -d N specify stringent qual diff score cutoff N > 20 (130)
- -l N specify min depth of coverage for repeats N > 20 (75)
- -m N specify amount of available memory in GB N >= 1 (1)
- -p N specify running pcap jobs in parallel N >= 0 (1)
- -s N specify adjusted overlap score cutoff N > 100 (4500)
- -t N specify overlap percent identity cutoff N > 75 (92)
- -v N specify program type: 1 for PCAP; 0 for PCAP.REP (1)
- -y N specify number of pcap jobs N >= 2 (2)

Input file format

- PCAP takes as input a number of pairs of gzip-compressed base and quality files in FASTA format,
- a file of read pairs,
- a file of all base file names without the gz suffix.

Output files

- Assembly results and statistics are in the following files:
  - *contigs.bases*: Contig base sequences in FASTA format.
  - *contigs.quals*: Contig quality scores in FASTA format.
  - *supercontigs*: Overview of supercontigs.
  - *reads.placed*: The positions of reads used in the assembly.
  - *reads.unplaced*: The names of reads that are not in the assembly.
  - *fofn.pcap.scaffold*.ace*: Ace files of contigs for the Consed assembly viewer and editor program.
  - *readpairs.contigs*: Major unused read pairs between contigs.
  - *readpairs.reads*: The positions of read pairs in the assembly.

- *fofn.con.pcap.results*: The status of read pairs.
- *fofn.con.pcap.sort.stat*: The distribution of read pair distances.
- *fofn.pcap.n50*: The length statistics of contigs and supercontigs.
- *fofn.pcap.contigs*.snp*: Alignment columns with potential SNPs. (pg.10)

faSize contigs.bases

- 2793272 bases (45 N's 2793227 real 2793227 upper 0 lower) in 1097 sequences in 1 files
- Total size: mean 2546.3 sd 5752.9 min 56 (Contig837.1) max 51430 (Contig0.1) median 479
- N count: mean 0.0 sd 0.3
- U count: mean 2546.2 sd 5752.9
- L count: mean 0.0 sd 0.0
- %0.00 masked total, %0.00 masked real
- Minimum size of contig – 56
- Maximum size of contig - 51K

- Total No. of reads: 391302
- Reads used - 353530
- Reads not used : 38869
- Total overlaps – 20648245 ~ 20.6M
- Unique overlaps – 18100563 ~ 18.1M
- [reads.placed file orientation 0 → given, 1 → reverse]
- [All other files : orientation 1 → given, 0 → reverse]

Assembly quality

- more fofn.pcap.bform.info
- The value for the -y option: 2
- No. of singlets used by bconsen or unused: 0
- Number of singlets: 38869
- Number of unused: 0
- Total No. of singlets: 38869
- Total No. of reads used: 352433
- Total No. of reads: 391302
- Number of scaffolds: 1097