Notes from Practical 1

Differences between algorithms:
1) boundary conditions
2) recursion relations
3) starting point for traceback

Needleman-Wunsch vs Smith-Waterman
Look at pairwise_durbin_ANNOTATED.xls

Dynamic programming widely used:

Tiling paths (clones, PCR products)
Hidden Markov Models (HMMs: Aylwyn)
Genefinding (Alastair)
Intron-sensitive mRNA/ genome alignments
(GI: splice motifs; FG: splice boundary/ exon discovery through RNA sequencing)
Protein/ DNA sequence alignments

Short read sequence aligners: 1

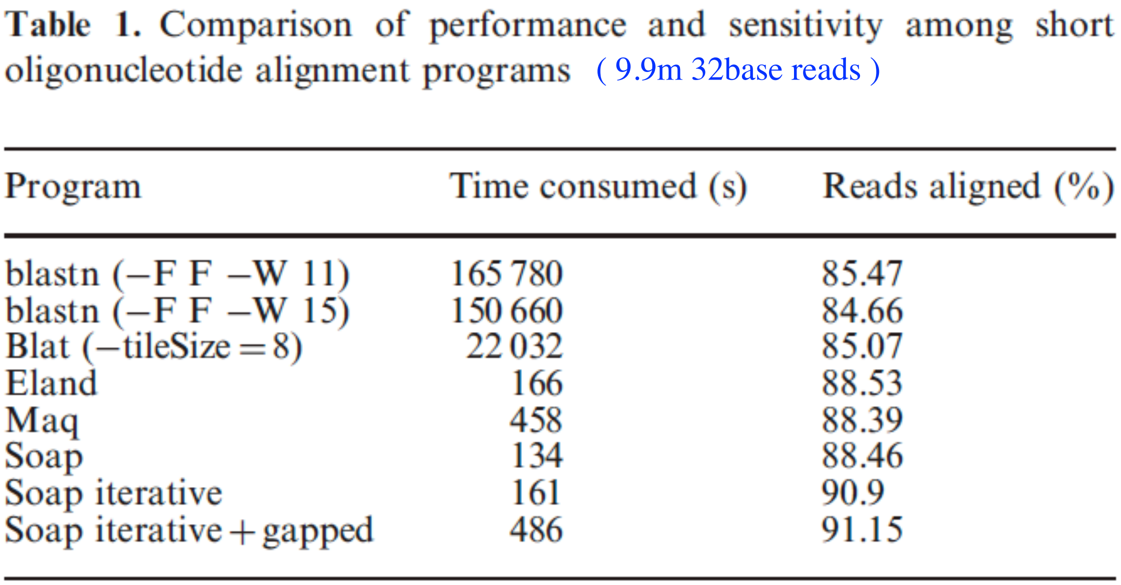


Table 1. Comparison of performance and sensitivity among short oligonucleotide alignment programs (9.9m 32base reads)

Program	Time consumed (s)	Reads aligned (%)
blastn (−F F −W 11)	165 780	85.47
blastn (−F F −W 15)	150 660	84.66
Blat (−tileSize = 8)	22 032	85.07
Eland	166	88.53
Maq	458	88.39
Soap	134	88.46
Soap iterative	161	90.9
Soap iterative + gapped	486	91.15

SOAP: short oligonucleotide alignment program.
Li R, Li Y, Kristiansen K, Wang J. Bioinformatics. 2008 Mar 1;24(5):713-4. PMID: 18227114

bowtie 200-600x faster than Soap

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.
Langmead B, Trapnell C, Pop M, Salzberg SL. Genome Biol. 2009;10(3):R25. PMID: 19261174

Short read sequence aligners: 2

Short read alignment is currently a very active field

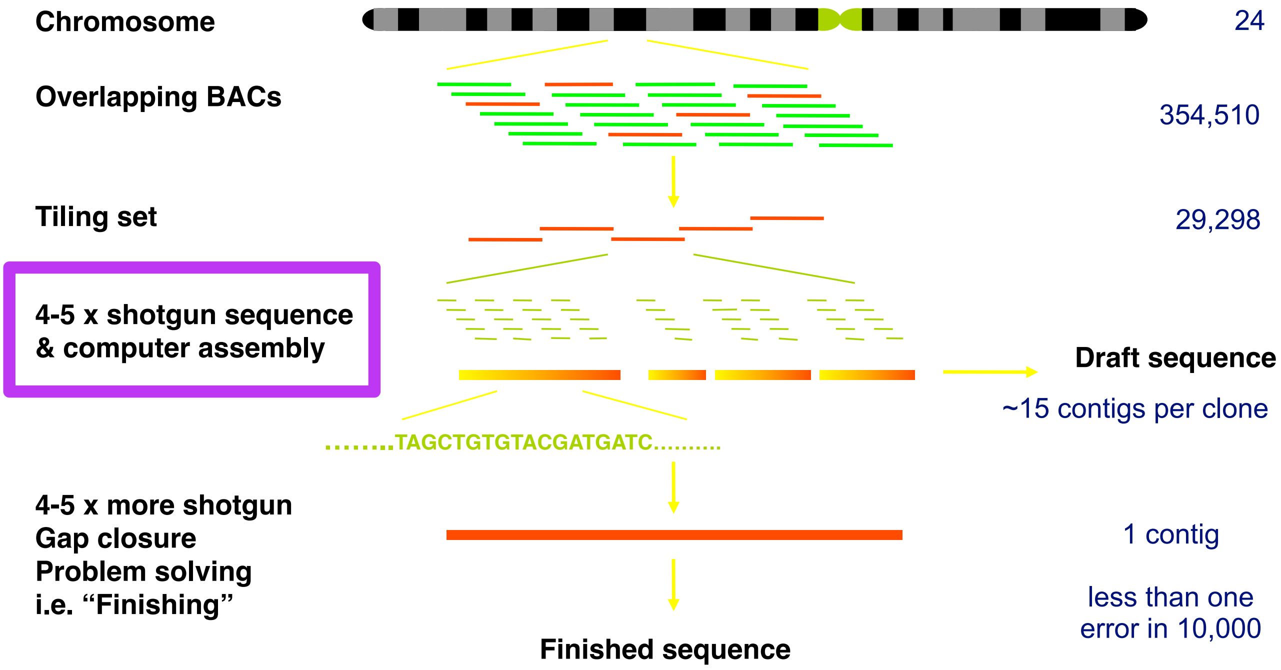
Soap: one of the first published and simplest to understand

Allows 1 or 2 mismatches, or a one gap of 1-3 bases with no flanking mismatches

Builds seed index table for database (e.g. genome sequence)
Then for each read: derive seeds
check index table for candidate hits
generate alignment

Pointers to lots more programs at Heng Li’s NGS aligner page:
http://lh3lh3.users.sourceforge.net/NGSalign.shtml

Sequence Assembly



Chromosome

24

Overlapping BACs

354,510

Tiling set

29,298

4-5 x shotgun sequence
& computer assembly

Draft sequence

~15 contigs per clone

........TAGCTGTGTACGATGATC..........

4-5 x more shotgun
Gap closure
Problem solving
i.e. “Finishing”

1 contig

less than one
error in 10,000

Finished sequence

BAC shotgun assembly 1

Starting material:

BAC clones: 100 - 150kb long

~2000 paired-end sequencing reads from ~2kb subclones

Process:

Check for repeat content

Pairwise sequence alignment: looking for overlaps
all vs all repeat-free sequences

Assemble highest scoring first

Assemble repeat-containing

Paired-end reads important for contigation

Generate consensus

Finishing: examine/ edit/ iterate

BAC shotgun assembly 2

Widely-used programs:

phrap: '**Ph**il Green's **r**apid **a**ssembly **p**rogram'

http://www.phrap.org/phredphrapconsed.html (+ consed)

Gap4:

http://staden.sourceforge.net/manual/gap4_unix_toc.html

Often phrap was used for assembly, gap4 for finishing

Sequencing reads have per-base quality values. These used to help distinguish between errors and repeats during assembly.

Quality values are used when calling final consensus which itself has per-base quality values: $-10\log(p_{error})$

Short read sequence assembly

40x coverage of human-scale genome is

$40 \times 3 \times 10^9$ bases = ~~10^{11} bases

This volume of data can be generated in
3-4 runs on 2010-generation machines

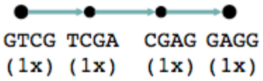
For 100 base reads, 10^{11} bases is ~10^9 reads
Brute force comparison of all vs all requires
~10^{18} comparison i.e. ~10^{22} operations
Forget it!

Short read sequence assembly toy problem

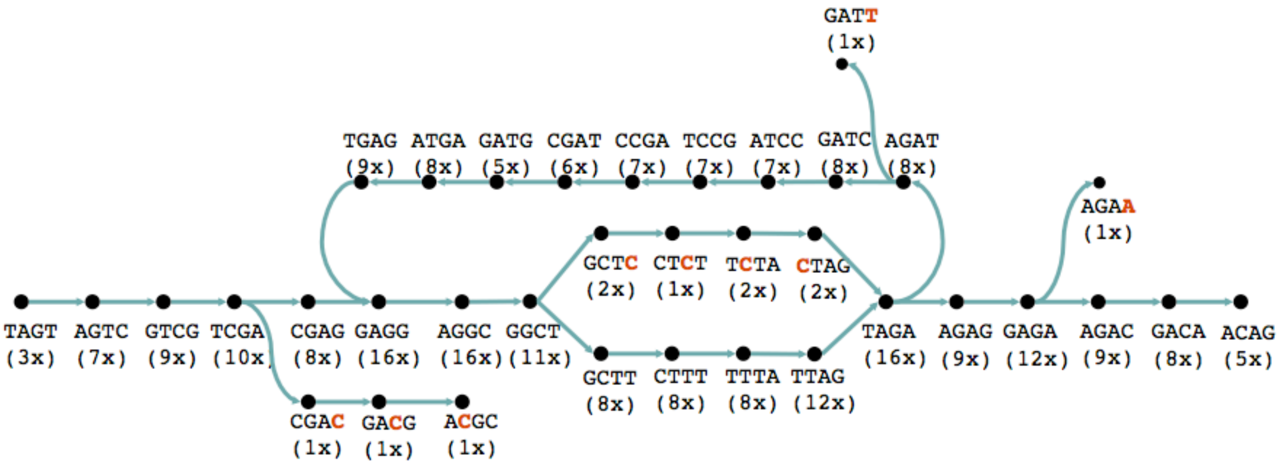
```
TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG

 AGTCGAG CTTTAGA  CGATGAG CTTTAGA
  GTCGAGG  TTAGATC  ATGAGGC    GAGACAG
     GAGGCTC   ATCCGAT AGGCTTT GAGACAG
 AGTCGAG    TAGATCC ATGAGGC  TAGAGAA
TAGTCGA  CTTTAGA CCGATGA    TTAGAGA
    CGAGGCT  AGATCCG TGAGGCT  AGAGACA
TAGTCGA GCTTTAG TCCGATG  GCTCTAG
   TCGACGC    GATCCGA GAGGCTT AGAGACA
TAGTCGA    TTAGATC GATGAGG TTTAGAG
  GTCGAGG TCTAGAT   ATGAGGC  TAGAGAC
      AGGCTTT  ATCCGAT AGGCTTT GAGACAG
 AGTCGAG   TTAGATT  ATGAGGC   AGAGACA
       GGCTTTA  TCCGATG    TTTAGAG
    CGAGGCT TAGATCC  TGAGGCT   GAGACAG
 AGTCGAG  TTTAGATC  ATGAGGC TTAGAGA
     GAGGCTT  GATCCGA GAGGCTT  GAGACAG
```

Velvet: de Bruijn graph based sequence assembly

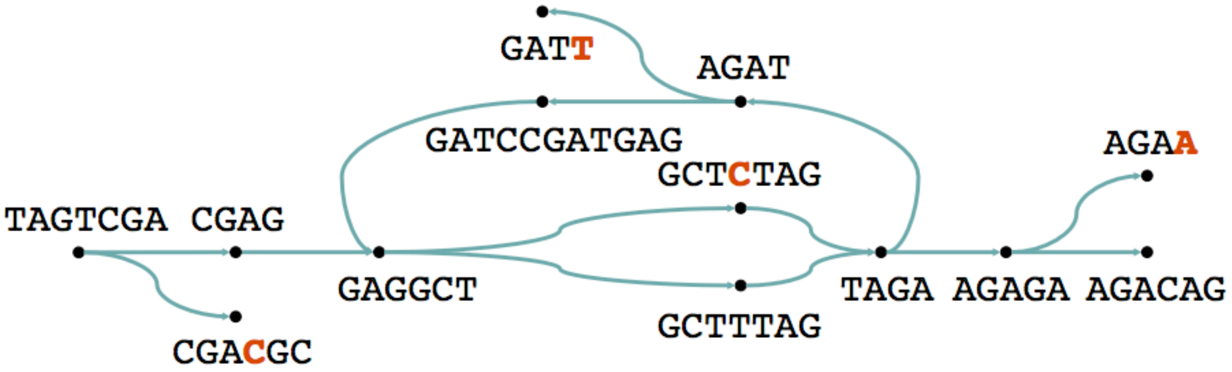
One read: GTCGAGG



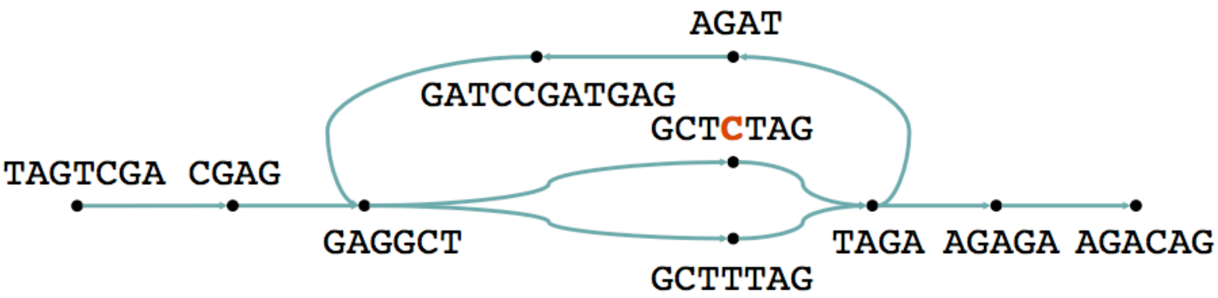
All the others…



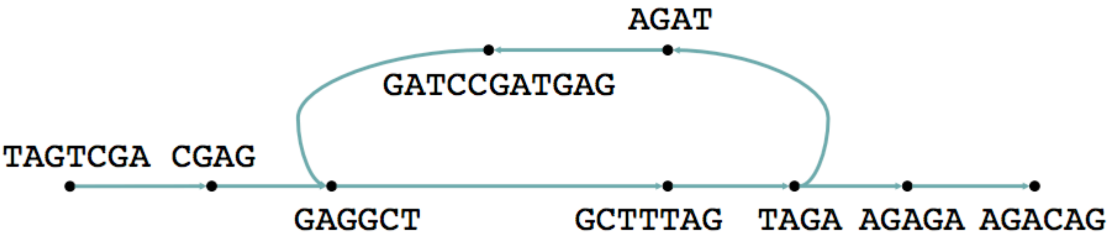
After simplification…



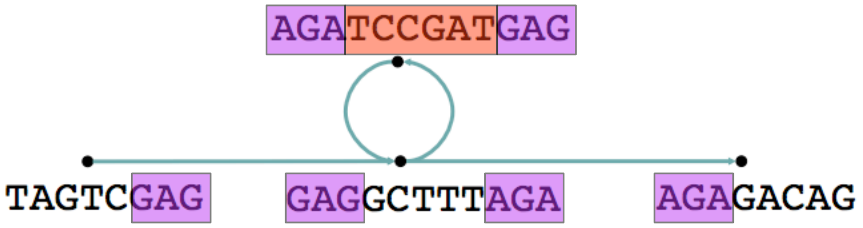
Tips removed…



Bubbles removed (Tour Bus)…



Final simplification…


AGATCCGATGAG

TAGTCGAG

GAGGCTTTAGA

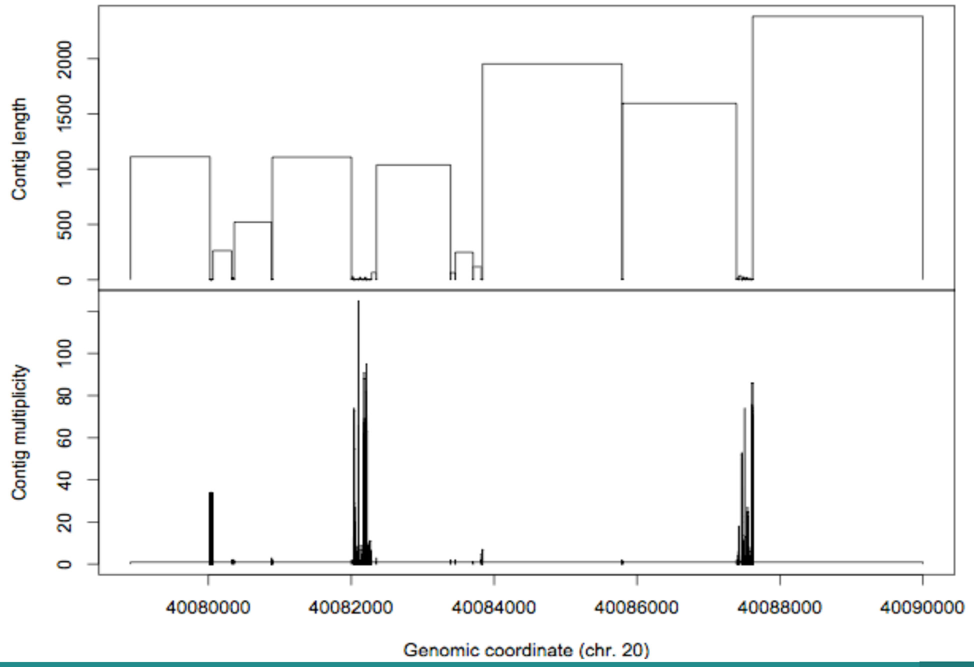
AGAGACAG

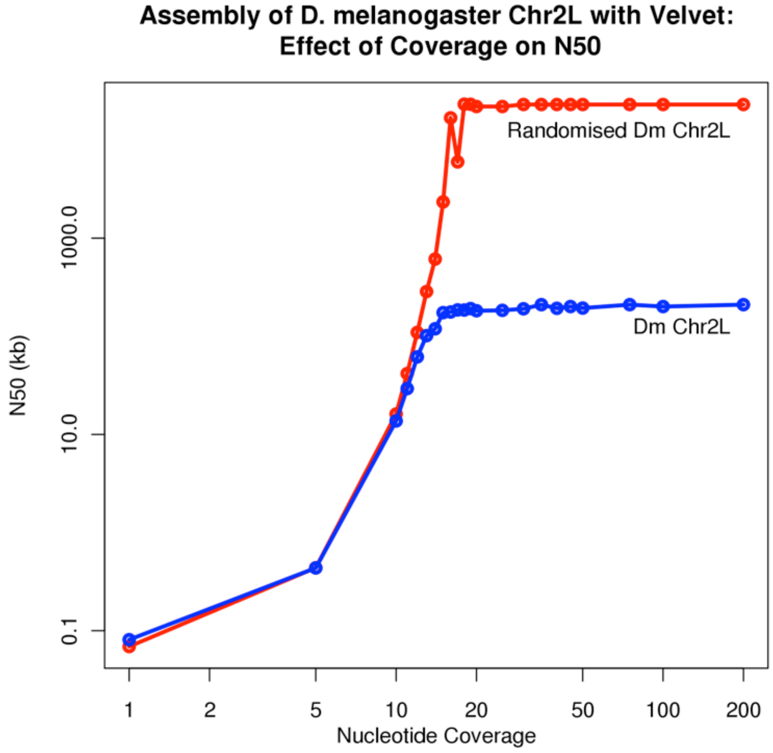
Target sequence:

TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG

Genome Res. 2008 May;18(5):821-9. **Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Zerbino DR, Birney E.** PMID: 18349386

Repetitive regions

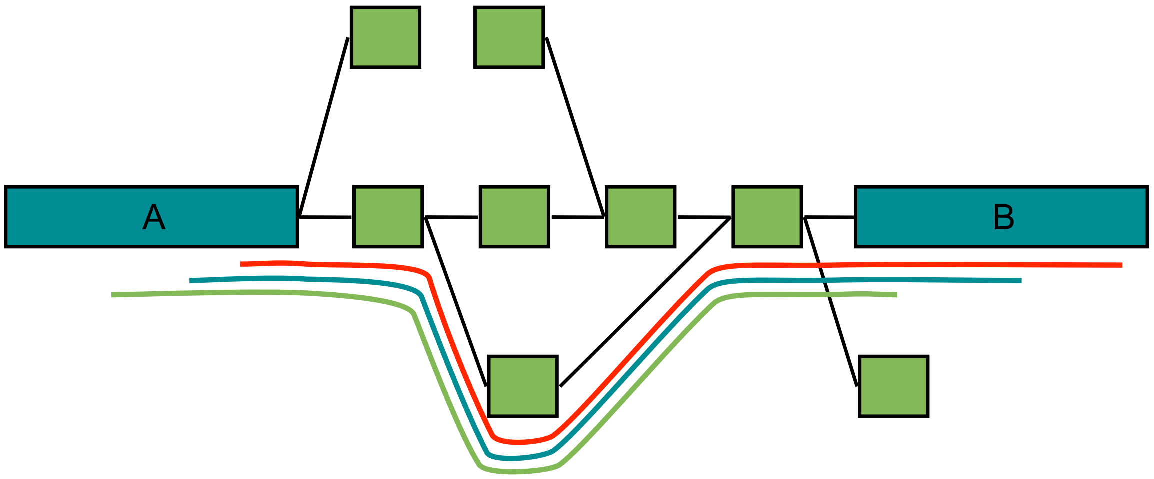




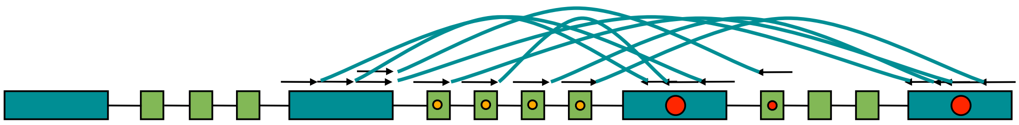
Read length 75bp, kmer 31bp

N50 is the contig size such that 50% of the assembly (or genome) is covered by contigs at least this long

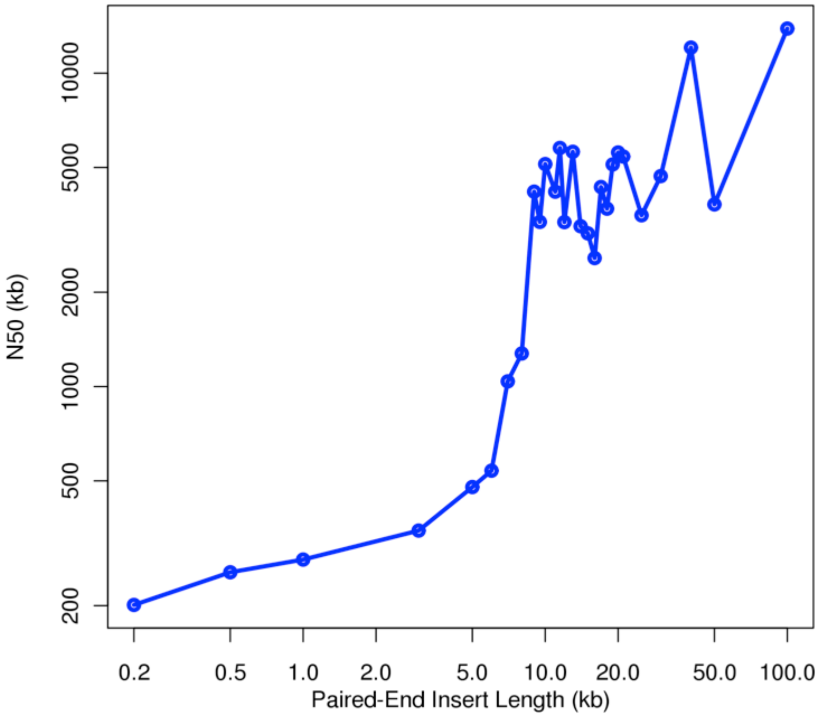
Rock Band: Using long and short reads together



Pebble: Handling paired-end reads (much improved version of breadcrumbs algorithm in the paper)



Assembly of D. melanogaster Chr2L with Velvet: Effect of paired-end insert length on N50



Read length 75bp, kmer 31bp, CV insert length 0.1