高通量生物学浅谈

Rui Tian, Ph.D.

April, 2017, Shanghai

Some concepts

- 生物信息学
- 计算生物学
- 高通量测序(**N**ext **G**eneration **S**equencing, NGS, 2nd generation sequencing)

3rd generation sequencing is already available

  - PacBio sequel
  - Oxford Nanopore
  - Around 15kb, 10-15% error rate

生物信息的核心问题

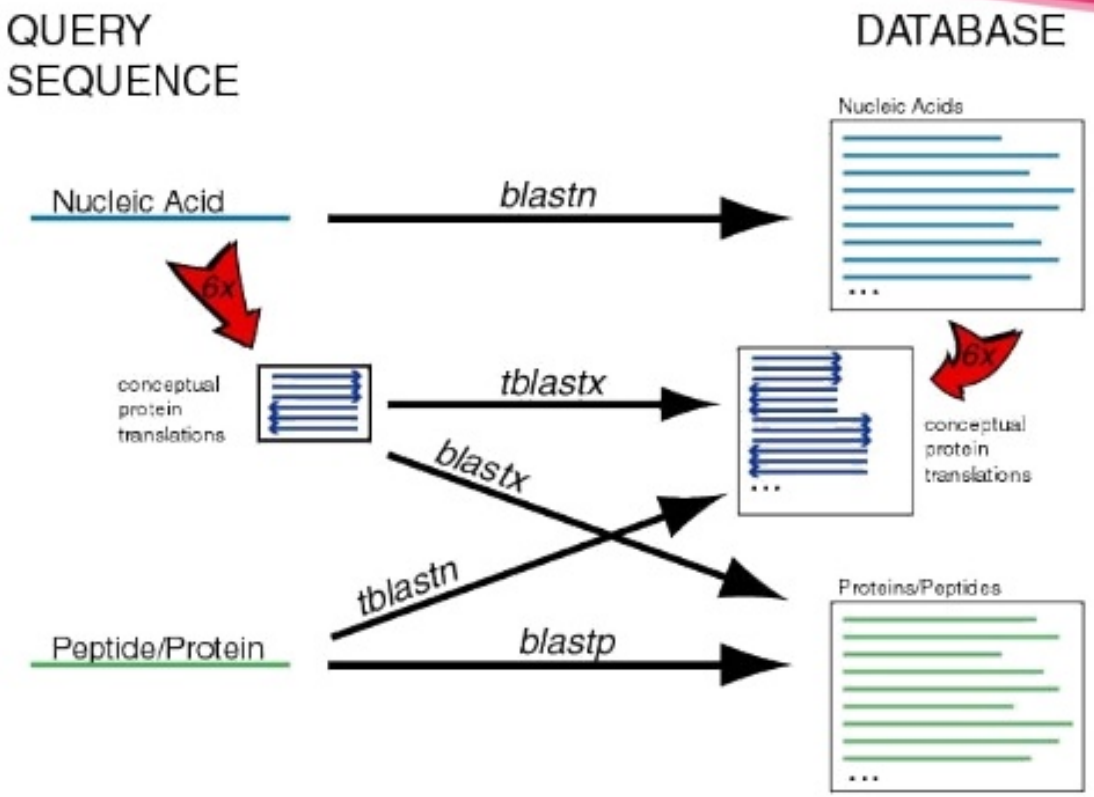
- 生物数据的格式化存贮，分享，提取和展示
  - IT层面，infrastructure
  - 数据库 ：Genbank, DDBJ, Swissprot, …
  - 数据仓库： GEO, SRA, ...
  - 数据门户：cBioportal, ICGC dataportal, ...
- 生物数据的算法问题
  - 序列比对, sequence alignment
  - 序列组装, sequencing assembly
  - 其他算法， 统计模型

Sequence Alignment

- Looking for similarities among sequences
  - query against databases
  - for a sequence of unknown origin, find its best hit
  - for a stretch of sequence, find its chromosomal coordinates
  - BLAST, blastn, blastp, blastx

1. blastp: compares a protein sequence against a protein sequence database.
2. blastn: compares a nucleotide sequence against a nucleotide sequence database.
3. blastx: compares a six frame translation of a nucleotide sequence against a protein database
4. tblastn: compares a protein sequence against a six frame translation of a nucleotide database
5. tblastx: compares a six frame translation of a nucleotide sequence against a six frame translation of a nucleotide database.

Diagrammatic View



https://www.slideshare.net/arundhatimehta50/blast-63513285

Sequence Alignment, How?

ATG
AATTACCGCT**ATG**T

Insertion, deletion
ACTGC
ACGC, AC-GC

complementary
ACTGC
GCAGT

- Imagine databases comprises tens of thousand of genomes
- each genome could be up to **3 billion** base pairs
- We need fast and efficient algorithms to speed the searching process!

BLAST

- BLAST, Basic Local Alignment Search Tool, similar to Simith-Waterman, but faster
- Global v.s. Local Alignment, 全局比对是找出两个序列全长的最优比对 (Not that useful)
- Designed by Altschul, Gish, Miller, Myers and Lipman at NIH in 1990, cited over 50,000 times, a tool becomes a verb in English
- BLAST was 50 times faster than previous method (Needleman-Wunsch, global, dynamic programming)
- I personally would count 1990 as the year of birth of Bioinfromatics

```
LGPSTKQFGKGSSSRIWDN
|       ||||    |   |  global alignment
LNQIERSFGKGAIMRLGDA
```

```
--------FGKG---------
        ||||  local alignment
--------FGKG---------
```

BLAST Algorithm

- Step1, remove low complexity region or repeats in query

- Step2, make k-letter hash of the query
  Example:
  QGEFGP
  1:QGE, 2:GEF, 3:EFG, 4:FGP

  for DNA, k=11

- Step3, List all possibly matched k-mers, score each letter using a Scoring Matrix

- Step4, Keep those k-mer match above certain threshold

- Repeat step 1-4 for every k-mer

- Seeding and extending (high-scoring segment pair, HSP)

Query sequence: R P P Q G L F

Database sequence: D P **P E G** V V

→Exact match is scanned.

Score: -2 **7 7 2 6 1** -1

→HSP

Optimal accumulated score = 7+7+2+6+1 = 23

BLAST Algorithm

- List all HSPs hit in the database above cutoff score (S)
  S can be set by observing random sequence scoring distribution, statistical significance can be gained

- BLAST results interpretation:

  Similarity score
  Query coverage
  Max identity
  E-value, the lower or the closer to zero, the more significant the match is

BLAST Lab 1

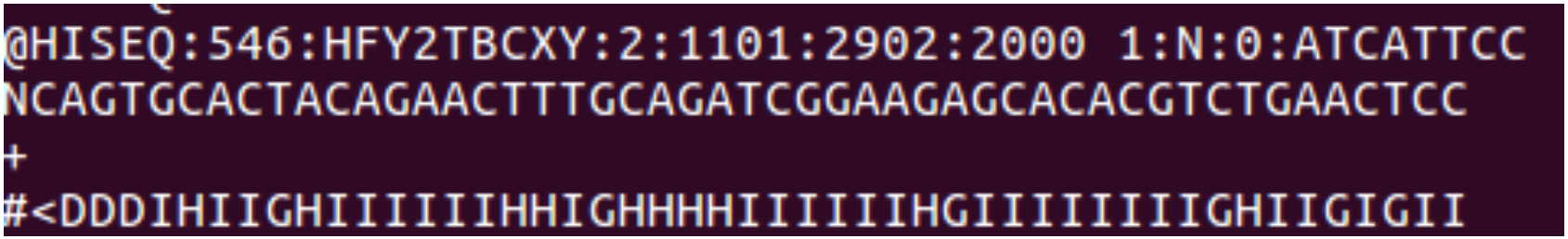
https://blast.ncbi.nlm.nih.gov/Blast.cgi

>seqX
AAAGCTTTTCTGTAAAGCTTATAAAAGCTAGGCCCGGACCCTTCTGTGGATAAGTGCC
TTTAGCCCTTG

Sequence Alignment in NGS

FASTQ file format

Line1: label
Line2: seq
Line3: +
Line4: Quality



From 30-50 bp Single End to 150 bp paired end, illumina/ solexa sequencing reads mapping (short read alignment against human genome, e.g., human, 3Gb)

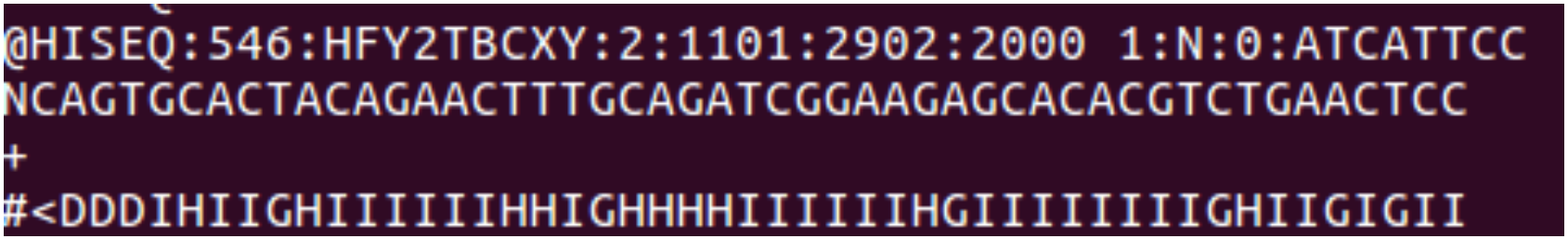
Even if each read cost 0.1 second, 10 million reads would cost 10*10^6/3600= 300 hours!!!

So BLAST can not handle NGS data!

FASTQ Format

FASTQ file format

Line1: label
Line2: seq
Line3: +
Line4: Quality



Q = -10 *$\log_{10}$p, p is sequencing error probability (base calling error)

Q value =30, means p is 0.001

Encoding: ASCII Q+33
American Standard Code for Information Interchange
ASCII (“D”)= 68, Q=68-32=36, meaning error prob is 10^-3.6

Short Reads Aligner

- Bowtie2
- BWA

Fast gapped-read alignment with Bowtie 2 : Nature Methods : Nature ...
www.nature.com › Home › archive › issue › Brief Communication ▾ 翻译此页
作者：B Langmead - 2012 - 被引用次数：5838 - 相关文章
The Bowtie 2 software achieves fast, sensitive, accurate and memory-efficient gapped alignment of sequencing reads using the full-text minute index and ...

Fast and accurate short read alignment with Burrows–Wheeler transform
https://www.ncbi.nlm.nih.gov › NCBI › Literature › PubMed Central (PMC) - 翻译此页
作者：H Li - 2009 - 被引用次数：10494 - 相关文章
2009年5月18日 - Results: We implemented Burrows-Wheeler Alignment tool (BWA), a new read alignment package that is based on backward search with ...

Bowtie Performance

Varying read length using Bowtie, Maq and SOAP

Length	Program	CPU time	Wall clock time	Peak virtual memory footprint (megabytes)	Bowtie speed-up	Reads aligned (%)
36 bp	Bowtie	6 m 15 s	6 m 21 s	1,305	-	62.2
	Maq	3 h 52 m 26 s	3 h 52 m 54 s	804	36.7×	65.0
	Bowtie -v 2	4 m 55 s	5 m 00 s	1,138	-	55.0
	SOAP	16 h 44 m 3 s	18 h 1 m 38 s	13,619	216×	55.1
50 bp	Bowtie	7 m 11 s	7 m 20 s	1,310	-	67.5
	Maq	2 h 39 m 56 s	2 h 40 m 9 s	804	21.8×	67.9
	Bowtie -v 2	5 m 32 s	5 m 46 s	1,138	-	56.2
	SOAP	48 h 42 m 4 s	66 h 26 m 53 s	13,619	691×	56.2
76 bp	Bowtie	18 m 58 s	19 m 6 s	1,323	-	44.5
	Maq 0.7.1	4 h 45 m 7 s	4 h 45 m 17 s	1,155	14.9×	44.9
	Bowtie -v 2	7 m 35 s	7 m 40 s	1,138	-	31.7

The performance of Bowtie v0.9.6, SOAP v1.10, and Maq versions v0.6.6 and v0.7.1 on the server platform when aligning 2 M untrimmed reads from the 1,000 Genome project (National Center for Biotechnology Information Short Read Archive: SRR003084 for 36 base pairs [bp], SRR003092 for 50 bp, and SRR003196 for 76 bp). For each read length, the 2 M reads were randomly sampled from the FASTQ file downloaded from the Archive such that the average per-base error rate as measured by quality values was uniform across the three sets. All reads pass through Maq's "catfilter". Maq v0.7.1 was used for the 76-bp reads because v0.6.6 does not support reads longer than 63 bp. SOAP is excluded from the 76-bp experiment because it does not support reads longer than 60 bp. Other experimental parameters are identical to those of the experiments in Table 1. CPU, central processing unit.

Maq & SOAP build hash table of locations of k-mers

2Million reads

Faster than

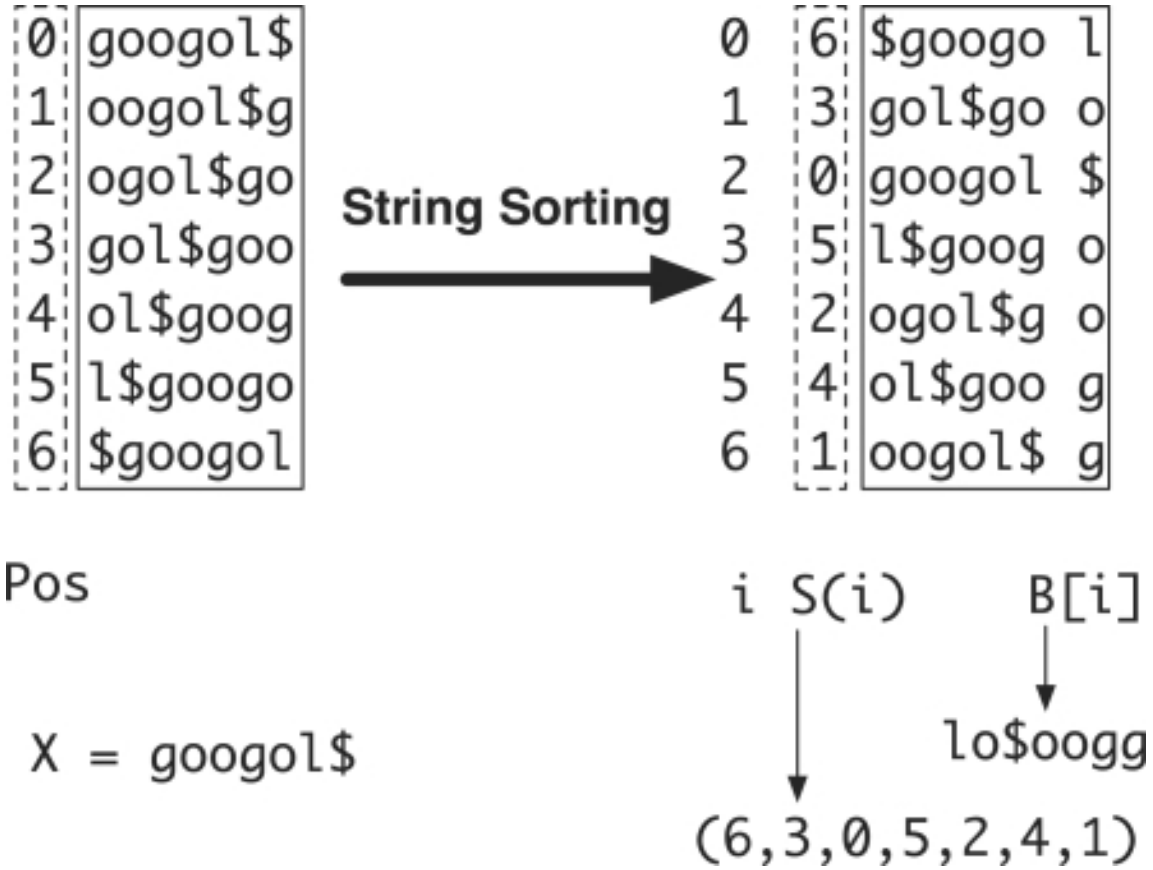
✓Maq: 20 times
(Li H. 2008, Genome Research)

✓SOAP: 700 times
(BGI, 2008)

The power of algorithm

Langmead et al. (2008)

Burrows-Wheeler Transform



Constructing suffix array and BWT string for X=googol$. String X is circulated to generate seven strings, which are then lexicographically sorted. After sorting, the positions of the first symbols form the suffix array (6, 3, 0, 5, 2, 4, 1) and the concatenation of the last symbols of the circulated strings gives the BWT string lo $oogg.

BWA lab 2

- Download and install bwa, only ok for Linux/Mac system

  https://sourceforge.net/projects/bio-bwa/files/
  bwa-0.7.15.tar.bz2
  Make
  Make install
- Download reference genome file (.fasta) and build bwa index

  http://hgdownload.cse.ucsc.edu/goldenpath/hg19/chromosomes/
  rsync -avzP rsync://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/ .

  Usage: bwa index [options] <in.fasta>
  For human genome, hg19
  bwa -a bwtsw path2dir/hg19.fa
- Short reads alignment (mapping reads back to the reference genome)

  *bwa mem *
  /home/tianr/02databases/bwa_indexes/hg19.fa in1.fq.gz in2.fq.gz
  >out.sam

SAM format

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                       ATAGCT..............TCAGC
-r003                             ttagctTAGGC
-r001/2                                        CAGCGGCAT
```

The corresponding SAM format is:[1]

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

SAM format

Field	Alignment 1	Alignment 2	Alignment 3
QNAME	1:497:R:-272+13M17D24M	19:20389:F:275+18M2D19M	19:20389:F:275+18M2D19M
FLAG	113	99	147
RNAME	1	1	1
POS	497	17644	17919
MAPQ	37	0	0
CIGAR	37M	37M	18M2D19M
MRNM/RNEXT	15	=	=
MPOS/PNEXT	100338662	17919	17644
ISIZE/TLEN	0	314	
SEQ	CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG	TATGACTGCTAATAATACCTACACATGTTAGAACCAT	GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT
QUAL	0;==-==9;>>>>>>=>>>>>>>>>>>>>>>>>>>>>>>	>>>>>>>>>>>>>>>>>>>> <<>>>><<>>4::>>:<9	;44999;499<8<8<<<8<<><<<<<><7<; <<<>>><
TAGs	XT:A:U NM:i:0 SM:i:37 AM:i:0 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37	RG:Z:UM0098:1 XT:A:R NM:i:0 SM:i:0 AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37	XT:A:R NM:i:2 SM:i:0 AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:1 XG:i:2 MD:Z:18^CA19

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	`[0,2`^{16}`-1]`	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	`[0,2`^{31}`-1]`	1-based leftmost mapping POSition
5	MAPQ	Int	`[0,2`^{8}`-1]`	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	`[0,2`^{31}`-1]`	Position of the mate/next read
9	TLEN	Int	`[-2`^{31}`+1,2`^{31}`-1]`	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

SAM format

FLAG: Combination of bitwise FLAGs.[5] Each bit is explained in the following table:

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

ReCap

Sequence Alignment:

BLAST
BWA, bowtie2
samtools, bedtools

Formats:
Fastq, SAM/BAM, bed

Sequence Assembly & Annotation

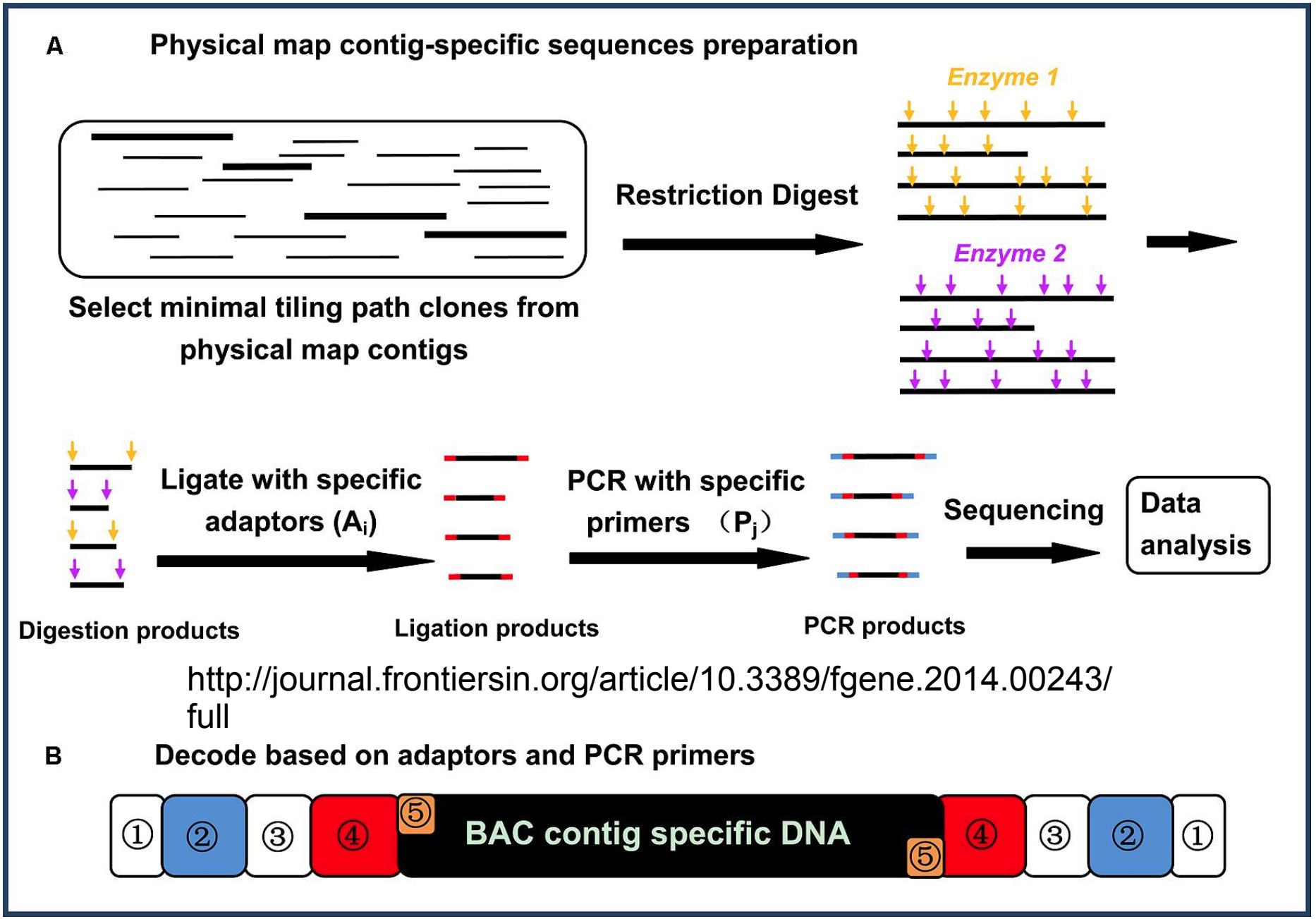
The Human Genome Project (1990 -2003)
NIH, USA
UK, Japan, France, Germany, Canada, China(%3???)

- map the nucleotides contained in a human haploid reference genome (more than three billion).

- The "genome" of any given individual is unique; mapping the "human genome" involved sequencing a small number of individuals and then assembling these together to get a complete sequence for each chromosome.

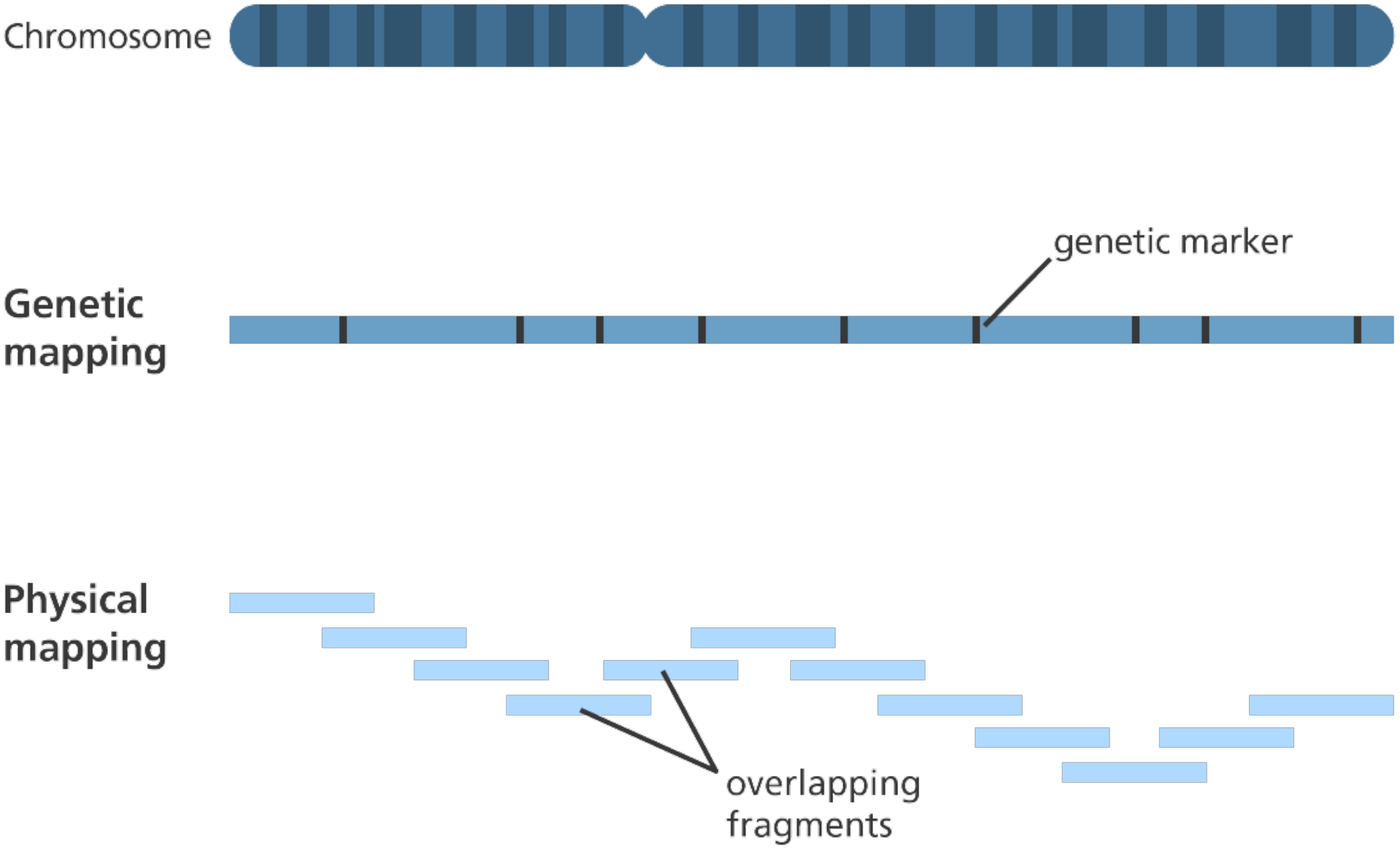
- The finished human genome is thus a mosaic, not representing any one individual.

How to sequence a entire genome?



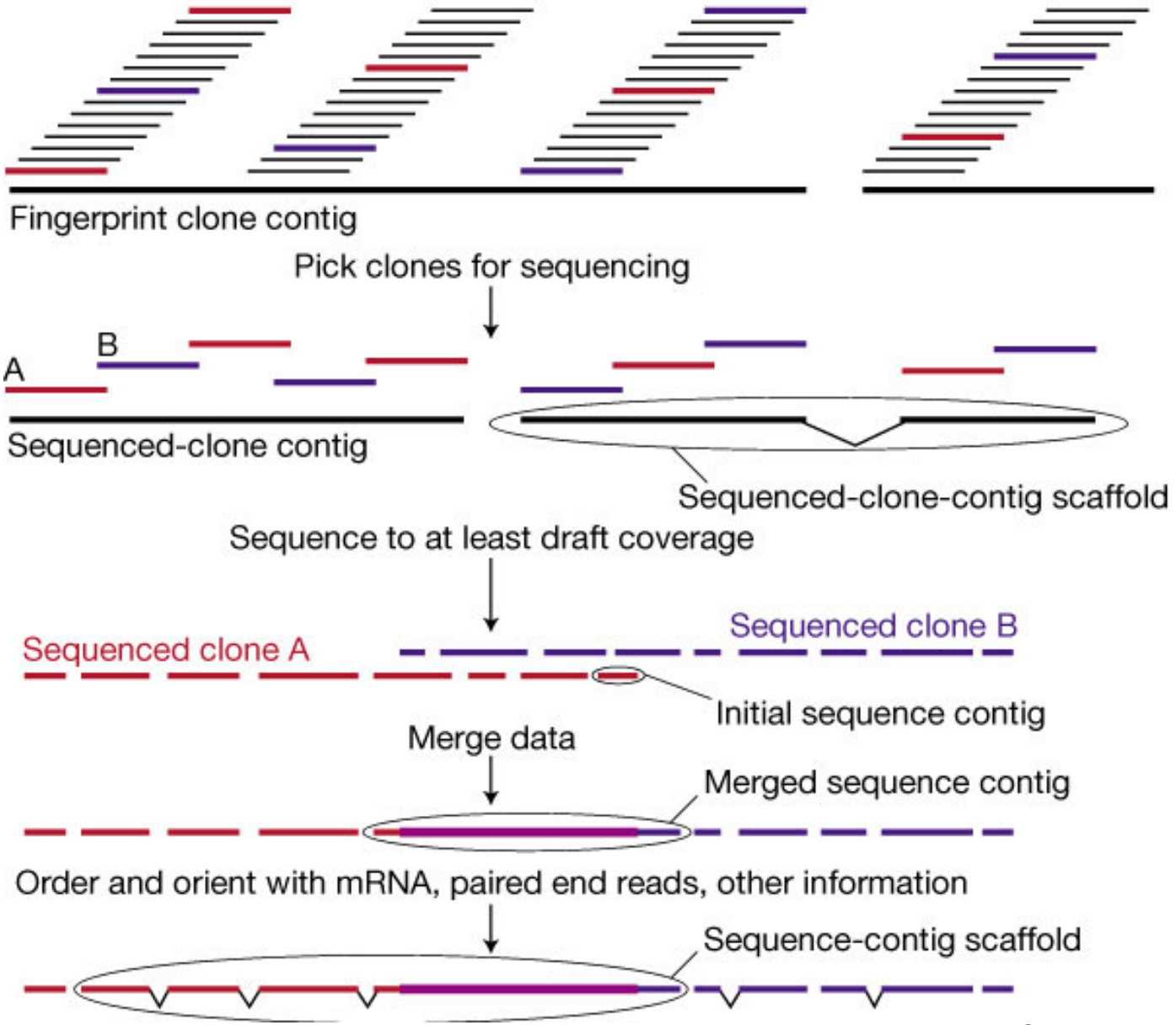
http://journal.frontiersin.org/article/10.3389/fgene.2014.00243/full

Genetic Map, Physical Map



http://www.yourgenome.org/facts/how-do-you-map-a-genome

Genetic Map, Physical Map



http://www.nature.com/nature/journal/v409/n6822/full/409860a0.html

Lab 3: hg18, hg19, hg38

- UCSC hg18, NCBI36, Mar. 2006
- UCSC hg19, GRCh37, Feb.2009
- UCSC hg38, GRCh38, Dec. 2013

http://asia.ensembl.org/Homo_sapiens/Tools/AssemblyConverter?db=core

https://genome.ucsc.edu/cgi-bin/hgTables

https://genome.ucsc.edu/cgi-bin/hgLiftOver?hgsid=585392609_XOAlORBlWpN5FHICO74JNXNfqWZd

De novo assembly

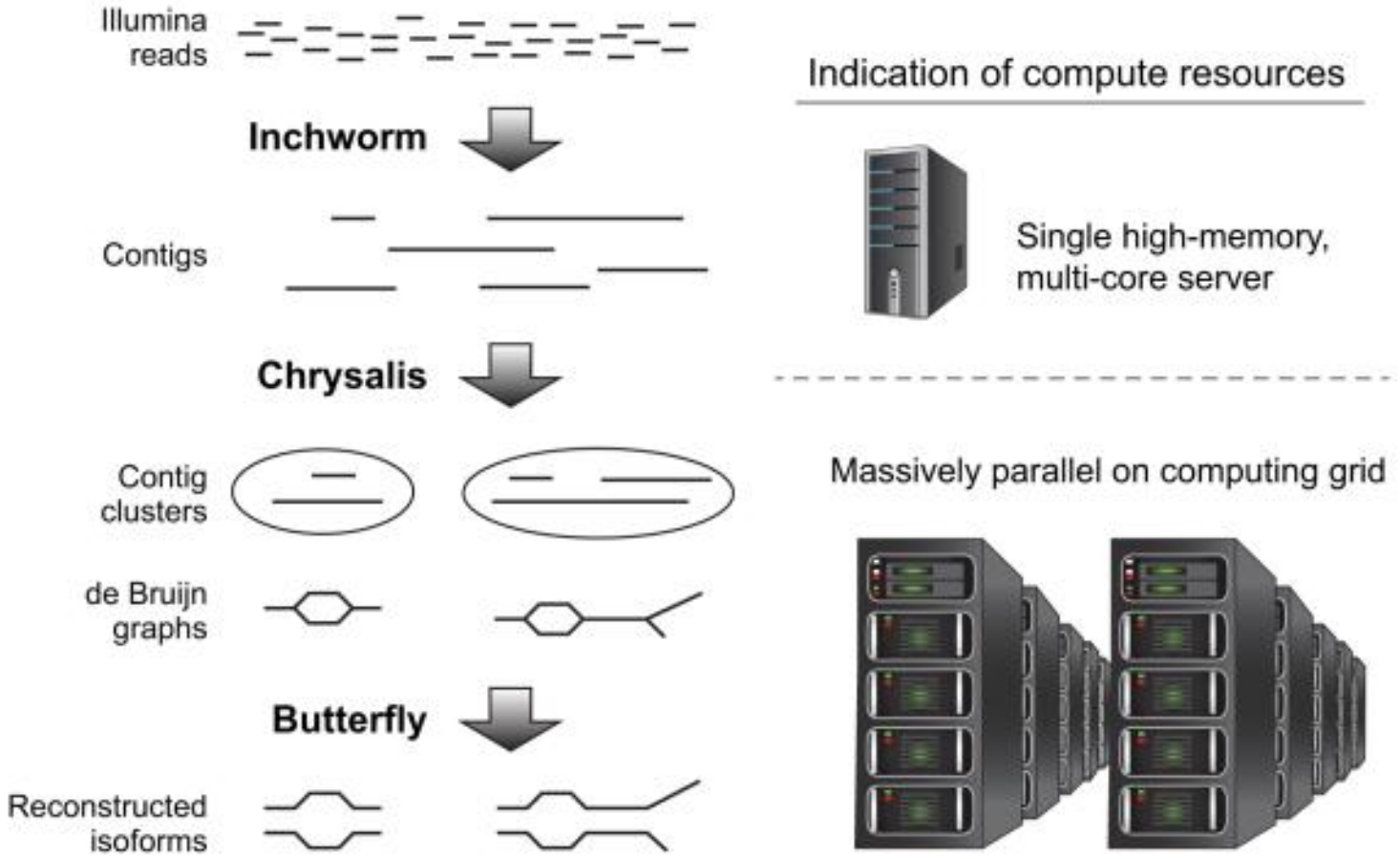
When no model reference genome is available:

Tritinity

De novo transcript sequence reconstruction from RNA-Seq: reference ...
https://www.ncbi.nlm.nih.gov › NCBI › Literature › PubMed Central (PMC) - 翻译此页
作者：BJ Haas - 2013 - 被引用次数：1094 - 相关文章
2013年7月11日 - In this protocol, we describe the use of the Trinity platform for de novo transcriptome assembly from RNA-Seq data in non-model organisms.

De novo assembly

Tritinity: RAM 256G-1T



ReCap

- Sequence Alignment
  - ✓ Blast, BWA
  - ✓ FASTQ, SAM, BED
  - ✓ Web utinities
- Sequence Assembly
  - ✓ Foundation of genome studies
  - ✓ Genome complexities

NGS based studies

1)Transcriptome

- ✓ Expression quantification
- ✓ Differentially expressed genes
- ✓ Alternative splicing
- ✓ Gene fusion
- ✓ mRNA, non-coding (microRNA, linc RNA, circle RNA)

NGS based studies

2)ChIP-seq

- TF ChIP-seq
- Histone mark ChIP-seq, H3K4me3, H3K27me3, H3K27me3, etc.
- ATAC-seq (Assay of Transposase Accessible Chromatin)
- Genome wide binding (signal distribution), cofactor finding, motif finding

NGS based studies

3)DNA methylation (WGBS)

- ✓ CpG methylation profile
- ✓ Hyper methylated genes
- ✓ Hypo methylated genes

NGS based studies

4)genomic variation (WEX, WGS)

- Population level, dozens of , hundreds of samples
- SNV, Indel
- CNV
- SV
- Tumor or genetic disease?

NGS based studies

5)genome de novo assembly

- ✓ Organisms of small genomes, microorganisms
- ✓ Model or non-model, Panda???

NGS based studies

6)Metagenomics

- ✓ Microorganisms, soil, water or human gut
- ✓ Strain and abundance of bacteria

Take home messages

- Blast algorithm, Blast web tool
- BWA algorithm, build index, alignment
- Genome assembly conversion web tool
- Major fields where NGS are used, questions addressed