

How to map billions of short reads onto genomes

Methods in Bioinformatics

Student Presentation Group 4

October 30th, 2013

Next-Generation Sequencing

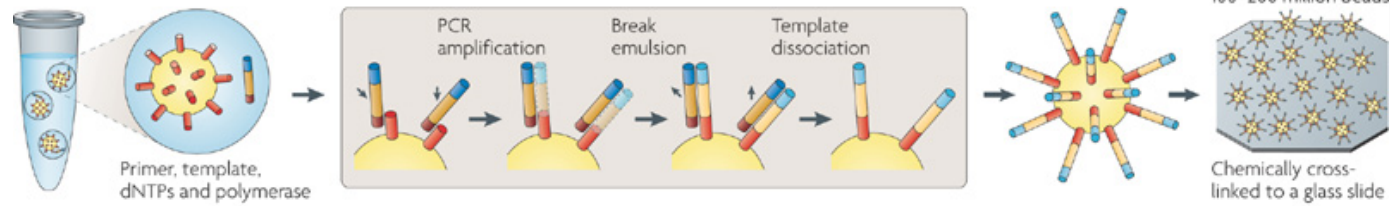
➤ A variety of sequencing-based assays:

- gene expression
- DNA-protein interaction
- human re-sequencing
- RNA splicing studies

➤ Examples: RNA-Seq, ChIP-Seq

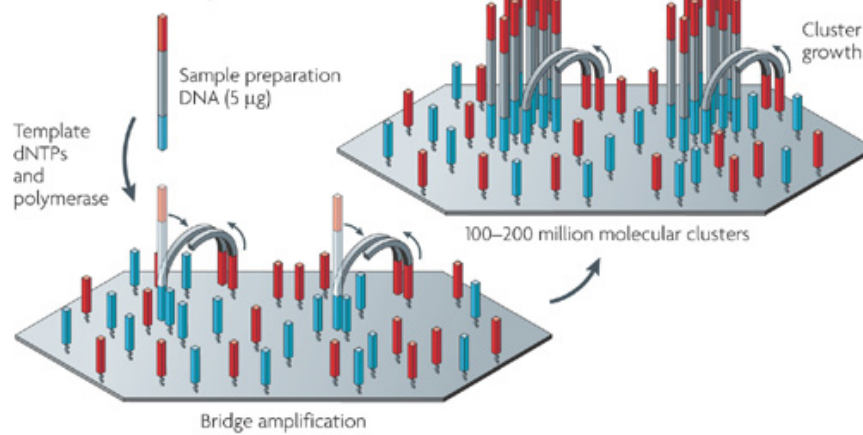
a Roche/454, Life/APG, Polonator
Emulsion PCR

One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion

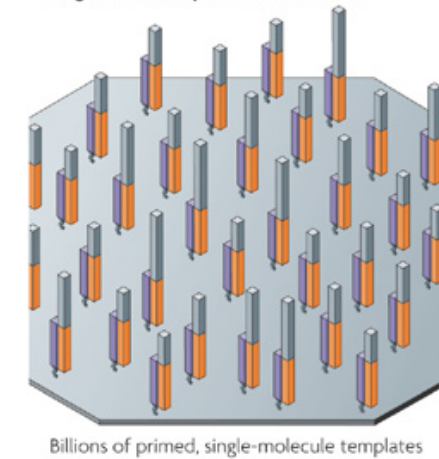


b Illumina/Solexa
Solid-phase amplification

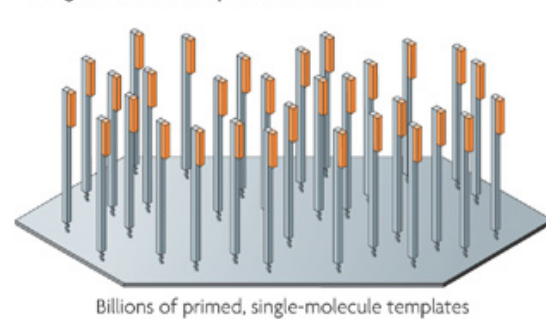
One DNA molecule per cluster



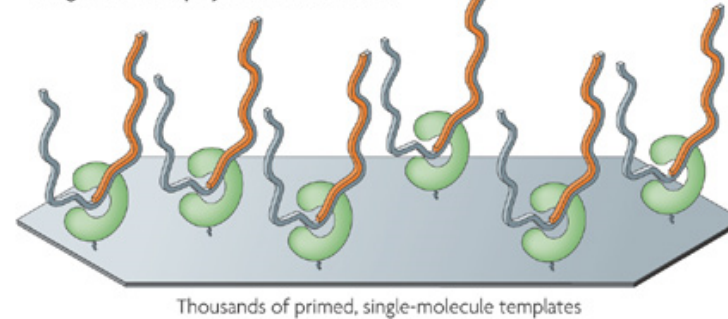
c Helicos BioSciences: one-pass sequencing
Single molecule: primer immobilized



d Helicos BioSciences: two-pass sequencing
Single molecule: template immobilized

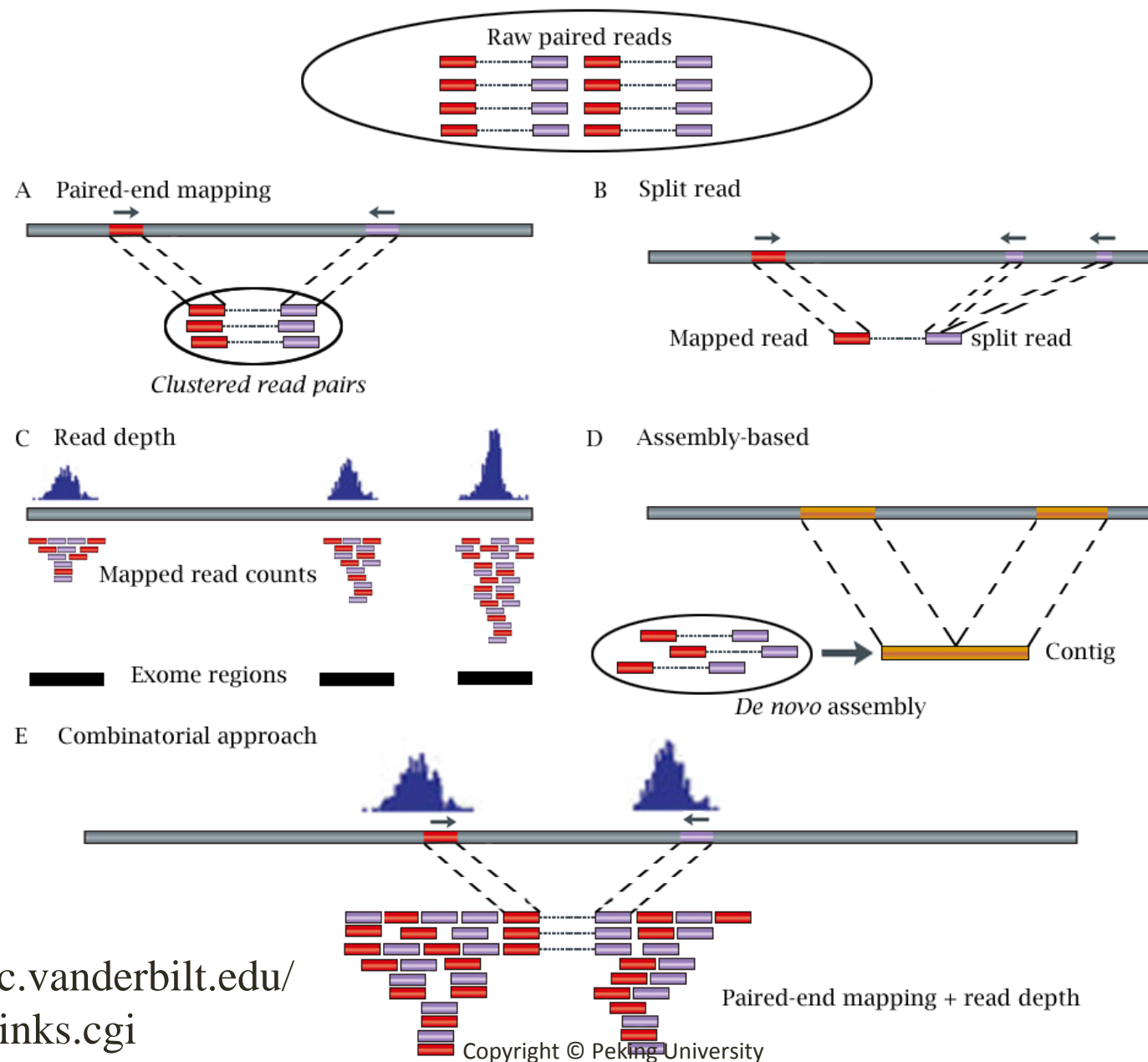


e Pacific Biosciences, Life/Visigen, LI-COR Biosciences
Single molecule: polymerase immobilized



Nature Reviews | Genetics

Genome Assembly



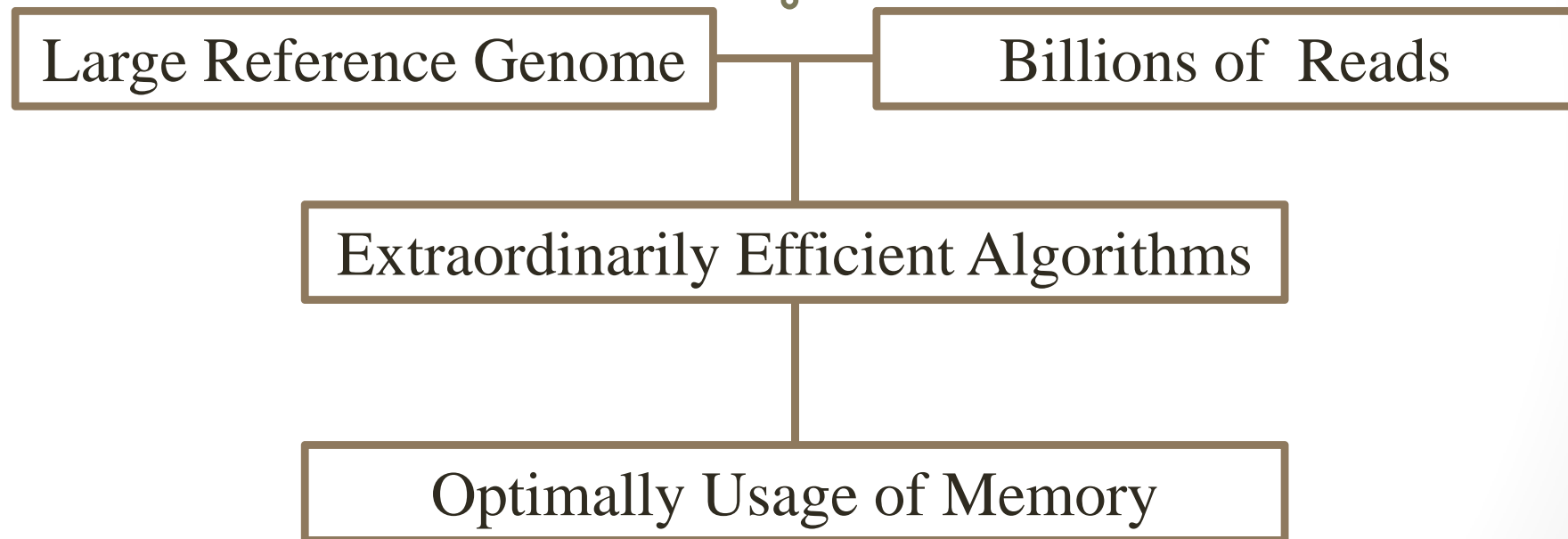
Scientific Question

- Short sequence fragments produced by next-generation sequencing platforms are quite large.
- Mapping the vast quantities of data is a challenge.
- ‘read mapping’ problem
- What programs are available and how do they work?

Challenges of mapping short reads

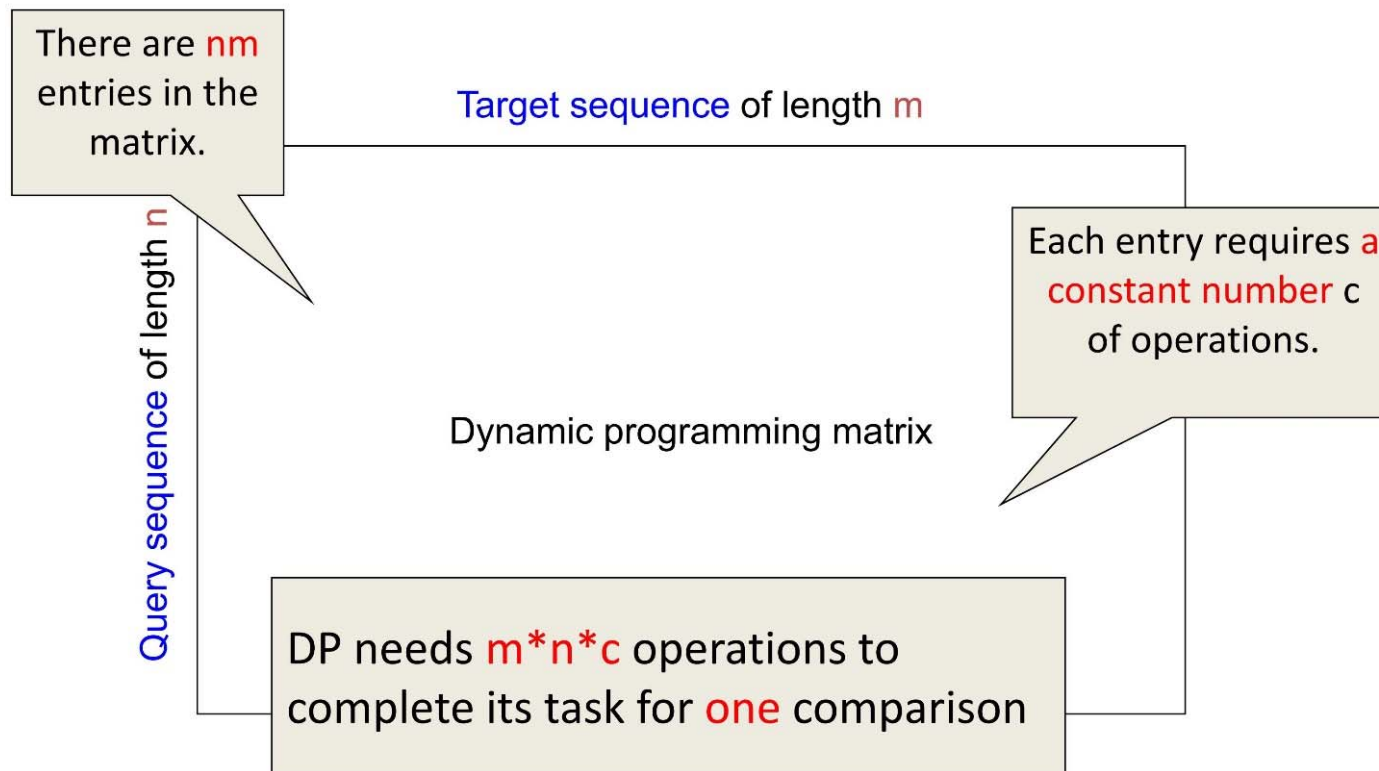
- The first challenge is a practical one:

How fast can
we align the
reads?



Genome Assembly & Mapping Problem

Sometimes, size does matter!



© Copyright 2012 Center for Bioinformatics, Peking University

Challenges of mapping short reads

- The second challenge is a strategic:

Which copy of
the reads
belongs to?

Repetitive Element Reads

Multiple possible locations

A Heuristic location

More Sequencing Errors or Variations

Spliced Mapping Problems

Alignment Programs

➤ Traditional alignment algorithms:

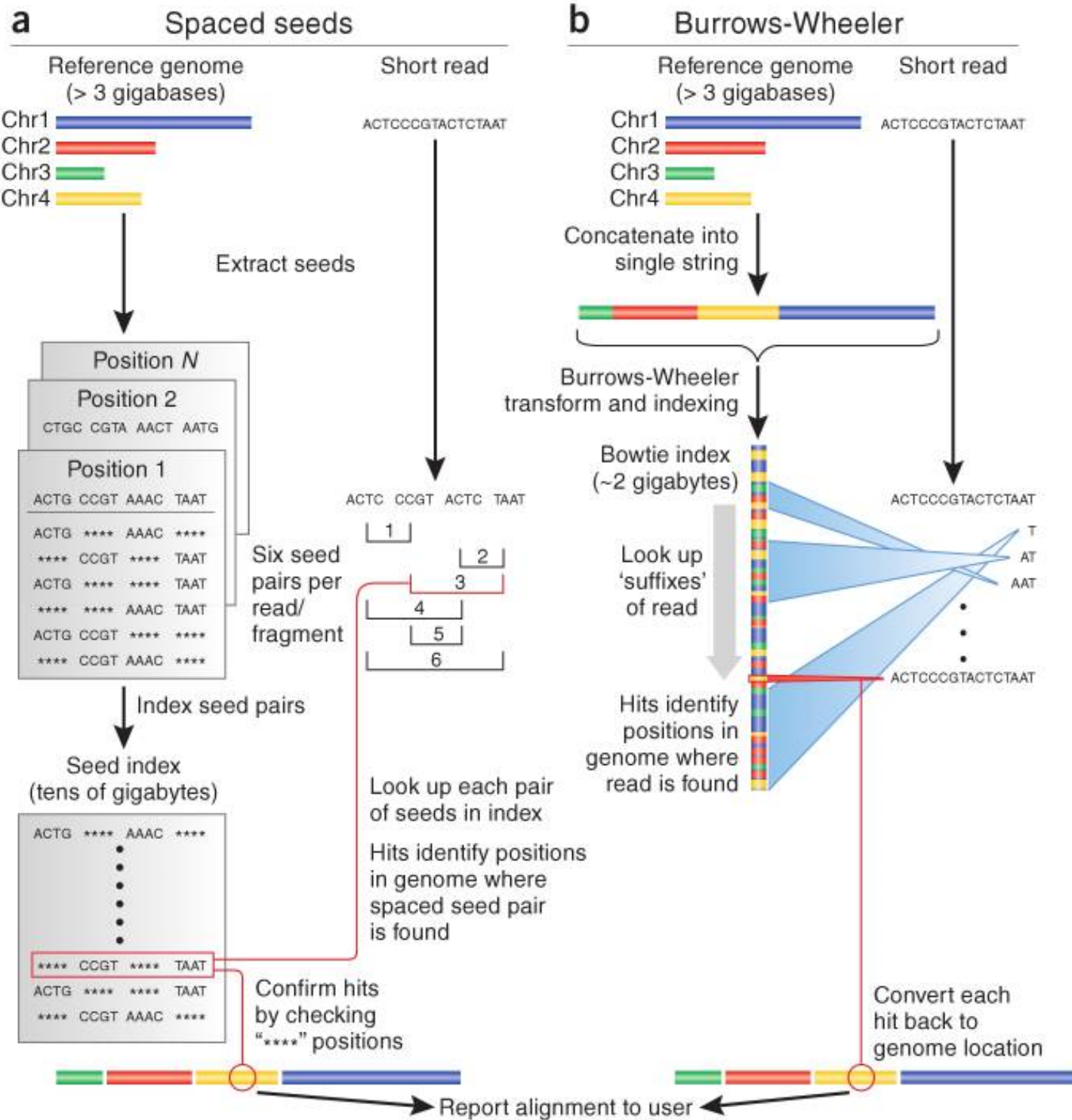
- BLAST (Basic Local Alignment Search Tool)
- BLAT (The BLAST-Like Alignment Tool)

➤ New generation alignment programs:

- E.g. ELAND program from Illumina

➤ Third-party software packages:

Program Website		Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbcb.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinfor.com	No	Yes	240



Limitations and Open Problems

- The current solutions for short-read mapping all have limitations.
- Many challenges and questions remain for developers of read mapping software:
 - Will the short-read mapping programs scale well as the reads get longer?
 - How should a program's parameters be adjusted, and can that adjustment happen automatically?
 - How useful is mapping quality in downstream analysis, and should it be computed while aligning reads, as Maq does, or later?

Acknowledgments

➤ All the members of Group 4:

- Xiqian Chen 陈西茜
- Dongqing Jiang 姜冬青
- Yuechen Liu 刘悦晨
- Yaping Wang 王雅萍
- Yixi Xu 徐毅曦
- Han Yan 闫晗
- Li Zhou 周莉
- Wenxiong Zhou 周文雄
- Yuxin Zhuang 莊宇新

➤ Our teacher:

- Ge Gao 高歌

➤ Our teacher assistant:

- Meng Wang 王萌