CEITEC

Central European Institute of Technology

BRNO | CZECH REPUBLIC

Introduction to Bioinformatics (LF:DSIB01)

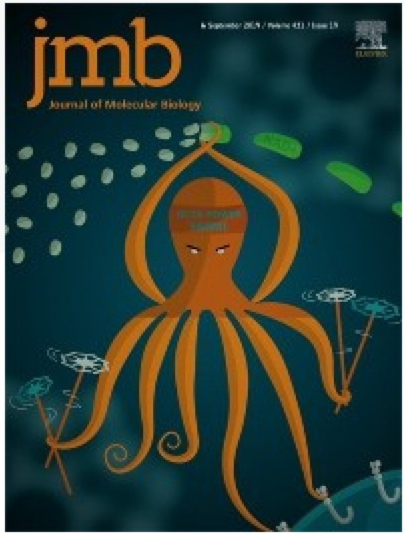
Week 5 : Sequence Alignment

BLAST - **Basic Local Alignment Search Tool**

- BLAST is a method for performing Local Alignment
- It uses a Seed query that must match perfectly to the reference, then builds around it
- The alignment ends when a score threshold is passed

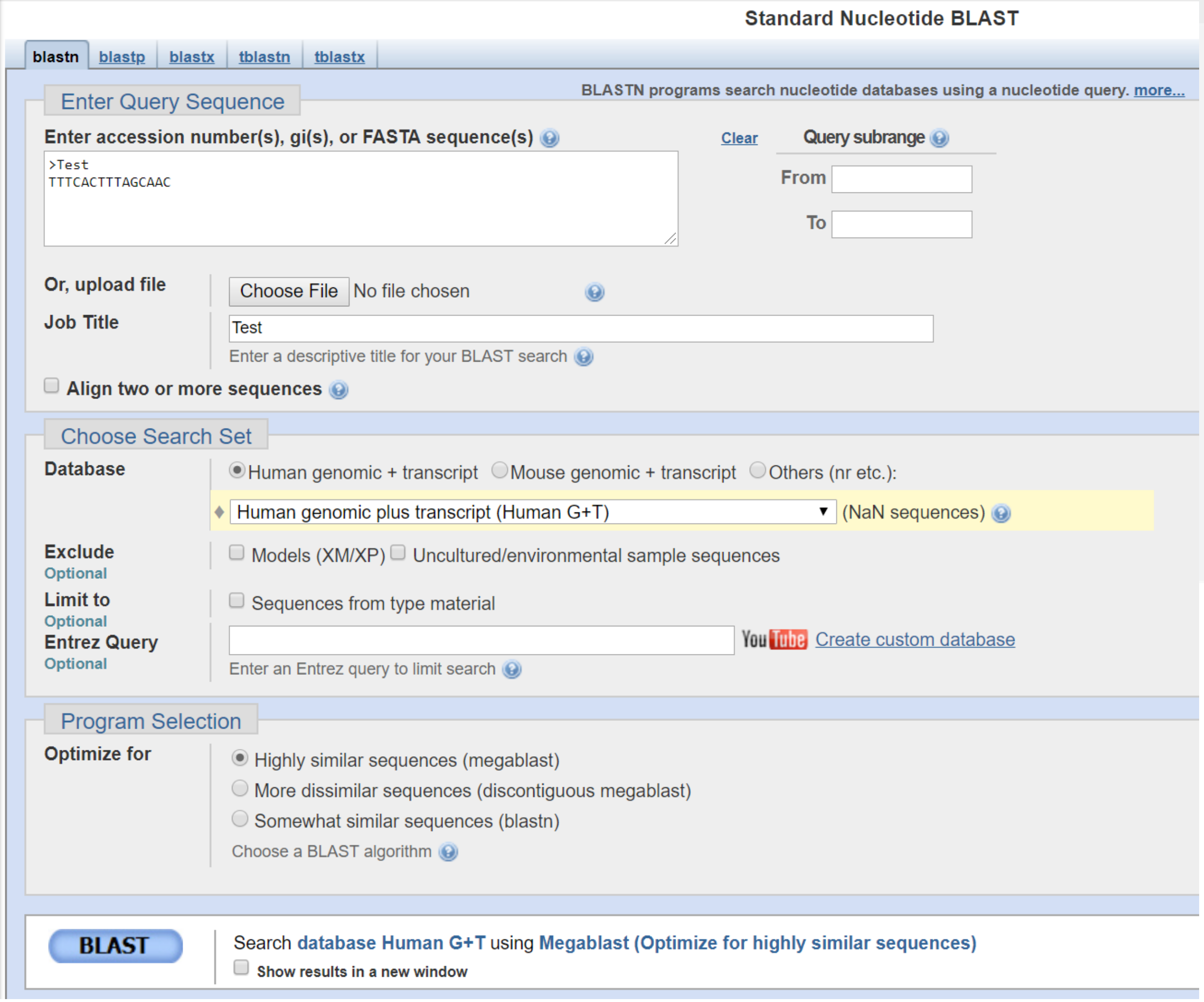
- Step 1: Break Query into short words of specific length W
- Step 2: Search for this sequence in a database of the reference
- Step 3: Keep seeds that pass Threshold and extend
- Step 4: Calculate E (Expected Value) – log10 chance that such alignment is found by luck

Published 1990
~80,000 citations



BLAST - **Basic Local Alignment Search Tool**

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



Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>Test
TTTCACTTTAGCAAC
```

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title: Test

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ● Human genomic + transcript ○ Mouse genomic + transcript ○ Others (nr etc.):

Human genomic plus transcript (Human G+T) (NaN sequences)

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to Optional: □ Sequences from type material

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Optimize for

● Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

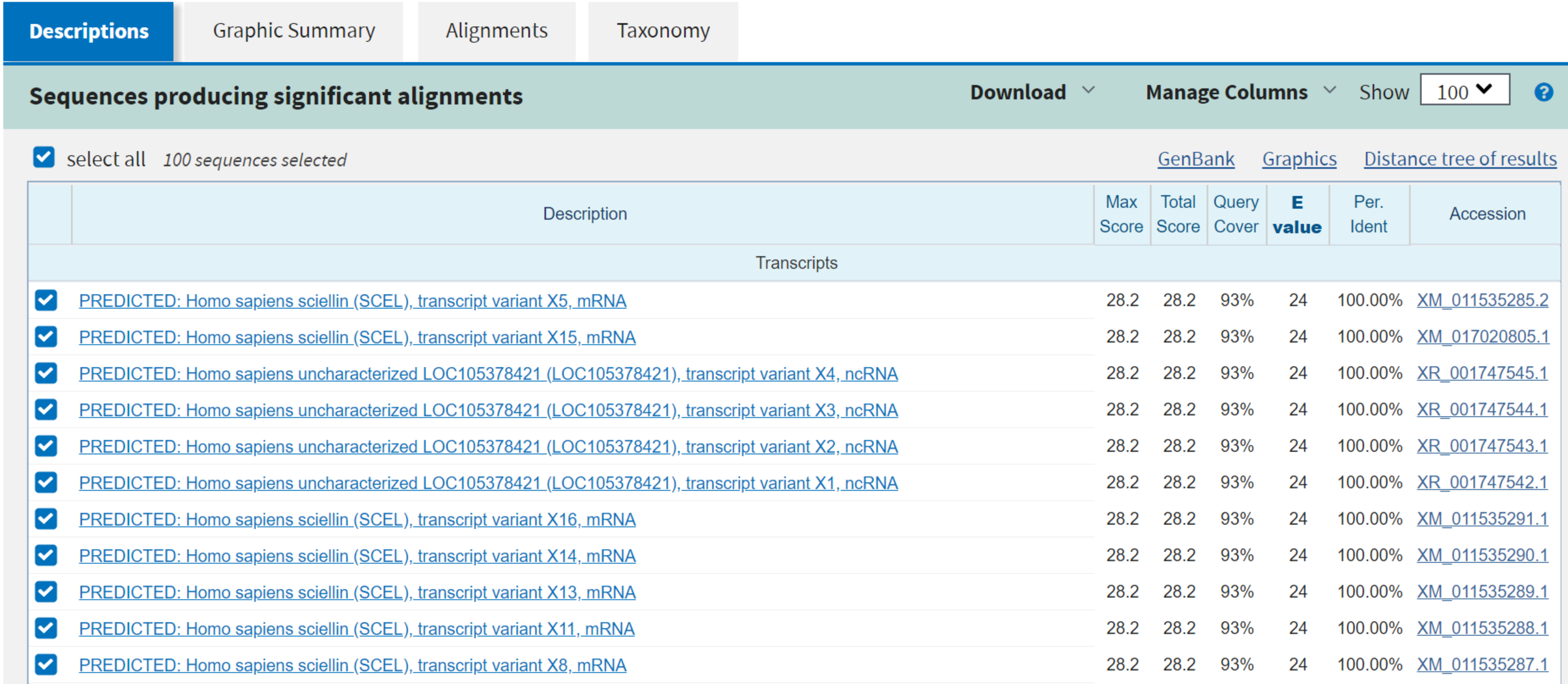
○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Search database Human G+T using Megablast (Optimize for highly similar sequences)

□ Show results in a new window

BLAST - **Basic Local Alignment Search Tool**



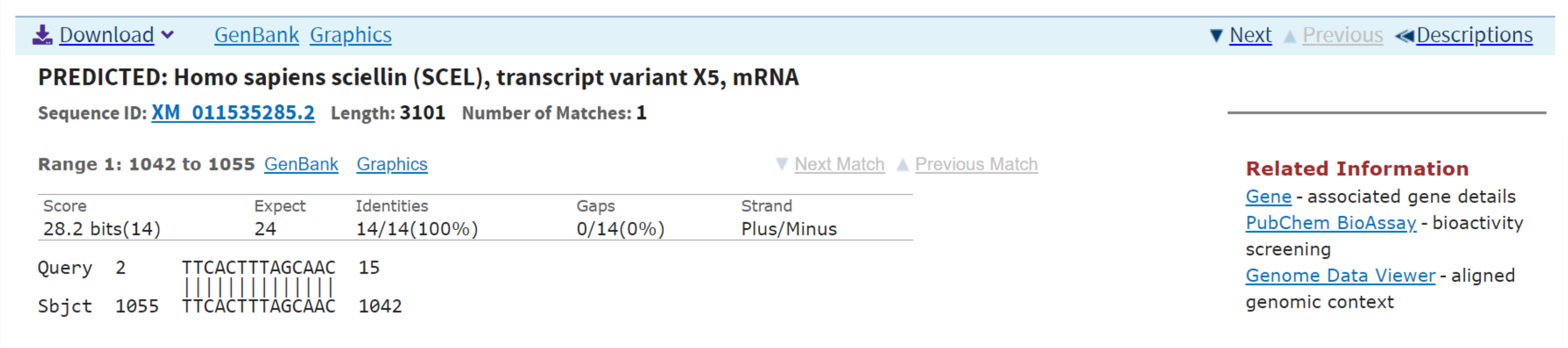
Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	**E value**	Per. Ident	Accession
	Transcripts						
☑	PREDICTED: Homo sapiens sciellin (SCEL), transcript variant X5, mRNA	28.2	28.2	93%	24	100.00%	XM_011535285.2
☑	PREDICTED: Homo sapiens sciellin (SCEL), transcript variant X15, mRNA	28.2	28.2	93%	24	100.00%	XM_017020805.1
☑	PREDICTED: Homo sapiens uncharacterized LOC105378421 (LOC105378421), transcript variant X4, ncRNA	28.2	28.2	93%	24	100.00%	XR_001747545.1
☑	PREDICTED: Homo sapiens uncharacterized LOC105378421 (LOC105378421), transcript variant X3, ncRNA	28.2	28.2	93%	24	100.00%	XR_001747544.1
☑	PREDICTED: Homo sapiens uncharacterized LOC105378421 (LOC105378421), transcript variant X2, ncRNA	28.2	28.2	93%	24	100.00%	XR_001747543.1
☑	PREDICTED: Homo sapiens uncharacterized LOC105378421 (LOC105378421), transcript variant X1, ncRNA	28.2	28.2	93%	24	100.00%	XR_001747542.1
☑	PREDICTED: Homo sapiens sciellin (SCEL), transcript variant X16, mRNA	28.2	28.2	93%	24	100.00%	XM_011535291.1
☑	PREDICTED: Homo sapiens sciellin (SCEL), transcript variant X14, mRNA	28.2	28.2	93%	24	100.00%	XM_011535290.1
☑	PREDICTED: Homo sapiens sciellin (SCEL), transcript variant X13, mRNA	28.2	28.2	93%	24	100.00%	XM_011535289.1
☑	PREDICTED: Homo sapiens sciellin (SCEL), transcript variant X11, mRNA	28.2	28.2	93%	24	100.00%	XM_011535288.1
☑	PREDICTED: Homo sapiens sciellin (SCEL), transcript variant X8, mRNA	28.2	28.2	93%	24	100.00%	XM_011535287.1

BLAST - **Basic Local Alignment Search Tool**

- Useful information about alignment(s)
- Works well for a single sequence



Download GenBank Graphics ▼ Next ▲ Previous ◀Descriptions

PREDICTED: Homo sapiens sciellin (SCEL), transcript variant X5, mRNA

Sequence ID: XM_011535285.2 **Length:** 3101 **Number of Matches:** 1

Range 1: 1042 to 1055 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
28.2 bits(14)	24	14/14(100%)	0/14(0%)	Plus/Minus

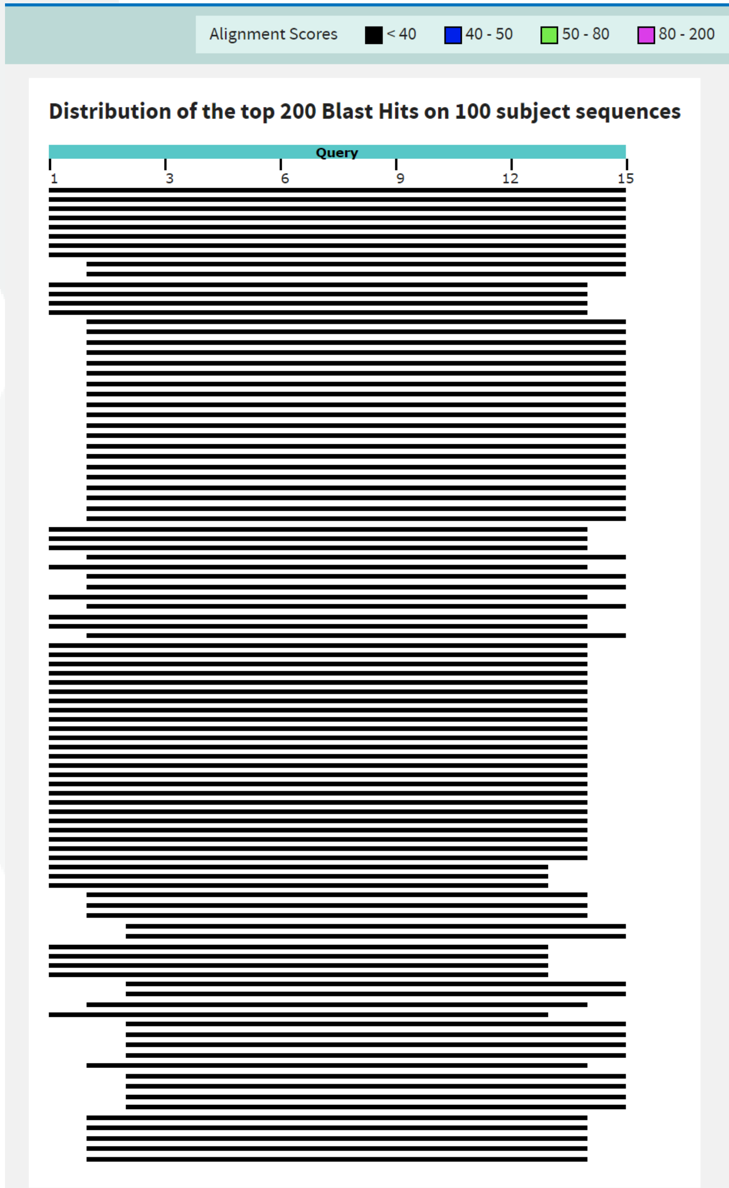
```
Query  2     TTCACTTTAGCAAC  15
             ||||||||||||||
Sbjct  1055  TTCACTTTAGCAAC  1042
```

Related Information

Gene - associated gene details

PubChem BioAssay - bioactivity screening

Genome Data Viewer - aligned genomic context



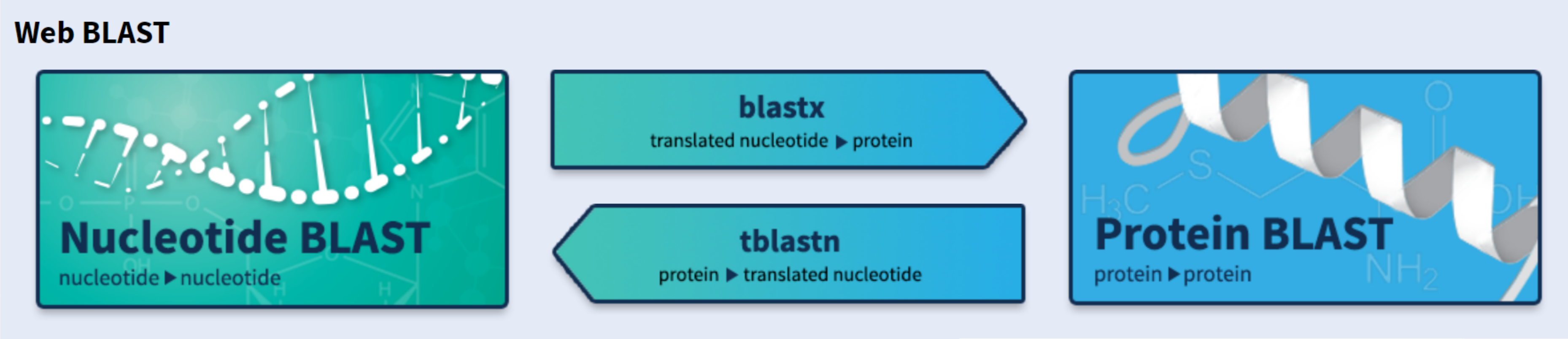
BLAST variations – old standalone programs

1999

- **BLASTN** - Compares a DNA query to a DNA database. Searches both strands automatically. It is optimized for speed, rather than sensitivity.
- **BLASTP** - Compares a protein query to a protein database.
- **BLASTX** - Compares a DNA query to a protein database, by translating the query sequence in the 6 possible frames, and comparing each against the database (3 reading frames from each strand of the DNA) searching.
- **TBLASTN** - Compares a protein query to a DNA database, in the 6 possible frames of the database.
- **TBLASTX** - Compares the protein encoded in a DNA query to the protein encoded in a DNA database, in the 6*6 possible frames of both query and database sequences (Note that all the combinations of frames may have different scores).
- **BLAST2** - Also called *advanced BLAST*. It can perform gapped alignments.
- **PSI-BLAST** - (Position Specific Iterated) Performs iterative database searches

BLAST variations – web services and API

2019



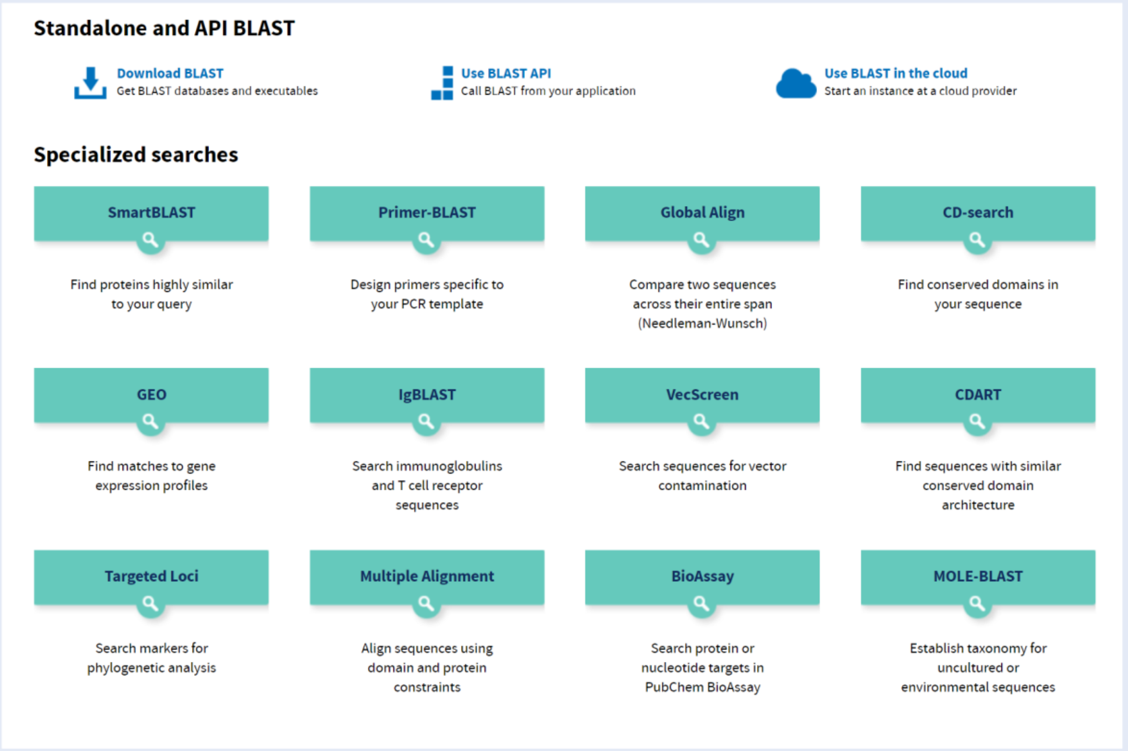
Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein



Standalone and API BLAST

Download BLAST
Get BLAST databases and executables

Use BLAST API
Call BLAST from your application

Use BLAST in the cloud
Start an instance at a cloud provider

Specialized searches

SmartBLAST
Find proteins highly similar to your query

Primer-BLAST
Design primers specific to your PCR template

Global Align
Compare two sequences across their entire span (Needleman-Wunsch)

CD-search
Find conserved domains in your sequence

GEO
Find matches to gene expression profiles

IgBLAST
Search immunoglobulins and T cell receptor sequences

VecScreen
Search sequences for vector contamination

CDART
Find sequences with similar conserved domain architecture

Targeted Loci
Search markers for phylogenetic analysis

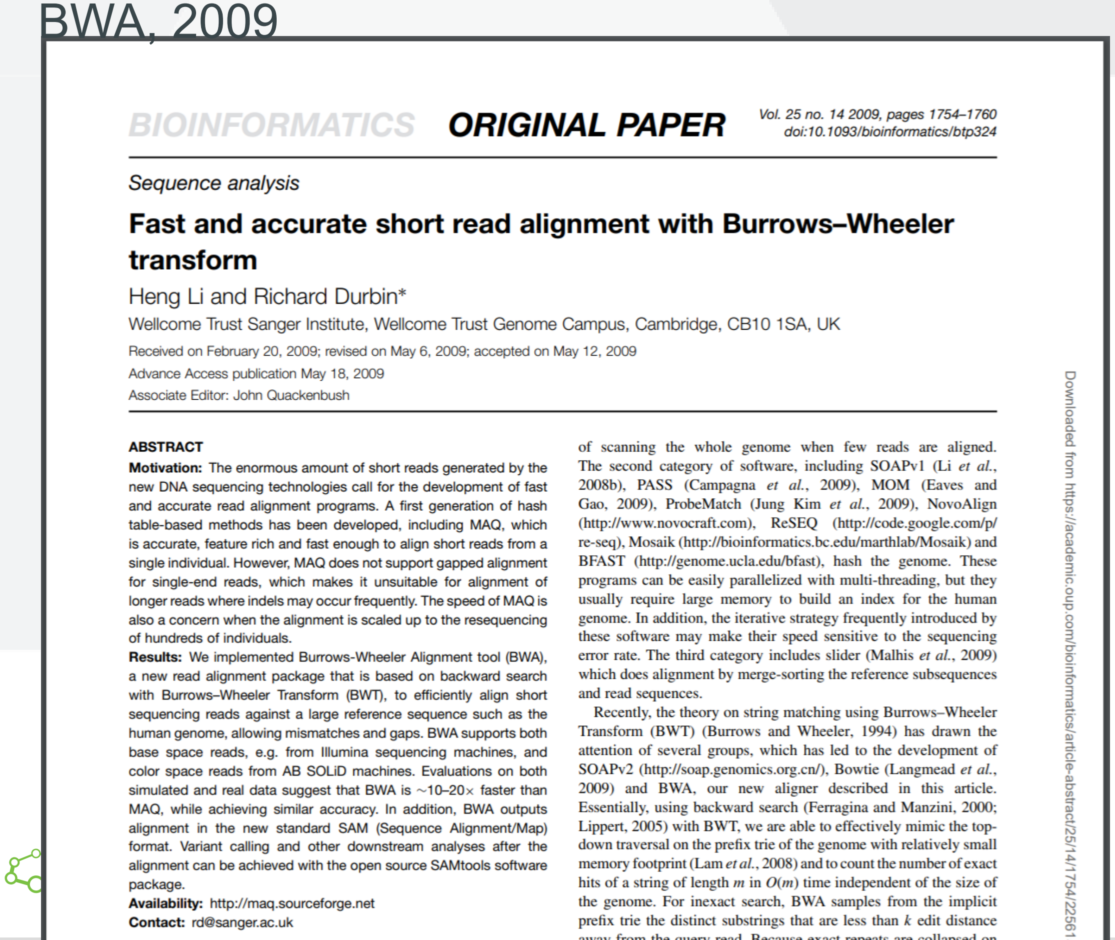
Multiple Alignment
Align sequences using domain and protein constraints

BioAssay
Search protein or nucleotide targets in PubChem BioAssay

MOLE-BLAST
Establish taxonomy for uncultured or environmental sequences

BWA - Burrows-Wheeler Aligner

BWA, 2009



BIOINFORMATICS **ORIGINAL PAPER**

Vol. 25 no. 14 2009, pages 1754–1760
doi:10.1093/bioinformatics/btp324

Sequence analysis

Fast and accurate short read alignment with Burrows–Wheeler transform

Heng Li and Richard Durbin*
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

Received on February 20, 2009; revised on May 6, 2009; accepted on May 12, 2009
Advance Access publication May 18, 2009
Associate Editor: John Quackenbush

ABSTRACT

Motivation: The enormous amount of short reads generated by the new DNA sequencing technologies call for the development of fast and accurate read alignment programs. A first generation of hash table-based methods has been developed, including MAQ, which is accurate, feature rich and fast enough to align short reads from a single individual. However, MAQ does not support gapped alignment for single-end reads, which makes it unsuitable for alignment of longer reads where indels may occur frequently. The speed of MAQ is also a concern when the alignment is scaled up to the resequencing of hundreds of individuals.

Results: We implemented Burrows-Wheeler Alignment tool (BWA), a new read alignment package that is based on backward search with Burrows–Wheeler Transform (BWT), to efficiently align short sequencing reads against a large reference sequence such as the human genome, allowing mismatches and gaps. BWA supports both base space reads, e.g. from Illumina sequencing machines, and color space reads from AB SOLiD machines. Evaluations on both simulated and real data suggest that BWA is ~10–20× faster than MAQ, while achieving similar accuracy. In addition, BWA outputs alignment in the new standard SAM (Sequence Alignment/Map) format. Variant calling and other downstream analyses after the alignment can be achieved with the open source SAMtools software package.

Availability: http://maq.sourceforge.net

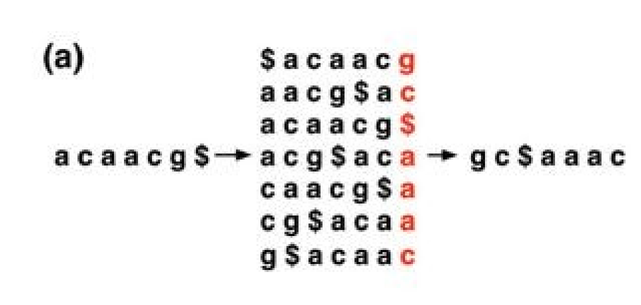
Contact: rd@sanger.ac.uk

of scanning the whole genome when few reads are aligned. The second category of software, including SOAPv1 (Li *et al.*, 2008b), PASS (Campagna *et al.*, 2009), MOM (Eaves and Gao, 2009), ProbeMatch (Jung Kim *et al.*, 2009), NovoAlign (http://www.novocraft.com), ReSEQ (http://code.google.com/p/re-seq), Mosaik (http://bioinformatics.bc.edu/marthlab/Mosaik) and BFAST (http://genome.ucla.edu/bfast), hash the genome. These programs can be easily parallelized with multi-threading, but they usually require large memory to build an index for the human genome. In addition, the iterative strategy frequently introduced by these software may make their speed sensitive to the sequencing error rate. The third category includes slider (Malhis *et al.*, 2009) which does alignment by merge-sorting the reference subsequences and read sequences.

Recently, the theory on string matching using Burrows–Wheeler Transform (BWT) (Burrows and Wheeler, 1994) has drawn the attention of several groups, which has led to the development of SOAPv2 (http://soap.genomics.org.cn/), Bowtie (Langmead *et al.*, 2009) and BWA, our new aligner described in this article. Essentially, using backward search (Ferragina and Manzini, 2000; Lippert, 2005) with BWT, we are able to effectively mimic the top-down traversal on the prefix trie of the genome with relatively small memory footprint (Lam *et al.*, 2008) and to count the number of exact hits of a string of length m in $O(m)$ time independent of the size of the genome. For inexact search, BWA samples from the implicit prefix trie the distinct substrings that are less than k edit distance away from the query read. Because exact repeats are collapsed on

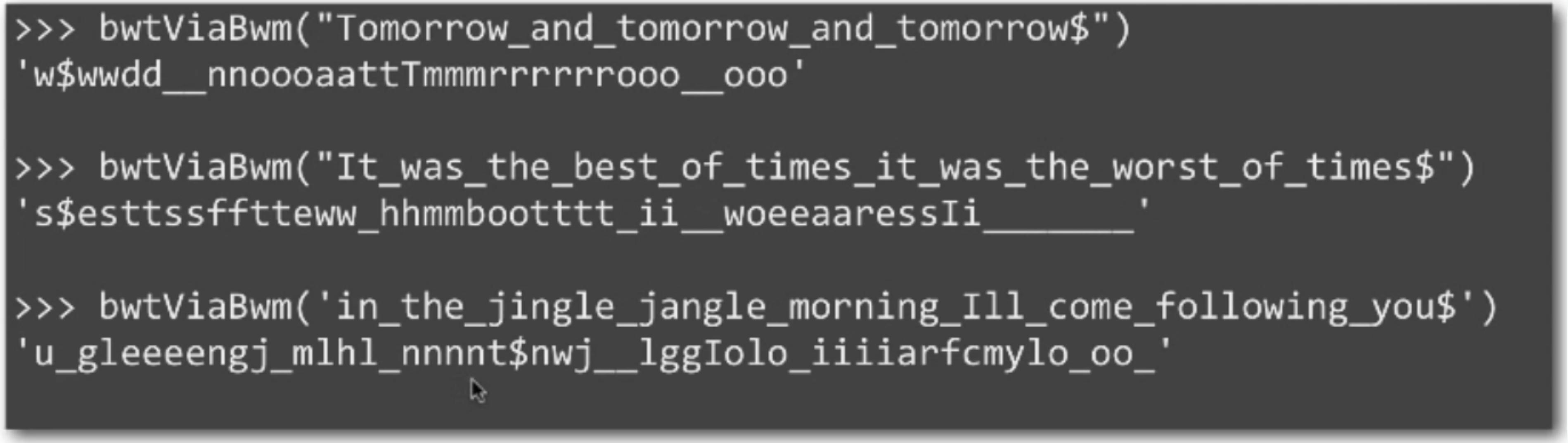
- BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome.
- Variants: short seq (up to 100bp), longer seq (70-1,000,000bp)
- Utilizes “Burrows Wheeler Transform” to make alignment faster

Burrows Wheeler Transform



a) Sort every permutation (sliding through) of the string and then take the last column = Transform

Essentially sorting by the character that comes AFTER this one (right-context)

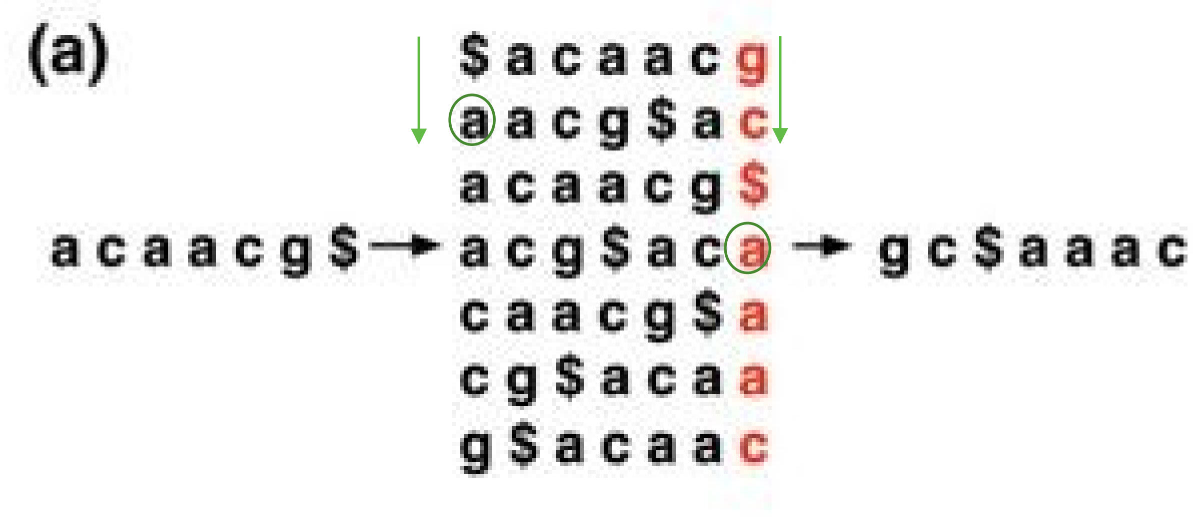


```
>>> bwtViaBwm("Tomorrow_and_tomorrow_and_tomorrow$")
'w$wwdd__nnoooaattTmmmrrrrrrooo__ooo'

>>> bwtViaBwm("It_was_the_best_of_times_it_was_the_worst_of_times$")
's$esttssfftteww_hhmmbootttt_ii__woeeaaressIi_______'

>>> bwtViaBwm('in_the_jingle_jangle_morning_Ill_come_following_you$')
'u_gleeeengj_mlhl_nnnnt$nwj__lggIolo_iiiiarfcmylo_oo_'
```

Burrows Wheeler Transform



Important Property:
LF (Last First)

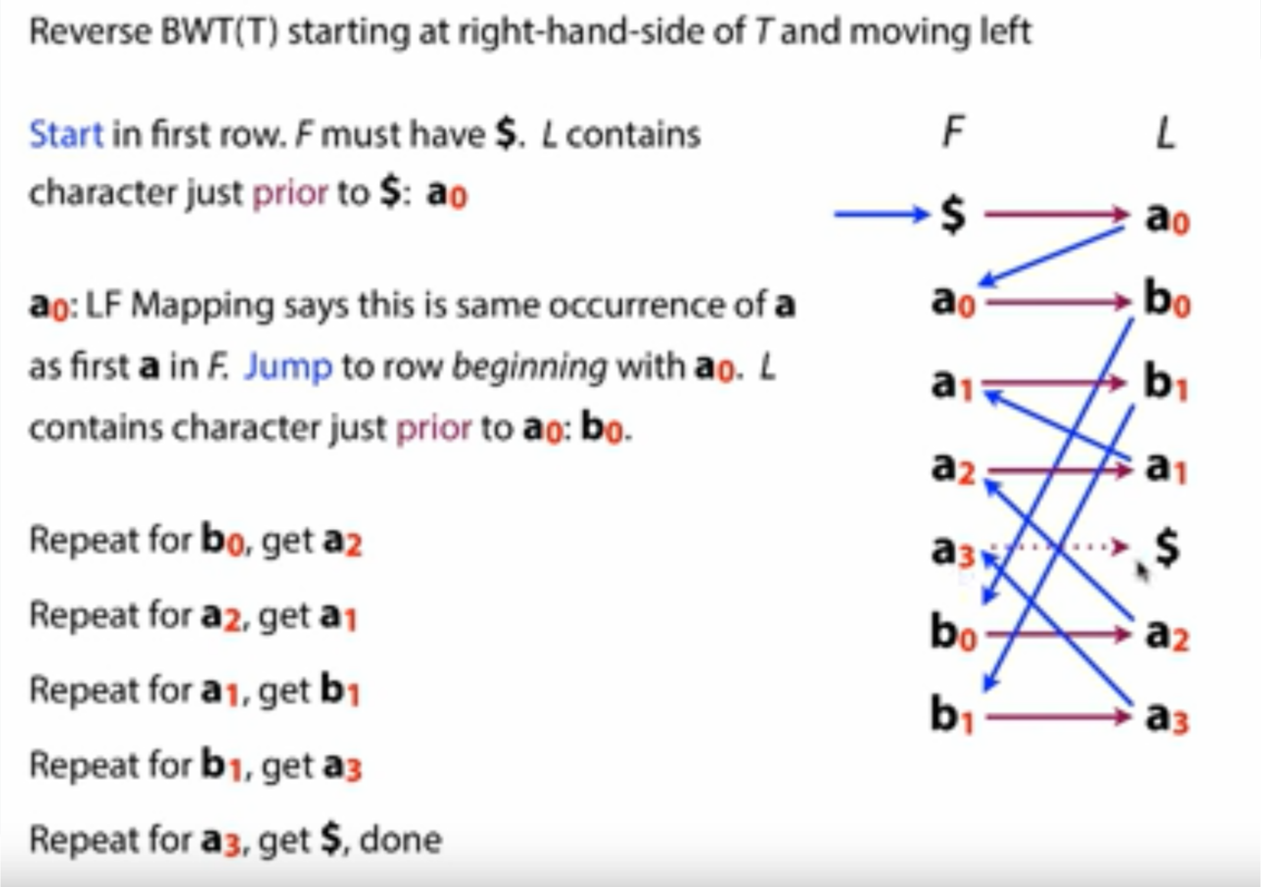
On the First and Last columns the order of same letters remains the same

Next Step:

Sort the table by order of letter AND index

$
A1
A2
A3
C1
C2
G1

Burrows Wheeler Transform - Reversing



Reverse BWT(T) starting at right-hand-side of *T* and moving left

Start in first row. *F* must have $\$$. *L* contains character just prior to $\$$: $\mathbf{a}_0$

$\mathbf{a}_0$: LF Mapping says this is same occurrence of **a** as first **a** in *F*. Jump to row *beginning* with $\mathbf{a}_0$. *L* contains character just prior to $\mathbf{a}_0$: $\mathbf{b}_0$.

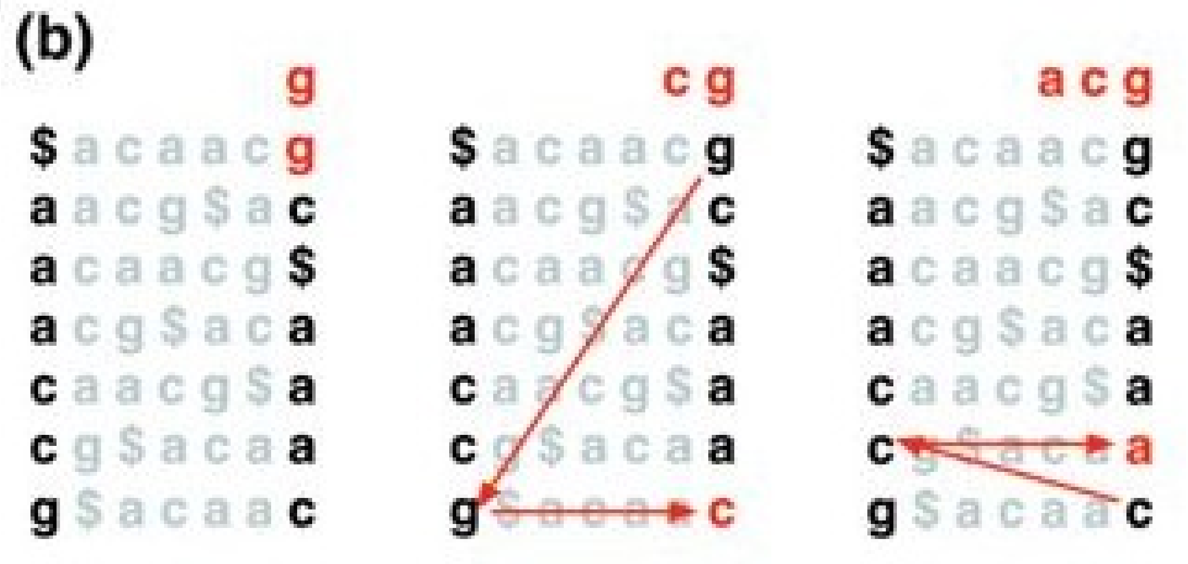
Repeat for $\mathbf{b}_0$, get $\mathbf{a}_2$

Repeat for $\mathbf{a}_2$, get $\mathbf{a}_1$

Repeat for $\mathbf{a}_1$, get $\mathbf{b}_1$

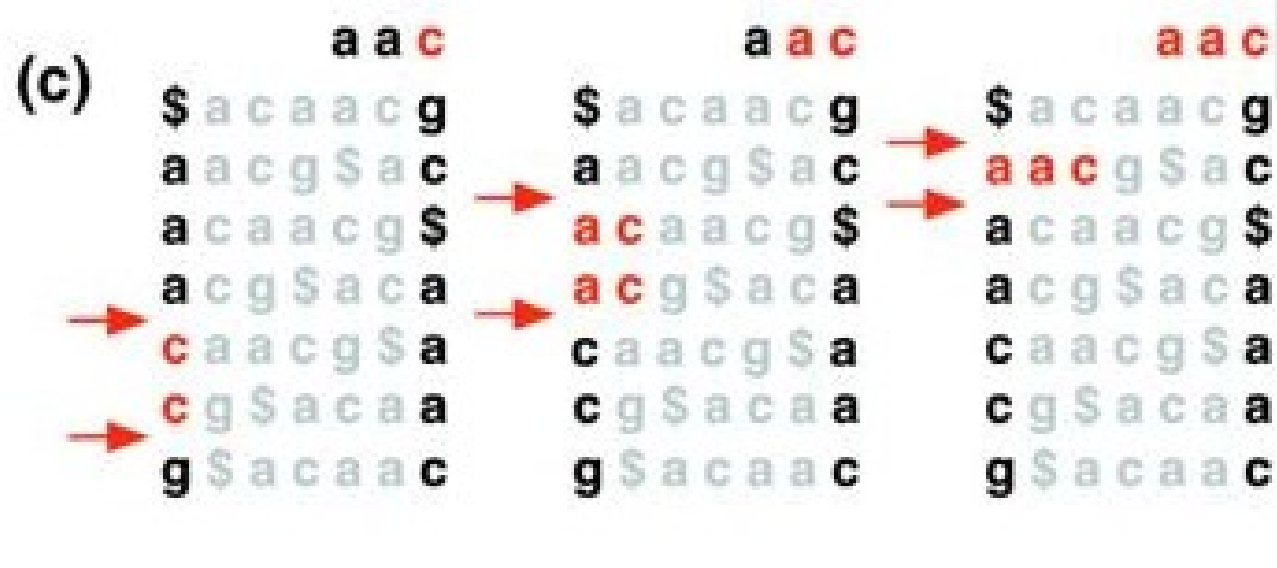
Repeat for $\mathbf{b}_1$, get $\mathbf{a}_3$

Repeat for $\mathbf{a}_3$, get $\$$, done



b) We can easily recreate the original sequence by working on a reverse order from any point in the Transform

Burrows Wheeler Transform



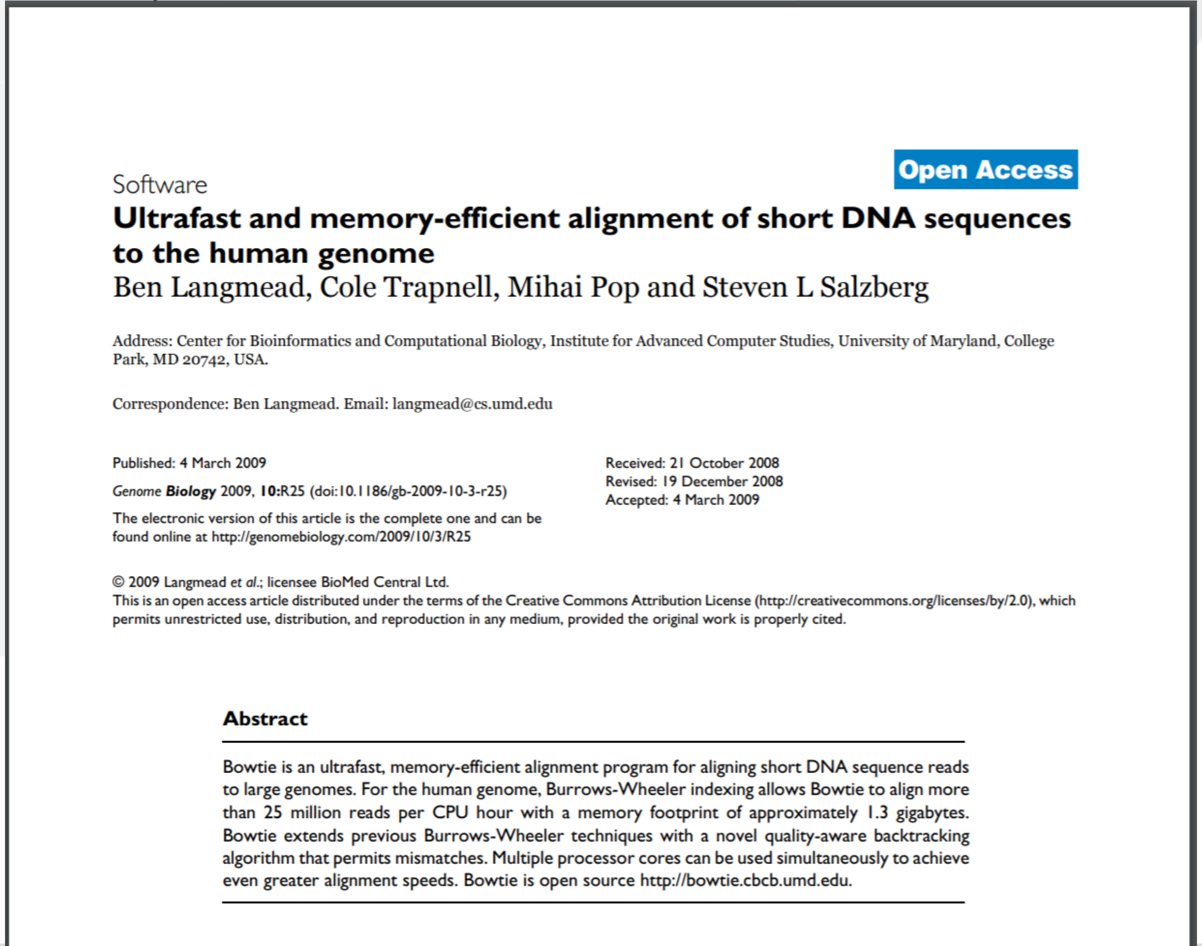
c) For every subsequence we can quickly find the location of all possible matches by working backwards in the transform

Using precalculation the complexity of the lookup can be reduced close to O(1)

Explanation of BWT for genomic indexing
by Ben Langmead (creator of bowtie)
https://www.youtube.com/watch?v=4n7NPk5lwbI
https://www.youtube.com/watch?v=kvVGj5V65io

Bowtie

Bowtie, 2009



Open Access

Software

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg

Address: Center for Bioinformatics and Computational Biology, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA.

Correspondence: Ben Langmead. Email: langmead@cs.umd.edu

Published: 4 March 2009

Genome **Biology** 2009, **10:**R25 (doi:10.1186/gb-2009-10-3-r25)

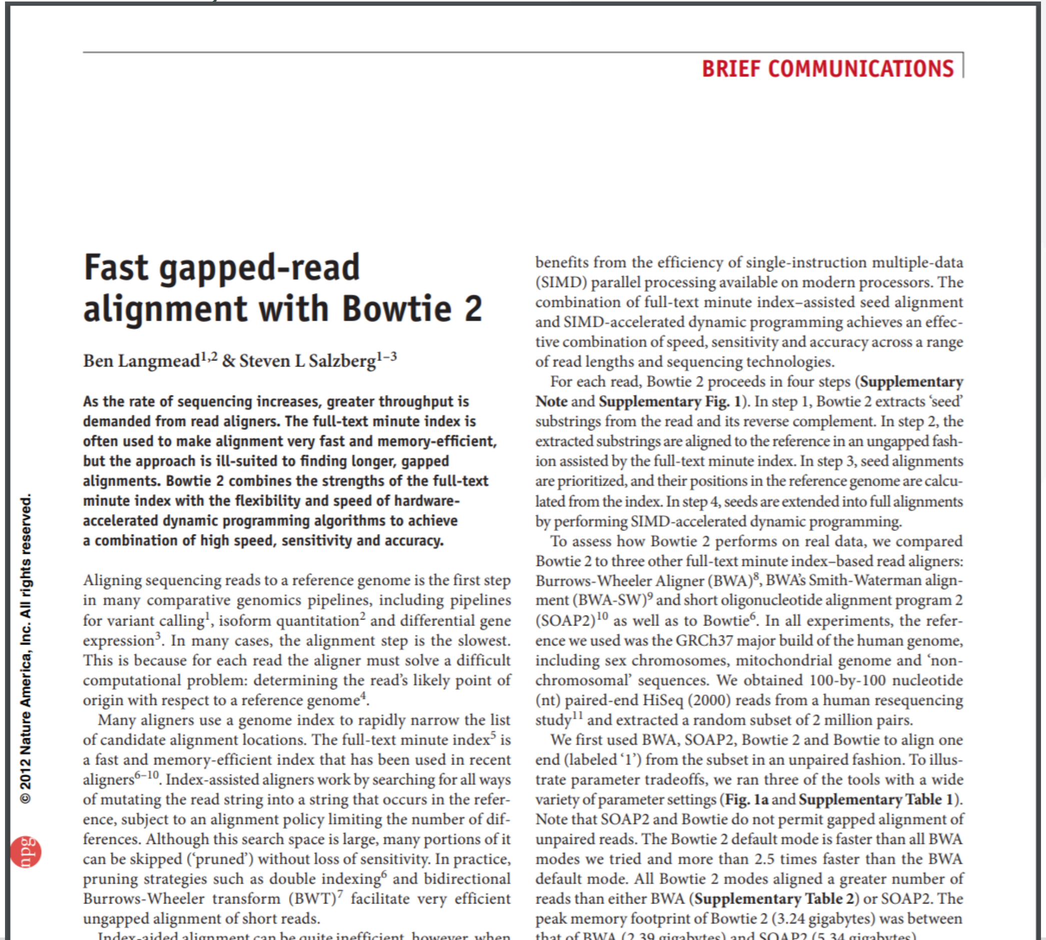
The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/3/R25

Received: 21 October 2008
Revised: 19 December 2008
Accepted: 4 March 2009

Abstract

Bowtie is an ultrafast, memory-efficient alignment program for aligning short DNA sequence reads to large genomes. For the human genome, Burrows-Wheeler indexing allows Bowtie to align more than 25 million reads per CPU hour with a memory footprint of approximately 1.3 gigabytes. Bowtie extends previous Burrows-Wheeler techniques with a novel quality-aware backtracking algorithm that permits mismatches. Multiple processor cores can be used simultaneously to achieve even greater alignment speeds. Bowtie is open source http://bowtie.cbcb.umd.edu.

Bowtie 2, 2012



BRIEF COMMUNICATIONS

Fast gapped-read alignment with Bowtie 2

Ben Langmead[1,2] & Steven L Salzberg[1–3]

As the rate of sequencing increases, greater throughput is demanded from read aligners. The full-text minute index is often used to make alignment very fast and memory-efficient, but the approach is ill-suited to finding longer, gapped alignments. Bowtie 2 combines the strengths of the full-text minute index with the flexibility and speed of hardware-accelerated dynamic programming algorithms to achieve a combination of high speed, sensitivity and accuracy.

Aligning sequencing reads to a reference genome is the first step in many comparative genomics pipelines, including pipelines for variant calling[1], isoform quantitation[2] and differential gene expression[3]. In many cases, the alignment step is the slowest. This is because for each read the aligner must solve a difficult computational problem: determining the read's likely point of origin with respect to a reference genome[4].

Many aligners use a genome index to rapidly narrow the list of candidate alignment locations. The full-text minute index[5] is a fast and memory-efficient index that has been used in recent aligners[6–10]. Index-assisted aligners work by searching for all ways of mutating the read string into a string that occurs in the reference, subject to an alignment policy limiting the number of differences. Although this search space is large, many portions of it can be skipped ('pruned') without loss of sensitivity. In practice, pruning strategies such as double indexing[6] and bidirectional Burrows-Wheeler transform (BWT)[7] facilitate very efficient ungapped alignment of short reads.

Index-aided alignment can be quite inefficient, however, when

benefits from the efficiency of single-instruction multiple-data (SIMD) parallel processing available on modern processors. The combination of full-text minute index–assisted seed alignment and SIMD-accelerated dynamic programming achieves an effective combination of speed, sensitivity and accuracy across a range of read lengths and sequencing technologies.

For each read, Bowtie 2 proceeds in four steps (**Supplementary Note** and **Supplementary Fig. 1**). In step 1, Bowtie 2 extracts 'seed' substrings from the read and its reverse complement. In step 2, the extracted substrings are aligned to the reference in an ungapped fashion assisted by the full-text minute index. In step 3, seed alignments are prioritized, and their positions in the reference genome are calculated from the index. In step 4, seeds are extended into full alignments by performing SIMD-accelerated dynamic programming.

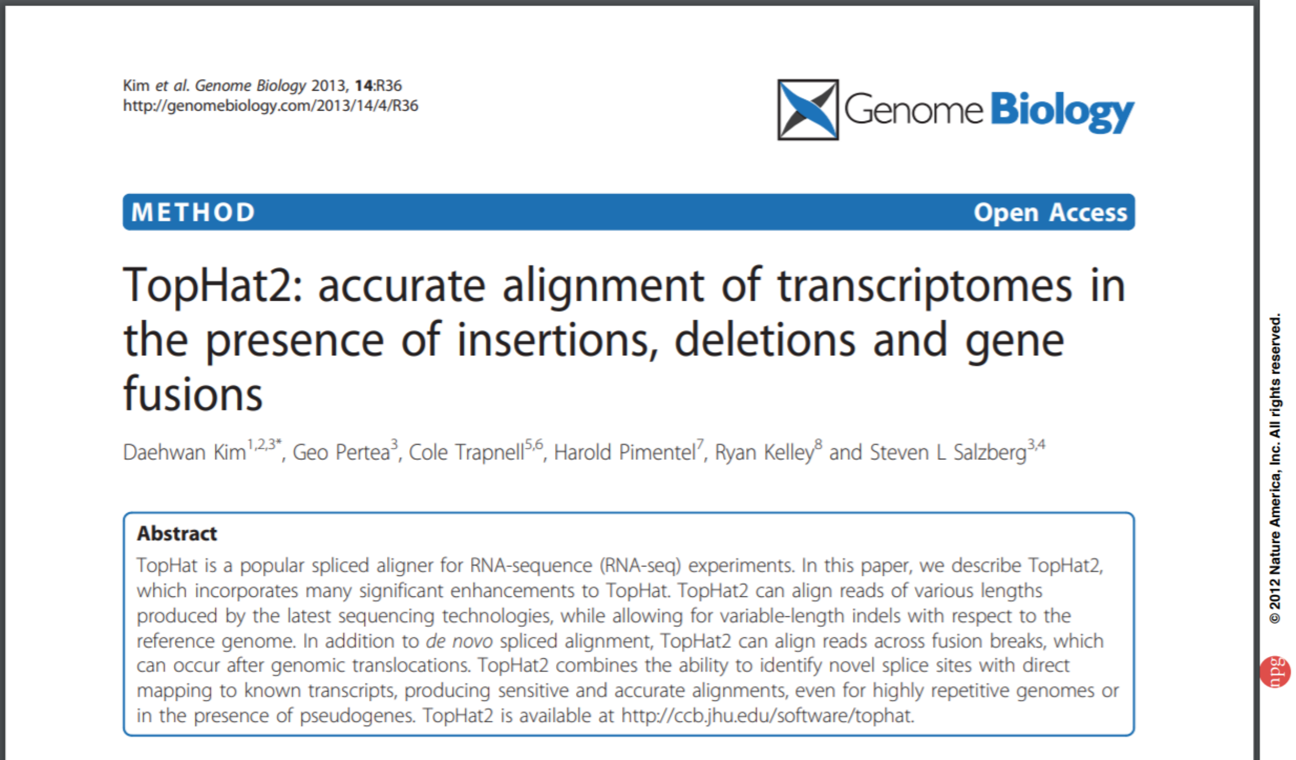
To assess how Bowtie 2 performs on real data, we compared Bowtie 2 to three other full-text minute index–based read aligners: Burrows-Wheeler Aligner (BWA)[8], BWA's Smith-Waterman alignment (BWA-SW)[9] and short oligonucleotide alignment program 2 (SOAP2)[10] as well as to Bowtie[6]. In all experiments, the reference we used was the GRCh37 major build of the human genome, including sex chromosomes, mitochondrial genome and 'nonchromosomal' sequences. We obtained 100-by-100 nucleotide (nt) paired-end HiSeq (2000) reads from a human resequencing study[11] and extracted a random subset of 2 million pairs.

We first used BWA, SOAP2, Bowtie 2 and Bowtie to align one end (labeled '1') from the subset in an unpaired fashion. To illustrate parameter tradeoffs, we ran three of the tools with a wide variety of parameter settings (**Fig. 1a** and **Supplementary Table 1**). Note that SOAP2 and Bowtie do not permit gapped alignment of unpaired reads. The Bowtie 2 default mode is faster than all BWA modes we tried and more than 2.5 times faster than the BWA default mode. All Bowtie 2 modes aligned a greater number of reads than either BWA (**Supplementary Table 2**) or SOAP2. The peak memory footprint of Bowtie 2 (3.24 gigabytes) was between that of BWA (2.39 gigabytes) and SOAP2 (5.34 gigabytes).

TopHat2

TopHat2 works well with gaps + introns

TopHat2, 2013



Kim *et al. Genome Biology* 2013, **14**:R36
http://genomebiology.com/2013/14/4/R36

Genome **Biology**

METHOD Open Access

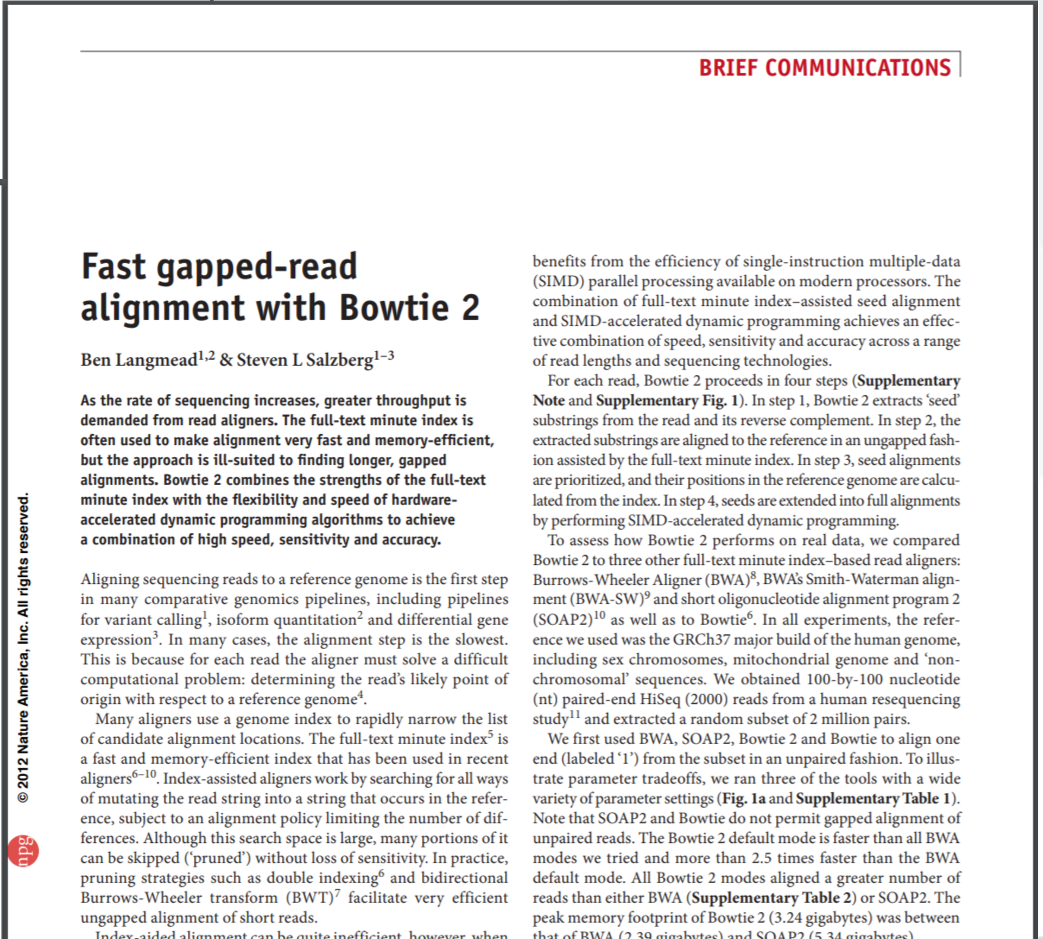
TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions

Daehwan Kim[1,2,3*], Geo Pertea[3], Cole Trapnell[5,6], Harold Pimentel[7], Ryan Kelley[8] and Steven L Salzberg[3,4]

Abstract

TopHat is a popular spliced aligner for RNA-sequence (RNA-seq) experiments. In this paper, we describe TopHat2, which incorporates many significant enhancements to TopHat. TopHat2 can align reads of various lengths produced by the latest sequencing technologies, while allowing for variable-length indels with respect to the reference genome. In addition to *de novo* spliced alignment, TopHat2 can align reads across fusion breaks, which can occur after genomic translocations. TopHat2 combines the ability to identify novel splice sites with direct mapping to known transcripts, producing sensitive and accurate alignments, even for highly repetitive genomes or in the presence of pseudogenes. TopHat2 is available at http://ccb.jhu.edu/software/tophat.

Bowtie 2, 2012



BRIEF COMMUNICATIONS

Fast gapped-read alignment with Bowtie 2

Ben Langmead[1,2] & Steven L Salzberg[1–3]

As the rate of sequencing increases, greater throughput is demanded from read aligners. The full-text minute index is often used to make alignment very fast and memory-efficient, but the approach is ill-suited to finding longer, gapped alignments. Bowtie 2 combines the strengths of the full-text minute index with the flexibility and speed of hardware-accelerated dynamic programming algorithms to achieve a combination of high speed, sensitivity and accuracy.

Aligning sequencing reads to a reference genome is the first step in many comparative genomics pipelines, including pipelines for variant calling[1], isoform quantitation[2] and differential gene expression[3]. In many cases, the alignment step is the slowest. This is because for each read the aligner must solve a difficult computational problem: determining the read's likely point of origin with respect to a reference genome[4].

Many aligners use a genome index to rapidly narrow the list of candidate alignment locations. The full-text minute index[5] is a fast and memory-efficient index that has been used in recent aligners[6–10]. Index-assisted aligners work by searching for all ways of mutating the read string into a string that occurs in the reference, subject to an alignment policy limiting the number of differences. Although this search space is large, many portions of it can be skipped ('pruned') without loss of sensitivity. In practice, pruning strategies such as double indexing[6] and bidirectional Burrows-Wheeler transform (BWT)[7] facilitate very efficient ungapped alignment of short reads.

Index-aided alignment can be quite inefficient, however, when

benefits from the efficiency of single-instruction multiple-data (SIMD) parallel processing available on modern processors. The combination of full-text minute index–assisted seed alignment and SIMD-accelerated dynamic programming achieves an effective combination of speed, sensitivity and accuracy across a range of read lengths and sequencing technologies.

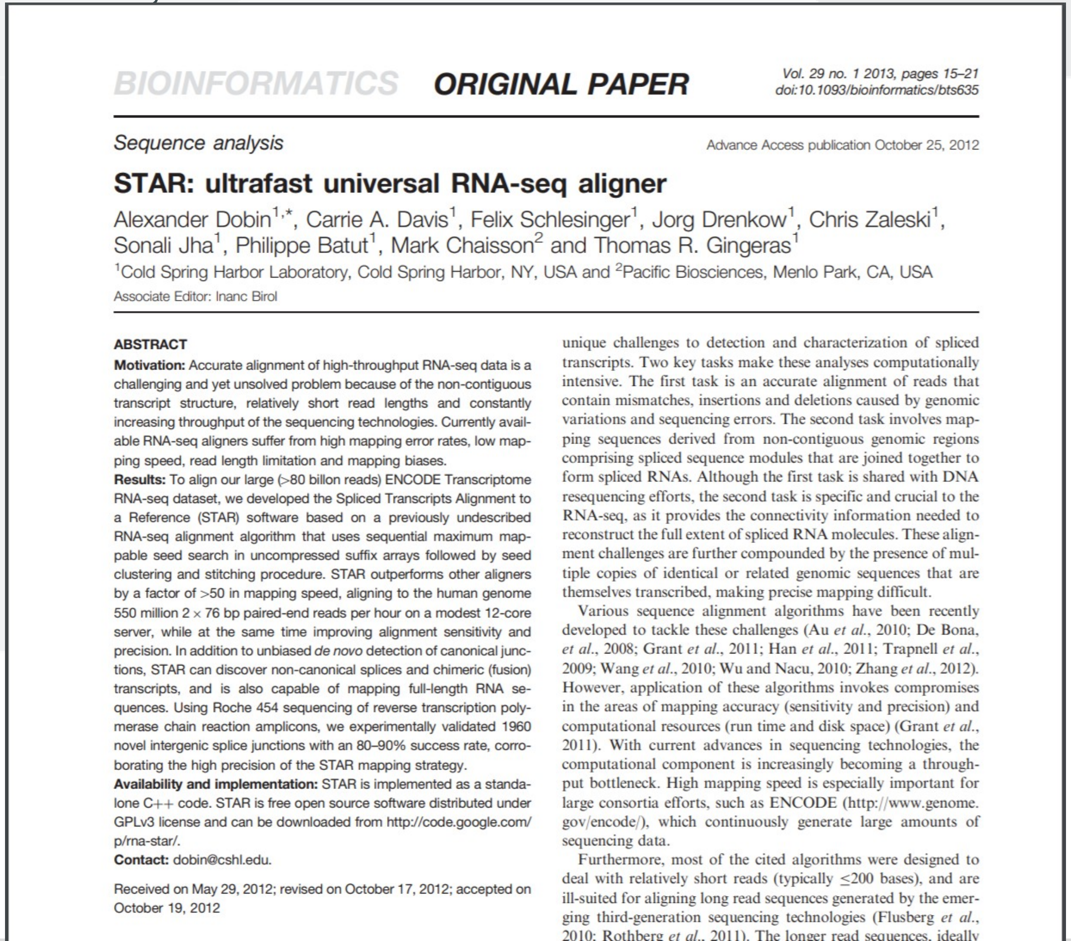
For each read, Bowtie 2 proceeds in four steps (**Supplementary Note** and **Supplementary Fig. 1**). In step 1, Bowtie 2 extracts 'seed' substrings from the read and its reverse complement. In step 2, the extracted substrings are aligned to the reference in an ungapped fashion assisted by the full-text minute index. In step 3, seed alignments are prioritized, and their positions in the reference genome are calculated from the index. In step 4, seeds are extended into full alignments by performing SIMD-accelerated dynamic programming.

To assess how Bowtie 2 performs on real data, we compared Bowtie 2 to three other full-text minute index–based read aligners: Burrows-Wheeler Aligner (BWA)[8], BWA's Smith-Waterman alignment (BWA-SW)[9] and short oligonucleotide alignment program 2 (SOAP2)[10] as well as to Bowtie[6]. In all experiments, the reference we used was the GRCh37 major build of the human genome, including sex chromosomes, mitochondrial genome and 'nonchromosomal' sequences. We obtained 100-by-100 nucleotide (nt) paired-end HiSeq (2000) reads from a human resequencing study[11] and extracted a random subset of 2 million pairs.

We first used BWA, SOAP2, Bowtie 2 and Bowtie to align one end (labeled '1') from the subset in an unpaired fashion. To illustrate parameter tradeoffs, we ran three of the tools with a wide variety of parameter settings (**Fig. 1a** and **Supplementary Table 1**). Note that SOAP2 and Bowtie do not permit gapped alignment of unpaired reads. The Bowtie 2 default mode is faster than all BWA modes we tried and more than 2.5 times faster than the BWA default mode. All Bowtie 2 modes aligned a greater number of reads than either BWA (**Supplementary Table 2**) or SOAP2. The peak memory footprint of Bowtie 2 (3.24 gigabytes) was between that of BWA (2.39 gigabytes) and SOAP2 (5.34 gigabytes).

STAR

STAR, 2013



BIOINFORMATICS **ORIGINAL PAPER**

Vol. 29 no. 1 2013, pages 15–21
doi:10.1093/bioinformatics/bts635

Sequence analysis

Advance Access publication October 25, 2012

STAR: ultrafast universal RNA-seq aligner

Alexander Dobin[1,*], Carrie A. Davis[1], Felix Schlesinger[1], Jorg Drenkow[1], Chris Zaleski[1], Sonali Jha[1], Philippe Batut[1], Mark Chaisson[2] and Thomas R. Gingeras[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA and [2]Pacific Biosciences, Menlo Park, CA, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Accurate alignment of high-throughput RNA-seq data is a challenging and yet unsolved problem because of the non-contiguous transcript structure, relatively short read lengths and constantly increasing throughput of the sequencing technologies. Currently available RNA-seq aligners suffer from high mapping error rates, low mapping speed, read length limitation and mapping biases.

Results: To align our large (>80 billon reads) ENCODE Transcriptome RNA-seq dataset, we developed the Spliced Transcripts Alignment to a Reference (STAR) software based on a previously undescribed RNA-seq alignment algorithm that uses sequential maximum mappable seed search in uncompressed suffix arrays followed by seed clustering and stitching procedure. STAR outperforms other aligners by a factor of >50 in mapping speed, aligning to the human genome 550 million 2 × 76 bp paired-end reads per hour on a modest 12-core server, while at the same time improving alignment sensitivity and precision. In addition to unbiased *de novo* detection of canonical junctions, STAR can discover non-canonical splices and chimeric (fusion) transcripts, and is also capable of mapping full-length RNA sequences. Using Roche 454 sequencing of reverse transcription polymerase chain reaction amplicons, we experimentally validated 1960 novel intergenic splice junctions with an 80–90% success rate, corroborating the high precision of the STAR mapping strategy.

Availability and implementation: STAR is implemented as a standalone C++ code. STAR is free open source software distributed under GPLv3 license and can be downloaded from http://code.google.com/p/rna-star/.

Contact: dobin@cshl.edu.

Received on May 29, 2012; revised on October 17, 2012; accepted on October 19, 2012

unique challenges to detection and characterization of spliced transcripts. Two key tasks make these analyses computationally intensive. The first task is an accurate alignment of reads that contain mismatches, insertions and deletions caused by genomic variations and sequencing errors. The second task involves mapping sequences derived from non-contiguous genomic regions comprising spliced sequence modules that are joined together to form spliced RNAs. Although the first task is shared with DNA resequencing efforts, the second task is specific and crucial to the RNA-seq, as it provides the connectivity information needed to reconstruct the full extent of spliced RNA molecules. These alignment challenges are further compounded by the presence of multiple copies of identical or related genomic sequences that are themselves transcribed, making precise mapping difficult.

Various sequence alignment algorithms have been recently developed to tackle these challenges (Au *et al.*, 2010; De Bona, *et al.*, 2008; Grant *et al.*, 2011; Han *et al.*, 2011; Trapnell *et al.*, 2009; Wang *et al.*, 2010; Wu and Nacu, 2010; Zhang *et al.*, 2012). However, application of these algorithms invokes compromises in the areas of mapping accuracy (sensitivity and precision) and computational resources (run time and disk space) (Grant *et al.*, 2011). With current advances in sequencing technologies, the computational component is increasingly becoming a throughput bottleneck. High mapping speed is especially important for large consortia efforts, such as ENCODE (http://www.genome.gov/encode/), which continuously generate large amounts of sequencing data.

Furthermore, most of the cited algorithms were designed to deal with relatively short reads (typically ≤200 bases), and are ill-suited for aligning long read sequences generated by the emerging third-generation sequencing technologies (Flusberg *et al.*, 2010; Rothberg *et al.*, 2011). The longer read sequences, ideally

- STAR works well with gaps + introns
- Preferable for mapping on RNA
- Faster than TopHat2

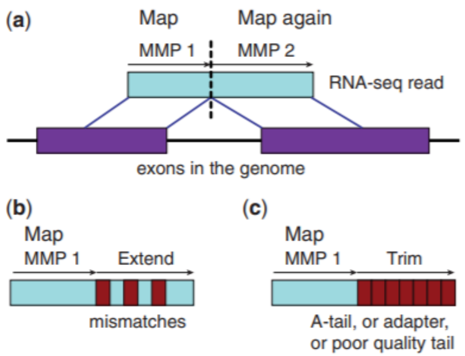
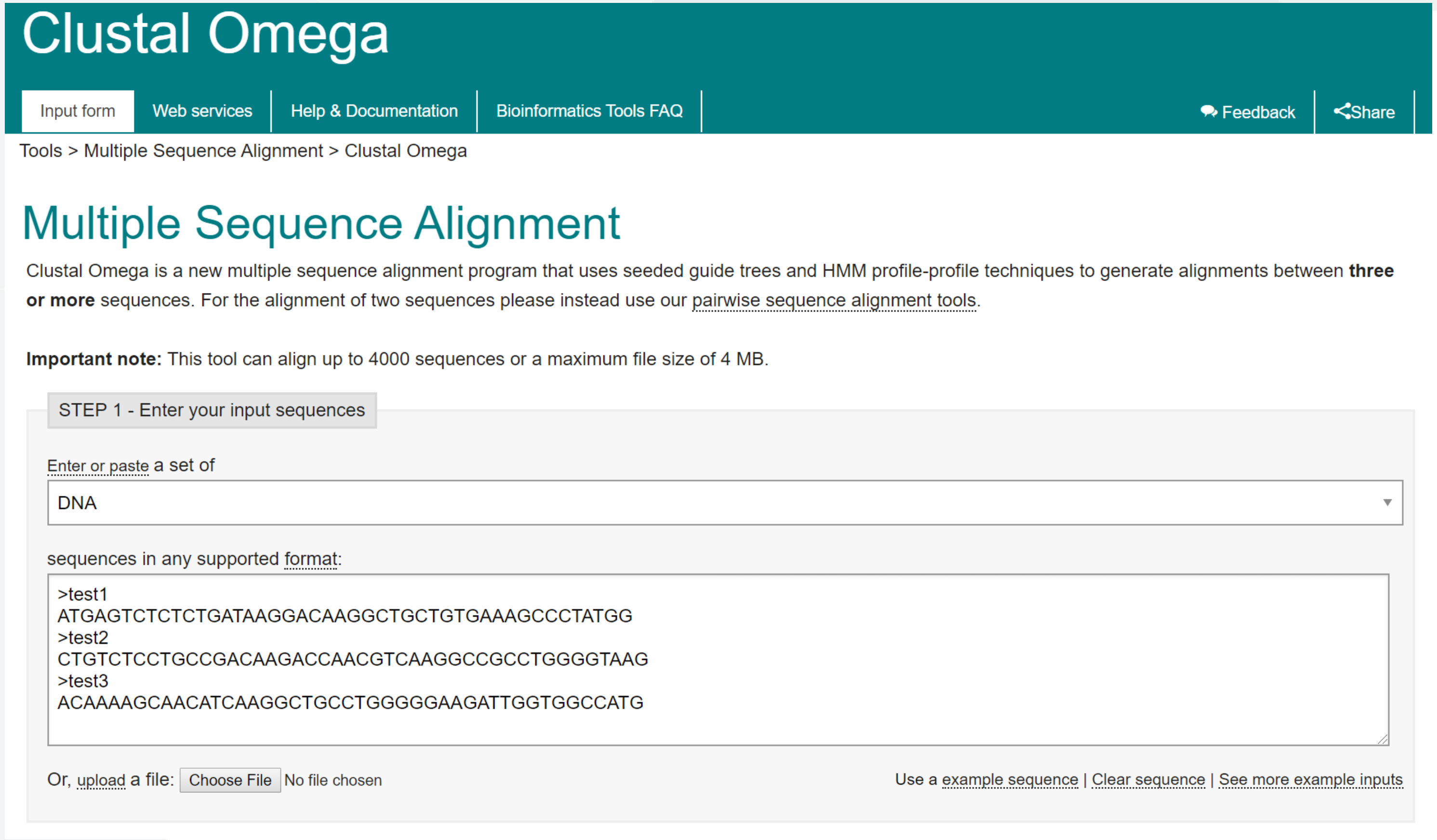


Fig. 1. Schematic representation of the Maximum Mappable Prefix search in the STAR algorithm for detecting **(a)** splice junctions, **(b)** mismatches and **(c)** tails

Clustal – Omega (Multiple Sequence Alignment)



Tools > Multiple Sequence Alignment > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

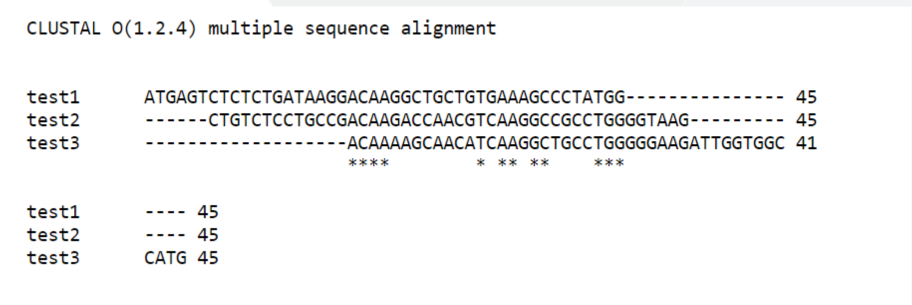
DNA

sequences in any supported format:

```
>test1
ATGAGTCTCTCTGATAAGGACAAGGCTGCTGTGAAAGCCCTATGG
>test2
CTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAG
>test3
ACAAAAGCAACATCAAGGCTGCCTGGGGGAAGATTGGTGGCCATG
```

Or, upload a file: Choose File No file chosen

Use a example sequence | Clear sequence | See more example inputs



```
CLUSTAL O(1.2.4) multiple sequence alignment


test1      ATGAGTCTCTCTGATAAGGACAAGGCTGCTGTGAAAGCCCTATGG--------------- 45
test2      ------CTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAG--------- 45
test3      -------------------ACAAAAGCAACATCAAGGCTGCCTGGGGGAAGATTGGTGGC 41
                              ****       * ** **    ***

test1      ---- 45
test2      ---- 45
test3      CATG 45
```

Exercise

- Global Alignment: Use Alignment Matrix

- Global Alignment: Align Position Weight Matrices

https://bit.ly/3427HeE

CEITEC

@CEITEC_Brno

Thank you for your attention!
60 minutes lunch break.