

An experiment on anti-academic research

Guillaume Filion
CRG (Barcelona)

What is the question?



A story from my blog...

The Grand Locus / *Life for statistical sciences*

June babies and bioinformatics

By Guillaume Filion, filed under extreme value fallacy, Gumbel, extreme value theory, bioinformatics.

• 06 April 2014 •



In July 1982, paleontologist **Steven Jay Gould** was diagnosed with cancer. Facing a median prognosis of only 8 months survival, he used his knowledge of statistics to prepare for the future. As he explains in **The Median Isn't the Message**, if half of the patients died of this rare case of **mesothelioma** within 8 months, those who did not had much better survival. Evaluating his own chances of being in the "survivor" group as high, he planned for long term survival and opted out of the standard treatment. He died 20 years later, from an unrelated disease.

If not the **median**, then what is the message? Statistics put a disproportionate emphasis on the typical or average behavior, when what matters is sometimes in the extremes. This general blindness to the extremes is responsible for a dreadful lot of confusion in the bio-medical field. One of my all time favorite traps is the **extreme value fallacy**. Nothing better than an example will explain what it is about.

... and the follow up

The Grand Locus / *Life for statistical sciences*

At that point, it took just a little push to assemble the pieces of the puzzle. Here is how the story continues.

“ *One day while doing the dishes it sort of occurred to me that what Karlin’s results would allow you to do was to put the seed-based heuristic methods on a sound statistical framework [...]. That way you could use a rapid heuristic but estimate with confidence your chances of missing something. The BLAST idea of generating all short kmers with score of $\geq s$ came quite quickly and the basic algorithm was set. Initially I’d been thinking of pre-indexing the database but Webb Miller – who’d agreed to program this up – pointed out that for the similarity levels we were searching for, pre-indexing wouldn’t be efficient. Within a week of explaining the idea he’d sent me a program which performed well. Stephen joined us to work out the application of Karlin’s statistical approach to the matches, Gene worked on efficient gapped extensions (we didn’t have good statistics for those at the time so that wasn’t part of the original program), and Warren Gish proposed a [Finite State Machine] for more efficiently finding the word matches.*

The people that David Lipman mentions are the authors of the original BLAST paper: Stephen Altschul, Warren Gish, Webb Miller and Gene Myers (see reference [2]).

That way you could use a rapid heuristic but estimate with confidence your chances of missing something.

Introducing mappers...



...and mapping quality


```
@SQ      SN:chr1 LN:249250621
@SQ      SN:chr2 LN:243199373
@SQ      SN:chr3 LN:198022430
@PG      ID:bwa  PN:bwa  VN:0.7.9a-r786  CL:bwa mem -t4 -k17 -B3 -r1.4 -T20 /mnt/shared/seq/bwa/GRCh
SRR037840.1  0      chr19  52745045      48      25M      *      0      0      GATATTGGTGC
SRR037840.2  16     chr4   148677601     48      25M      *      0      0      AGCACTGTTAG
SRR037840.3  0      chr7   152605924     24      25M      *      0      0      TAGGAATATAC
SRR037840.5  16     chr13  69336181      24      25M      *      0      0      GGAAACTGTGT
SRR037840.6  16     chr6   147663686     28      25M      *      0      0      GCTGTGTATAC
SRR037840.9  16     chr5   129100767     30      25M      *      0      0      ACAATCAGAGC
SRR037840.12 0      chr5   159081219     27      25M      *      0      0      AAATTAGAGGA
SRR037840.13 0      chr5   159081219     27      25M      *      0      0      AAATTAGAGGA
SRR037840.16 0      chr2   99889540      0       25M      *      0      0      AAAGGTGTCCA
```

All the mappers are heuristics, so they do not always get it right. The fifth column of the .sam format is mapping quality. It gives the **probability that the location is wrong**.

MAPQ 10: 1 error per 10 reads
MAPQ 20: 1 error per 100 reads
MAPQ 30: 1 error per 1000 reads

...

Quite unpopular score...



THOUGHTS ON BIOLOGY, GENOMICS, AND THE ONGOING THREAT TO HUMANITY
FROM THE BOGUS USE OF BIOINFORMATICS ACRONYMS. BY KEITH BRADNAM

ABOUT BLOG CONTACT

Understanding MAPQ scores in SAM files: does 37 = 42?

December 18, 2014

The official specification for
stored in each column of this
file stores MAPping Quality I


MAPQ: MAPPING Quality
rounded to the nearest
not available.

Biofinysics

Wednesday, May 21, 2014

How does bowtie2 assign MAPQ scores?

by *Unknown*



Outline:

- I. Introduction
- II. A little bit to know about bowtie2
 - A. How bowtie2 scores a mismatch
 - B. How bowtie2 decides if an alignment is valid
 - C. How bowtie2 describes its alignments
- III. Experiments to look at how BT2 assigns MAPQ scores
 - A. Methods
 - B. Experiments, Results, and Conclusions
- IV. An English translation of how bowtie2 assigns MAPQs from the code and experiments
- V. Python code for calculating bt2-style MAPQs

Introduction:
Bowtie2 is an ultra-fast program for aligning next generation sequence reads to large genomes written by Ben Langmead and which happens to be my aligner of choice. RTFM [here](#). Download it [here](#) or learn how to get it on your Mac OS X with "homebrew" from one of my previous posts, "The Craft of Homebrew-ing for Mac OS".



THOUGHTS ON BIOLOGY, GENOMICS, AND THE ONGOING THREAT TO HUMANITY
FROM THE BOGUS USE OF BIOINFORMATICS ACRONYMS. BY KEITH BRADNAM

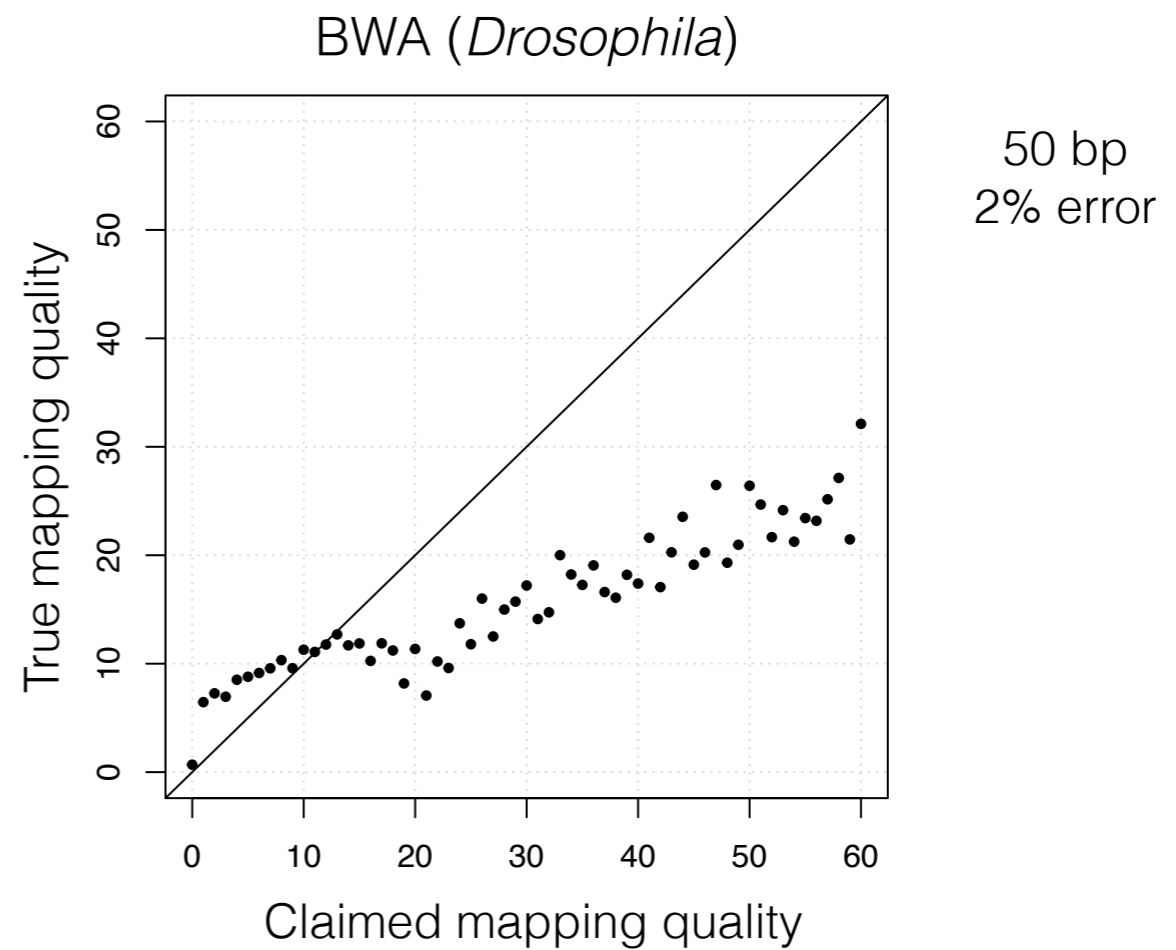
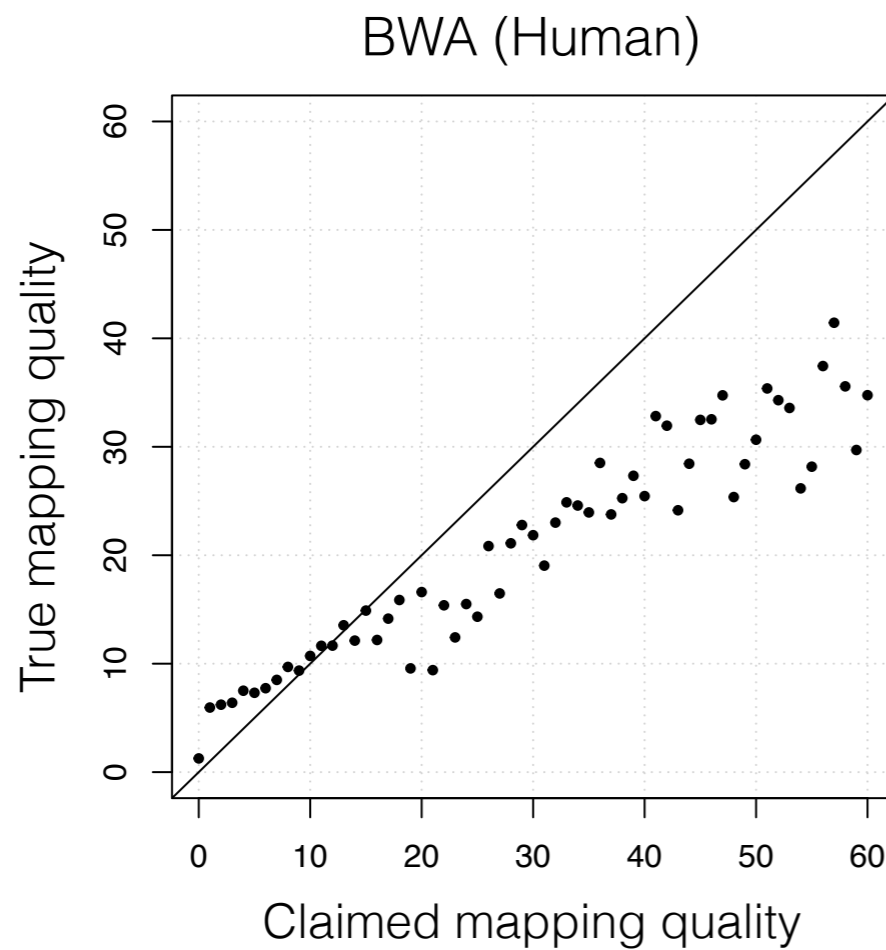
ABOUT BLOG CONTACT

More madness with MAPQ scores (a.k.a. why bioinformaticians hate poor and incomplete software documentation)

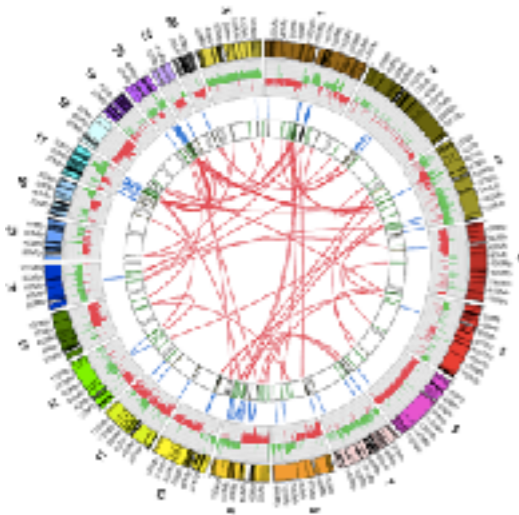
March 17, 2015



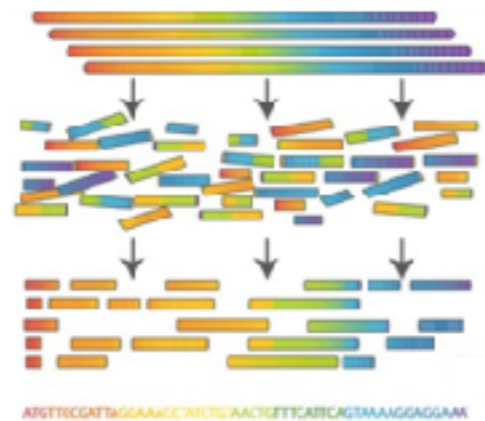
... and quite unreliable



Why is this important?



Cancer sequencing
(mutation calling)



De novo assembly



Contaminated
samples



"Universal"
sequencing

“ One day while doing the dishes it sort of occurred to me that what Karlin’s results would allow you to do was to put the seed-based heuristic methods on a sound statistical framework [...]. That way you could use a rapid heuristic but estimate with confidence your chances of missing something.

David Lipman about BLAST

BWA spends 6x more time on reads with mapping quality 0. This time is wasted.

What is the problem?



What is the chance that we have at least 17 correct nucleotides in a row (a seed)?

Unmapped read

What is the chance that all seeds are wrong?

Wrongly mapped read

How information flows

coursera

Explore Catalog ▾ Degrees ▾ Certificates ▾ | For Enterprise

What do you want to learn?



LOG IN

Join for Free



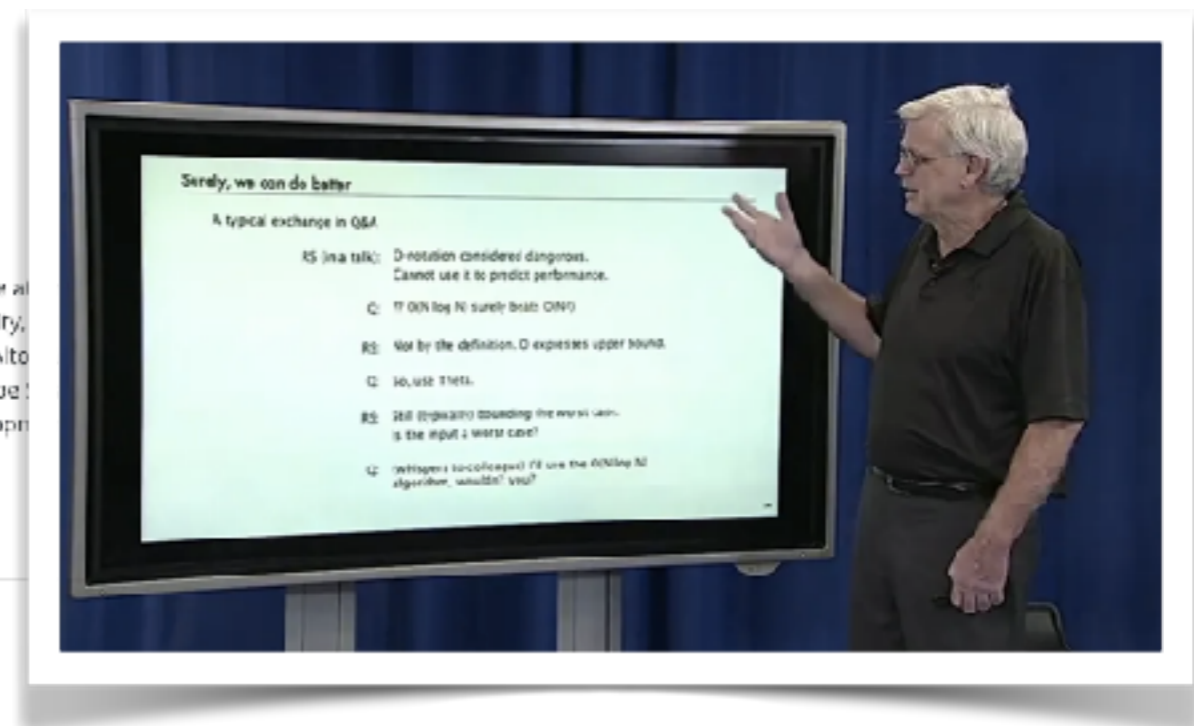
Robert Sedgwick

William O. Baker *29 Professor of Computer Science

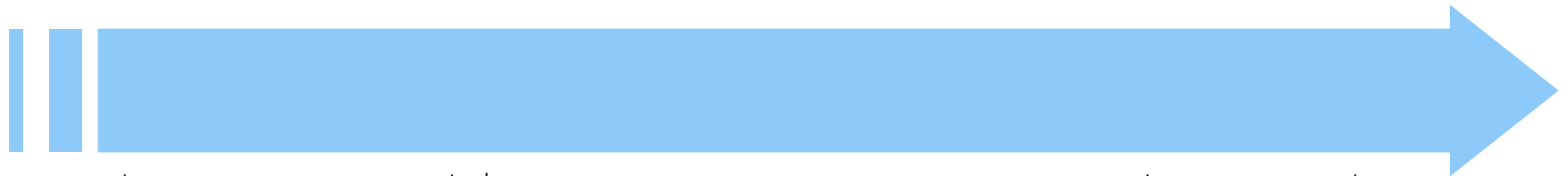
[Princeton University](#)

Bio

Robert Sedgwick is the William O. Baker Professor of Computer Science at Princeton University. He received the Ph.D. degree from Stanford University, Palo Alto, California, and has held visiting research positions at Xerox PARC, Palo Alto, California, and INRIA, Rocquencourt, France. He is a member of the board of directors of Adobe Systems, Inc. His research interests are in algorithm design, the scientific analysis of algorithms, curriculum development, and the design of algorithms. He has published widely in these areas and is the author of several books.



Analytic combinatorics



Donald Knuth
analyses an
algorithm
(1963)



Robert Sedgewick
defends PhD
(1975)



Sedgewick
meets Flajolet
(1977)



The
textbook
(2009)



Flajolet dies
(2011)

Philippe Flajolet , Robert Sedgewick, Analytic Combinatorics ...

<https://dl.acm.org/citation.cfm?id=1506267>

by P Flajolet - 2009 - Cited by 3087 - Related articles

Analytic Combinatorics is a self-contained treatment of the mathematics underlying the analysis of discrete structures, which has emerged over the past several ...

The scientific process



Problem

Solution

Our world



Translation

Analysis

The "other"
world

Projections
Fourier space
Simulations
Model organism
etc.

Analytic combinatorics



Problem



**Generating
functions**

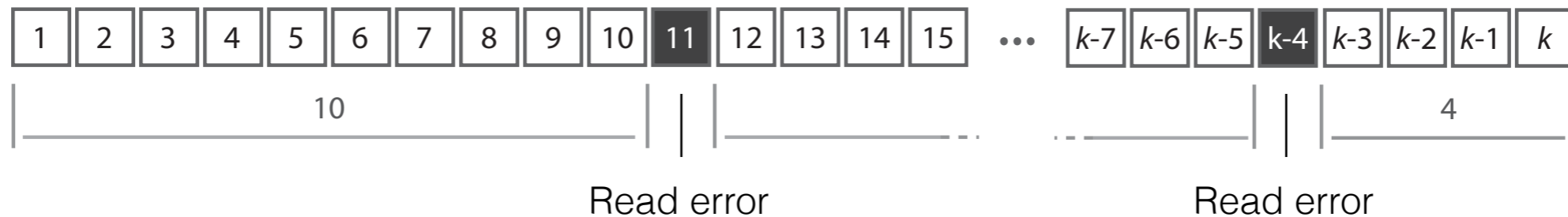
Combinations

Singularity
analysis

Solution

A construction game

What is the chance that a read has a seed of size 17?



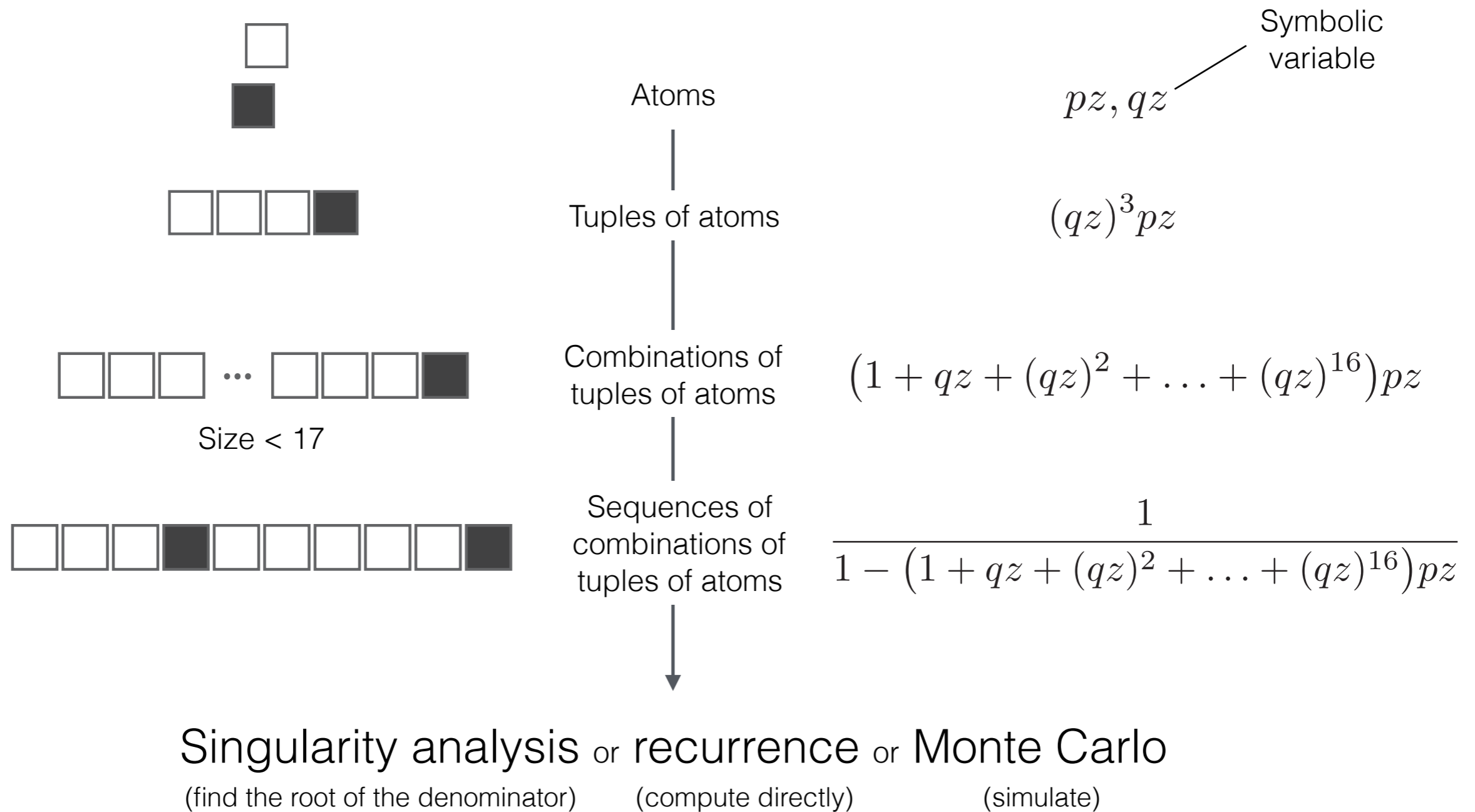
Option #1

$$\begin{aligned}
 \langle n \rangle &= \frac{1}{2\kappa} \left[\frac{\gamma_{0 \rightarrow 2} - \gamma_{1 \rightarrow 2} + 2\gamma_{2 \rightarrow 0} - 2\gamma_{2 \rightarrow 1}}{\gamma_{0 \rightarrow 2} + \gamma_{1 \rightarrow 2} + 2\gamma_{2 \rightarrow 0} + 2\gamma_{2 \rightarrow 1}} \left\{ \left(\frac{\gamma_{0 \rightarrow 2}^2}{4} - \frac{\gamma_{1 \rightarrow 2}^2}{4} \right) \frac{\kappa}{2g^2} + \gamma_{2 \rightarrow 0} + \gamma_{2 \rightarrow 1} \right\} \right. \\
 &\quad \left. - \frac{(\gamma_{0 \rightarrow 2} + \gamma_{1 \rightarrow 2})^2}{4} \frac{\kappa}{2g^2} + \gamma_{2 \rightarrow 1} - \gamma_{2 \rightarrow 0} \right] \\
 &= \frac{1}{2\kappa} \left[\frac{\Gamma_1 - 2\Gamma_2}{\Gamma_1 + 2\Gamma_2} \left\{ \frac{\Gamma_1^2 \kappa \cos \theta}{8g^2} + \Gamma_2 \right\} \cos \theta - \frac{\Gamma_1^2 \kappa}{8g^2} + \Gamma_2 \cos \theta \right] \\
 &= \frac{\Gamma_1}{2\kappa} \left[\frac{1}{1 + (\Gamma_1/2\Gamma_2)} \cos \theta - \left(1 + \frac{1 - (\Gamma_1/2\Gamma_2)}{1 + (\Gamma_1/2\Gamma_2)} \cos^2 \theta \right) \frac{\Gamma_1^2 \kappa}{8g^2} \right].
 \end{aligned}$$

Option #2



The method (example)



How information overflows

The image shows a screenshot of a MathOverflow question titled "Diagonal asymptotics of integer compositions". The question asks for the asymptotics of $\kappa(n, j, k)$ as $k \rightarrow \infty$ with $n \sim \lambda k$. A comment provides a solution using the bivariate generating function $K(x, y) = 1/(1 - yf_j(x))$ and the standard smooth point formula for Riordan arrays.

Question: Diagonal asymptotics of integer compositions
 Asked 4 years, 10 months ago · Active 4 years, 10 months ago · Viewed 222 times

A (weak) composition of a positive integer n into k parts is an ordered sequence of integers (a_1, a_2, \dots, a_k) such that $\sum_{i=1}^k a_i = n$. I am interested in the case where bounded: $a_i \in \{0, 1, \dots, j-1\}$. The number of such compositions satisfies the recurrence

$$\kappa(n, j, k) = \sum_{l=0}^{j-1} \kappa(n-l, j, k-1)$$

and can be expressed as

$$\kappa(n, j, k) = \sum_{s=0}^k (-1)^s \binom{k}{s} \binom{k-sj+n-1}{k-1}$$

[R. P. Stanley, Enumerative Combinatorics, Vol. 1, p. 307]. Does someone know how to find the asymptotics of $\kappa(n, j, k)$ when $k \rightarrow \infty$, $n \sim \lambda k$, $\lambda \in (0, j-1)$, and j is fixed? In particular, I would like to determine

$$\lim_{k \rightarrow \infty, n \sim \lambda k} (\kappa(n, j, k))^{1/k} = ?$$

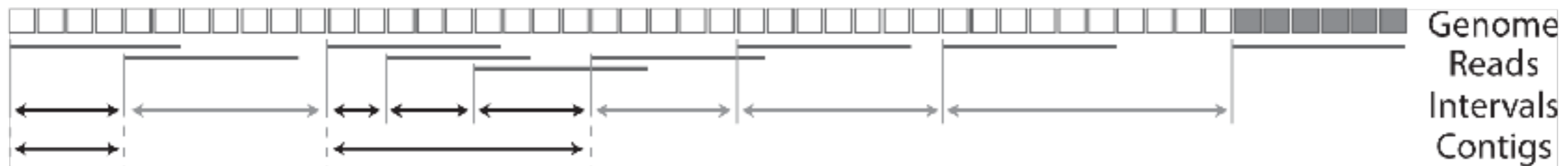
Answer: Such problems can be solved routinely using the methods explained in the book by me and Robin Pemantle, *Analytic Combinatorics in Several Variables* (preprint version online at our websites, book published by Cambridge in 2013). The bivariate generating function for $\kappa(n, j, k)$ is $K(x, y) = 1/(1 - yf_j(x))$ where $f_j(x) = \sum_{i=0}^{j-1} x^i$. This simplifies to a nice rational function - use the standard smooth point formula for Riordan arrays (Section 12.2 as I recall) to get the desired results. I won't work out all details, because I am on vacation, but this is an absolutely standard application of our basic theory.

answered Dec 28 '14 at 3:27
 Mark C. Wilson
 128 ● 7

add a comment

Book Cover: *Analytic Combinatorics in Several Variables* by Robin Pemantle and Mark C. Wilson. Cambridge Series in Advanced Mathematics, 140.

A first manuscript



GUILLAUME FILION

CRG, Barcelona

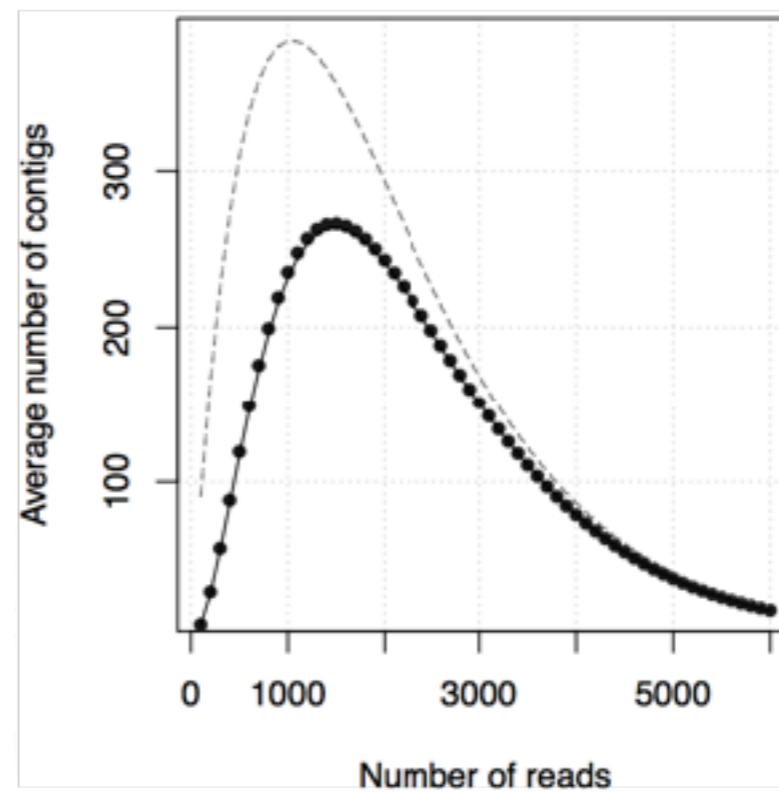
August 4, 2018

Abstract

Here is the text of the abstract. I need to type more in order to see this goes beyond the two columns of the text.

1 Introduction

How many reads should I sequence? How long should they be? With minimum overlap? These questions were first addressed in this context by Lander and Waterman in a landmark study that defined the "classic" de novo assembly [4], giving estimates of the number of contigs. At the time, genome assembly was a long term endeavour and it made sense to gather information about the progress of the project. The Lander-Waterman estimators are accurate for intermediate stages where the number of contigs is high, but quality drops when the assembly nears completion.



A second manuscript



Analytic combinatorics for bioinformatics I: seeding methods

Guillaume J. Filion^{1,2}

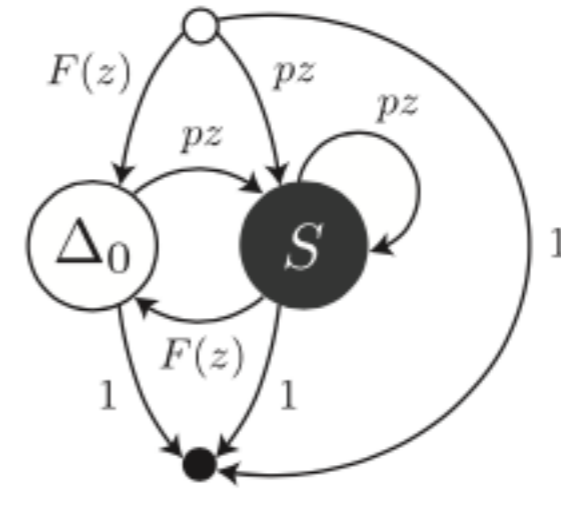
¹Genome Architecture, Gene Regulation, Stem Cells and Cancer Programme, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain.

²University Pompeu Fabra, Doctor Aiguader, 08003 Barcelona, Spain.

February 7, 2019

Abstract

Seeding heuristics are the most widely used strategies to speed up sequence alignment in bioinformatics. Such strategies are most successful if they are calibrated, so that the speed-versus-accuracy trade-off can be properly tuned. In the widely used case of read mapping, it has been so far impossible to predict the success rate of competing seeding strategies for lack of a theoretical framework. Here I present an approach to esti-

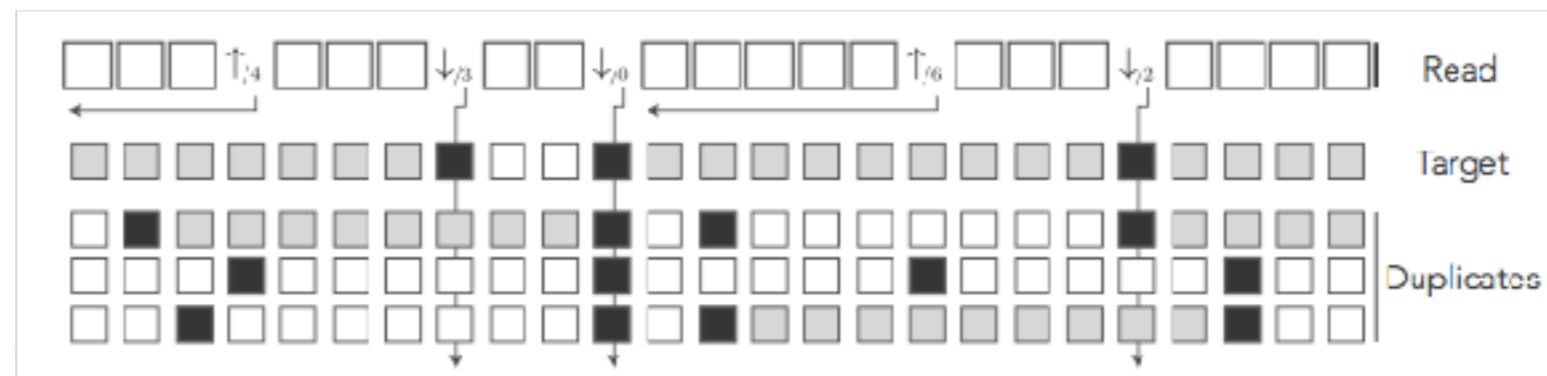


A third manuscript

Calibrating seed-based heuristics to map short DNA reads

Guillaume J. Filion^{1,2}, Ruggero Cortini¹, and Eduard Zorita¹

¹Center



need

reference sequence, typically a genome. Modern mappers rely on heuristics to gain speed at a reasonable cost for accuracy. In the seeding approach, short matches between the reads and the genome are used to narrow the search to a set of candidate locations. Several seeding variants used in modern mappers show good empirical performance but they are difficult to calibrate or to optimize for lack of theoretical results. Here we develop a theory to estimate the probability that reads are mapped to a wrong location due to limitations at the seeding step. We describe the properties of simple exact seeds, skip-seeds and MEM seeds (Maximal Exact Match). The main innovation of this work is to use concepts from analytic com-

A fourth manuscript

Faithful short-read mapping with Sesame

Ruggero Cortini^{1,✱}, Eduard Valera Zorita^{1,✱}, Guillaume J. Filion^{1,2,3,✱}

¹Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain; ²University Pompeu Fabra (UPF), Barcelona, Spain; ³present address: Department of Biological Sciences, University of Toronto Scarborough, Toronto, ON, Canada; ✱ equal contributions.

Received January 1, 2009; Revised February 1, 2009; Accepted March 1, 2009

ABSTRACT

Abstract will come later.

INTRODUCTION

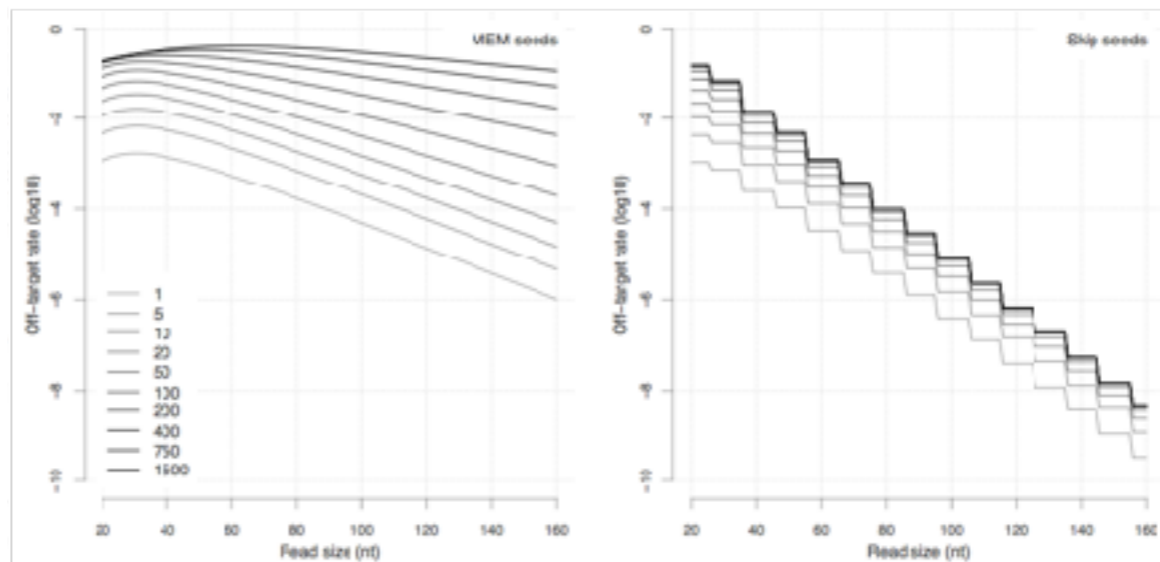
High throughput DNA sequencing is now a standard technology in the academia and in the industry, with countless applications in diagnosis, forensics, surveillance and research (1). The Illumina short-read technology currently dominates the market of DNA sequencing, making the associated software an important target for optimization.

The standard way to identify a read is to map it to a known reference sequence, typically a genome. Work on short-read mappers has been traditionally focused on improving the speed, reducing the memory footprint and increasing the accuracy. Another important focus has been to develop variations to address the actual needs of the user, as different problems often entail different read types (e.g., compare ChIP-seq and Hi-C).

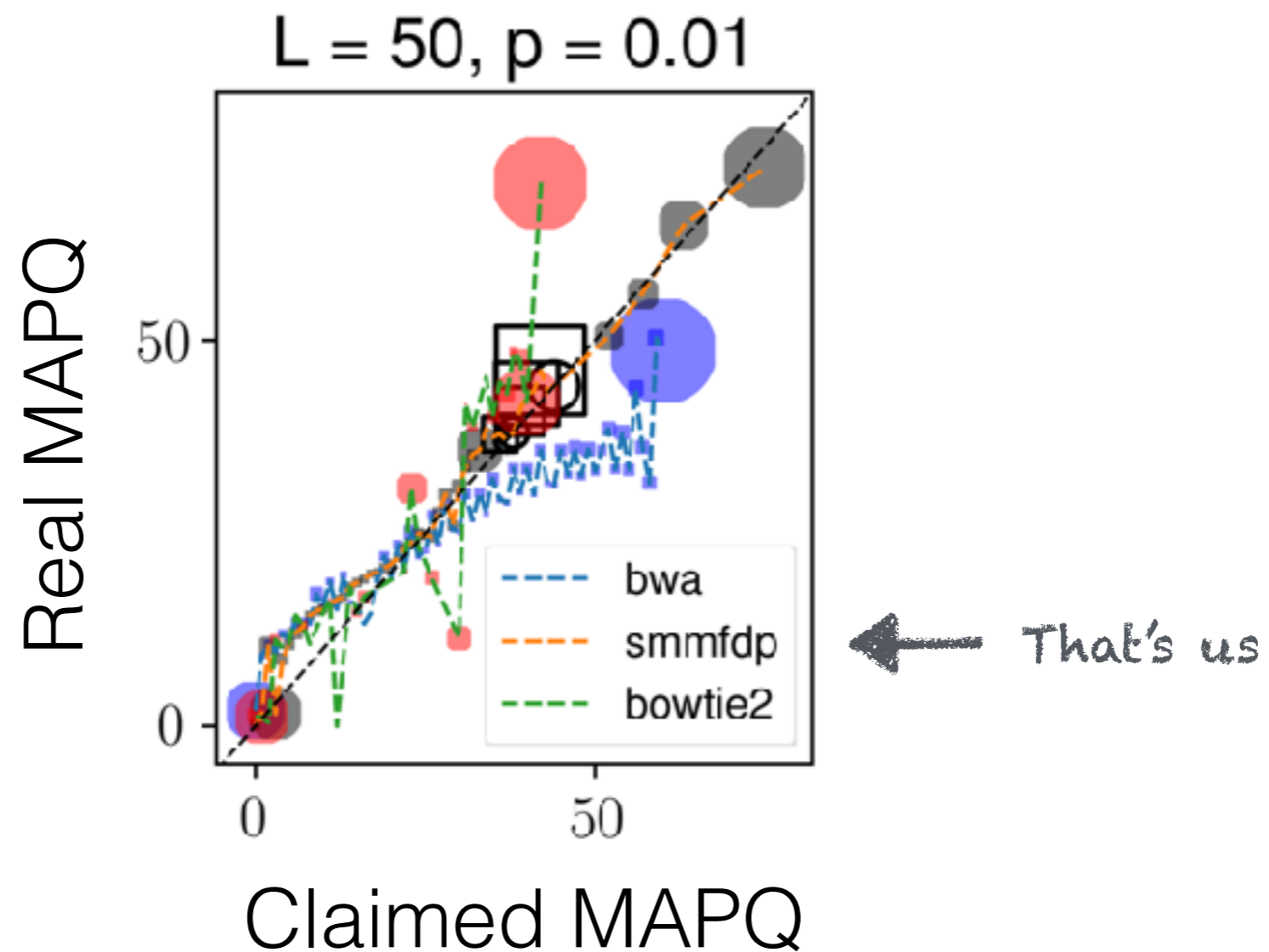
Meanwhile, progress has been more limited on another aspect of the mapping process: *faithfulness*. Mapping algorithms are heuristics, so there is always a chance that the proposed location of a read is wrong. Faithfulness is the capacity to correctly estimate this probability. Importantly, one can reduce the probability of errors without having to measure it, so accuracy and faithfulness are usually unrelated.

in many applications it is essential to know the risk associated with every read (e.g., contaminated material such as ancient DNA); *ii* a mapper is easier to use if error rates are what they claim to be; *iii* good calibrations open opportunities to improve the speed-accuracy tradeoff.

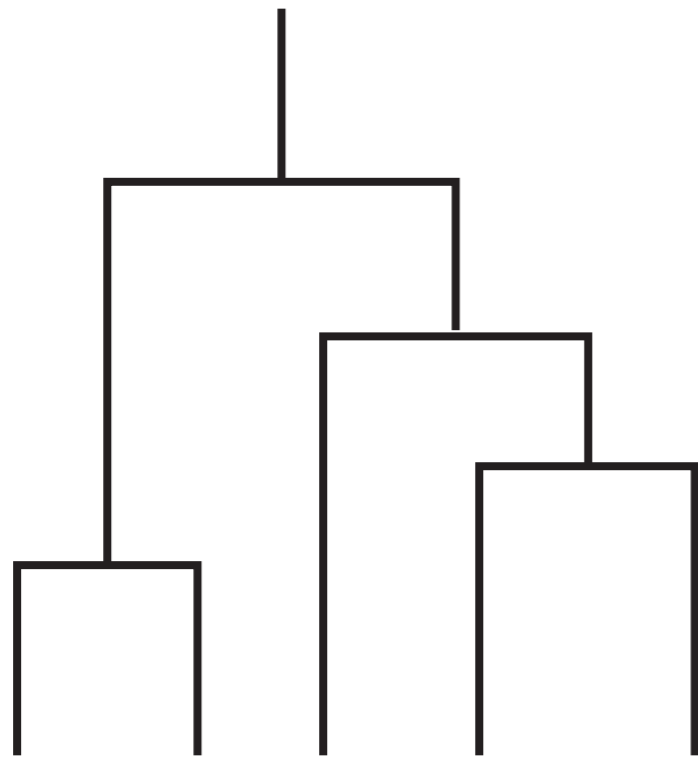
We recently proposed a strategy to compute the error rate of different seeding strategies used in short-read



Where we are now



What is next?



Duplicates evolve as a tree, their sequences are not statistically independent. We need a special method to count them.

Some perspective

This line of research is entirely in the academic 'underworld'. It started on my blog, used information from Coursera, Stack Overflow, and Roman's podcast.

Connections between the ideas are illogical.

The drive is just my own taste.

This is 7 years of research.

Acknowledgements



Eduard
Valera



Ruggero
Cortini

Thanks for your attention!