SPAAM

Standards, Precautions & Advances in Ancient Metagenomics

Date : 22 Sept. 2020
Chairs: Alex Hübner
Anna Fotakis

Session 5: Reuse the Refuse

Applying new analytical methods beyond current practices

Session Scope

- New and exciting methods and approaches for pushing the boundaries of metagenomics analysis
- Ancient metagenomics ‘adopting’ new protocols for genome-resolved metagenomics and functional analysis with the challenges of ancient datasets
- Encourage novel approaches in data analysis but with caution on the potential pitfalls and limitations when designing and embarking on such a study

SPAAM

A brief summary of concepts

What is genome resolved metagenomics?

Shotgun sequencing

Assembly part 1

Downstream analysis

Binning into MAGs (metagenome assembled genomes)

Assembly part 2 - contigs

Downstream analysis

SPAAM

In other words...

SPAAM

SPAAM

MENU

nature
International journal of science

Article | OPEN | Published: 11 February 2019

A new genomic blueprint of the human gut microbiota

Alexandre Almeida, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley & Robert D. Finn

Nature **568**, 499–504 (2019) | Download Citation

New Results

Comment on this paper

The genomic and proteomic landscape of the rumen microbiome revealed by comprehensive genome-resolved metagenomics

Robert D. Stewart, Marc D. Auffret, Amanda Warr, Alan W. Walker, Rainer Roehe, Mick Watson

doi: https://doi.org/10.1101/489443

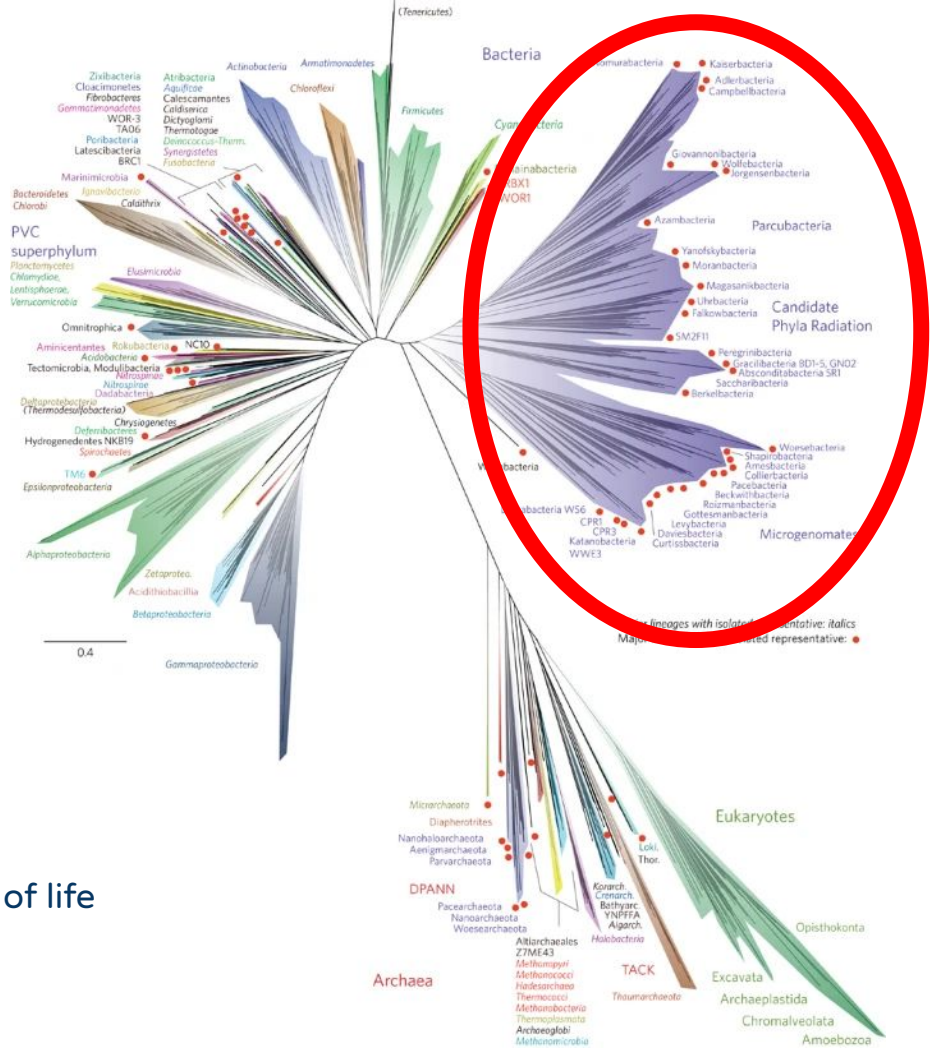
This article is a preprint and has not been peer-reviewed [what does this mean?].

RESOURCE | VOLUME 176, ISSUE 3, P649-662.E20, JANUARY 24, 2019

Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle

Edoardo Pasolli • Francesco Asnicar [8] • Serena Manara [8] • ... Christopher Quince • Curtis Huttenhower • Nicola Segata [9] • Show all authors • Show footnotes

Open Access • Published: January 17, 2019 • DOI: https://doi.org/10.1016/j.cell.2019.01.001 •



A new view on the tree of life

Hug *et al* 2016
https://doi.org/10.1038/nmicrobiol.2016.48

Some considerations

- ❏ Study design (including laboratory strategy, sequencing depth)
- ❏ Validations
- ❏ Quality of data (soil vs dental calculus vs bone etc) and the limitations (a good assembly is favoured by long good quality reads)
- ❏ Question and application for your study

Further applications

- ❏ Novel genome discovery
- ❏ Functional analysis
- ❏ Pangenomics
- ❏ Phylogenetics
- ❏ And more...

In summary

- Experimental set-up
- Data quality (combining both assembly and assembly free methods)
- Computational power
- Validating your steps
- Know your samples/data!!!!

http://www.poorlydrawnlines.com/comic/reason/

Ice breaker speakers

Maxime Borry (PhD - MPI for the Science of Human History) - “A new take on ancient DNA metagenomics assembly and validation”

Antonio Fernandez-Guerra (Assistant Professor, Section for GeoGenetics, Globe Institute, UCPH) - “Microbial Ancient Metagenomics: Beyond who is there”

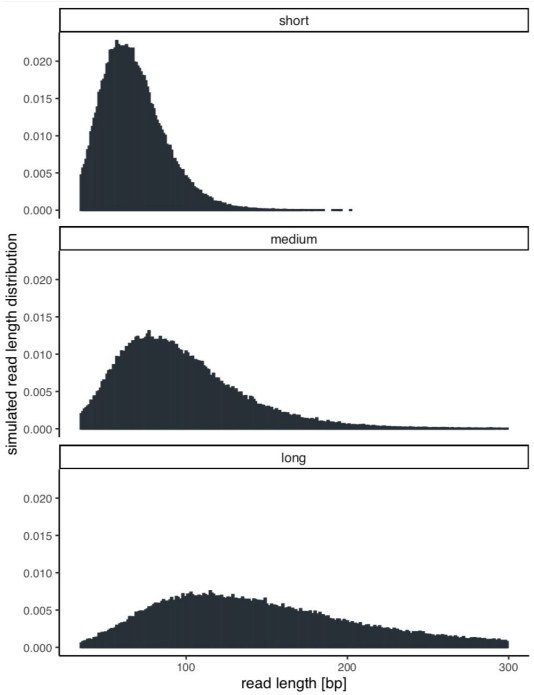
SPAAM

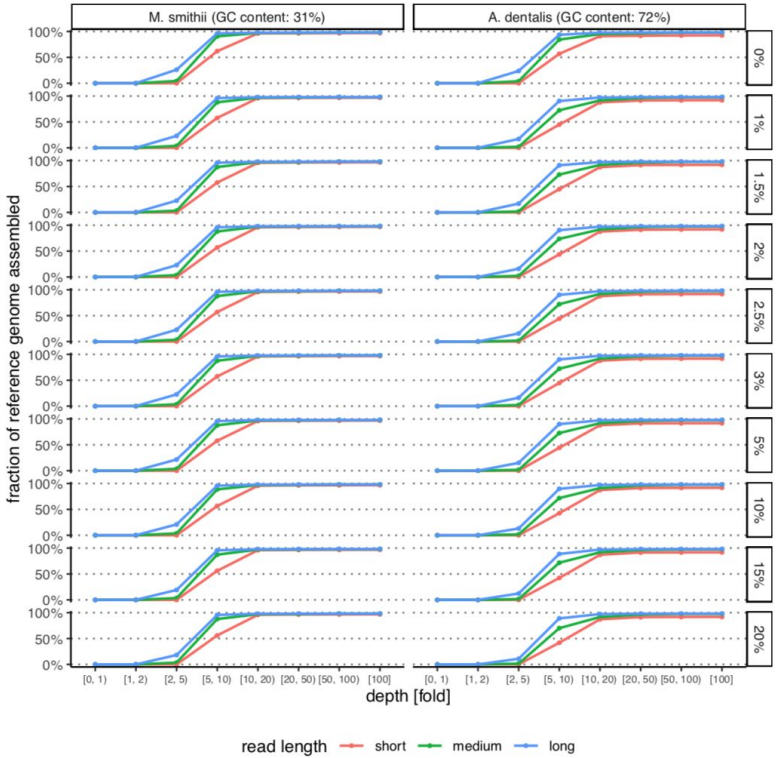
Standards, Precautions & Advances in Ancient Metagenomics

A new take on ancient DNA metagenomics assembly validation

Maxime Borry

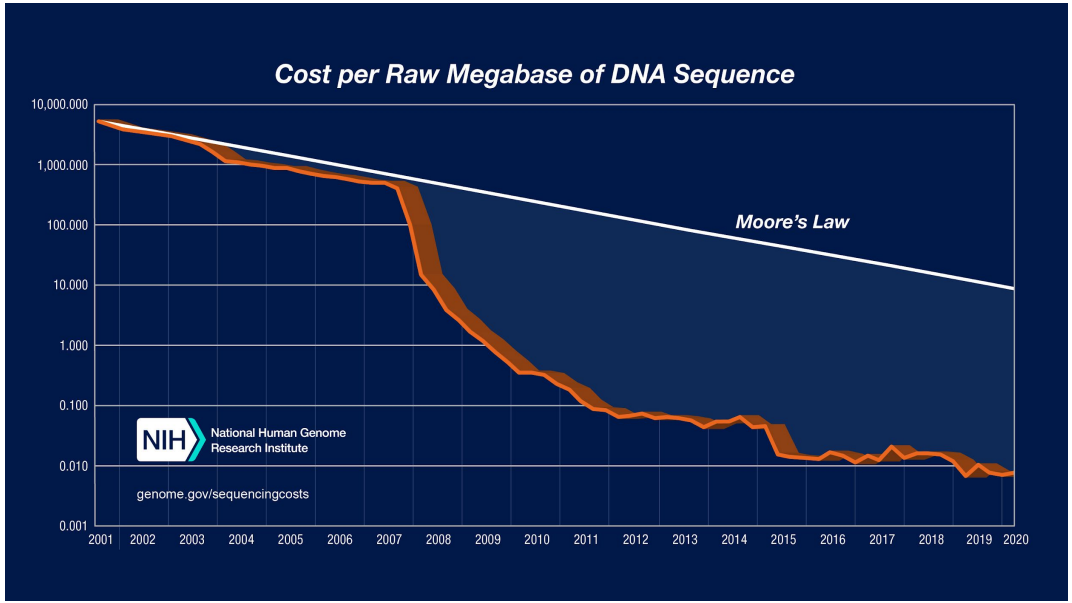
aDNA de novo assembly: is it even possible ?





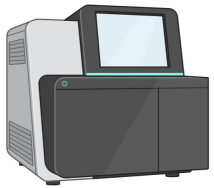
Coverage is key

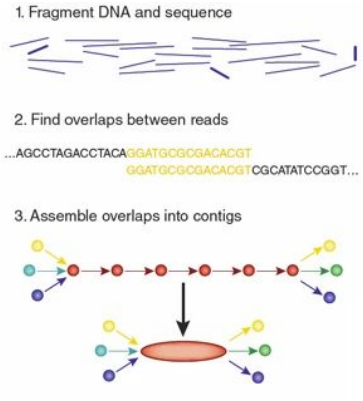
And you will always get more for you money

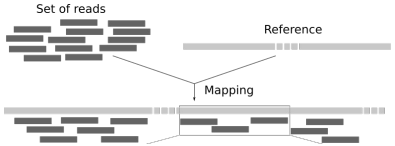


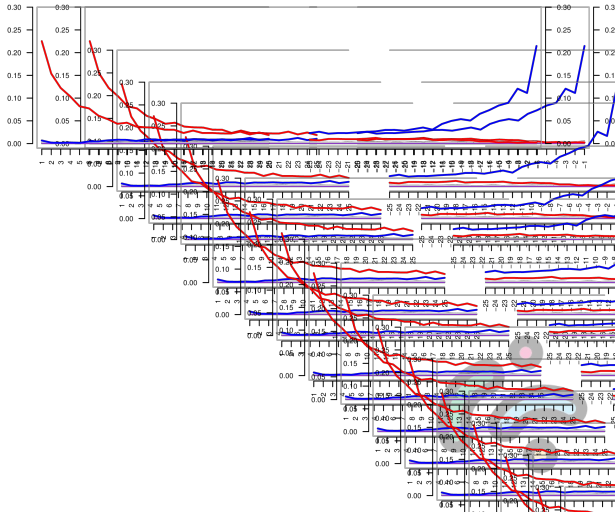
Scaling up damage profile inspection

aDNA validation of sequences often involves damage profiles



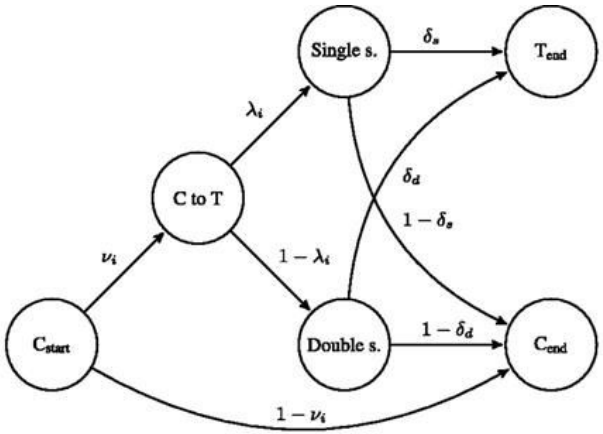






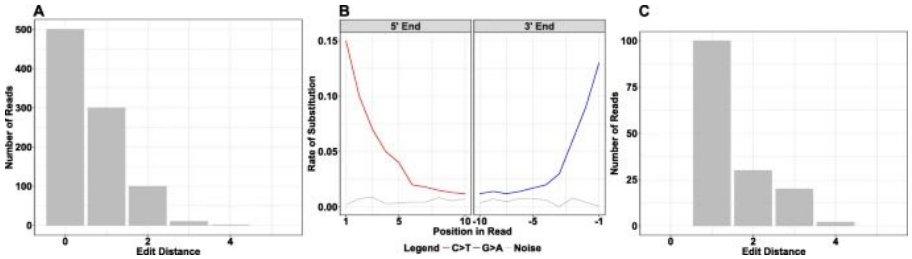
How to programatically check for aDNA deamination ?

The mapdamage2 approach



The HOPS approach

1 CtoT in first 10 bp = ancient



More accurate but **Slower**

Jónsson et al. 2013

Fast but very **Approximate**

Hübler et al. 2019

SPAAM

PYDAMAGE

$\hat{\theta}$

github.com/maxibor/pydamage

Dr. Alexander Huebner

Dr. Adam Ben Rohrlach

What next ?

- Damage as extra metric/dimension for binning

- Other aDNA contig validation criteria

SPAAM

Standards, Precautions & Advances in Ancient Metagenomics

Microbial Ancient Metagenomics: Beyond who is there

Antonio

- **On the complexity of assembling ancient metagenomes**
  - very short reads
  - low coverage
  - damage

- **Give me functions!**
  - damage and the amino acid space
  - substitution matrices

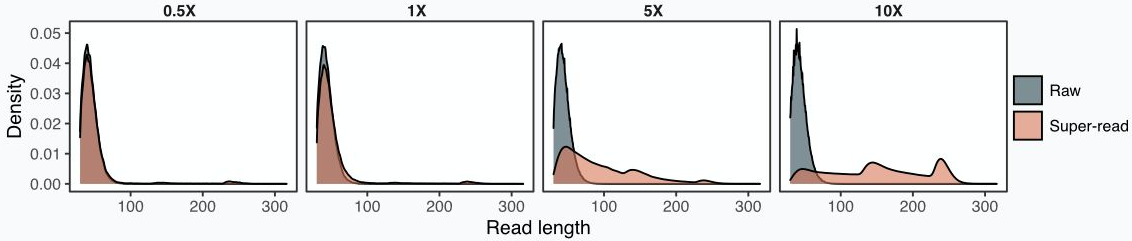
SPAAM

Let’s keep it simple

exploring one genome at a time

Synthetic data from *Yersinia pestis* A1122

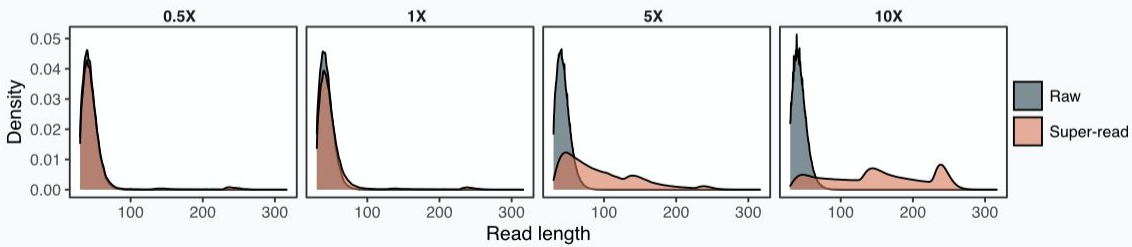
- different coverage
- mode read length = 40 nt
- with and without damage



0.5X: 52,342 | **1X**: 104,684 reads | **5X**: 523,421 reads | **10X**: 1,046,843 reads

Synthetic data from *Yersinia pestis* A1122

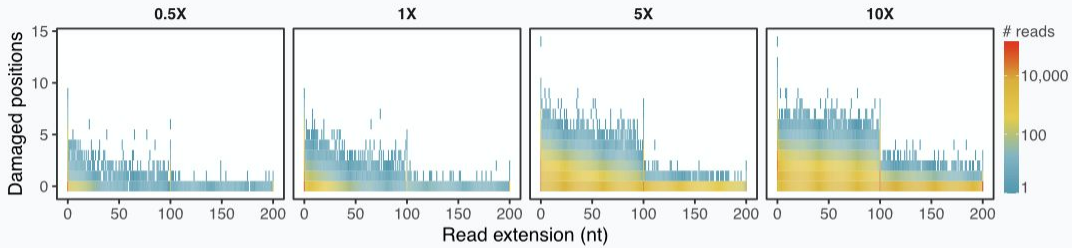
- different coverage
- mode read length = 40 nt
- with and without damage



0.5X: 52,342 | **1X**: 104,684 reads | **5X**: 523,421 reads | **10X**: 1,046,843 reads

Super-reads?

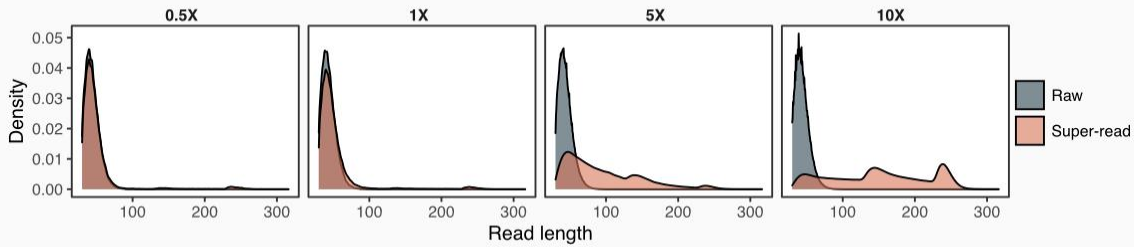
- careful extension of reads from both ends
- we can gain extra bits of information



MaSuRCA: Zimin et al., 2013

Synthetic data from *Yersinia pestis* A1122

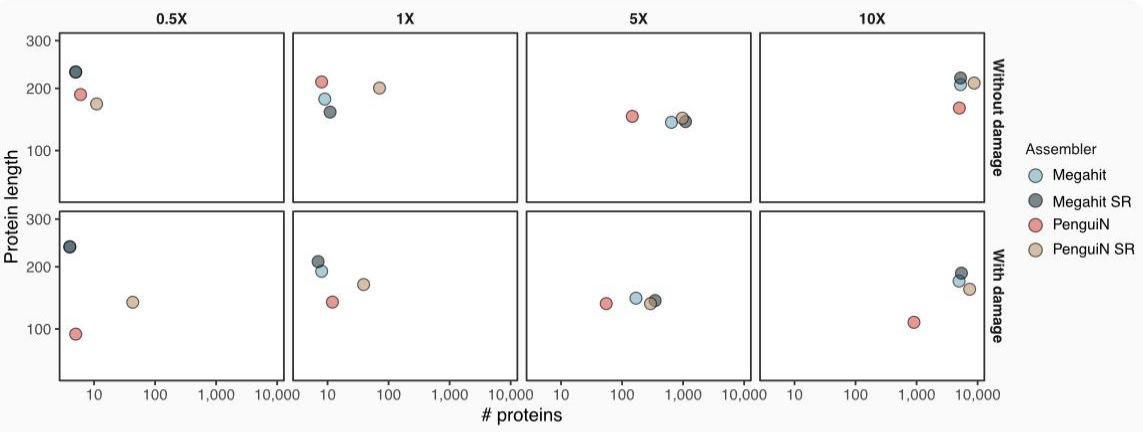
- different coverage
- mode read length = 40 nt
- with and without damage



0.5X: 52,342 | **1X**: 104,684 reads | **5X**: 523,421 reads | **10X**: 1,046,843 reads

How can I assemble my ancient metaG?

- Which options do I have?
- How can I recover more genes?
- An amino acid centric approach



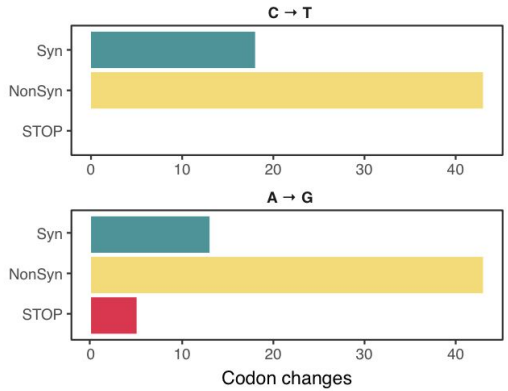
MaSuRCA: Zimin et al., 2013 | MEGAHIT: | PenguiN: Steingger et al. unpublished

Assembling your ancient metaG will not always be possible

- in many cases assemblies will be very bad (fragmented)
- assembly only recovers a small fraction
- very low coverage + short reads = worst combination

The coding sequence space of ancient metagenomes

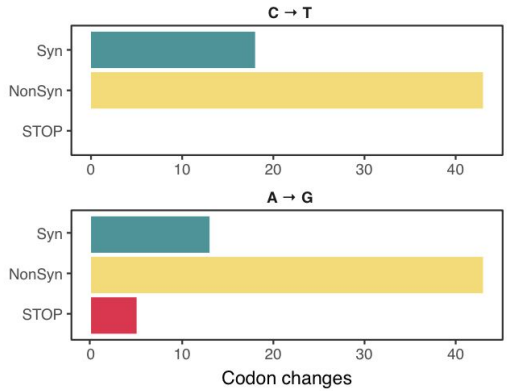
First Position	Second Position: T	Second Position: C	Second Position: A	Second Position: G	Third Position
T	TTT, TTC: Phe TTA, TTG: Leu	TCT, TCC, TCA, TCG: Ser	TAT, TAC: Tyr TAA: STOP TAG: STOP	TGT, TGC: Cys TGA: STOP TGG: Trp	T C A G
C	CTT, CTC, CTA, CTG: Leu	CCT, CCC, CCA, CCG: Pro	CAT, CAC: His CAA, CAG: Gln	CGT, CGC, CGA, CGG: Arg	T C A G
A	ATT, ATC, ATA: Ile ATG: Met	ACT, ACC, ACA, ACG: Thr	AAT, AAC: Asn AAA, AAG: Lys	AGT, AGC: Ser AGA, AGG: Arg	T C A G
G	GTT, GTC, GTA, GTG: Val	GCT, GCC, GCA, GCG: Ala	GAT, GAC: Asp GAA, GAG: Glu	GGT, GGC, GGA, GGG: Gly	T C A G



Codons and damage

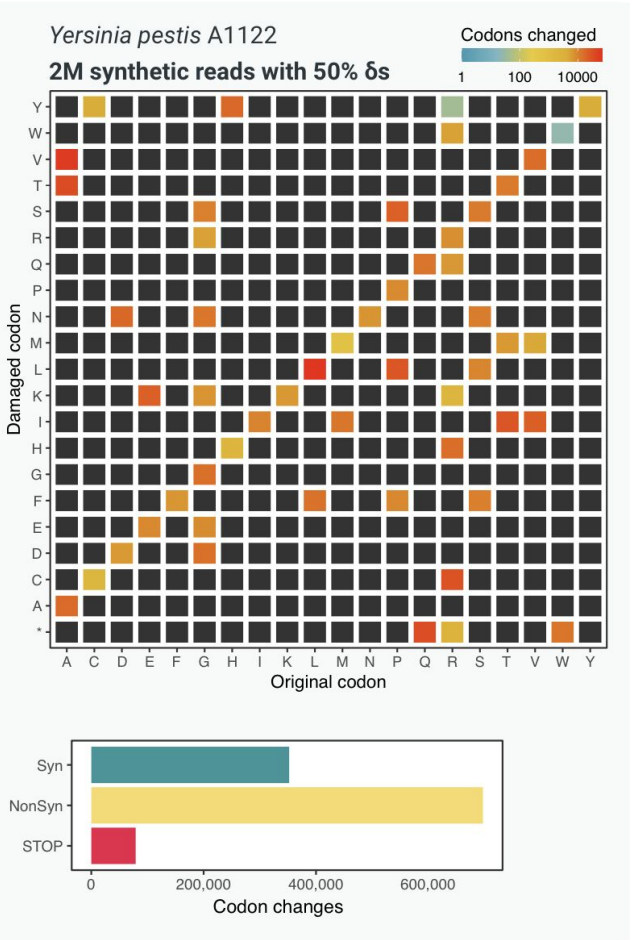
what to expect when looking for function

First Position	Second Position: T	Second Position: C	Second Position: A	Second Position: G	Third Position
T	TTT, TTC Phe; TTA, TTG Leu	TCT, TCC, TCA, TCG Ser	TAT, TAC Tyr; TAA STOP; TAG STOP	TGT, TGC Cys; TGA STOP; TGG Trp	T C A G
C	CTT, CTC, CTA, CTG Leu	CCT, CCC, CCA, CCG Pro	CAT, CAC His; CAA, CAG Gln	CGT, CGC, CGA, CGG Arg	T C A G
A	ATT, ATC, ATA Ile; ATG Met	ACT, ACC, ACA, ACG Thr	AAT, AAC Asn; AAA, AAG Lys	AGT, AGC Ser; AGA, AGG Arg	T C A G
G	GTT, GTC, GTA, GTG Val	GCT, GCC, GCA, GCG Ala	GAT, GAC Asp; GAA, GAG Glu	GGT, GGC, GGA, GGG Gly	T C A G



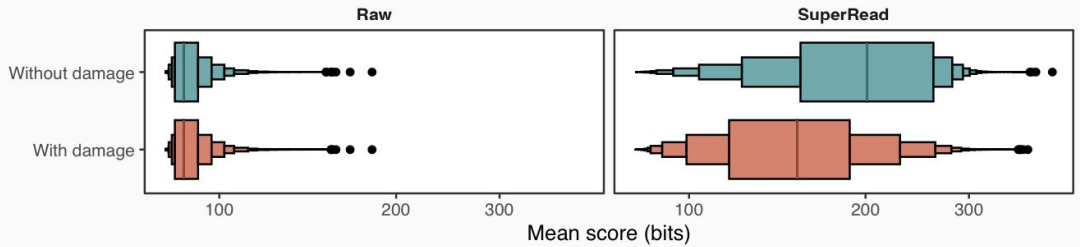
Codons and damage

what to expect when looking for function



On the importance of choosing the right

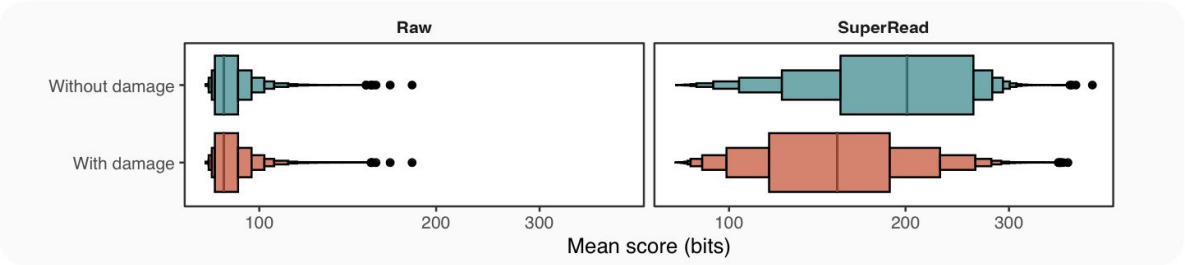
substitution matrix

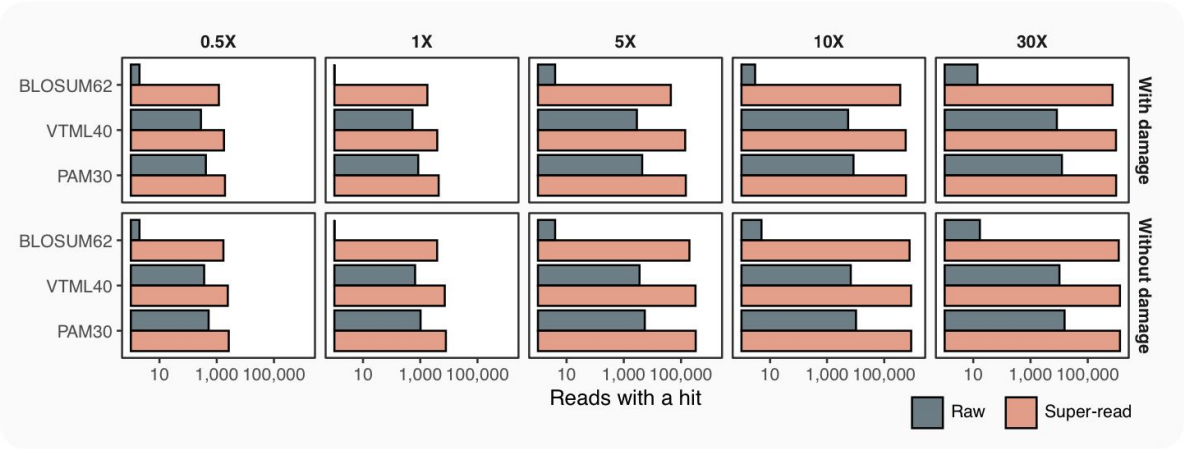


Selecting the Right Similarity-Scoring Matrix: Pearson, 2013

On the importance of choosing the right

substitution matrix





	Gap penalty	Similarity (%)	Bits/pos.	50 bit length
BLOSUM80	10/1	32.0	0.48	104
BLOSUM62	11/1	28.9	0.40	125
VTML140	10/1	28.4	0.44	114
VTML120	11/1	32.1	0.54	93
VTML80	10/1	40.5	0.74	68
VTML40	13/1	64.7	1.92	26
VTML20	15/2	86.1	3.30	15
VTML10	16/2	90.9	3.87	13
PAM70	10/1	33.9	0.58	86
PAM30	9/1	45.9	0.90	56

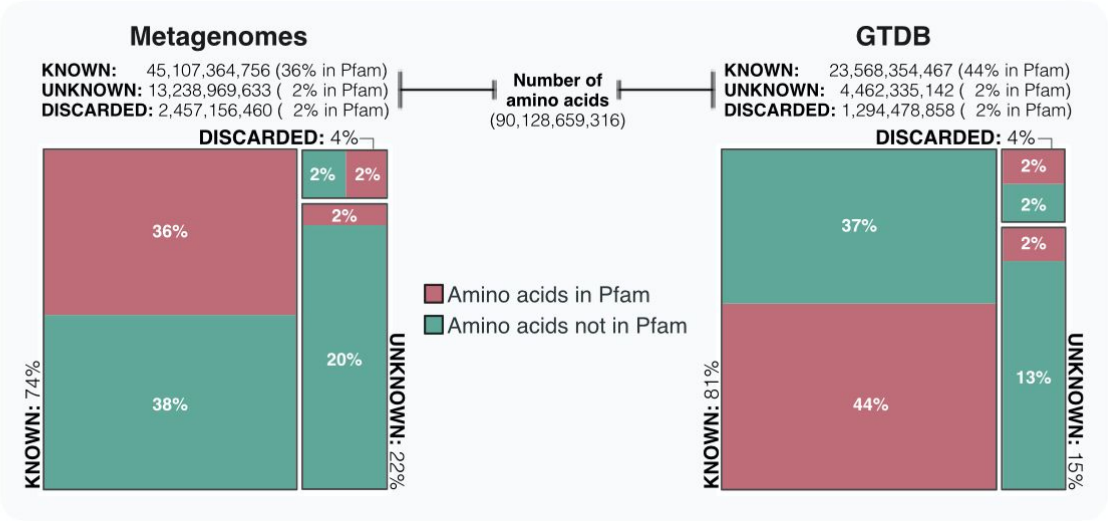
Yersinia pestis A1122 **2M synthetic reads with 50% δs** vs *Yersinia pestis* A1122 **proteome**

Selecting the Right Similarity-Scoring Matrix: Pearson, 2013

A word of **caution**

when using modern methods

if you use **very short reads**...



AGNOSTOS: Vanni et al., 2020 :: http://merenlab.org/2020/07/01/dark-side/

Discussion Points:

- Ancient de novo assembly validation - what we should look out for?
- How to go about designing a de novo assembly study
- Considerations for functional annotations
- Potential of Capture for targeted de novo applications
- Broader future applications - including pangenomics or even strain resolution