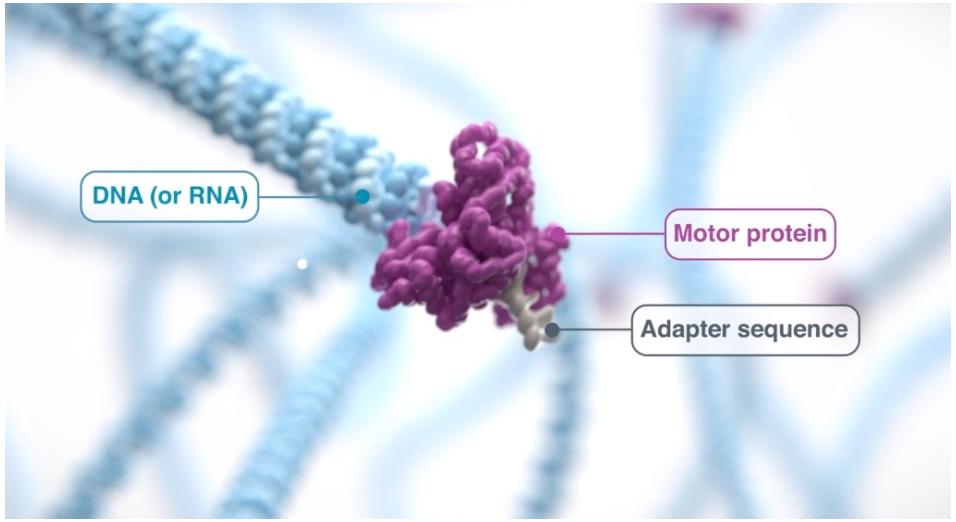
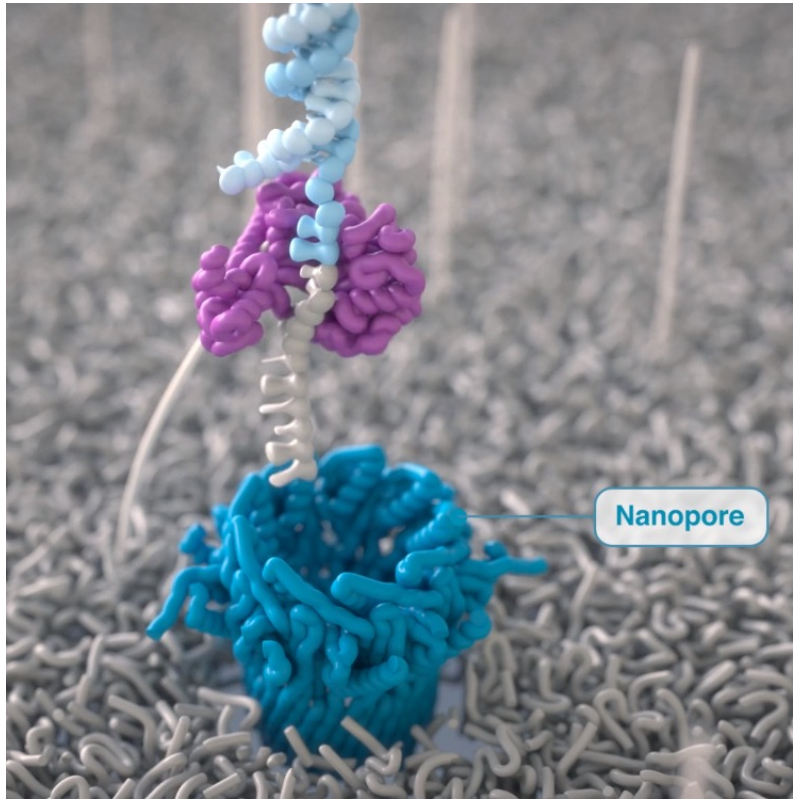
Nanopore workshop

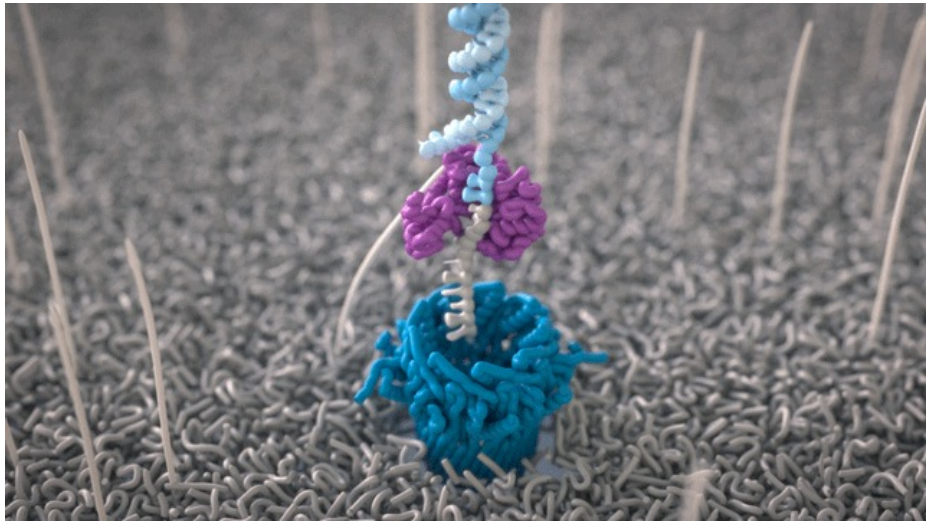
Chandigarh April 2022

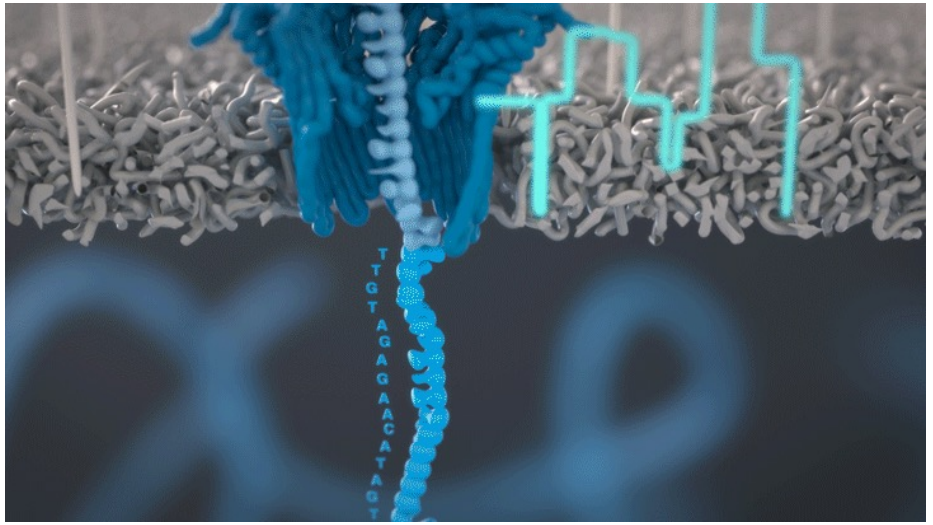
@vwaschulin

How it works


DNA (or RNA)
Motor protein
Adapter sequence

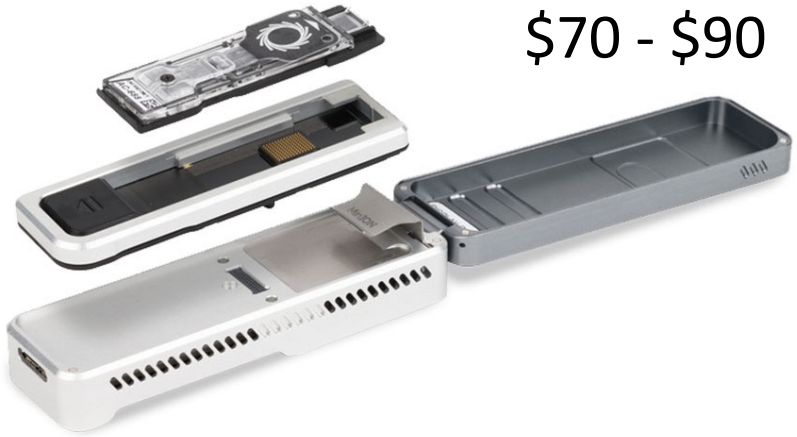

Nanopore




TGTAGAGAACATAGT

Devices

Flongle flow cells
1-3 Gb (i.e. 1-2 strains; plasmids, etc)
$70 - $90



MinION flow cells
10-30 Gb (24+ strains)
$500 - $900

+ other devices using MinIon flow cells

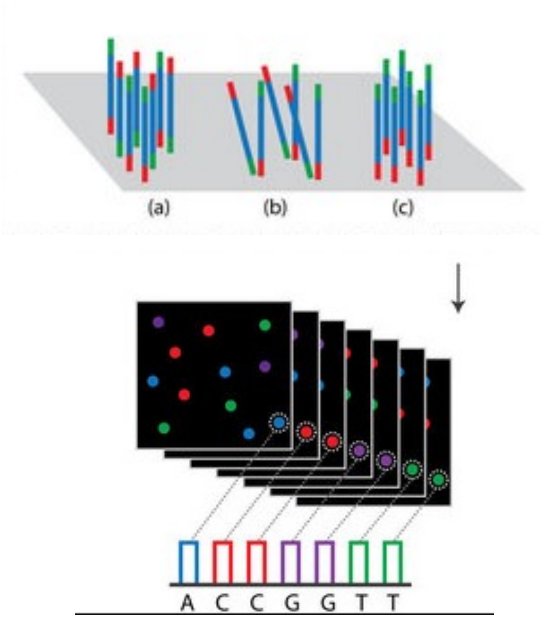
Promethion flow cells
50-200 Gb (high throughput)
$700 - $2000

Key Differences to other platforms

Sequencing by synthesis

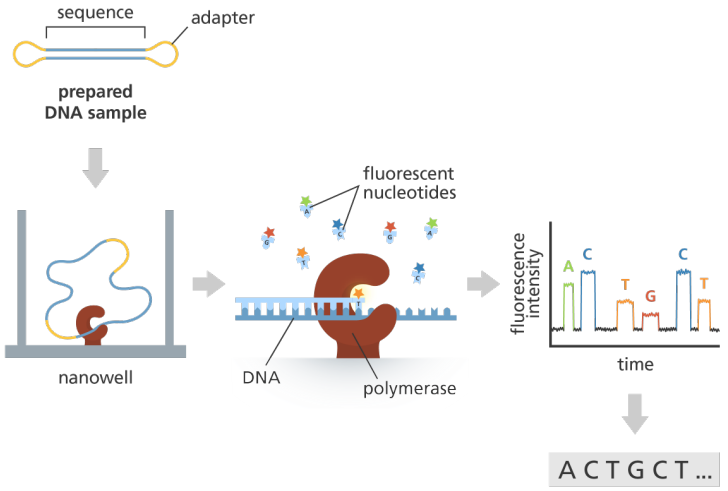
“Sequencing by observation”

Illumina
massive parallel sequencing



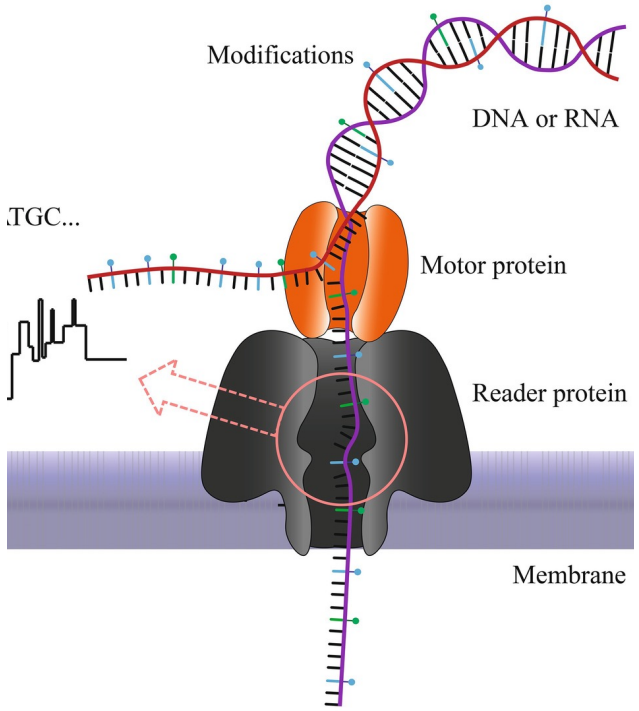
75-350 bp
>99.9% accuracy (Q30+)

PacBio HiFi
single-molecule real-time sequencing



up to c. 20 kb (= 20 000 bp)
>99.9% accuracy (Q30+)

Nanopore sequencing



up to 2 Mb = 2 000 000 bp)
>99.9% accuracy (Q30+)

Advantages

- Read length (theoretically) limited only by molecule length
- Methylation
- Direct RNA sequencing
- Rapid “click chemistry” sequencing
- Portable and “field kits” available --> sequencing in a tent
- Sequencing process and data processing can be optimized for speed or accuracy

Advantages in metagenomics

- Allows assembly of complete, circular genomes
- **16S genes on MAGs**
  - Combine 16S and MAG studies
- Lower coverage needed for assembly
- Assembles BGCs

Letter | Open Access | Published: 10 February 2020

Complete, closed bacterial genomes from microbiomes using nanopore sequencing

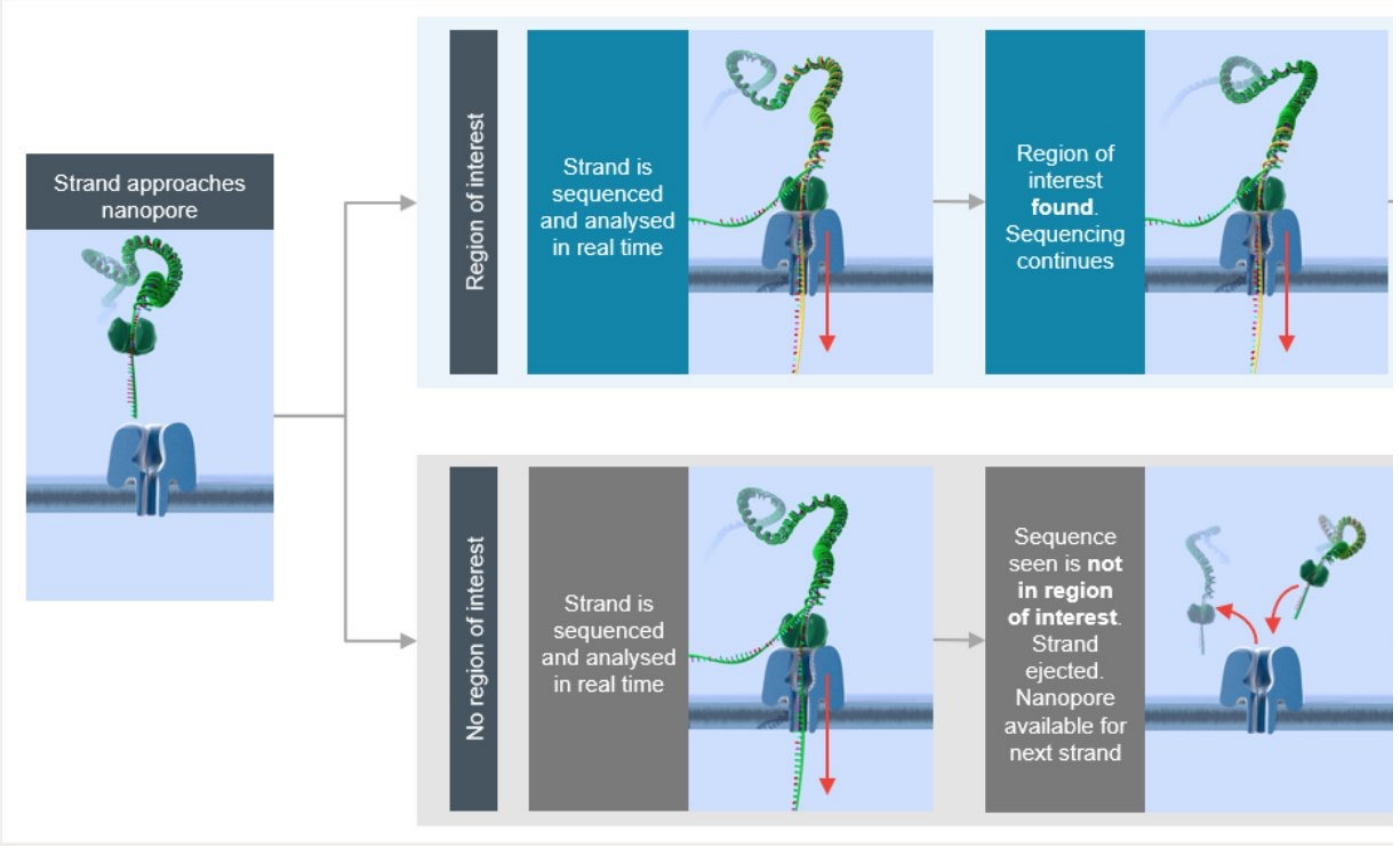
Eli L. Moss, Dylan G. Maghini & Ami S. Bhatt

Article | Open Access | Published: 31 March 2021

Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing

Caitlin M. Singleton, Francesca Petriglieri, Jannie M. Kristensen, Rasmus H. Kirkegaard, Thomas Y. Michaelsen, Martin H. Andersen, Zivile Kondrotaite, Søren M. Karst, Morten S. Dueholm, Per H. Nielsen & Mads Albertsen

Adaptive sequencing

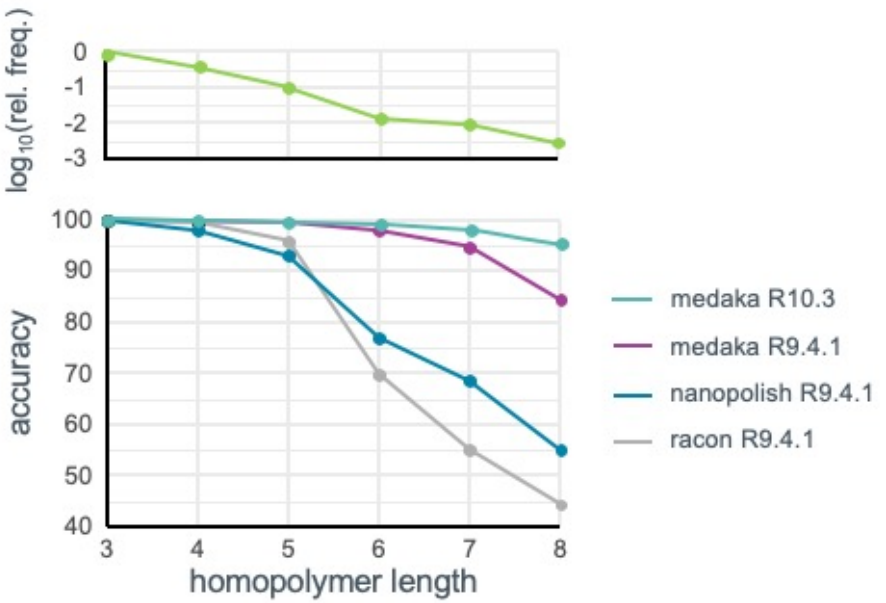


Uses

- Even sample coverage when barcoding
- Sequence rare members of microbiome
- Selective sequencing of e.g. exons

Biggest disadvantage: Accuracy

- Raw read accuracies 90-99% (kit dependent)
- **Total error = random error + biased error**
- Each “squiggle” represents several bases → long homopolymers (CCCCCCCCCC, GGGGGGGG, etc) give difficult signal



Insertions/deletions (indels) lead to frameshifts in assemblies

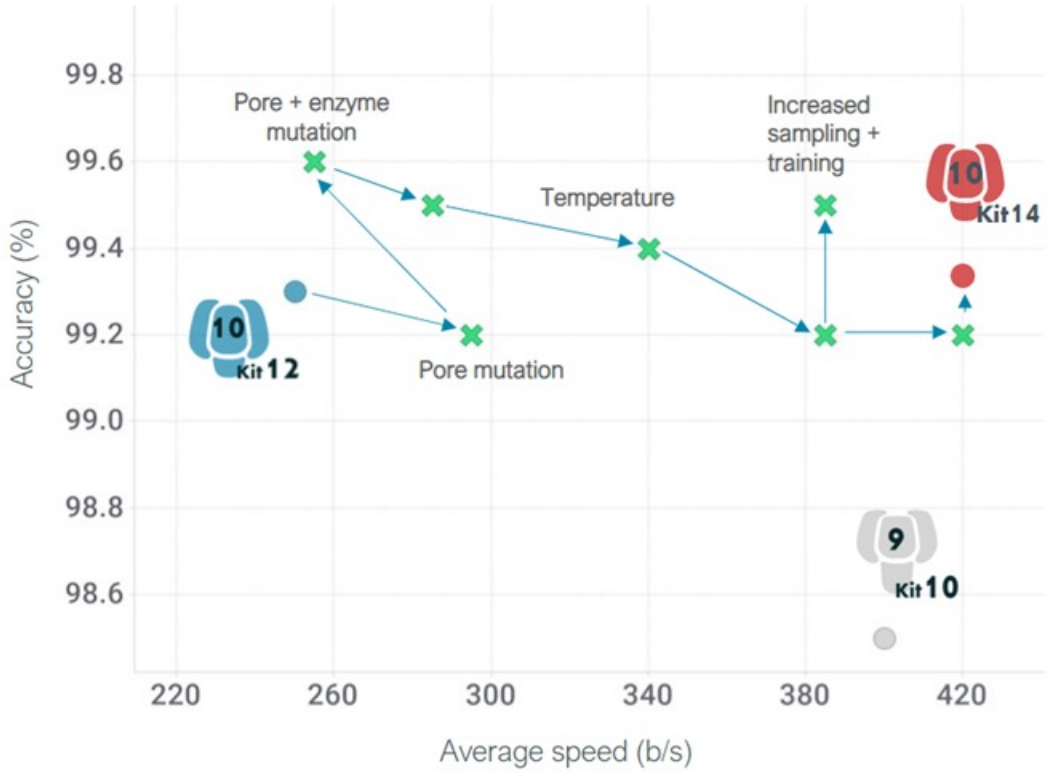
How to fix accuracy

User actions:

- Increase coverage
  - Bias errors not fixed by coverage
- Polish with Illumina reads
- Choose more accurate kit

ONT product improvements:

- New chemistry (Q20+)
- New pore architecture
- Changed temperature
- Improved bascalling/error correction



Other disadvantages

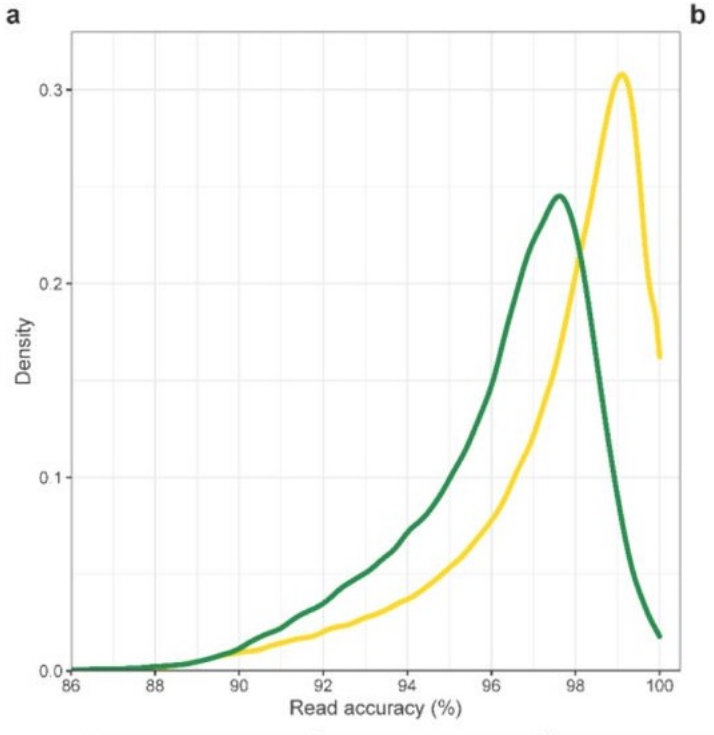
- Sequencing native DNA requires a lot of DNA (100s of ng)
- DNA needs to be HMW for long reads to be obtained
- Rapid kits = low yield (do not use)
- Not all DNA is equally well basecalled: Bias through machine learning algorithms
- Tools & software out of date within months
- Computing power needed (GPUs)
- Storage needed for raw files

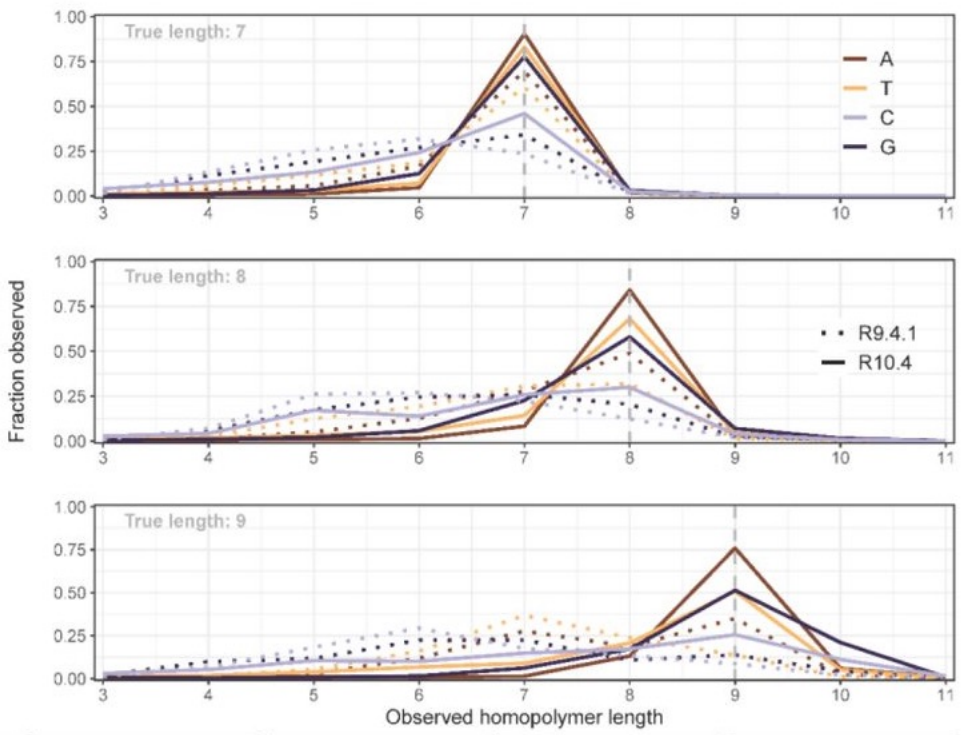
Addressed by devices
e.g. MinION Mk1c

In a few months

- Kit 14 chemistry + R10.4 pore:
  - Q20+ (99%) read modal accuracy
  - Q30 for duplex (99.9%)

- MAG assembly comparable to PacBio
- Coverage > 40x needed
- Rare homopolymers 8+ still a big issue

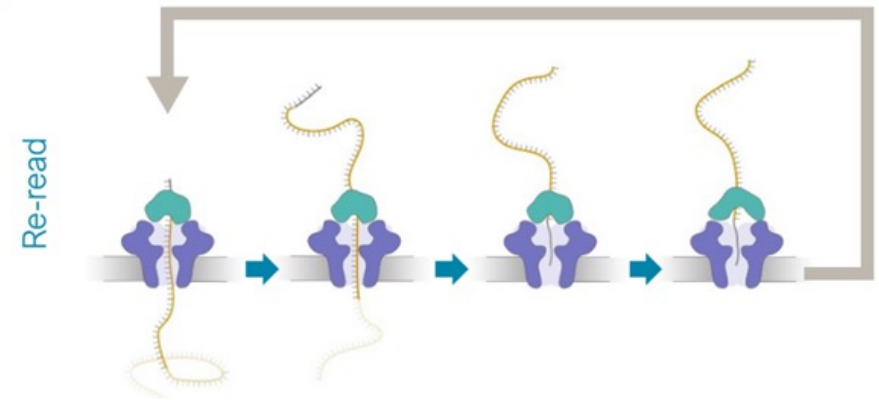




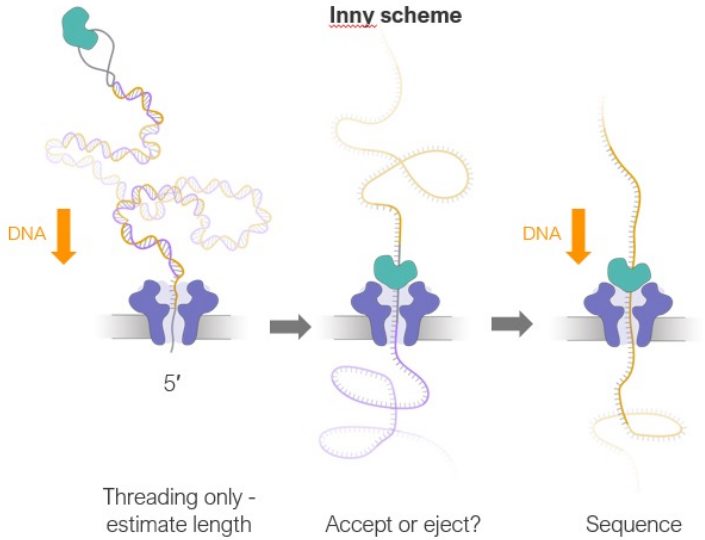
Sereika et al 2021
doi: https://doi.org/10.1101/2021.10.27.466057

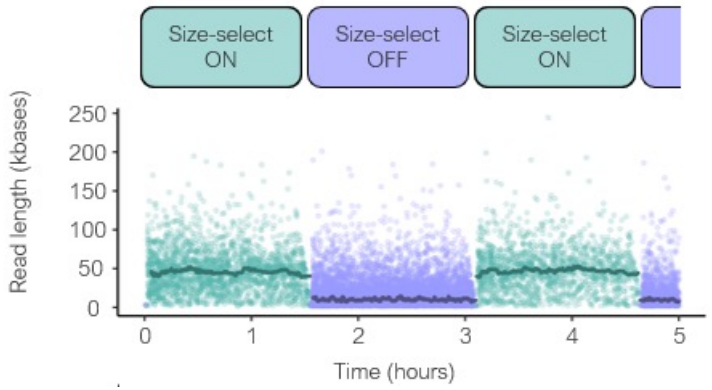
Exciting developments

Adaptive sampling: re-read one strand many times



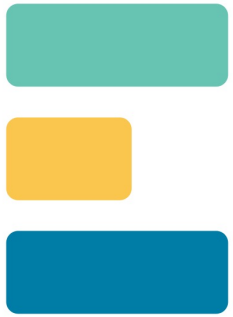
Size selection before sequencing





Ready-made EPI2ME pipelines

- 16S sequencing
- Plasmid verification
- Centrifuge classification of metagenomes
- RNA seq

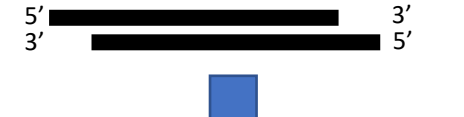


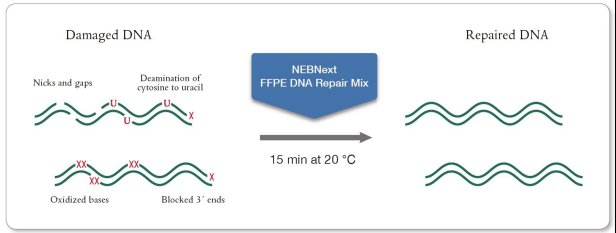
EPI2ME

Practical aspects

Library prep (genomes)

HMW DNA extraction

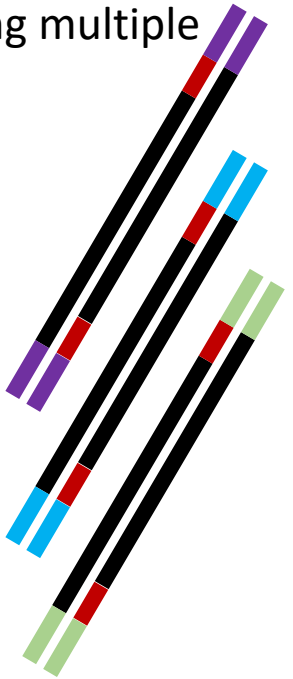




Blunt ending
& nick repair

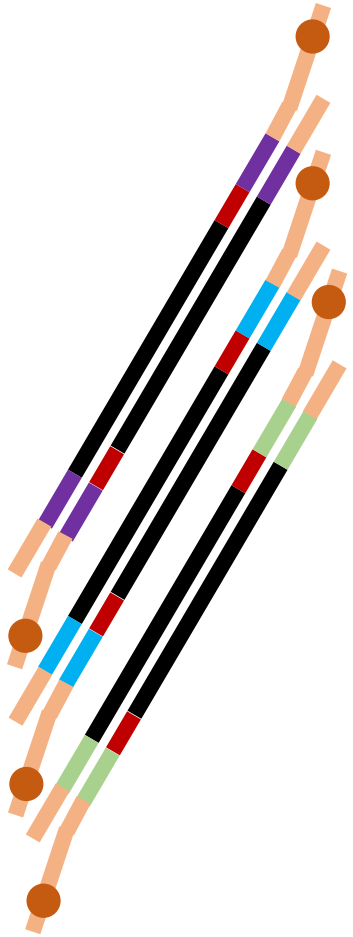
Necessary only if sequencing multiple samples

Cleanup



Barcoding

Cleanup

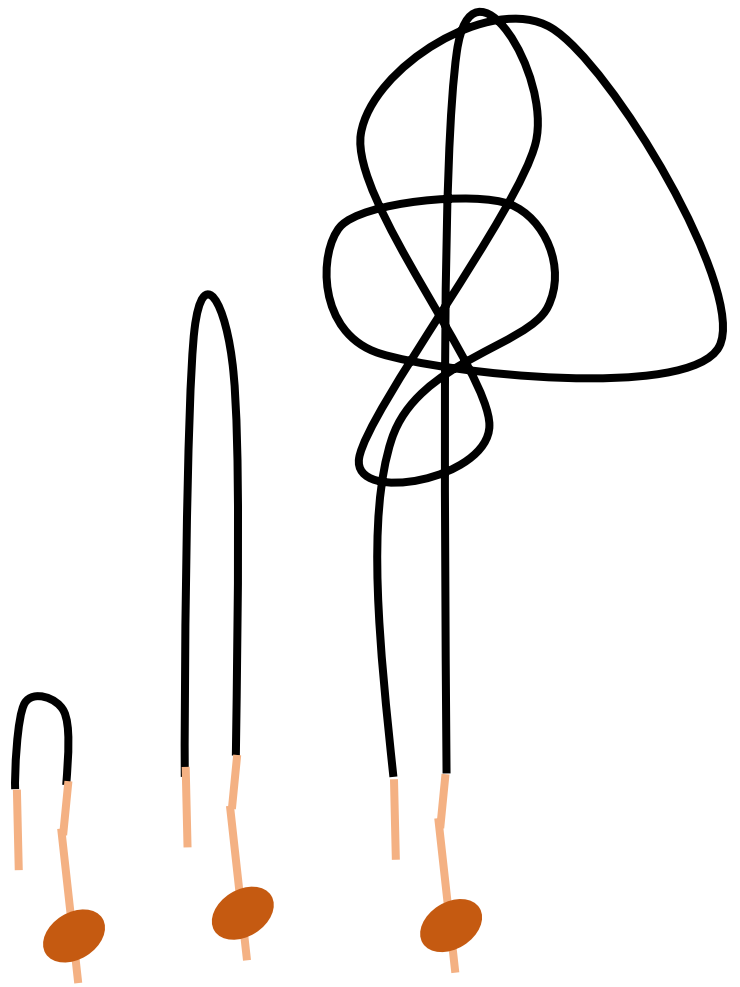


Sequencing adapter ligation

Cleanup

Understanding read length

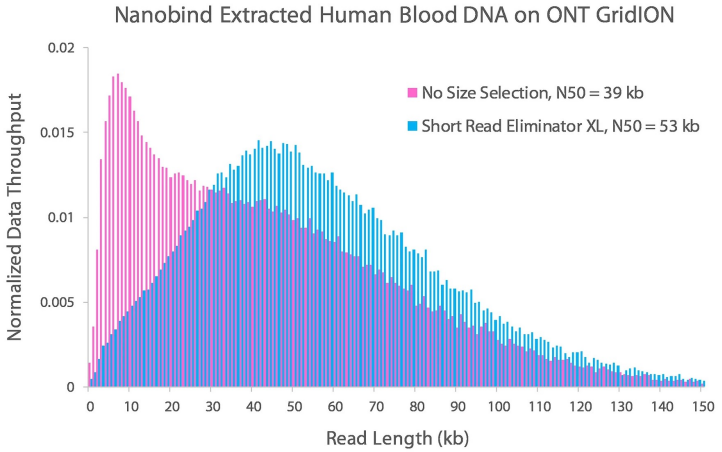
- Shorter reads = worse assembly
- Each DNA end has equal chances of being sequenced → flow cell load best measured in fmol, not ng
- Shorter fragments → more ends
- Longer fragments → fewer ends
- **Therefore,** even in HMW preparations, lots of sequenced fragments will be short.
- Difficult to balance multiplexed samples



Improving read length

HMW DNA “spooling”

Bluepippin automatic size selection



Circulomics short read eliminator (PEG precipitation)

Qiagen genomic tip

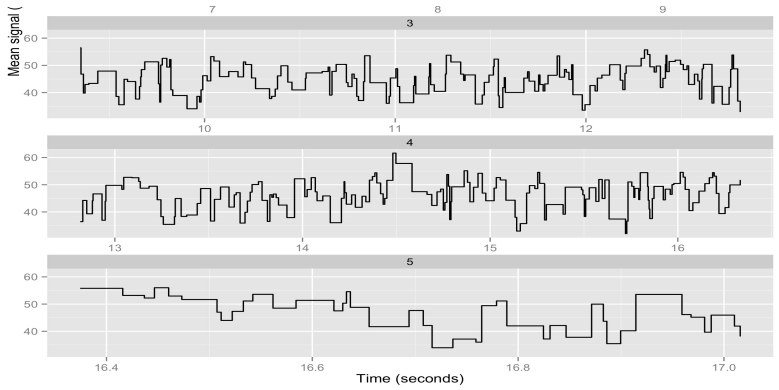
“whale watching”

Ultra high molecular weight → few ends → reduced throughput

In-silico selection of long reads to optimize assembly (if coverage high enough)

Sequence processing (genomes)

fast5



Basecalling

Machine learning
(neural networks)
guppy

fastq

```
@HWI-EAS440_0386:1:23:17547:1423#0/1
TACGNAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTTTAAGT
+
hhhdHddddddddfehhfhhhghggfhhhfhhgggfhhgfgdfcfhehfdgfhggfggfggffgc

@HWI-EAS440_0386:1:23:14818:1533#0/1
CCCCNCAGCGGCAAAAATTAAAATTTTTACCGCTTCGGCGTTATAGCCTCACACTCAATCTTTTA
+
US[QCQSZPHNYIa[^aaccaac[]accBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBE

@HWI-EAS440_0386:1:23:14401:1629#0/1
TACGNAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGCGGACGCTTAAGT
+
fffbF`bbWZZlZggggfgcgggggddggdgggggdbgdddgbgeggg`ffdggeabba`aaad]
```

Assembly
flye

fasta

```
>contig_1
GGTGCTCGTACCAGCGGCGCGTCTGCTCGCGCAGCTCGTCGATGATCCCCGGGGTGATGACGTATC
TACGAGGAAGGCGCGTCGGCGGGCTCGGGCTTCTCCACCAGGCCCGTGATGCGCAGCGTGCCGCCG
GACGATCGGCACGCCGTACCGCTGGGACTCACCCGGTCCATCGGCATCAGGGCCAGCACCGGGCAG
GGCACGGATCAACTGCTGGGCGCGGGGACTTCGGCGACGAACACGTCGTCCGGCCAGAGCACCAGC
GATGTTGCGGGCCGCGTTGAGCACCGGGGTGCCGTTGCCGTACGGACCGTGCTGGTCGAGGTAGGT
GGGCCAGCTCGGCGACCTCCTCGACGGCGTCCGCGTAGGCGGTCTTTGCCGGCGGCCCGCAGTTGC
GCTGGGCCGGAAGTGGTCCTGGATCAGGGACTTGCCGCCGGAGACGACGATGGTGATGTCGGTGAT
GCTCGCGCACCGTGTGCTCGATGACCGGCTTGTCGCCGACCGGCAGCATCTCCTTCGGGGTGGCCT
GCCGGGAGCCGAGTCCGGCGGCGGGGATCACGGCCCTGCGGATCGTCGGGGACATCACGTCTTCCT
ACCGCGCCGGCCGGCACGGCACGATCCGACGCTACCAACGCGGTCGGCGGCGGCAGCGGCGCCAGG
ACGGCTTCGGCCTCCCGGCCCGCCCCGGACCCCCTCCGGGTGCGGCCGGTGACCGCTCAGGACTGG
CTGCTGCTGCGGCTGCCCGTAGGCCTGAGCCCCGTGGGCGCCCGCGGCCTGGGCCCTGCAACTGCT
TTCAGCTTCGCTTCCGCACCGCAGAACGTGCACTGGGTGCTGTGGCTGGTCGAGATCGGGAACAGC
GGTGAACTTCGTGACCCGCTTGCGGAGCGTGTGCGCGGCGGGGGTTGCCGCAGCGGCTGCACACCA
AGCTGGTACAGGTAGCCCCTGGTACCAAAGATGATCATATTGTCGTCCTTCCCGGGTCGGTGCTGT
CGGGCGGCGATGCGATCACTGACTTCGGGTCTCAGCTCTCCAGCGCCTGCCGCAGGCCGGGGTAGC
TCGATGCCGACCATCCGCATGATGCGTACGACGGATTCCCGCGCCCCGGCCAGGCGCAGCCAGCCG
GGCGTGGTGAGCGGCGATCAGCGCGTTGATGCCGCTGGAGTCCATGAACGTCACACCGCTGAGGTC
GCCGGGCGGCGTCGCCGGGCGGCAGCAGGGCCTGGCCGAGGTGCTCGCGGATCGTAGTCGAACTCG
```

Error correction

Machine learning
(neural networks)
medaka

Corrected assembly

```
>contig_1
GGTGCTCGTACCAGCGGCGCGTCTGCTCGCGCAGCTCGTCGATGATCCCCGGGGTGATGACGTATC
TACGAGGAAGGCGCGTCGGCGGGCTCGGGCTTCTCCACCAGGCCCGTGATGCGCAGCGTGCCGCCG
GACGATCGGCACGCCGTACCGCTGGGACTCACCCGGTCCATCGGCATCAGGGCCAGCACCGGGCAG
GGCACGGATCAACTGCTGGGCGCGGGGACTTCGGCGACGAACACGTCGTCCGGCCAGAGCACCAGC
GATGTTGCGGGCCGCGTTGAGCACCGGGGTGCCGTTGCCGTACGGACCGTGCTGGTCGAGGTAGGT
```

Short read
polishing

Final assembly

```
>contig_1
GGTGCTCGTACCAGCGGCGCGTCTGCTCGCGCAGCTCGTCGATGATCCCCGGGGTGATGACGTATC
TACGAGGAAGGCGCGTCGGCGGGCTCGGGCTTCTCCACCAGGCCCGTGATGCGCAGCGTGCCGCCG
GACGATCGGCACGCCGTACCGCTGGGACTCACCCGGTCCATCGGCATCAGGGCCAGCACCGGGCAG
GGCACGGATCAACTGCTGGGCGCGGGGACTTCGGCGACGAACACGTCGTCCGGCCAGAGCACCAGC
GATGTTGCGGGCCGCGTTGAGCACCGGGGTGCCGTTGCCGTACGGACCGTGCTGGTCGAGGTAGGT
```

Necessary? Depends on sequencing kit & coverage

16S sequencing using nanopore

Pros:

- Full 16S sequence allows for better identification
- Practical and immediately doable

Cons

- Low read accuracy
- Few tools available and they’re all bad
- Expensive per basepair

Short-reads workflows Doable for nanopore?

1. Quality control
2. OTU generation – how to distinguish sequencing errors from true variants?
   1. Clustering with similarity cut-off (97%)
   2. Denoising with dada2
3. Taxonomic assignment

Illumina reads are bad at the ends, nanopore reads are “bad” throughout

This method obscures natural variation
nanoCLUST implementation unreliable

Difficult to implement for uncorrected reads with 90-95% accuracy!
Dadasnake only works with ~98% accuracy reads

Better than illumina!

My own implementation

Lobna Hudifa

- Target: Isolation of secondary metabolite producing Acidobacteria from Antarctic soil

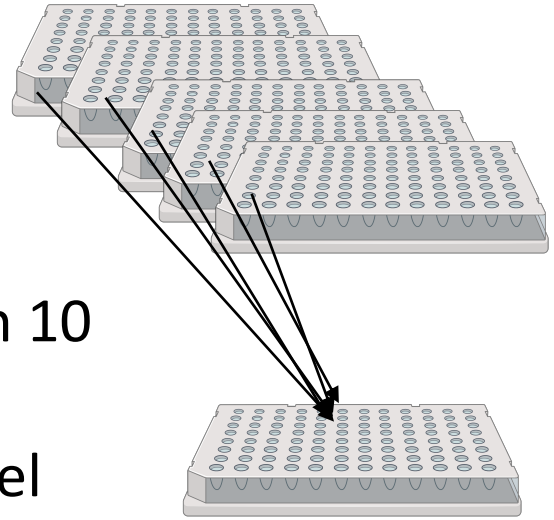
Problem:

- 900 isolates
- 16S PCR & bidirectional Sanger sequencing > £3500 & a lot of work
- MiSeq not practical – isolates slowly die!

Solution:

- 96 well pooled amplicon sequencing with nanopore:
- Reagents £120/plate (had them in freezer anyways)
- 2h of flow cell time
- Results immediate

Workflow



Barcode & normalise

Wet lab

1. Amplify full-length 16S in 10 plates
2. Check amplification on gel
3. Pool amplicons in 2 plates → 5 isolates per well
4. Barcode by PCR with kit
5. Normalise with SequalPrep plate
6. Sequence on MinION

Sequence processing

1. BLASTn all reads against 16S database with 1000bp length cutoff and keep best hit
    1. GTDB dada2 classifier file (“quick & dirty”: not very extensive, but GTDB taxonomy)
    2. NCBI 16S database (larger)
2. Compile into table & summarise genera per barcode
3. Remove genera representing <1% of barcode reads

Results

barcode	P_gtdb	C_gtdb	O_gtdb	F_gtdb	G_gtdb	taxon_read_	avg_similarit	barcode_read_	taxon_pct
barcode48_f	Actinobacter	Actinobacter	Actinomycet	Micrococcac	Arthrobacter_I	44	93.6970455	1084	4.05904059
barcode48_f	Actinobacter	Actinobacter	Actinomycet	Micrococcac	Pseudarthrobacter_A	351	94.937094	1084	32.3800738
barcode48_f	Actinobacter	Actinobacter	Streptomyce	Streptomyce	Streptomyces	142	93.5093662	1084	13.099631
barcode48_f	Proteobacter	Gammaprote	Pseudomona	Pseudomona	Pseudomonas_E	512	95.0897852	1084	47.2324723

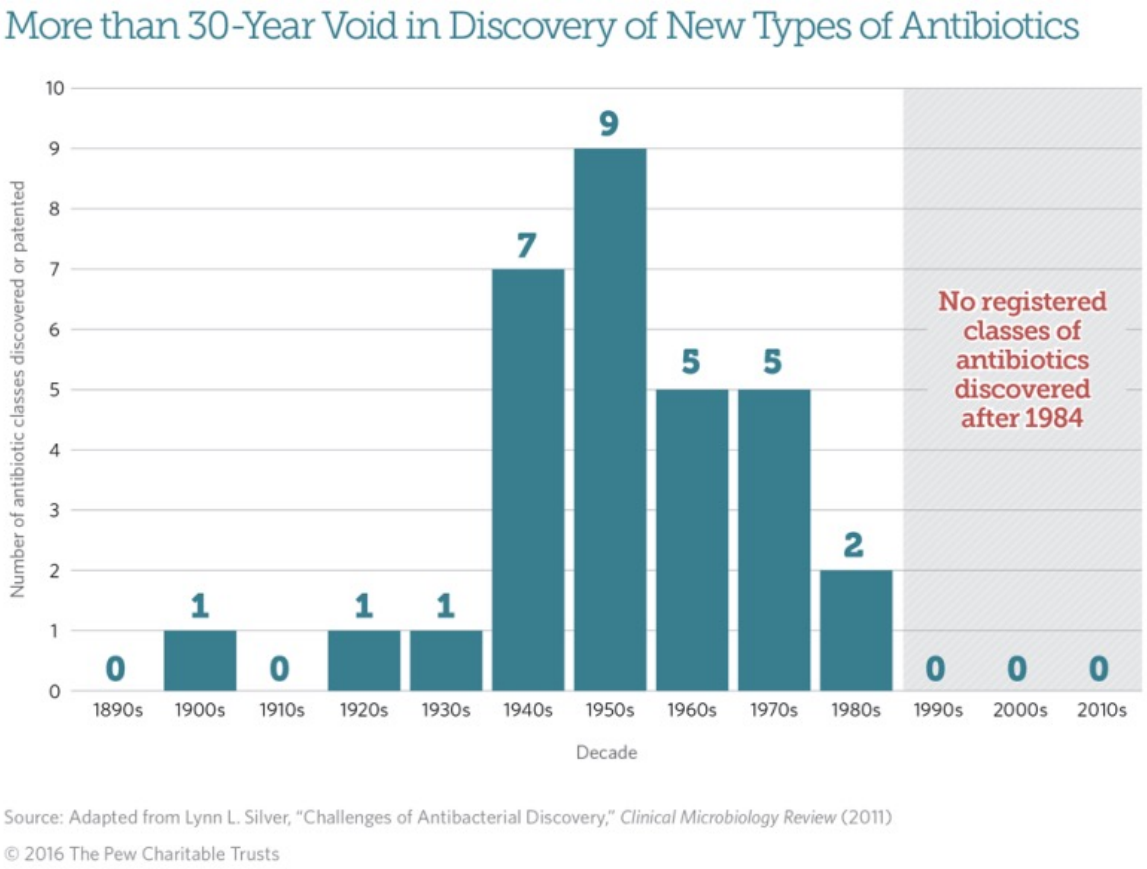
- Taxons “split” into different genera due to noisy reads (Arthrobacter_I vs. Pseudarthrobacter_A
- Difficult to distinguish low read count taxa from artifacts
- Computationally very slow
- **Enabled targeted isolation**

Considerations for usage

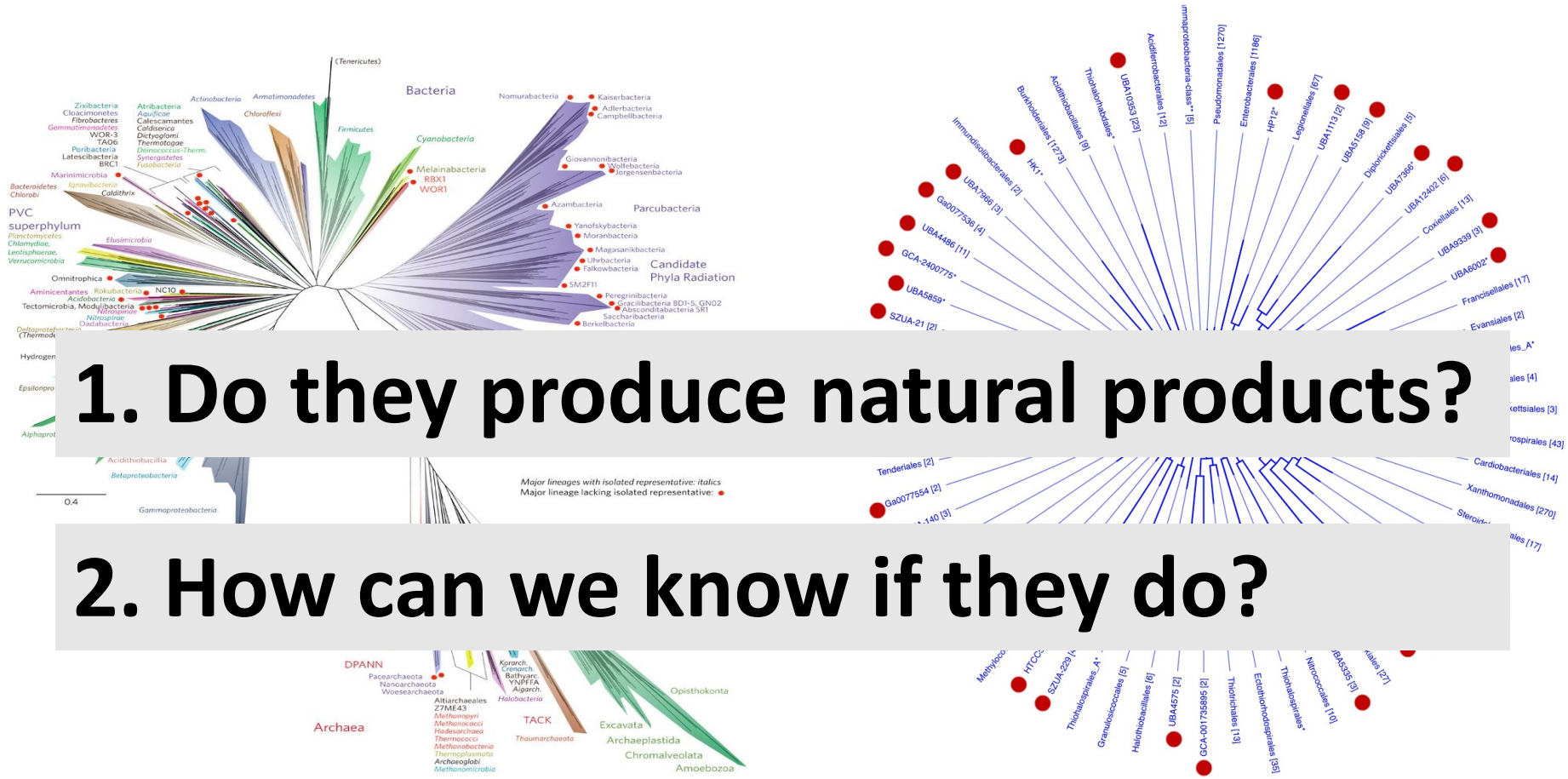
- Q20+ chemistry should make it **much more accurate** (15 errors per 1500 bp read instead of 75– 150) which *might work* with dada2?
- Probably there'll be more tools developed
- Database alignment approach can be refined & sped up (Seb?)
- Epi2me "what's in my pot pipeline" – ready-made, similar approach – but data format questionable
- MiSeq might be more appropriate for many applications still

My own work:
Long reads for metagenomic BGC assembly

New antibiotics – where from?



Uncultured Bacteria at all taxonomic levels



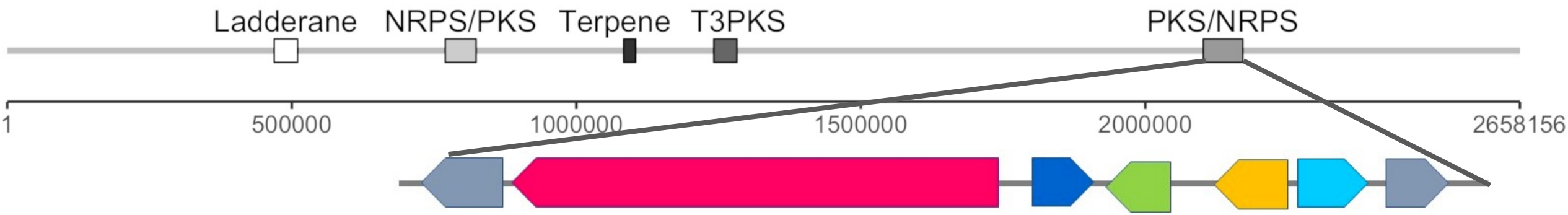
1. Do they produce natural products?

2. How can we know if they do?

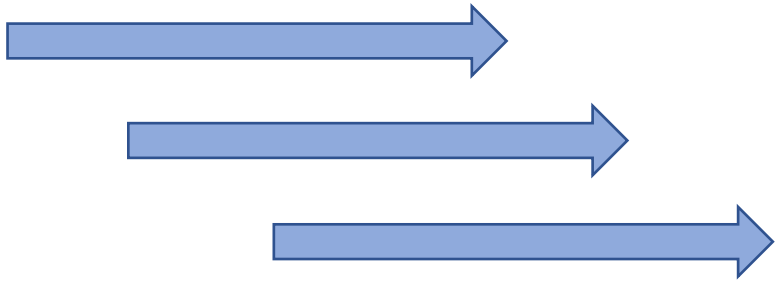
Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. (April 2016). "A new view of the tree of life".

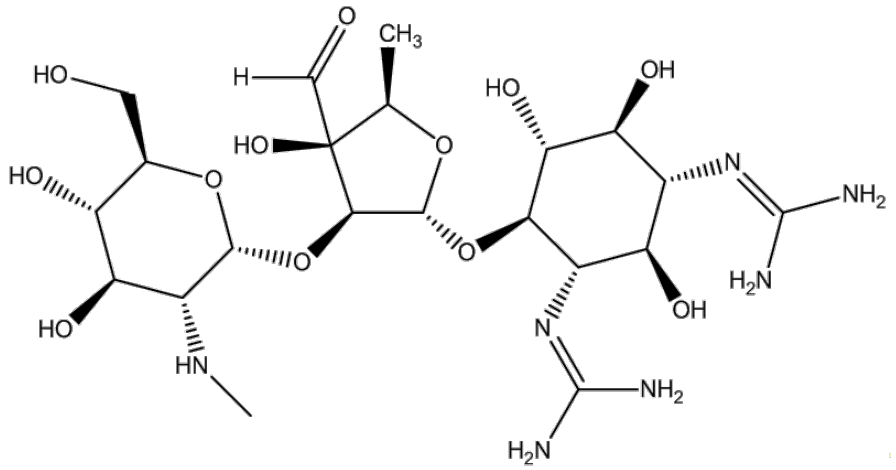
GTDB/AnnoTree: Gammaproteobacteria

Biosynthetic Gene Clusters (BGCs)



Cluster size up to >100.000 bp





BGCs from short-read metagenomes

Uncharted biosynthetic potential of the ocean microbiome

Lucas P
Alessio Mil
Dylan R. C
Georg Zell

doi: https:

**Insights into rumen microbial biosynthetic gene cluster diversity throu
genome-resolved metagenomics**

Christopher L. Anderson, Samodha C. Fernando

doi: https://doi.org/10.1101/2020.05.19.105130

Large-Scale Metagenome Assembly Reveals Novel Animal-Associated Microbial Genomes, Biosynthetic Gene Clusters, and Other Genetic Diversity

Nicholas D. Youngblut, Jacobo de la Cuesta-Zuluaga, Georg H. Reischer, Silke Dauser, Nathalie Schuster, Chris Walzer, Gabrielle Stalder, Andreas H. Farnleitner, Ruth E. Ley

Jack A. Gilbert, Editor

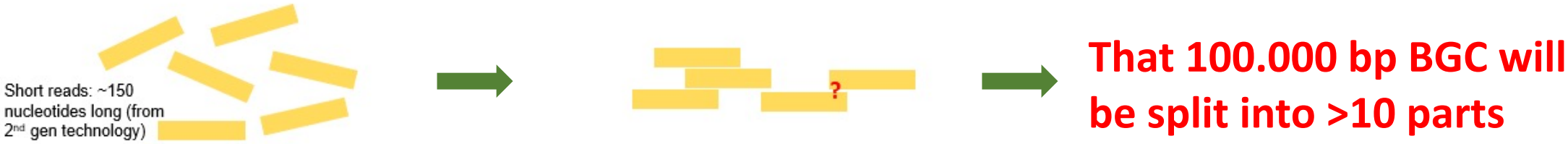
Front. Microbiol., 21 August 2020 | https://doi.org/10.3389/fmicb.2020.01950

DOI: 10.1128/mSystems.01(

Discovery of an Abundance of Biosynthetic Gene Clusters in Shark Bay Microbial Mats

Ray Chen[1,2], Hon Lun Wong[1,2], Gareth S. Kindler[1,2], Fraser Iain MacLeod[1,2], Nicole Beaud[1], Belinda C. Ferrari[1,2] and Brendan P. Burns[1,2*]

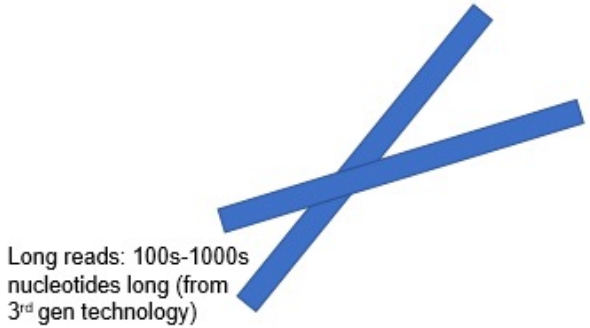
- What papers tell you: High diversity
- What they don't tell you:
  - **Huge sampling & sequencing effort needed**
  - **Almost all BGCs fragmented → analysis difficult, cloning impossible**



That 100.000 bp BGC will be split into >10 parts

Long reads for BGC assembly: Theory

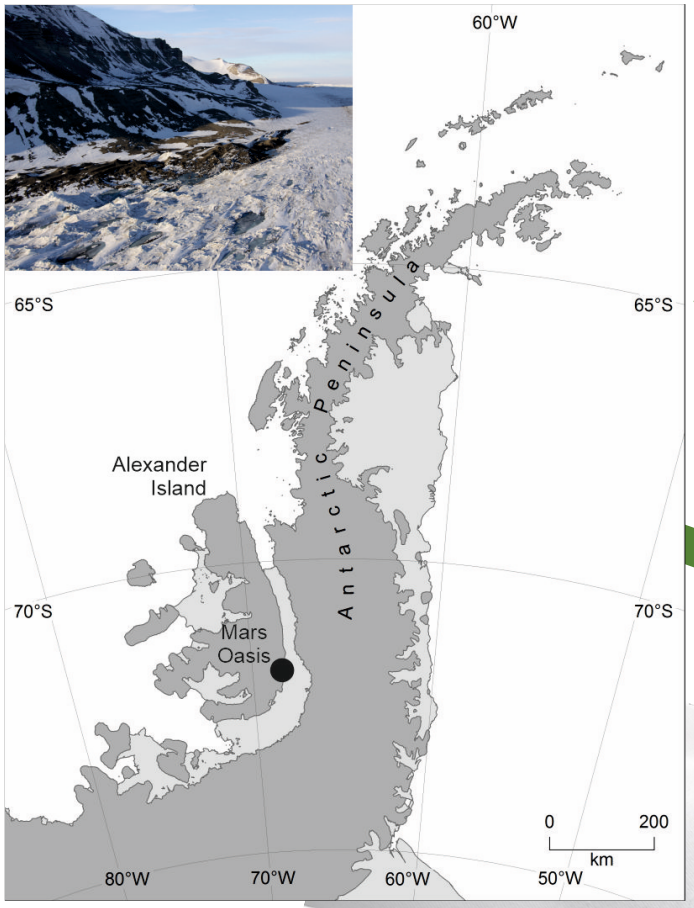
That 100.000 bp BGC will be split into >10 parts



Complete BGCs

Fewer samples & less coverage needed

Long vs short reads for BGC assembly: Results



28 Gb Illumina 150bp PE

Assembly size: circa 3 Gb
N50: 430 bp
Max: 344,761 bp

anti SMASH

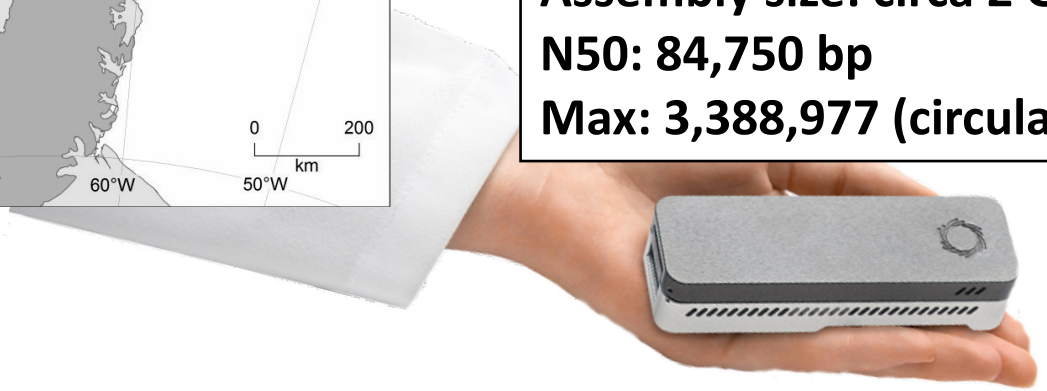
430 clusters
Mostly fragmented

44 Gb Nanopore (read N50 10kb)
(+28 Gb Illumina 150bp PE)

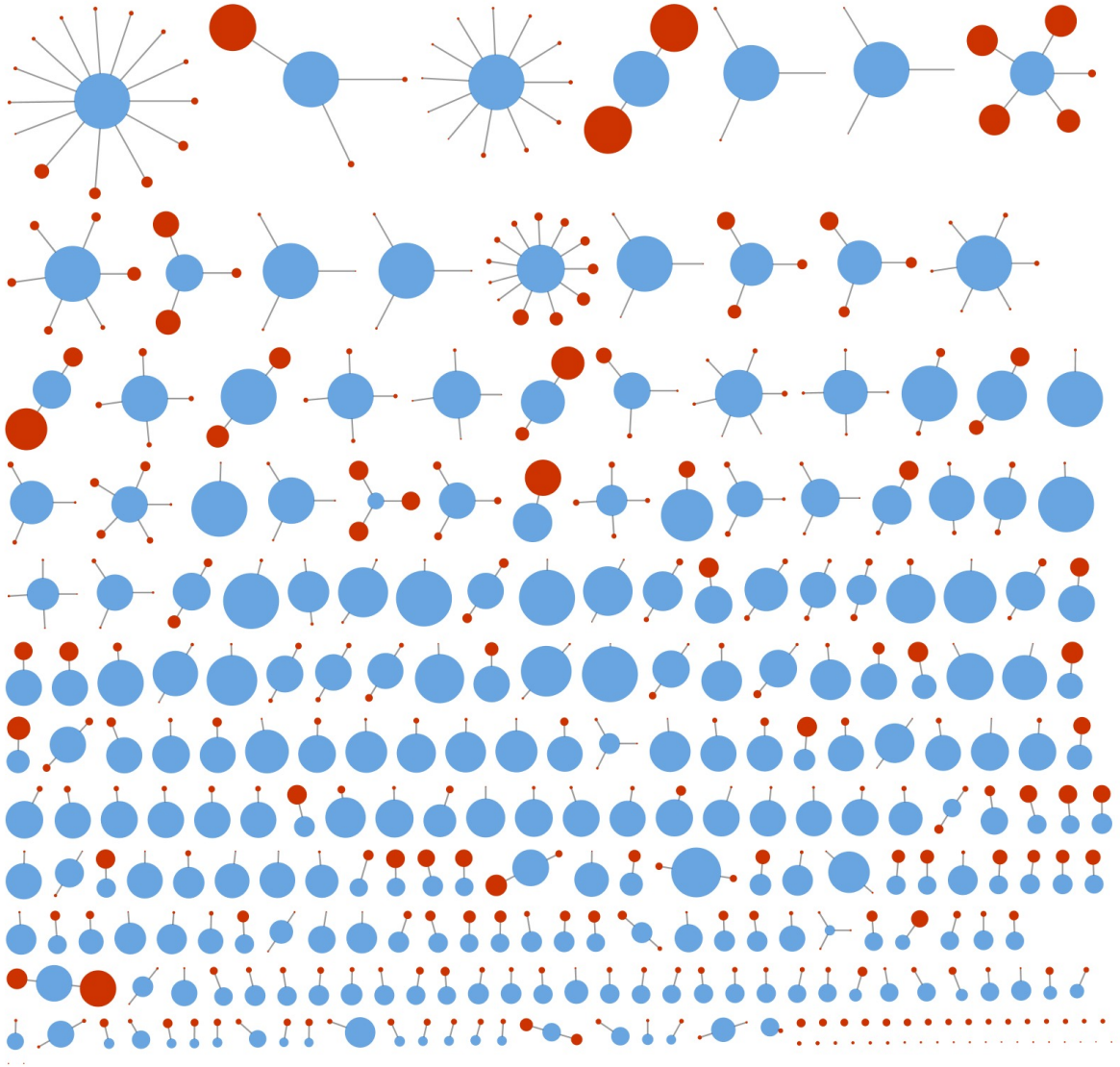
Assembly size: circa 2 Gb
N50: 84,750 bp
Max: 3,388,977 (circular)

anti SMASH

1417 clusters
>60% full length
Size up to 125kb



Long reads drastically improve BGC assembly



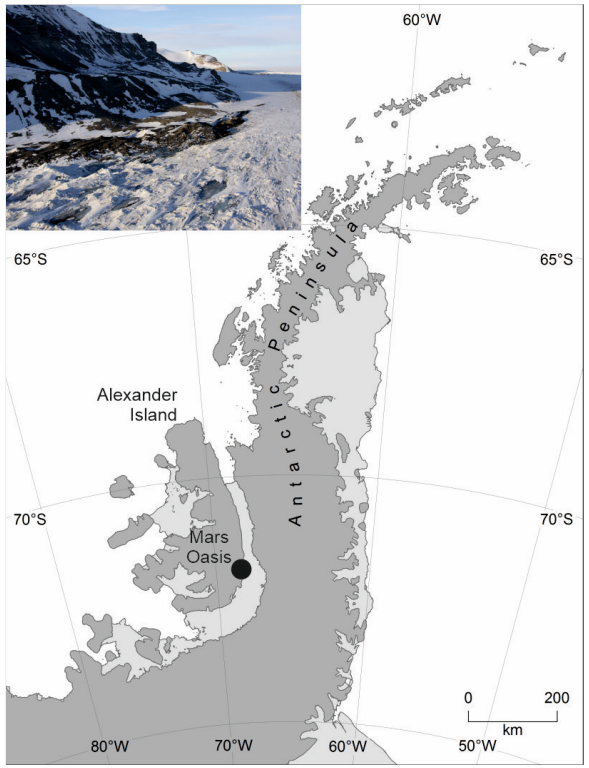
Blue = Long-read assembly BGCs

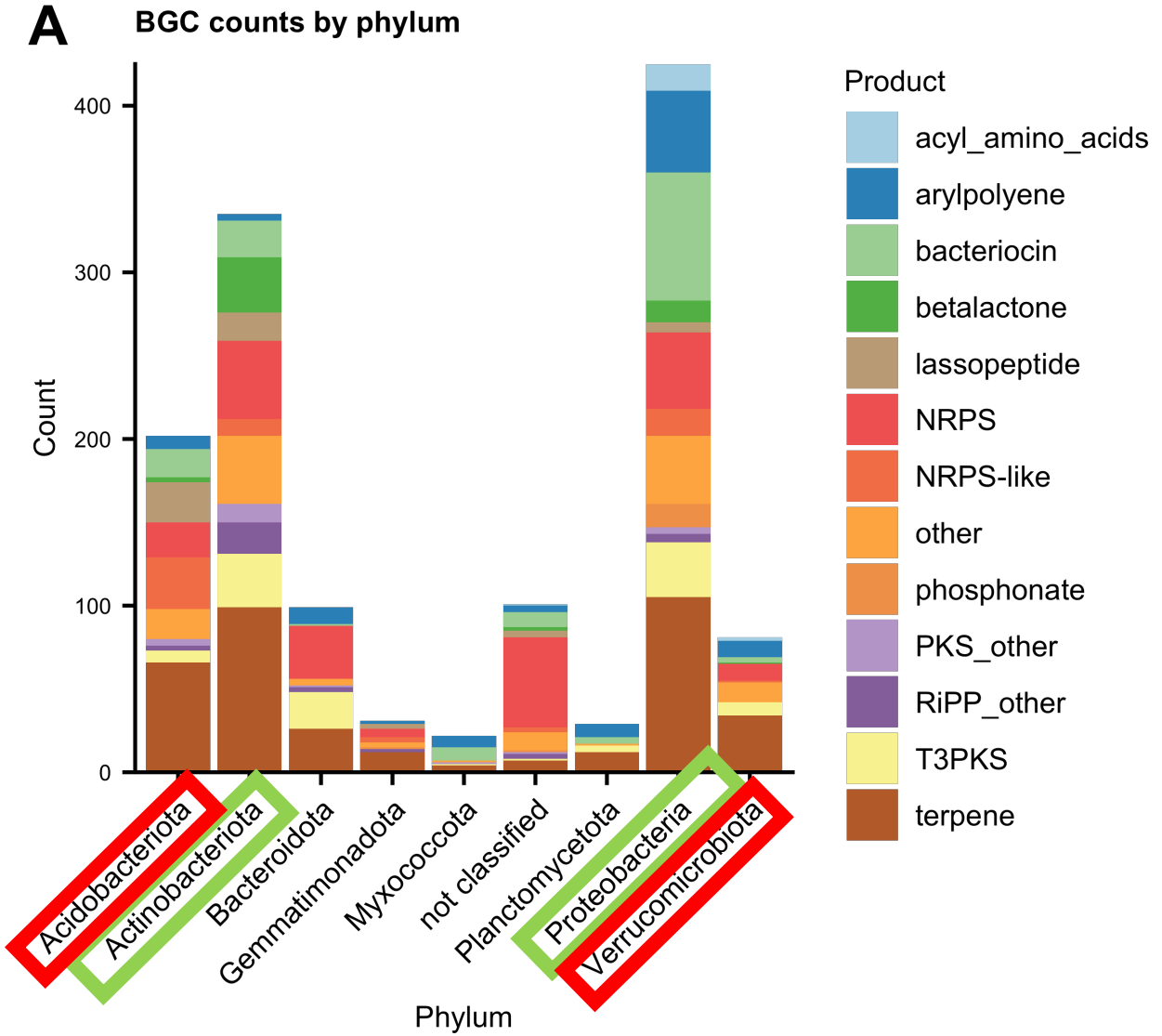
Red = short-read assembly BGCs

Size = length in bp

Connection = BLAST match

Nanopore metagenomics recovers BGCs from diverse phyla

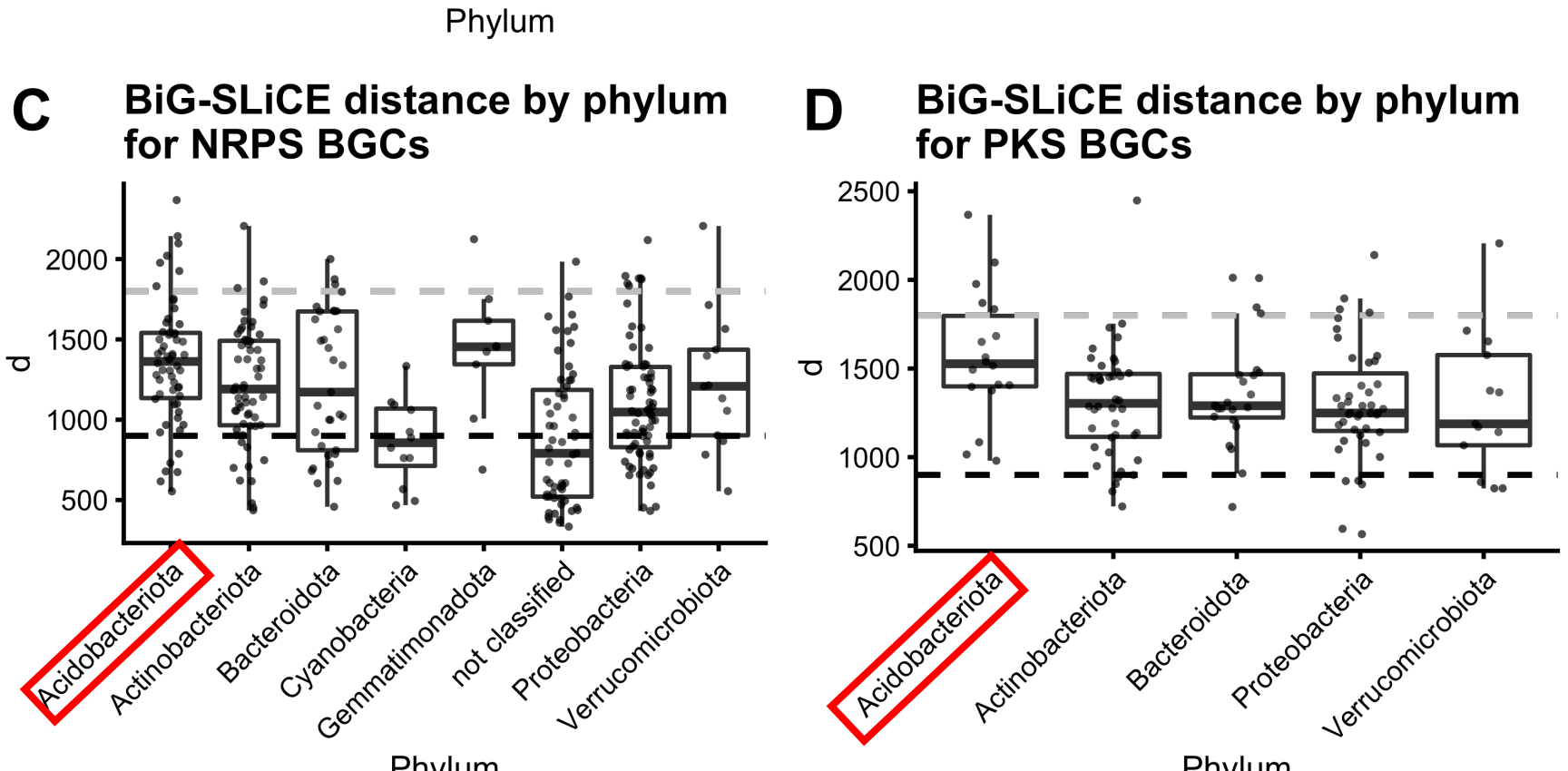




Are their BGCs any different?

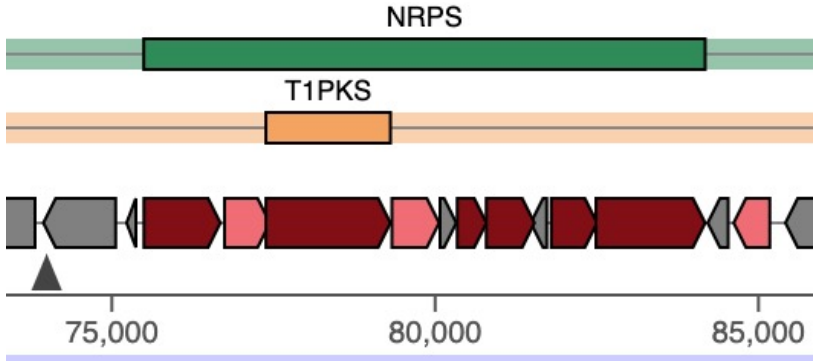
BGCs from mostly underexplored phyla are different from known BGCs

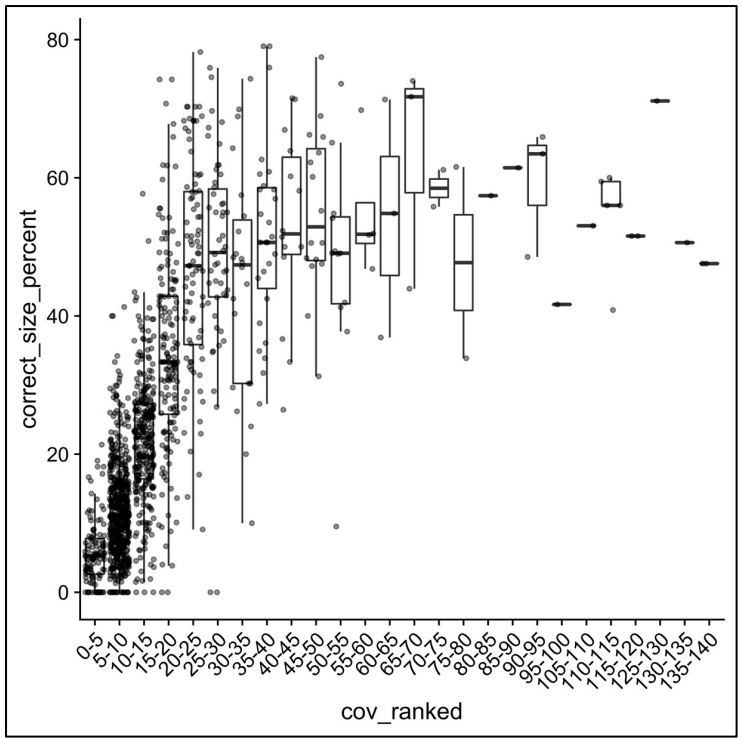
- BiG-SLiCE: Estimates distances between query BGCs and known families of BGCs
- The higher the distance d, the more different the BGC is

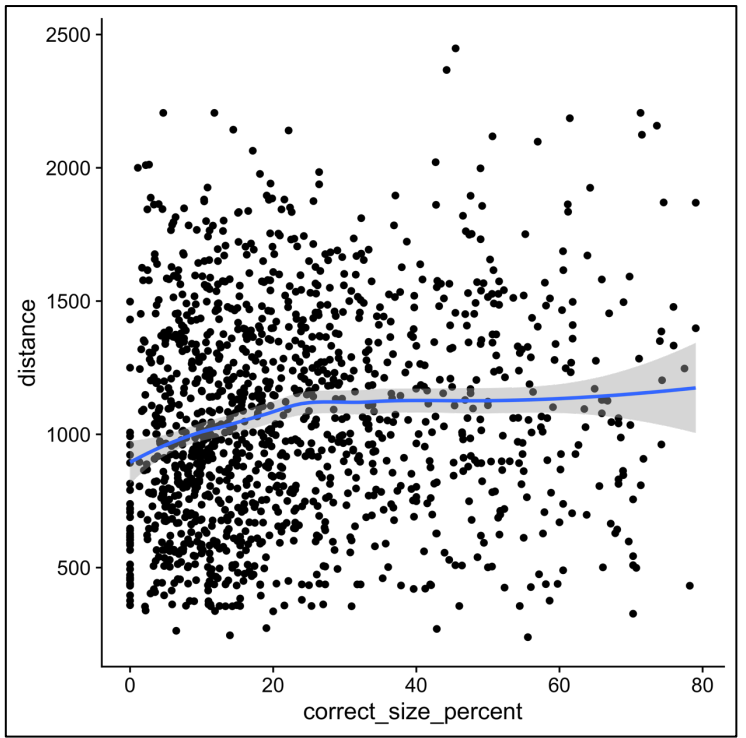


Issues with nanopore

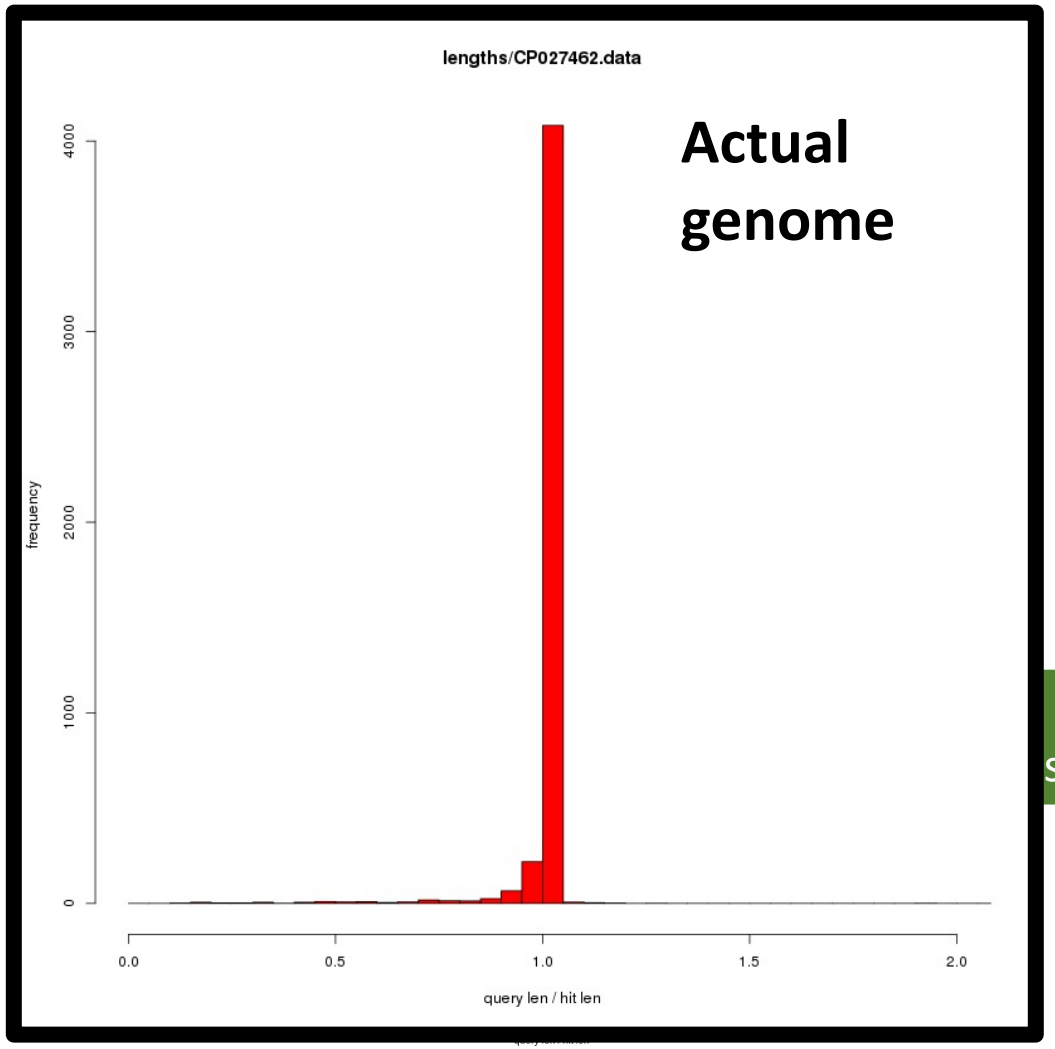
- Shallow sequencing depth in combination with old LSK109 kit leads to high error rate
  - Frameshifted proteins: >50% of proteins in assembly were split in 2 or more fragments
  - Issues for cloning
  - **Under**estimates novelty with BiG-SLiCE

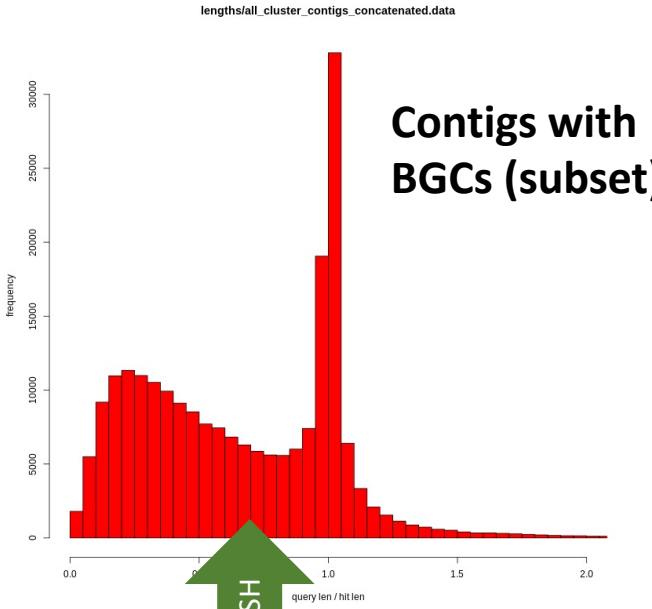




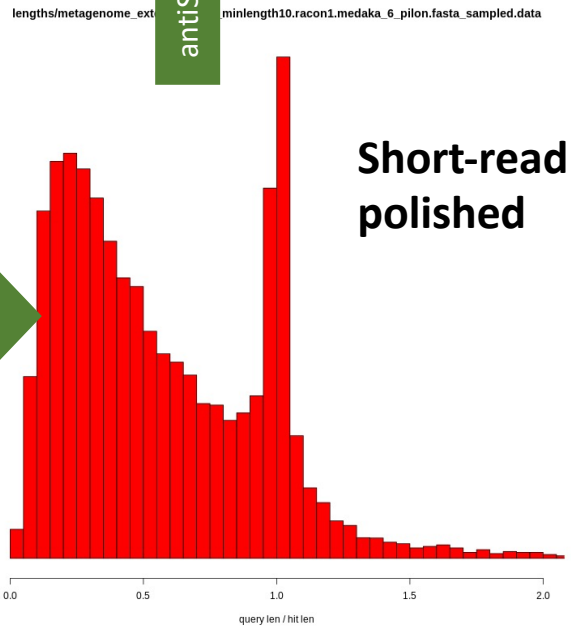


Assessing indels with ideel

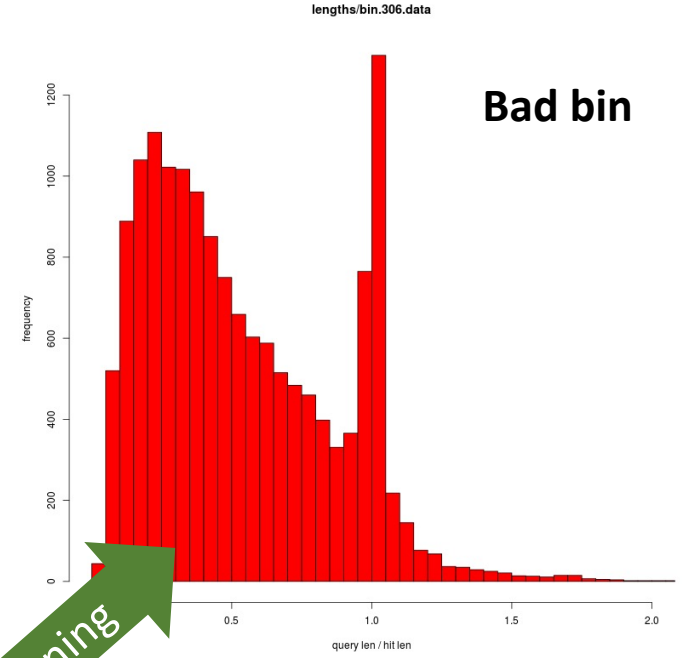




antiSMASH

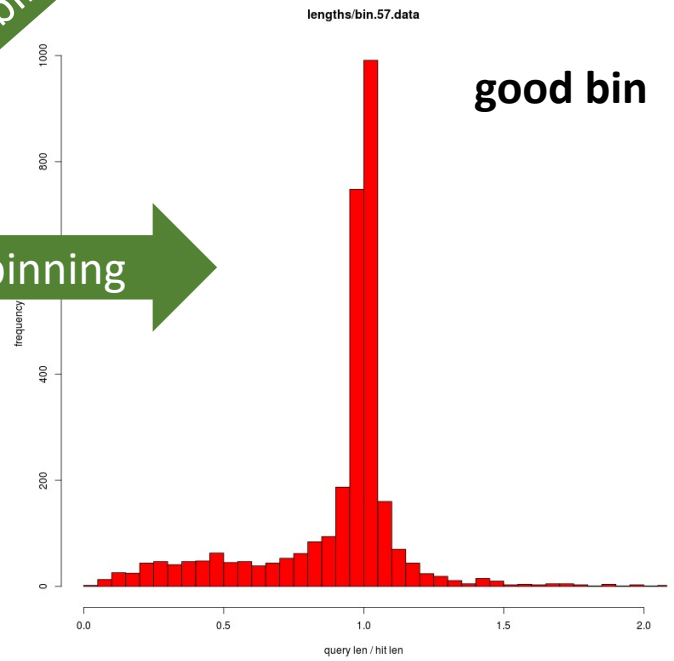


s



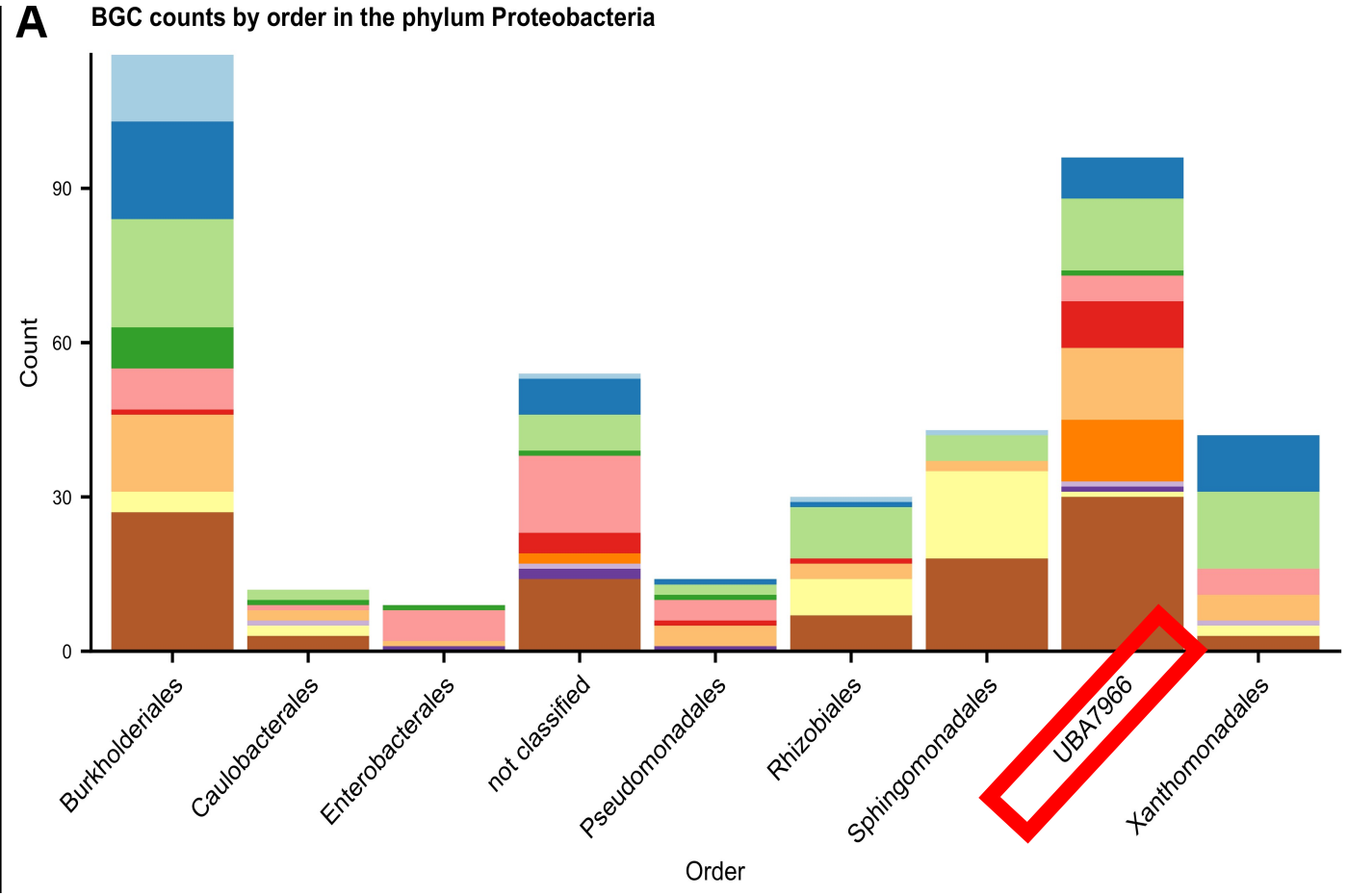
binning

binning

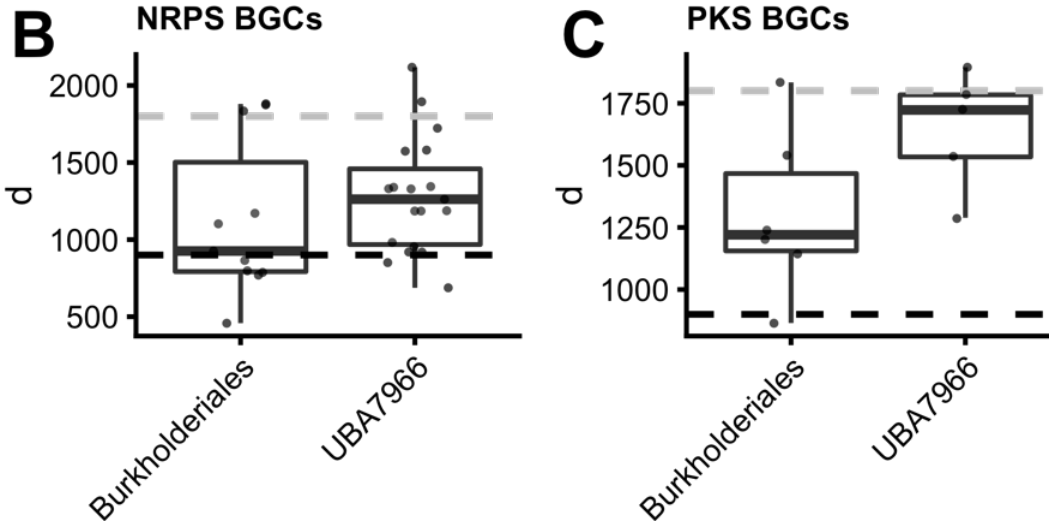


Mick Watson https://github.com/mw55309/ideel

UBA7966: Uncultured proteobacterial order



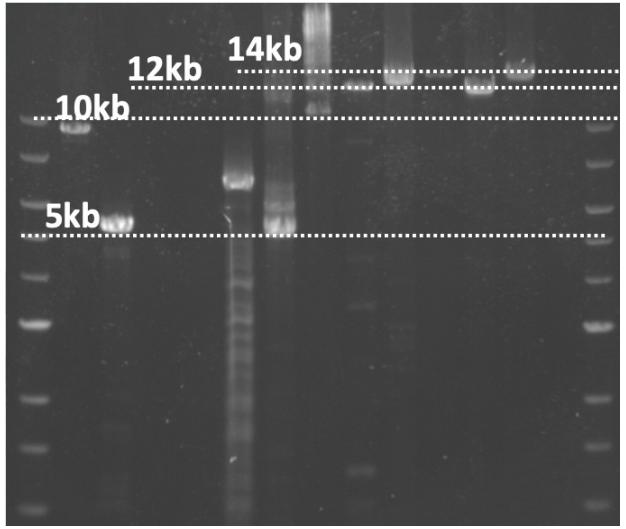
- Highly abundant
- Uncultured
- Methanotrophic (USCy?)



Heterologous expression

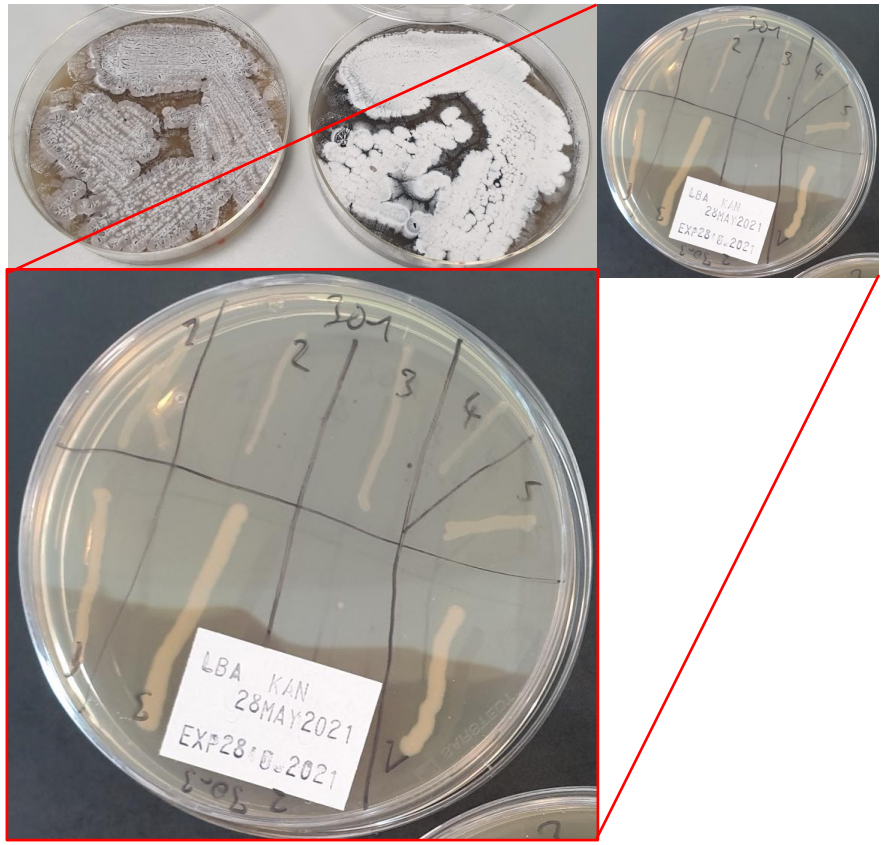
BGC amplification & cloning with SLIC (= cheap HiFi)

- 13 BGCs successful

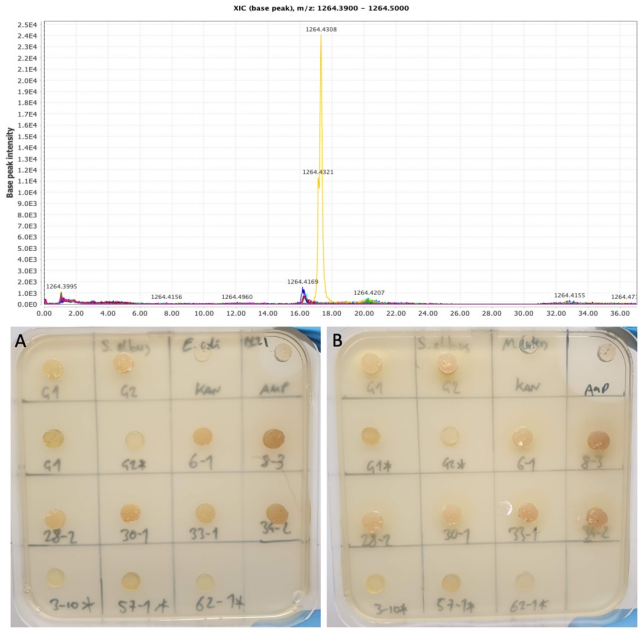


Transformation & conjugation

- *Streptomyces coelicolor*
- *Streptomyces albus*
- *Pseudomonas putida*

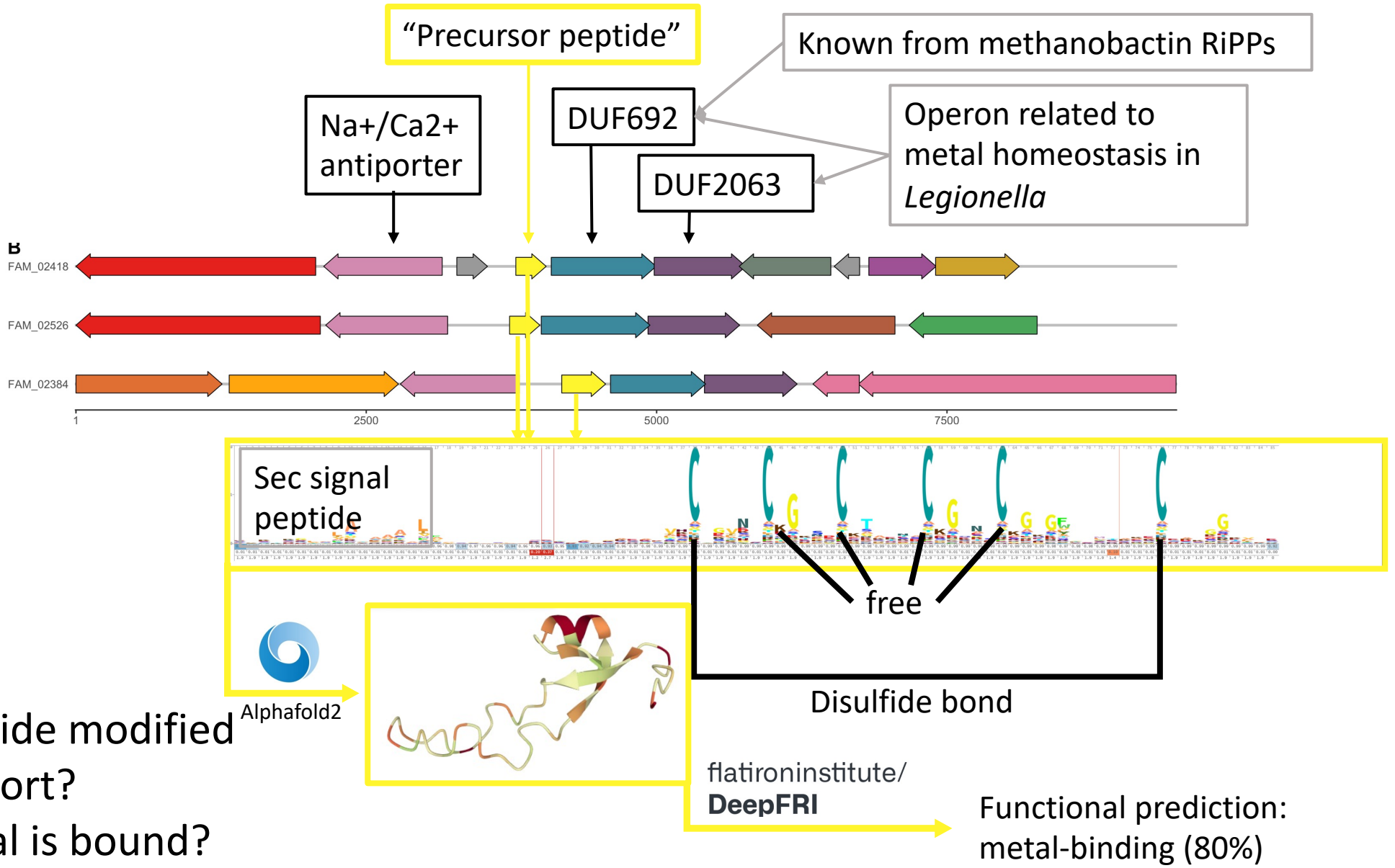


LC-MS & antimicrobial assays



- **No compounds or activities found ☹**
- **Streptomyces deleted many BGCs**
- **Constitutive promoters at fault?**

UBA7966: Potentially novel bioactive peptides



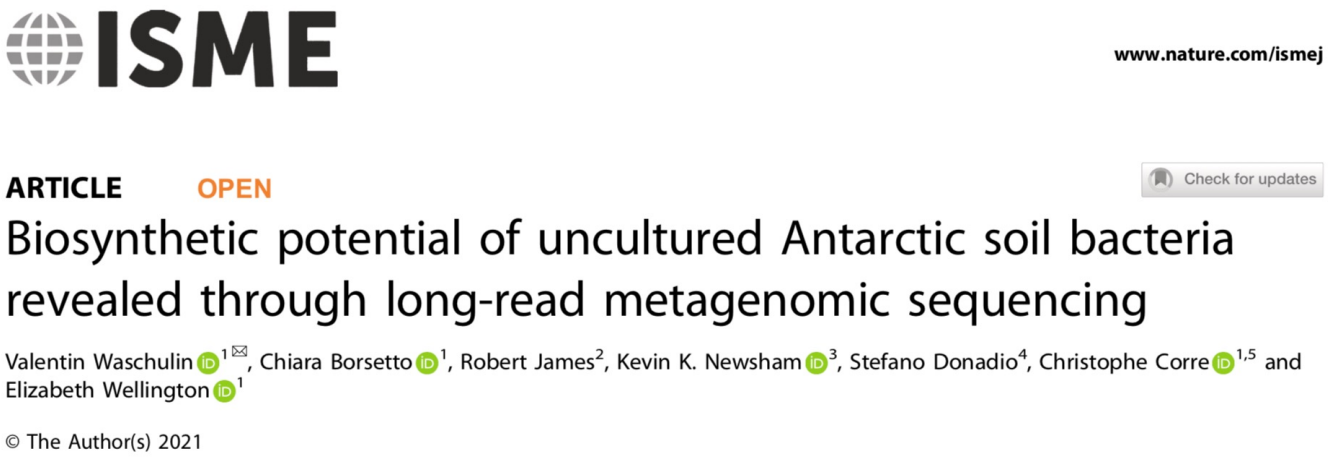
Questions:

- Is the peptide modified before export?
- What metal is bound?

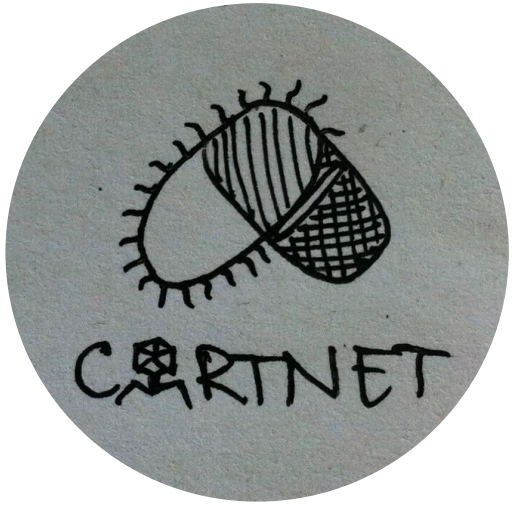
Conclusions

1. Long reads: Great for metagenomic BGC analysis
2. Novelty found on all levels
3. Potential metal-binding peptides in UBA7966 methanotrophs

Future work: Expression & analysis of the UBA7966 peptides



Thanks! Questions?



@vwaschulin

"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 765147".

PRACTICAL NOTES FOR TOMORROW

1. Workshop will place on virtual machines on Seb's server. If using windows, please download PuTTY or another ssh client. Mac & Linux: no client needed

2. Two groups possible:
    1. Workshop option 1: Nanopore WGS assembly, polishing, QC with Valentin
    2. Workshop option 2: Genome-resolved metagenomics with Seb