

Enabling SMEs to harness the power of big data

From data to real world products

22nd April 2021

Bioinformatics For BioBusiness

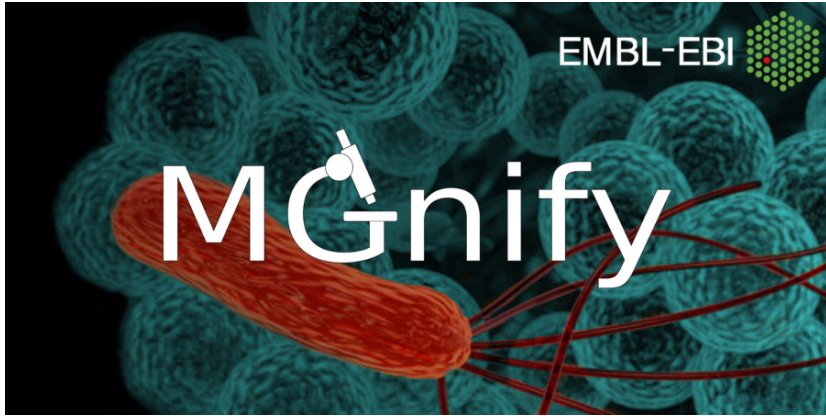
Rob Finn (rdf@ebi.ac.uk, @robdfinn)

European Molecular Biology Laboratory

European Bioinformatics Institute (EMBL-EBI)

Broad range of microbiomes studied



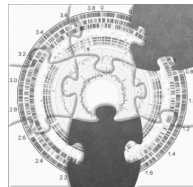


A **free** to use resource for the archiving, assembly, analysis, & browsing of microbiome data

Data archiving



Assembly



Analysis

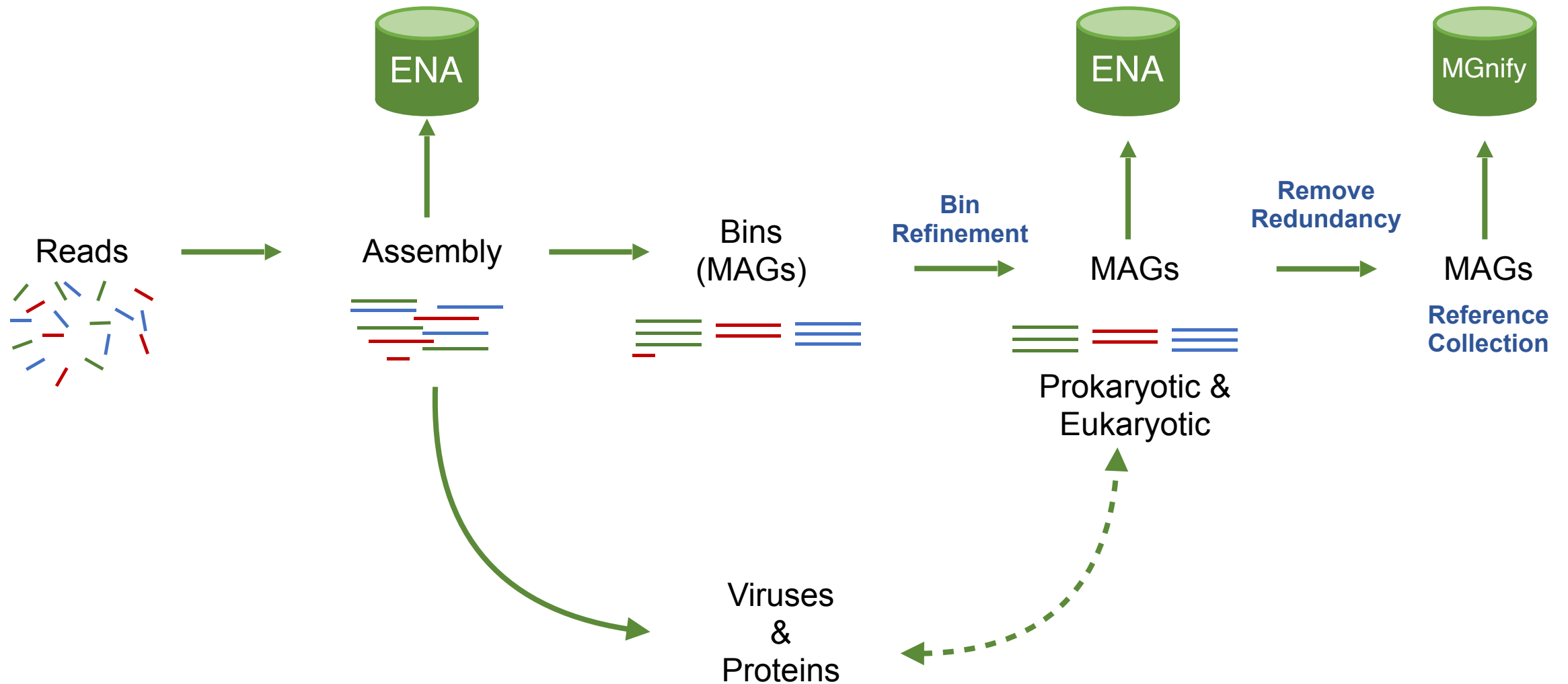


A screenshot of the MGnify website's search interface. It features a search bar at the top with the text "Submit, analyse, discover and compare microbiome data". Below the search bar, there are navigation tabs for "Overview", "Submit data", "Text search", "Sequence search", "Browse data", "API", "About", "Help", and "Login". The main content area is titled "Search by" and offers options to search by "Name, biome, or keyword" (Text search) and "Sequence similarity" (Sequence search). It also provides "Or by data type" with a list of categories: amplicon (101207), assemblies (6733), metabarcoding (1601), metagenomes (17545), and metatranscriptomes (1727). There are also options to search by "selected biomes" with icons for Human (49667), Digestive system (30766), Aquatic (11542), Plants (10395), Soil (10142), Marine (8725), Digestive system (6592), Skin (4337), Wastewater (2472), and Food production (1245).

A screenshot of the MGnify search results interface. It shows a search bar with "Search" and "Clear all" buttons. Below the search bar, there are tabs for "Studies", "Samples", and "Analyses". The results section displays a table with columns for "Analysis", "Pipeline Version", "Sample", "MGnify ID", and "Experiment Type". The table shows results for a search of "Temperature (°C)" and "Depth (meters)". Below the table, there are two charts: "Reads length histogram" and "Reads GC distribution". The "Reads length histogram" shows the number of reads versus sequence length (bp) with a peak around 400 bp. The "Reads GC distribution" shows the number of reads versus GC content (%) with a peak around 45%.

A screenshot of the GO Terms annotation interface. It displays a summary of Gene Ontology (GO) terms derived from InterPro matches. The interface includes a "Switch view" dropdown and an "Export" button. Below these, there are three pie charts: "Biological process", "Molecular function", and "Cellular component". The "Biological process" chart shows categories like "metabolic process", "transport", and "cellular metabolic process". The "Molecular function" chart shows categories like "nucleic acid binding", "catalytic activity", and "protein binding". The "Cellular component" chart shows categories like "membrane", "cytoplasm", and "ribosome".

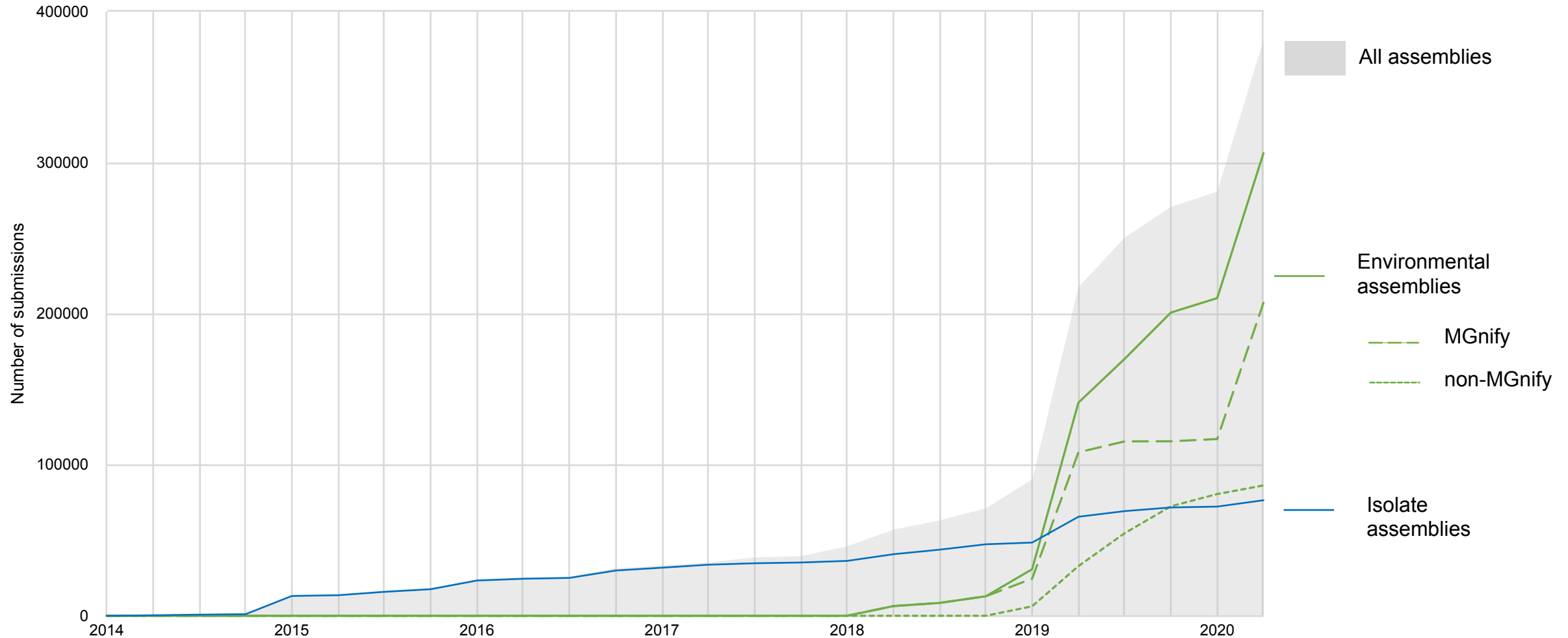
From raw metagenome reads to assemblies to genomes



MAGs = Metagenome Assembled Genomes

What is big data to me?

Cumulative number of assembly submissions to ENA dependent on 'assembly type'



Knowledge from this big data?

NEWS

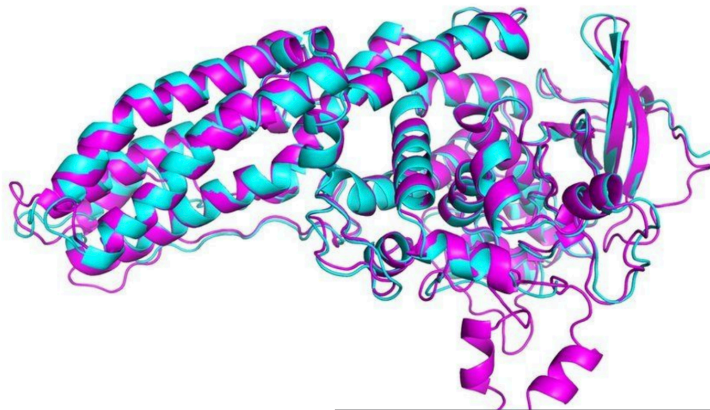
Home | Coronavirus | Brexit | UK | World | Business | Politics | Tech | Science | Health | Family & Education

Science & Environment

One of biology's biggest mysteries 'largely solved' by AI

By Helen Briggs
BBC science correspondent

30 November 2020



CASP/DEEPMIND/VTAGLIABRACCIDTOMCHICK,UT SOUTHWESTE

A DeepMind model of a protein from the Legionnaire's disease bacteria (Casp-14)

One of biology's biggest mysteries has been solved using artificial intelligence, experts have announced.

<https://www.bbc.co.uk/news/science-environment-55133972>

- AlphaFold2 storms to CASP14 victory
 - Shows potential for ML & big data
- Why?

Knowledge from this big data?

NEWS

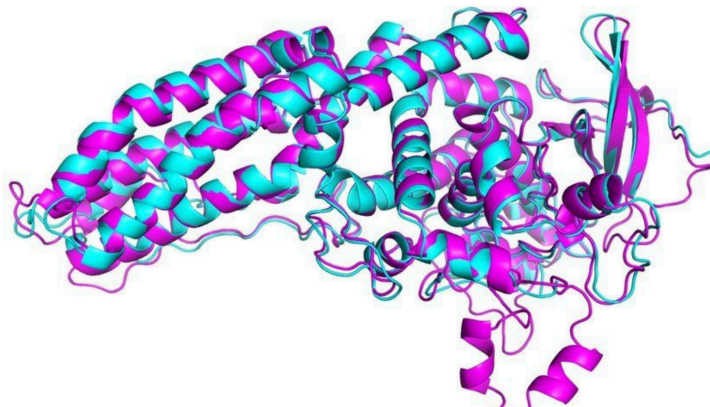
Home | Coronavirus | Brexit | UK | World | Business | Politics | Tech | Science | Health | Family & Education

Science & Environment

One of biology's biggest mysteries 'largely solved' by AI

By Helen Briggs
BBC science correspondent

30 November 2020



CASP/DEEPMIND/VTAGLIABRACCIDTOMCHICK,UT SOUTHWESTE

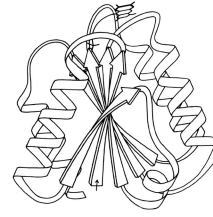
A DeepMind model of a protein from the Legionnaire's disease bacteria (Casp-14)

One of biology's biggest mysteries has been solved using artificial intelligence, experts have announced.

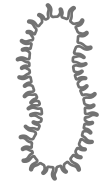
<https://www.bbc.co.uk/news/science-environment-55133972>

- AlphaFold2 storms to CASP14 victory
 - Shows potential for ML & big data
- Why?
 - MGnify's billions of proteins

Types of industrial interactions with MGnify



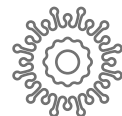
Proteins



Bacterial Genomics



Eukaryotic Genomics



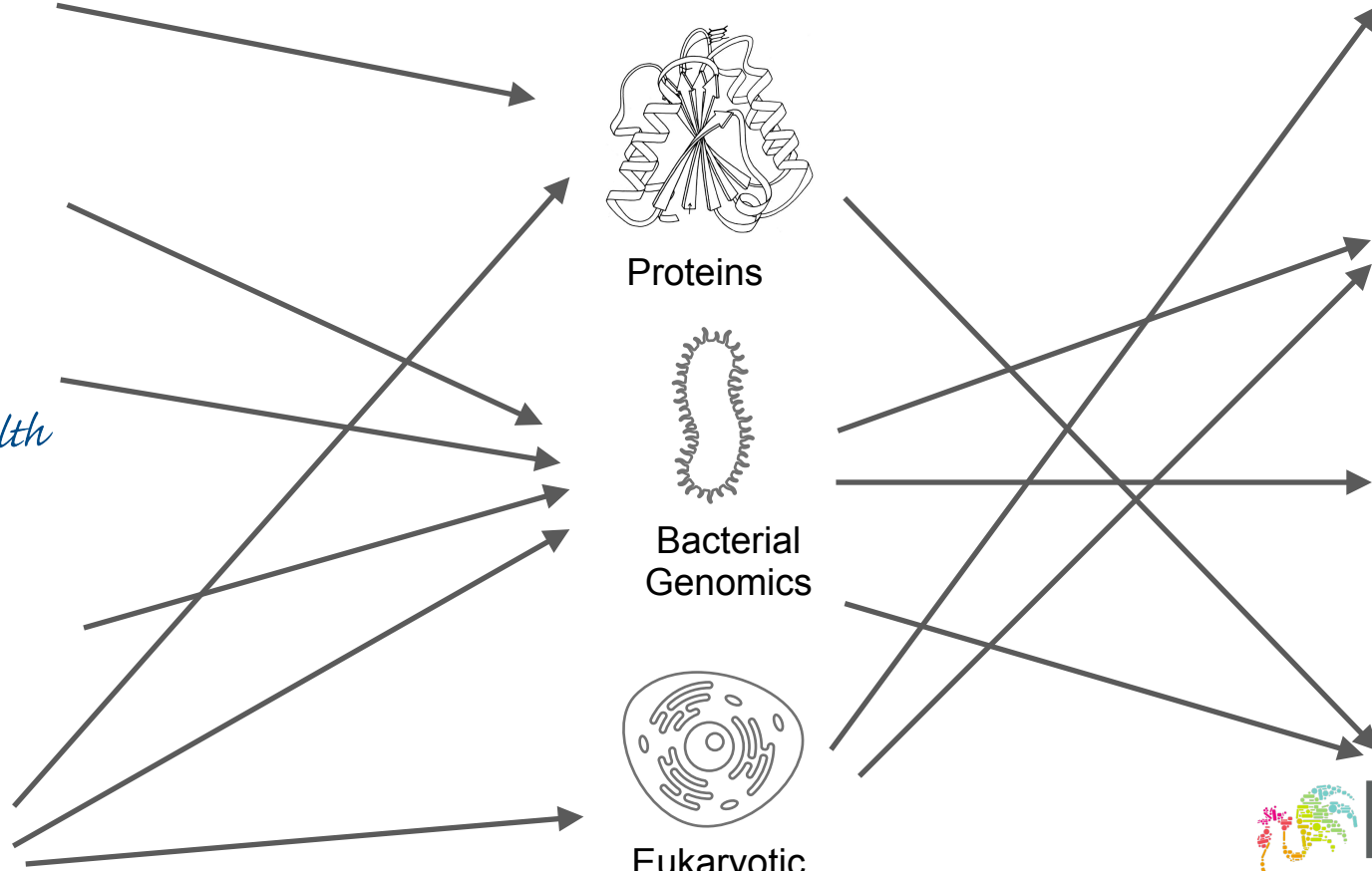
Viruses & Phage

EMBL-EBI Industry

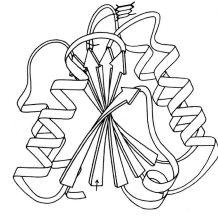
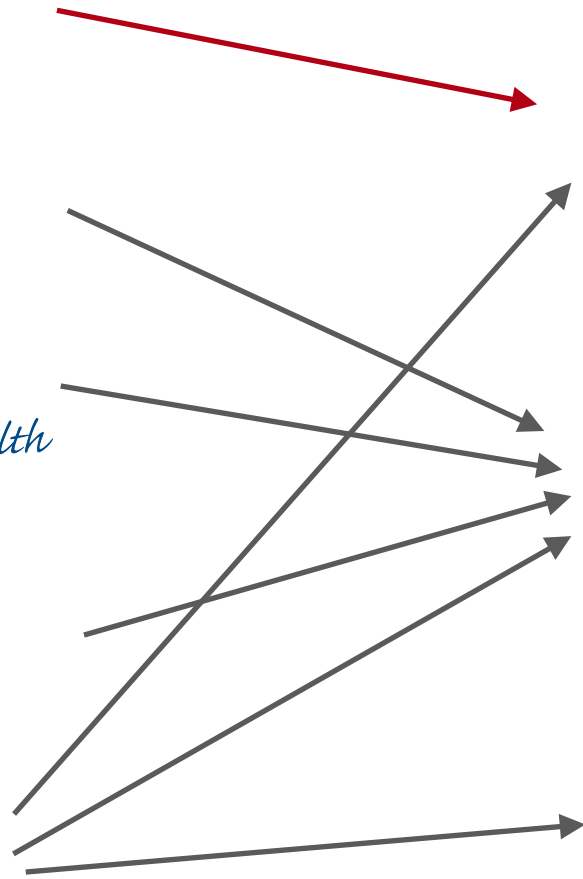
EMBLEM Project

Funded Internship

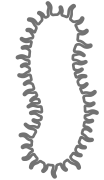
Grant Funding



Types of industrial interactions with MGnify



Proteins



Bacterial Genomics



Eukaryotic Genomics



Viruses & Phage

EMBL-EBI Industry

EMBLEM Project

Funded Internship

Grant Funding





Speciality
Enzyme Range
(Kg – Tonnes)



Customised Enzyme
Development &
Manufacture

Project Outcomes

- Database of ~ 300 million proteins derived from metagenomic samples provided by MGnify group
- Sequences linked to sample metadata enabling sequence search result refinement (e.g., by temperature, pH, etc)
- Coupled to in-house data and selection system as part of Biocatalysts' MetXtra platform

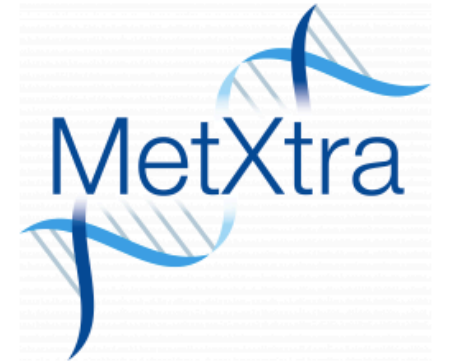
Biocatalysts Ltd and EMBL-EBI Develop New Discovery Platform for Rapid, Rational Enzyme Selection

Only a tiny fraction (~1%) of all microorganisms can be grown in the laboratory, so the potential of 99% of the world's microorganisms has been locked away. Now, metagenomics techniques make it possible to sequence entire environmental samples at once, avoiding the need for laboratory culturing before DNA sequencing. This opens the doors to a treasure trove of genetic sequence data. But because of the very large volumes of data produced, analysis is the major bottleneck to exploiting metagenomic collections for enzyme discovery.

Biocatalysts Ltd have worked with The European Bioinformatics Institute (EMBL-EBI) to develop a novel, unique software platform for data analysis called MetXtra™, launched at Food Ingredients Europe 2017 in Frankfurt, Germany. MetXtra™ is a system for identifying completely novel enzymes rapidly from large metagenomic DNA sequence libraries. This allows food industry researchers, among others, to discover unique enzymes quickly and rationally for specific applications. Food enzymes are used to improve many food production processes, for example protein hydrolysis, carbohydrate modification, flavour generation and many more.

Biocatalysts Ltd coupled their proprietary metagenomic libraries with open datasets in EBI Metagenomics (<https://www.ebi.ac.uk/metagenomics/>), a public resource for metagenomics data developed at EMBL-EBI. Mining this data resulted in a collection of over 111 million unique protein sequences, which function in diverse environmental niches (e.g. hot/cold, acid/alkaline, saline). But analysing such a huge collection manually was not practical.

To solve the problem, Innovate UK supported a collaboration between Biocatalysts Ltd and EMBL-EBI – a publicly funded intergovernmental research institute and data provider – to develop new technology for searching and analysing very large datasets automatically. The new proprietary platform, MetXtra™, enables Biocatalysts to screen metagenomic libraries for enzymes in minutes, rather than days.



Case Study - From data to real world solutions

[Back](#)

Identification of Improved Nitrilase Enzymes from Metagenomes

OVERVIEW: This project was in collaboration with [Chemoxy International Limited](#), [CPI](#) & [Northumbria University](#). Whilst this case study is focussed on a speciality chemical application, a similar approach can be taken to discover and develop enzymes for food and other non-food applications.

CUSTOMER CHALLENGE: A chemical waste stream from adiponitrile manufacture is used to produce a speciality chemical. This conversion is achieved with a nitrilase enzyme. An early feasibility study demonstrated that the conversion is complex and not sufficiently efficient for commercialisation. It was decided to probe Biocatalysts' metagenomic libraries with the aim of discovering a commercially suitable nitrilase enzyme.

APPROACH: 400 metagenomes were analysed and 1,209 hits to the probe sequence were isolated through Biocatalysts' bespoke metagenomics analysis software to produce a final set of 328 candidates. Biocatalysts' 'Design for Manufacture' (DFM) principles are incorporated to maximise the probability that can be produced in any of Biocatalysts' recombinant expression platforms and scaled up to large-scale.



MGnify
Submit, analyse, discover and compare microbiome data

Overview Search Results Help Contact

PHMMER Results [Search Again](#)

Score Download

Distribution of Significant Hits [?](#)

« First « Previous Page 1 of 170 Next » Last »

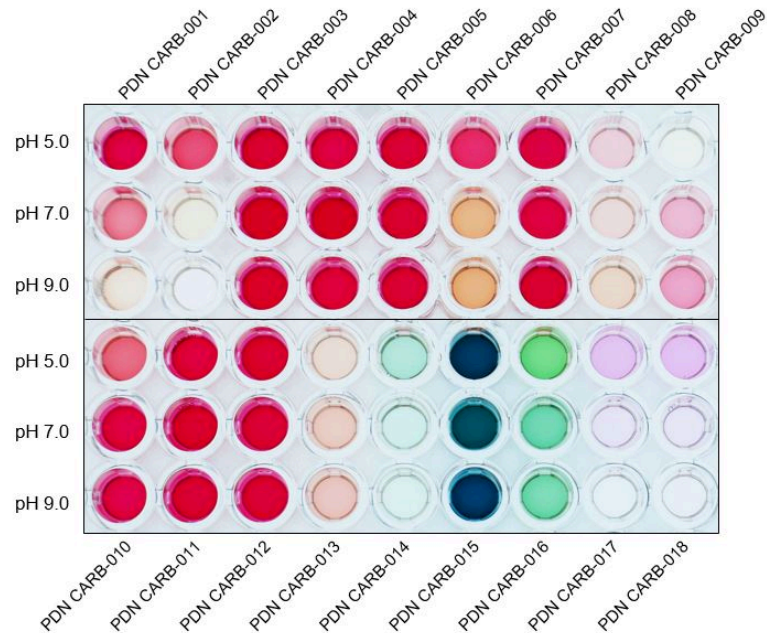
Significant Query Matches (14809) in full (v.2018_06) [Customise](#)

	Target	Run & Sample IDs	UniProt matches	E-value
>	MGYP000462796542	SRR1950710		1.9e-85
>	MGYP000014382218	SRR7054527 SRR7054529		2.1e-82
>	MGYP000267865566	SRR5754713		2.2e-82

Enzymes in the food industry

- Chemically synthesised artificial sweeteners have been used since 1950s
- Market estimated worth **\$10 billion** by 2025
- Within the “healthy-carbohydrate” market, enzymes used to breakdown complex carbohydrate, modify or build up sugars

- Xylanases
- Pectinases
- Mananases



Recognition in the field

- MetXtra platform achieves the Queen's Award for Enterprise
 - Faster response to customers
 - Cheaper as harnessing Nature's protein engineering
 - Broader spectrum of targets
 - Influence public MGnify data bundles

FOLLOW US: [f](#) [in](#)

Search the Site


Cosmetics & Toiletries
The Definitive Peer-Reviewed Cosmetic Science Resource

[Home](#) **News** > [Company News](#) [People News](#) [Event Coverage](#)

Fit for the Queen: Biocatalysts Receives Enterprise Award for Innovation

April 24, 2019 | [Contact Author](#) | Eden Stuart

[f](#) [t](#) [p](#) [in](#) [e](#)



Biotechnology company Biocatalysts Ltd. received the Queen's Award for Enterprise in the category of Innovation.

The company, which has produced specialty enzymes for a variety of industries for more than 35 years, was one of 201 companies selected out of "thousands" of applicants.

Among the Biocatalysts innovations: MetXtra, a bespoke software system capable of screening millions of sequences to identify new enzymes within hours. The program reduces the early stages of enzyme discovery from years to weeks, enabling the supply of laboratory-grade enzyme samples to customers for as little as £1,000.

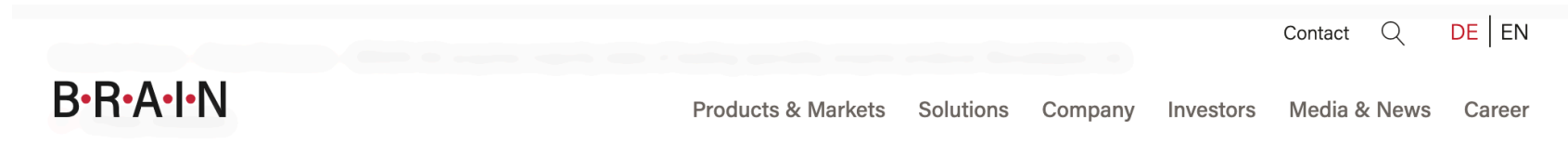
MOST POPULAR IN COMPANY NEWS

- #1 [Ipsy Introduces Ingredient Ban List](#)
- #2 [\[correction\] Kumar Organic Products Awarded for Green Isodecyl Oleate Production](#)
- #3 [The Estée Lauder Cos. Joins the Global Shea Alliance](#)
- #4 [Tata Chemicals Wins Product Innovator of](#)

<https://www.cosmeticsandtoiletries.com/networking/news/company/Fit-for-the-Queen-Biocatalysts-Receives-Enterprise-Award-for-Innovation--509014061.html>

Since the end of the project

■ Company buyout by BRAIN AG



17 March 2018, Zwingenberg (Germany), Cardiff (United Kingdom)

BRAIN AG acquires majority stake of leading speciality enzyme producer Biocatalysts Ltd.

- Strengthening of BRAIN's BioIndustrial segment through expanding access to attractive speciality enzyme markets and cutting edge enzyme production facilities
- Widening of commercial opportunities through excellent complementary portfolio and technology fit including access to BRAIN's unique BioArchive and Biocatalyst's MetXtra metagenomic library
- More effective targeting of speciality enzymes markets and growth of the global distribution network

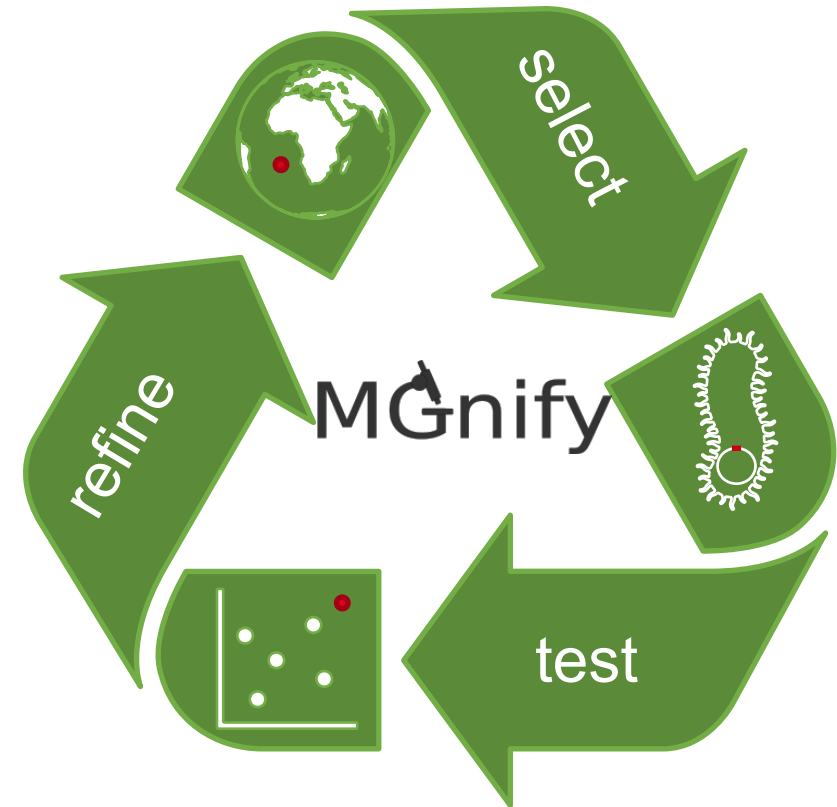
Collaborative Press Release



<https://www.brain-biotech.com/press/20180317-brain-ag-acquires-majority-stake-of-leading-speciality-enzyme-producer>

Ongoing/future academic and industrial partnerships

- Huge wealth of **novelty** arising from metagenomics
- Collaborations enables feedback from screening
 - Analysis of additional relevant datasets, e.g. extreme temperature environments
 - Expanding or refining target range
- Harness the information from protein fragments
- Data organisation and access is complex
 - Requires intimate knowledge of the data
 - Computational resources
 - Input from the community



Acknowledgements



Lorna Richardson

Alex Almeida

Germana Baldi

Ales Escobar

Martin Beracochea

Danilo Horta

Juan Caballero

Sara Kashaf

Santiago Fragoso

Felix Langer

Tanya Gurbich

Paul Saary

Varsha Kale

Gustavo Salazar

Kate Sakharova

Guy Cochrane

Alex Mitchell

Tony Burdett

& past team members

Josie Burgin

ENA Team



Mark Blight

Andrew Ellis



Birgit Kerber



The HoloFood project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 817729

