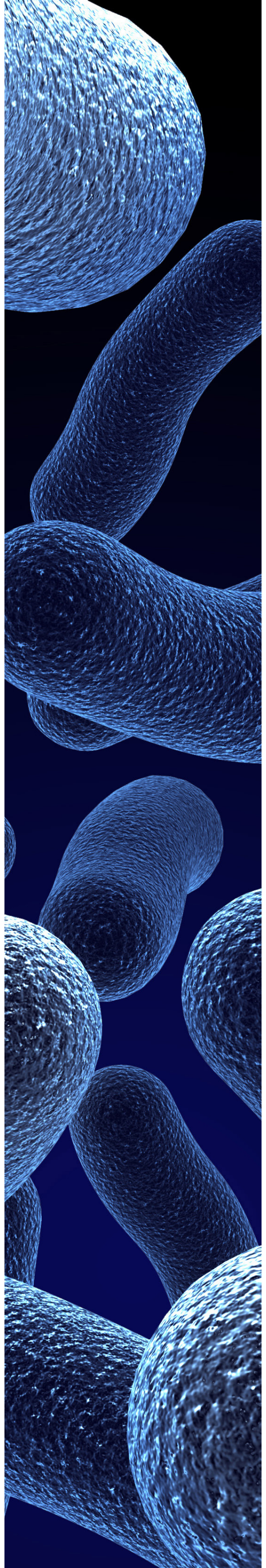
eagle® genomics

data - science - life



eagle®
genomics

Considerations & Challenges in Building an End-to-End Microbiome Workflow

Many of the data management and analysis challenges in microbiome research are shared with genomics and other life-science big-data disciplines. However there are aspects that are specific: some are intrinsic to microbiome data, some are related to the maturity of the field, with others related to extracting business value from the data.

In this white paper, Considerations and Challenges are discussed for large-scale microbiome data management and analysis. An industrialized End to End Microbiome Workflow solution is described that is already being used successfully within the life sciences industry.

data - science - life

Contents

data - science - life

eagle® genomics

1. Considerations

The considerations for successful microbiome data management and analysis fall into three main areas:

- Intrinsic factors associated with complexity
- The relatively early maturity of the field
- Calculation of business value

Intrinsic

Microbiomes are complex and diverse biological systems. Description at a molecular level includes genomes, genes, proteins and their chemical products alongside the pathways they are involved in. In terms of total information content, metagenomes are typically more complex than the genomes of their hosts in terms of length of unique DNA sequence (several-fold larger), number of unique genes (possibly a hundred-fold larger) and so on. Scale issues apart, the key intrinsic difference between metagenomic and genomic data is taxonomic diversity; a microbiome may contain thousands of different organisms. Microbiome-specific experimental workflows have been developed to profile this diversity, interactions between these organisms, and with the host. These workflows both consume and produce microbiome-specific data types. Microbiome experiments are complex with lots of variables and things to go wrong ; appropriate study design and metadata collection are crucial for meaningful analysis of microbiome data and control of batch effects.

Maturity

Next-generation technologies such as NGS that are driving the utility of microbiome research are also transforming its data landscape. Metagenomic methods that are displacing 16S-rRNA for taxonomic profiling can generate 100-fold increased data volumes. With sequence files of over 10Gb for single sample, even modest studies can generate terabytes of data to be processed, analyzed and archived. In contrast with human genomics, reliable reference data for the microbiome is limited leading to an ongoing requirement for de-novo assembly and annotation. Metagenomics is fast developing; much microbiome research software is less polished and easy to use than in more mature fields. Some fundamental steps of the data analysis workflow such as data normalization and assembly are still evolving; future data re-analysis with improved methods must be considered.

Business value

Commercial exploitation of the microbiome is in its infancy, and still at an exploratory phase. Many research/findings remain controversial and some early claims on the gut microbiome have not been replicated. The true potential of microbiome data is still to be tapped, and some novel experimental designs are being developed that result in maximum value being generated.

> **"Unilever's digital data program now processes genetic sequences twenty times faster - without incurring higher compute costs. In addition, its robust architechture supports ten times as many scientists, all working simultaneously."**
>
> Pete Keeley, eScience Technical Lead - R&D IT at Unilever

2. Challenges

The analysis-at-scale of microbiomics data faces several challenges:

- Ineffective analysis due to poor metadata management
- Large data volume and compute resources necessary for storage and analysis
- Immaturity of analysis algorithms and statistical techniques

Metadata Management

This is crucial to being able to analyze the experiment meaningfully, but is often incomplete (or in the worst case lost entirely). Scientists can find managing metadata difficult, and it is often lost and/or corrupted in the process! Eagle has developed simple methods for efficient tagging of data to ensure the correct metadata is gathered.

In addition, there is a problem of making sure the metadata has all the information that is needed to check for batch effects. Here is a list of just some of the things that can influence microbiome results: http://www.opiniomics.org/the-unbearable-madness-of-microbiome/

For this reason, one needs to be cautious about combining separate microbiome studies - it is an easy way to end up with a batch effect rather than a real result.

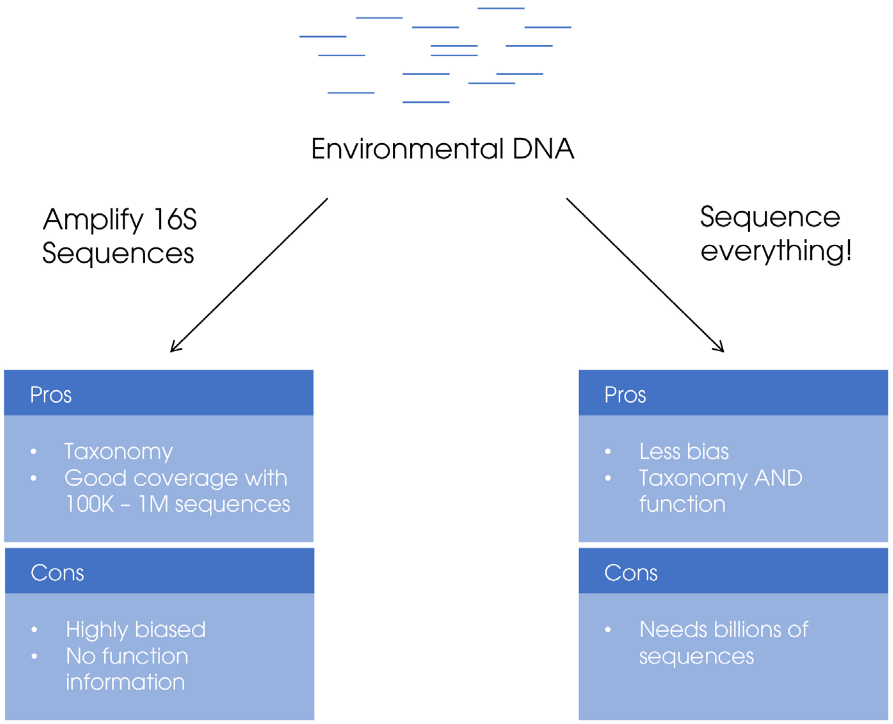
Data Storage and Compute Resources

The shotgun metagenomics study we have recently analyzed was 1.7 billion paired-end Illumina reads, 266.5 Gb. This is a bigger data volume than the human genome at 30x coverage. It is very likely that future studies will be even bigger. Storing that data volume can be challenging, but even more challenging is the data analysis. Analyses need to be selected that run at a reasonable time and cost on this volume of data. Fortunately, there are analyses that can do this e.g. Kraken for taxonomy assignment and HUMAnN2 for functional assignment. However these analyses may still require very large instances, e.g. to run Kraken optimally with its default taxonomy database requires an instance with > 100 Gb of memory. Cloud deployments with large scale architectures are recommended, such as AWS and Azure.

Statistical Analysis

Based on publications, there is little to no consensus on the best way to normalize data. Some papers say rarefaction is the best approach, others insist on normalization, some say that log transformation is the way to go, others say that you should never use this. This also applies to the statistical analysis, with some publications recommending compositional methods, whilst others recommend very different approaches. This variation in approaches leads to confusion and sub-optimal designs, which often have a significant negative effect on analysis results. Furthermore, different types of experiment may well need different analysis strategies, i.e. what is appropriate for amplicon studies may not suit shotgun studies. As shotgun becomes more popular, it is important to compare the benefits of each approach, as shown in the figure below.

Amplicon (old) vs Shotgun (new)



Environmental DNA

Amplify 16S Sequences

Sequence everything!

Pros

- Taxonomy
- Good coverage with 100K – 1M sequences

Cons

- Highly biased
- No function information

Pros

- Less bias
- Taxonomy AND function

Cons

- Needs billions of sequences

3. An End-to-End Industrialized Microbiome Workflow

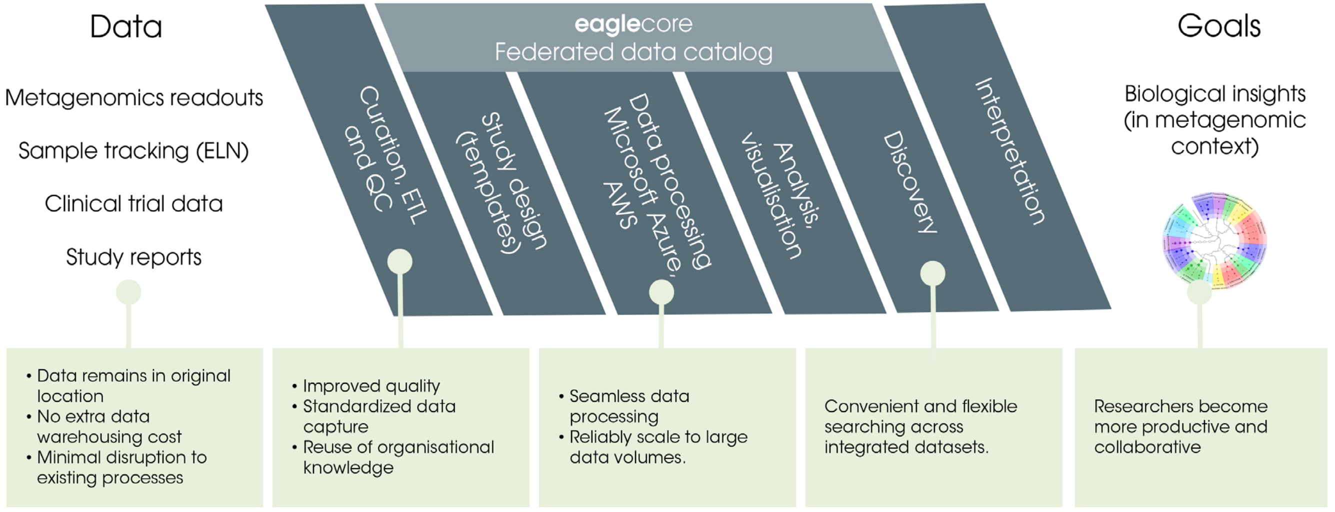
Eagle Genomics leads the industry in configuring and deploying best-in-class microbiome workflow solutions. These run large-scale parallelized pipelines cost-efficiently, with seamless connectivity between the source data and the scientists. The figure below shows Eagle's high level architecture.

Eagle Genomics has addressed these considerations and challenges in building an architecture for large scale microbiome analysis. For a free technical review of your environment and the opportunity for low-risk, cost-efficient scale-out, please reach out to us at sales@eaglegenomics.com.

BENEFITS

- DATA REUSE
- DESIGN BETTER EXPERIMENTS
- ANALYSE BILLIONS OF SEQUENCES
- COST-EFFICIENT EXPERIMENTS
- SUPPORT MULTIPLE PRODUCT DEV PIPELINES

MICROBIOMICS SOLUTION ARCHITECTURE



Data

Metagenomics readouts

Sample tracking (ELN)

Clinical trial data

Study reports

eaglecore
Federated data catalog

Curation, ETL and QC

Study design (templates)

Data processing Microsoft Azure, AWS

Analysis, visualisation

Discovery

Interpretation

Goals

Biological insights (in metagenomic context)

- Data remains in original location
- No extra data warehousing cost
- Minimal disruption to existing processes

- Improved quality
- Standardized data capture
- Reuse of organisational knowledge

- Seamless data processing
- Reliably scale to large data volumes.

Convenient and flexible searching across integrated datasets.

Researchers become more productive and collaborative