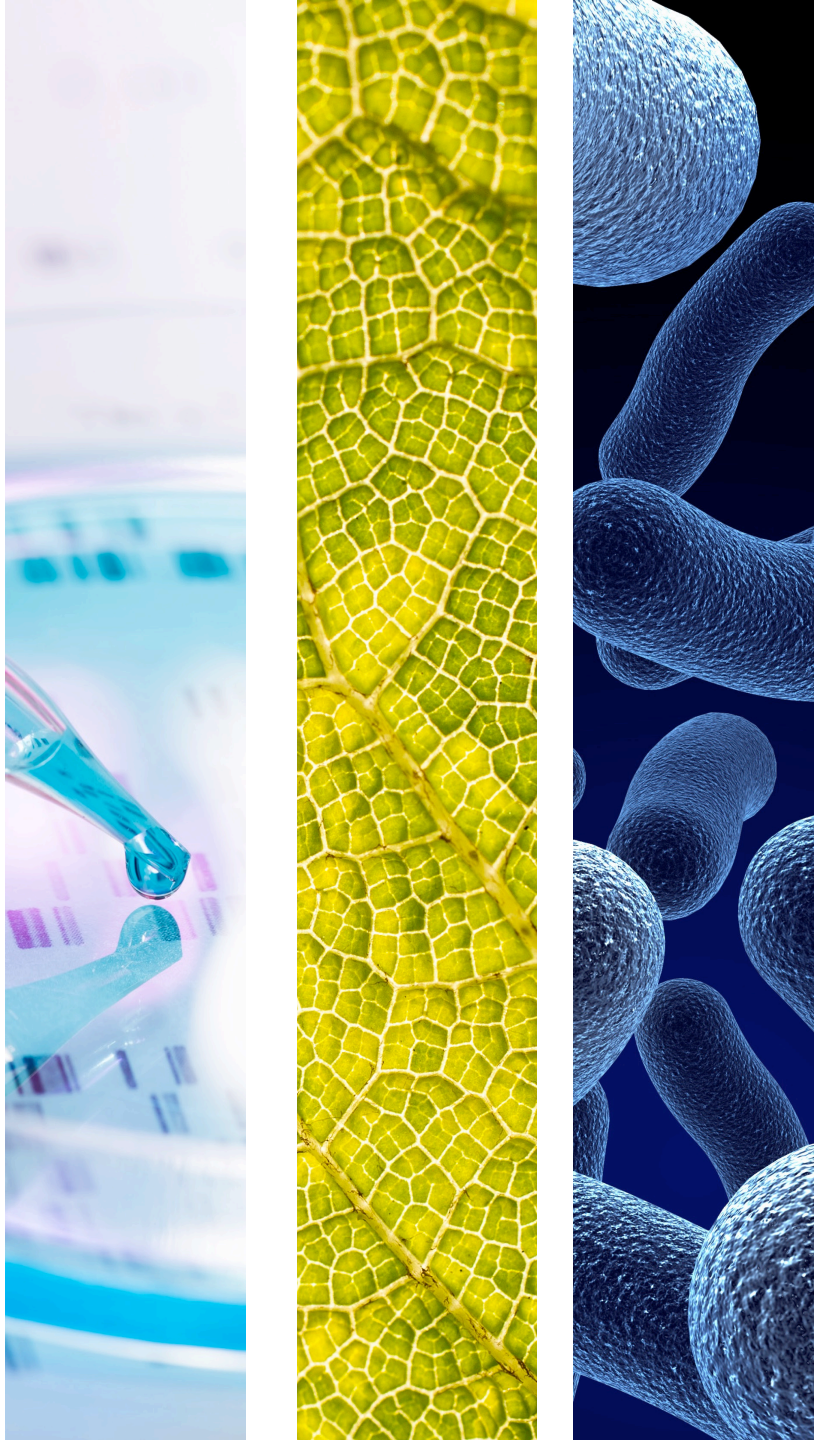
eagle genomics

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

data - science - life



Presenters

Dr Craig McAnulla

Senior Consultant Bioinformatics at Eagle Genomics

Craig earned his Ph.D. in Microbology at the University of Warwick, he then completed his Postdoctoral research at the John Innes Centre before joining EBI and developing their bioinformatics capability. Dr McAnulla has extensive experience in developing large scale pipelines for analyzing microbiome datasets. His skills cover both software and biology and he often provides deep insights to industry experts on these topics.

Dr Raminderpal Singh

Vice President Business Development at Eagle Genomics

Raminderpal earned his PhD in the semiconductor industry where he spent a decade building advanced technologies for semiconductor modelling and data mining. In 2012, Dr Singh took the helm as the Business Development Executive for IBM Research's genomics program, where he developed and led the execution of the go-to-market and business plan for the IBM Watson Genomics program. In addition to many published papers, Dr Singh has two books and twelve patents.

data - science - life

eagle® genomics

Please Tweet your questions **@eaglegen** or enter them into GoTo Webinar

Agenda

1. Considerations

2. Challenges

3. Industrialized Microbiome Workflow & Collaboration Environment

4. Deployment of a Collaboration System

1. Considerations

Considerations for successful microbiome data management and analysis:

1. Intrinsic factors
2. The relatively early maturity of the field
3. Calculation of business value

1. Considerations - Intrinsic

Intrinsic

- Microbiomes are complex and diverse biological systems
- Many components – organisms, genes, pathways, metabolites etc
- More DNA/genes than human genome
- Taxonomic diversity
- Many sources of error
- Appropriate study design and metadata collection are crucial.

1. Considerations - Maturity

Maturity

- Changing sequencing technologies
- Move from amplicon-based to shotgun
- Rapidly increasing data volumes
- Limited reference data
- Fast-moving software development
- Future data re-analysis with improved methods must be considered.

1. Considerations – Business value

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.

2. Challenges

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

- Metadata management
- Compute resources necessary for storage and analysis
- Statistical techniques

Metadata Management

- Crucial to analysis
- Often done poorly
- Needs to encompass all relevant information
    - Check for batch effects

2. Challenges

Data Storage and Compute Resources

- Shotgun metagenomics example study - 1.7 billion paired-end Illumina reads, 266.5 Gb.
    - Bigger data volume than the human genome at 30x coverage.
    - Study size is increasing

How to store this?

- Analysis selection critical
    - Analyses MUST work with these data volumes
    - eg Kraken, Humann2
- Compute resources
    - Eg Kraken, multicore machine > 100Gigs
- Solution – the cloud
    - AWS, Azure

2. Challenges

Statistical Techniques

- Data normalization
  - rarefaction, normalization, log transformation?
- Statistical techniques
  - parametric, nonparametric?
- Confusion
- No correct answer
- Depends on the data
  - function vs taxonomy
  - shotgun vs amplicon

Amplicon (old) vs Shotgun (new)

Environmental DNA

Amplify 16S Sequences

Pros

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

Cons

- Highly biased
- No function information

Sequence everything!

Pros

- Less bias
- Taxonomy AND function

Cons

- Needs billions of sequences

3. Industrialized Microbiome Workflow & Collaboration Environment

- 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.

“Unilever’s digital data program now processes genetic sequences twenty times faster - without incurring higher compute costs. In addition, its robust architecture supports ten times as many scientists, all working simultaneously.”

Pete Keeley, eScience Technical Lead - R&D IT at Unilever

3. Industrialized Microbiome Workflow & Collaboration Environment

MICROBIOMICS SOLUTION ARCHITECTURE

Data

Metagenomics readouts

Sample tracking (ELN)

Clinical trial data

Study reports

eaglecatalog
Federated data catalog

Curation, ETL and QC

Study design (templates)

Data processing Microsoft Azure, AWS

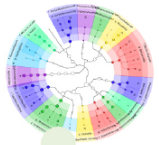
Analysis, visualization

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 organization knowledge

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

Convenient and flexible searching across integrated datasets.

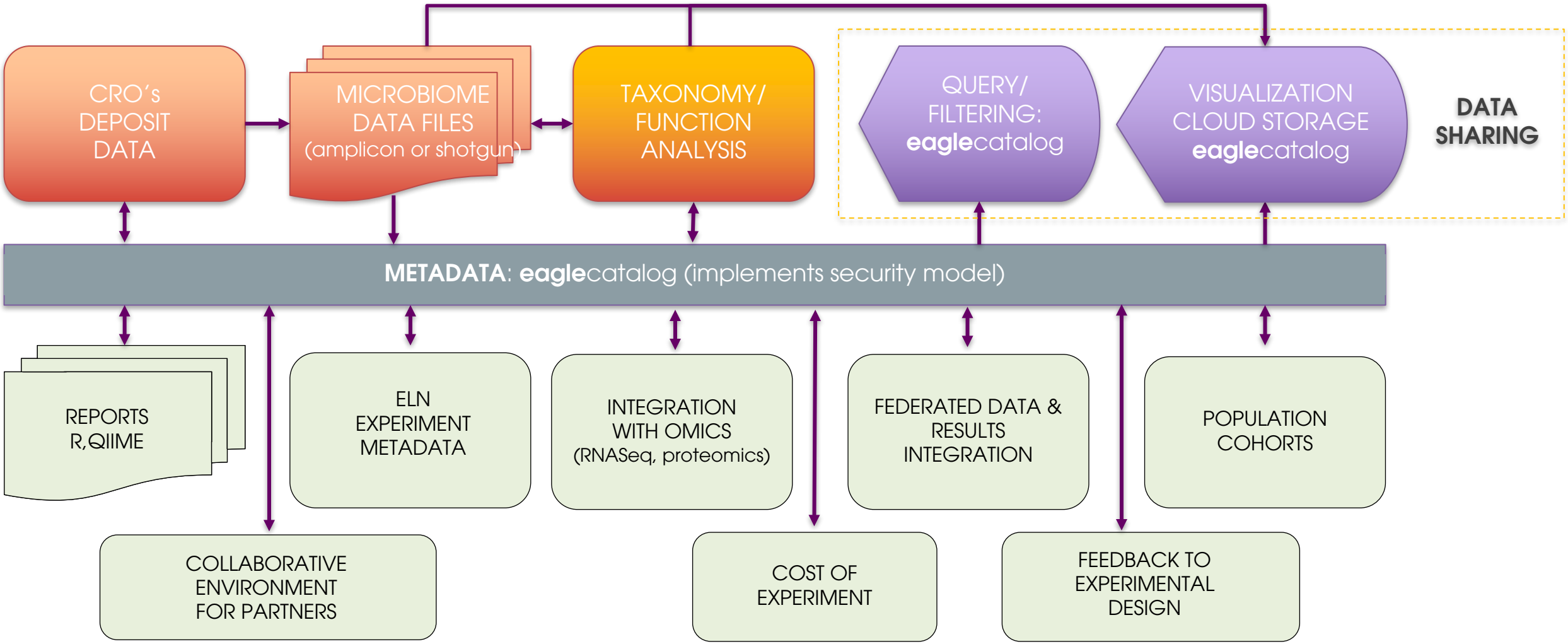
Researchers become more productive and collaborative

data - science - life

eagle genomics

DATA COLLABORATION ARCHITECTURE

Solving the metadata problem



data - science - life

eagle genomics

4. Deployment of a Collaboration System

Eagle Genomics Proposal

Accelerating metadata adoption and scale up

- Offering a new rapid deployment for microbiome R&D organizations looking to cost-efficiently build world-class integrated workflows.
- Set up of an initial collaboration system for your team.
- 1 to 4 weeks
- Enabling self-load your data to the catalog – leading to a standardized and streamlined workflow, which is ready for future scale out
- Managing internal data, partner data, or public data

BENEFITS

Current Situation	Desired State	Negative Consequences of Non-action	Positive Business Impact, using Eagle software
Data in silos, not reused	Easily searchable, discoverable, accessible and intuitive portal by researchers to all data and results	Significant researcher time spent in "data wrangling". Breakage pending, based on projected bottleneck	Improved time and ability to insight through radically improved productivity of researchers
No correlation between experiments	Quick and easy exploratory analyses across experiments and data	Missed opportunities for insights. Failure to leverage data as an asset	Faster identification and generation of successful product innovation and insight, leading to reduced time to market
No unified view of data	Data and analysis results stored in a unified catalog, with standardized & repeatable access using best-in-class technology	Data (internal, external) and the associated learning not available or exploited as necessary	Competitive advantage through cutting edge research, driving innovation in the product pipeline
Lack of governance and data provenance	Secure authenticated access with relevant data provenance of data sets; simple data sharing with vendors/collaborators	Costly, inefficient and limited data sharing with vendors; risk of inappropriate data sharing	Efficient compliant, documented, evidenced, and enforced data sharing process; effective (time & cost) collaboration between vendors and internal teams
Data is in-effect archived	Catalog integrated in the scientific workflow, used continuously and keeping pace with innovation in the industry	Missed insights; silo-ed research activities, with high barriers to collaboration	Systematic data sharing and re-use, driving faster unique cross-functional insights and reducing waste in spend

In Summary, Considerations & Challenges in Building an End-to-End Microbiome Workflow

data - science - life

eagle® genomics

For more info, contact:

Raminderpal Singh
raminderpal.singh@eaglegenomics.com