Closing Wrap Up

Mathematical Frameworks for Integrative Analysis of Emerging Biological Data Types

June 15 - 19, 2020

Zoom from Banff International Research Station, Canada

Aedin Culhane (Dana-Farber Cancer Institute, Harvard TH Chan School of Public Health)
Elana Fertig (John Hopkins University)
Kim-Anh Lê Cao (University of Melbourne)

Banff International Research Station
for Mathematical Innovation and Discovery

#BIRSBioIntegration

Goals of this workshop

Transparency

Collaboration

Open science

Fairness

Inclusion

Multi-omics integration of single cell data

- is an active and emerging field
- May provide insight that cannot be obtained from single datasets
- **lacks established performance benchmarks,**
- gold standard datasets, assessment standards.

Bring together **interdisciplinary computational scientists**

- to examine cutting edge techniques for integrative analysis of diverse multi-omics.
- Provide & assess **open source resources** for multi-platform analysis
- Formulate **goals and future directions** to advance multi-omics analysis

Products: Guidelines, build collaboration, code & datasets, a **white paper**

#BIRSBioIntegration

#BIRSBiointegration Community

3 challenging data challenges

16 contributed talks focusing on analysis
5 keynotes
9 Brainstorming sessions

Data and GitHub code shared

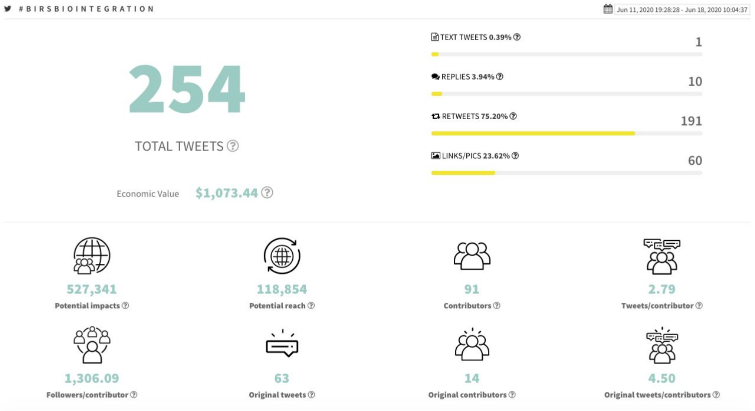
339 Commits to manubot

156 Members, 16 Active Channels on Slack

#BIRSBioIntegration

Outreach Beyond Banff

#BIRSBioIntegration



# B I R S B I O I N T E G R A T I O N

Jun 11, 2020 19:28:28 - Jun 18, 2020 10:04:37

254

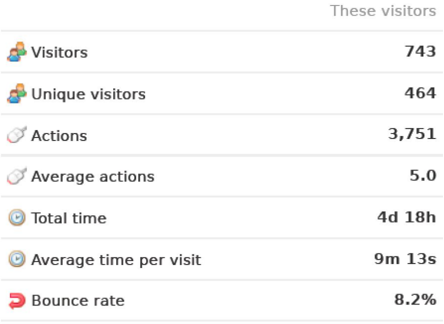
TOTAL TWEETS ?

Economic Value $1,073.44 ?

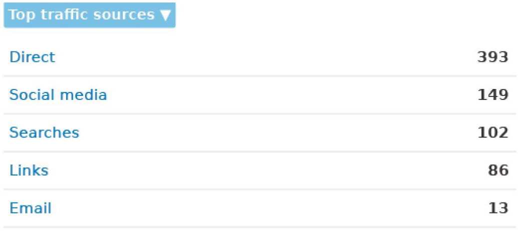
TEXT TWEETS 0.39% ?	1
REPLIES 3.94% ?	10
RETWEETS 75.20% ?	191
LINKS/PICS 23.62% ?	60

527,341 Potential impacts ?	118,854 Potential reach ?	91 Contributors ?	2.79 Tweets/contributor ?
1,306.09 Followers/contributor ?	63 Original tweets ?	14 Original contributors ?	4.50 Original tweets/contributors ?

Live Stream http://www.birs.ca/live



	These visitors
Visitors	743
Unique visitors	464
Actions	3,751
Average actions	5.0
Total time	4d 18h
Average time per visit	9m 13s
Bounce rate	8.2%



Top traffic sources ▼	
Direct	393
Social media	149
Searches	102
Links	86
Email	13

https://twitter.com/hashtag/BIRSBiointegration

Emerging Research: Five keynote speakers

Prof. GC Yuan

Dana-Farber Cancer Institute, Harvard TH Chan School of Public Health

Mon

Prof. Bernd Bodenmiller

University of Zurich

Tues

Prof. Oliver Stegle

German Cancer Research Center & EMBL

Wed

Prof. Susan Holmes

Stanford University

Thurs

Prof. Vincent Carey

Harvard Medical School, Brigham & Women’s Hospital

Fri

#BIRSBioIntegration

Contributed talks from hackathon participants

sc seq-FISH

Alexis Coullomb

Hang Xu

Dario Righelli

Amrit Singh

Joshua Sodicoff

sc Targ Proteomics

Yingxin Lin

Chen Meng

Pratheepa Jeganathan

Kris Sankaran

Lauren Hsu

Duncan Forster

scNMT-seq

Al J Abadi

Joshua Welch

Arshi Arora

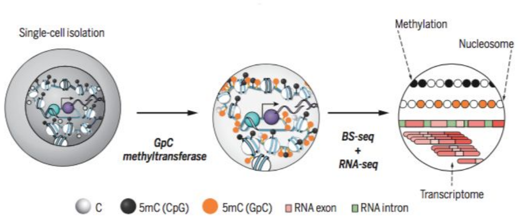
Wouter Meuleman

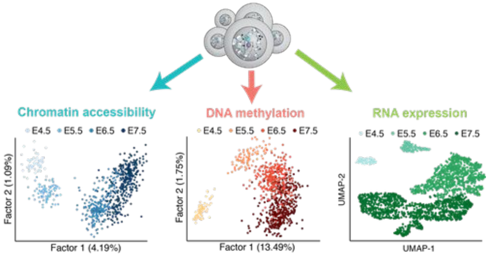
Slides from Brainstorming sessions available, see on Slack #information

3 Hackathon Challenges

Gastrulation (scNMT)

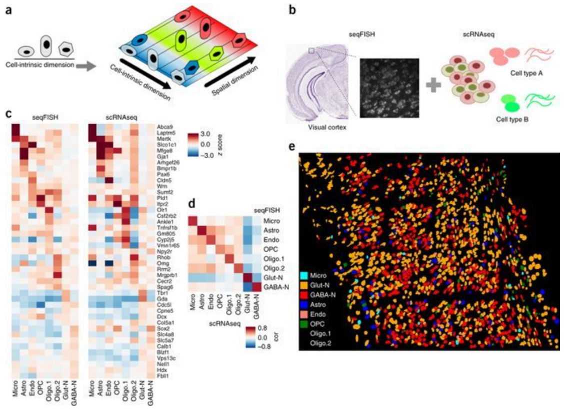
826 cells matching across all data sets (transcriptome, DNA accessibility and DNA methylation) after quality control and filtering.





Adult mouse visual cortex seqFISH, scRNAseq

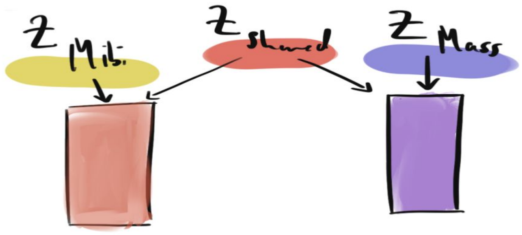
- seqFISH - 1,597 single cells x 125 genes mapped (Zhu *et al* 2018)
- scRNA-seq. ~1,600 cells (Tasic *et al* 2016)



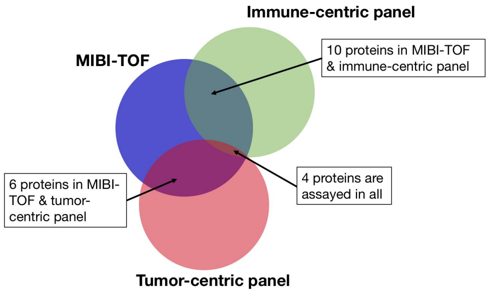
Breast Cancer sc Proteomics

Non-overlapping patients

MIBI 40 TN, Mass Tag 7 TN



… with 20 overlapping proteins



Immune-centric panel

MIBI-TOF

10 proteins in MIBI-TOF & immune-centric panel

6 proteins in MIBI-TOF & tumor-centric panel

4 proteins are assayed in all

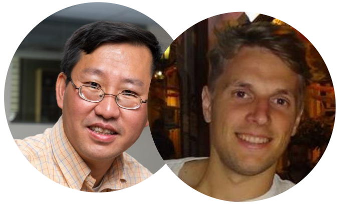
Tumor-centric panel

Hackathon Challenge Brainstorms

Spatial Fish	**Targeted Proteomics**	**RNA - DNA**	**Summary**
Expt design, Platform Specific bias, Inclusion of spatial information	Normalisation, Partial feature overlap **Non-overlapping cells** Integrating by phenotype Inherent spatial nature of biologial data,	Binary data **Transfer learning** or imputation using other atlases, Non-linear integration	**Summary of common challenges**: Non-overlapping features and/or cells, from data-driven towards mechanistic driven,
Objective Assessment,	Scale/metrics from single cell to cell communities	DNA features summary,	Generic towards context specific methods
	Annotation Atlases and maps for benchmarking	Annotation of histone db	Incorporate prior knowledge

#BIRSBioIntegration

9 Brainstorming sessions



seqfish_theme

**Guo-Cheng Yuan &
Ruben Dries**

Dana-Farber Cancer Institute, Harvard TH Chan School of Public Health & Boston University

sc_targ_proteomics_theme

**Aedin Culhane &
Olga Vitek**

Dana-Farber Cancer Institute, Harvard TH Chan School of Public Health & Northeastern University

scNMT-seq_theme

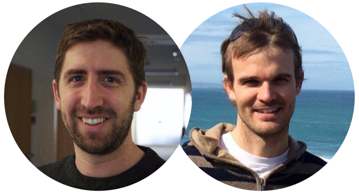
**Ricard Arguelaget &
Oliver Stegle**

German Cancer Research Center & EMBL

summary_analyses_theme

**Kim-Anh Lê Cao &
Casey Green**

University of Melbourne & Uni Pennsylvania



benchmark_theme

**Mike Love &
Matt Ritchie**

University of North Carolina-Chapel Hill & Walter and Eliza Hall Institute

Susan Holmes

Stanford University

interpretation_theme

Vincent Carey

Harvard Medical School and Brigham & Women's Hospital

software_theme

Elana Fertig

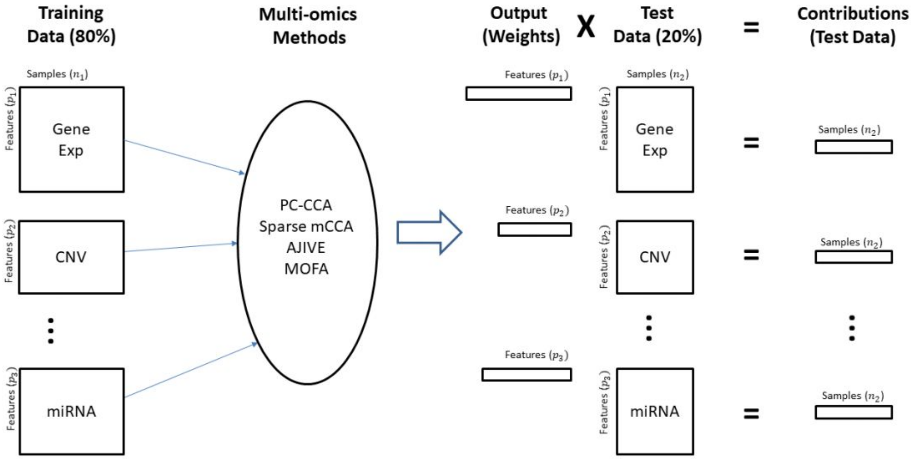
Johns Hopkins University

future_theme

Benchmarking	Interpretation	Software	Future
Establish performance **benchmarks** and assessment standards	Issue of benchmarking datasets immunology gated descrete	**Representation** mutli-view data Spatial Modality Colocation eQTL	High cell/large tissue (HCA, Allen, HTAN)
Assessment metrics Datasets benchmarks Deliver open source resources for multi-platform analysis (data wrangling) Awesome-multi-omics	Vocabulary for inside data science versus towards biologists Glossary for paper (appendix) Figures and visualization for communication versus discovery.	Annotation 4D, blueprint -Cell State-Cell State. Dropouts Scalability - containers Connecting to consoritums Color blind standard (import for UMAP)	Need pertubations/ dynamic datsets Data sharing Molecular coverage Deeper sampling Which data for which question **Training** on model

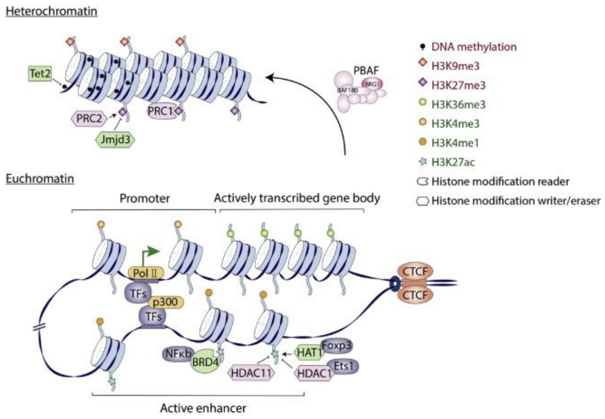
Community Coordination & Communication

- Representations
- Scale
- Metrics
- **Unified language**
- Annotation, ontology resources
- Leverages skills in other disciplines (spatial)
- Training - across disciplines
-
- Benchmarking dataset - ground truth
  - **What would be most interesting?**



DNA “accessible” for gene expression?

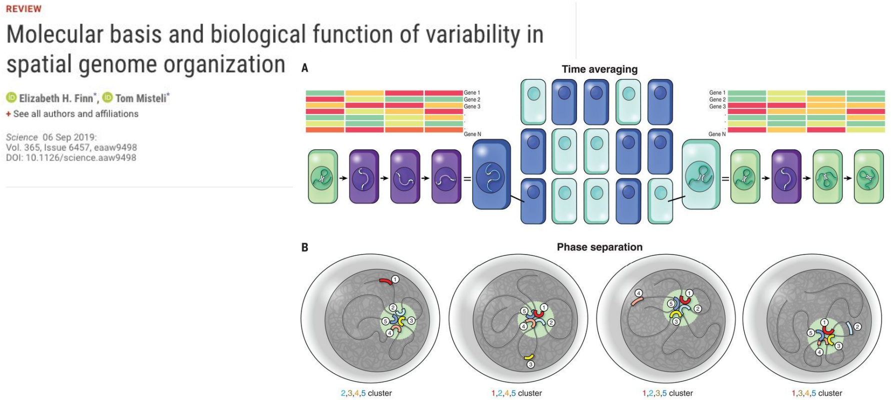
- DNA ->Regulation -> RNA -> Protein-> Regulation
- heterochromatin v euchromatin (silent v active) DNA defines the genome accessible for transcription
- Genome organization variability in cell types, states, (differentiation, development, stress, disease) unknown
- If regions are expected background off and other expected “accessible” (within a expt negative control?)



<u>Using the Genome in experimental design</u>

Which chromatin features under selection (active) and which features are evolutionary silent (historical)?

How precisely can chromatin define normal cell types



REVIEW

Molecular basis and biological function of variability in spatial genome organization

Elizabeth H. Finn*, Tom Misteli*

+ See all authors and affiliations

Science 06 Sep 2019:
Vol. 365, Issue 6457, eaaw9498
DOI: 10.1126/science.aaw9498

“Stable functional states and cell populations can be generated by two mechanisms: time- or population averaging of gene activity (**Fig. 4A**) or the formation of functionally equivalent but morphologically diverse cellular structures (**Fig. 4B**).”

The accessible genome “open” for gene expression

Bulk RNAseq normalization approaches assumed 50% genes silent in sample

>50% RNAseq in single cells are silent?

Impact on DE gene expression analysis of scRNAseq if the

Heterchromatin ∋ G p(E) =0
Euchromatin ∋ G p(E) >0

(imputation, dropout..)

Predicting # functional mRNA molecules

Delineate heterochromatin and transcriptional silencing.

- Histone marks, Methylation of promoter/enhancers
- Transcription bursts (3 state model)
- Nascent mRNA, half life (cap/tail)
- miRNA
- How do we distinguish cause vs effect of interactions?

*Activity dependent on functional network of gene

- Protein complexes
- Activation enzyme (precursor -> active form cleavage)
- Post -translational modification
- Co-localization

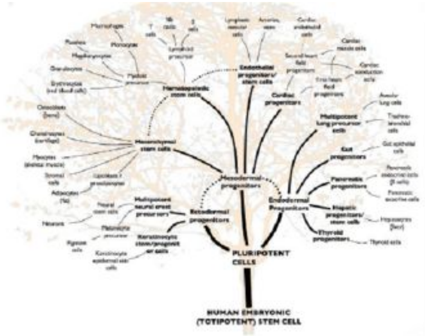
Requires Multi-omics *activity can be measured with proteins or inferred by expression of downstream targets

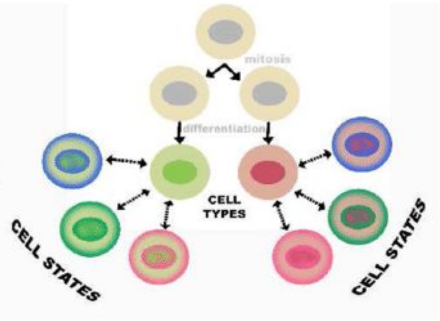
bulk - single cell

BULK → **sc**

Qualitative assessments of cell identity → Quantitative, high-resolution cell atlases

Cell lineage -> Cell Type ≠ Cell State

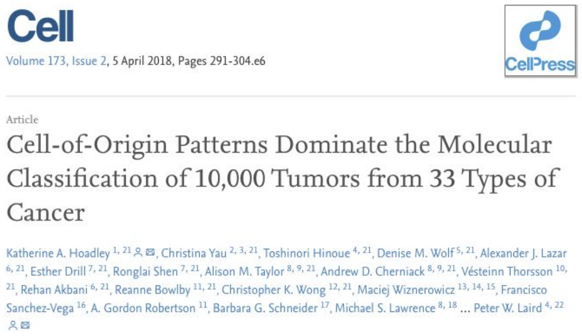




Cell State - dependent on local autocrine, paracrine, community signalling. More dynamic/variant.

Cell Type - relatively stable except for chromatin reorganization (stress/CNV/ dev)

=> Would predict bulk RNAseq captures



Cell

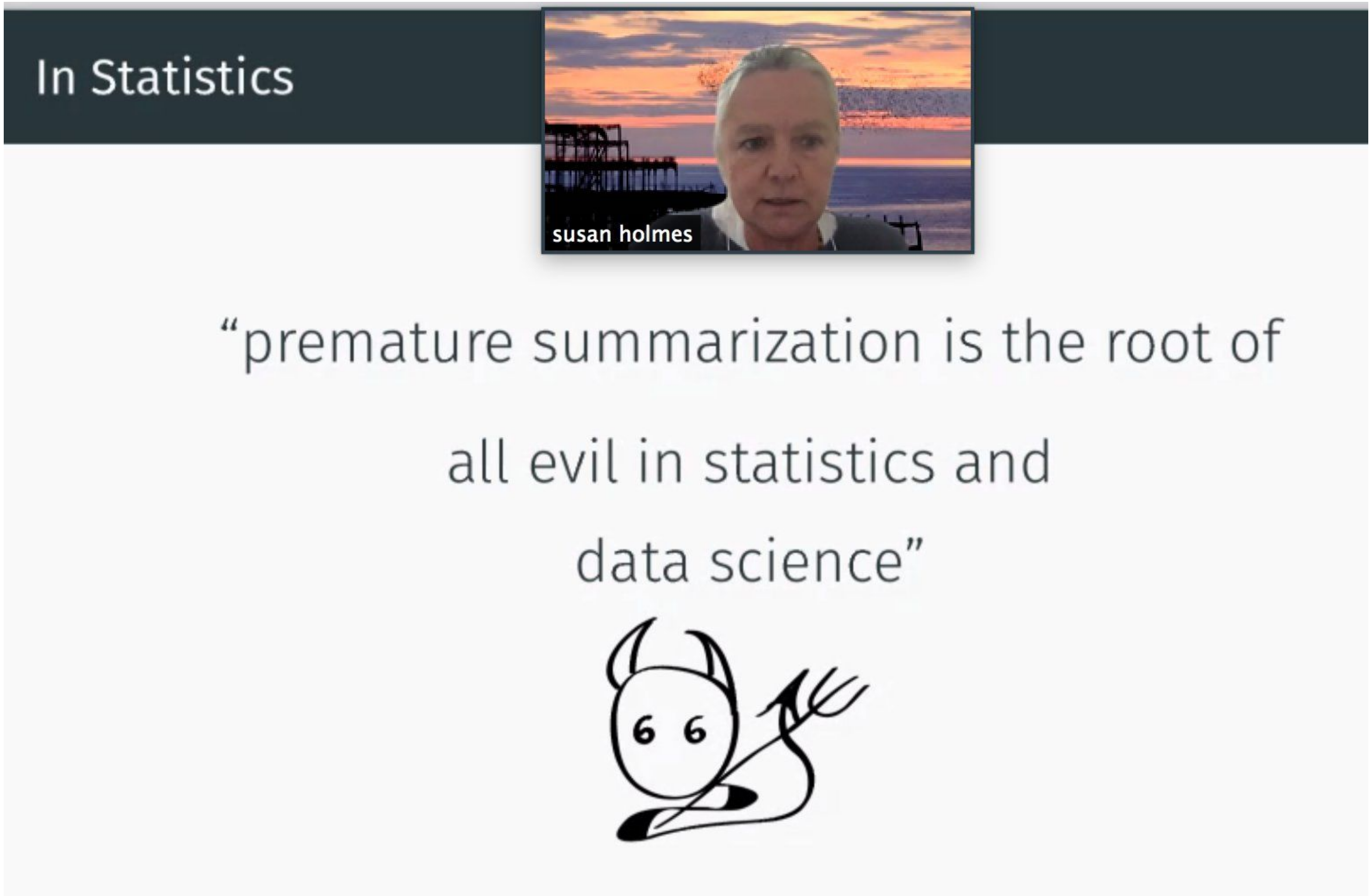
Volume 173, Issue 2, 5 April 2018, Pages 291-304.e6

CellPress

Article

Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer

Katherine A. Hoadley [1, 21], Christina Yau [2, 3, 21], Toshinori Hinoue [4, 21], Denise M. Wolf [5, 21], Alexander J. Lazar [6, 21], Esther Drill [7, 21], Ronglai Shen [7, 21], Alison M. Taylor [8, 9, 21], Andrew D. Cherniack [8, 9, 21], Vésteinn Thorsson [10, 21], Rehan Akbani [6, 21], Reanne Bowlby [11, 21], Christopher K. Wong [12, 21], Maciej Wiznerowicz [13, 14, 15], Francisco Sanchez-Vega [16], A. Gordon Robertson [11], Barbara G. Schneider [17], Michael S. Lawrence [8, 18] ... Peter W. Laird [4, 22]



In Statistics

“premature summarization is the root of

all evil in statistics and

data science”

Single Cells -> Communities -> Phenotype

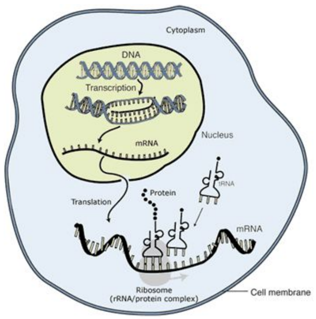
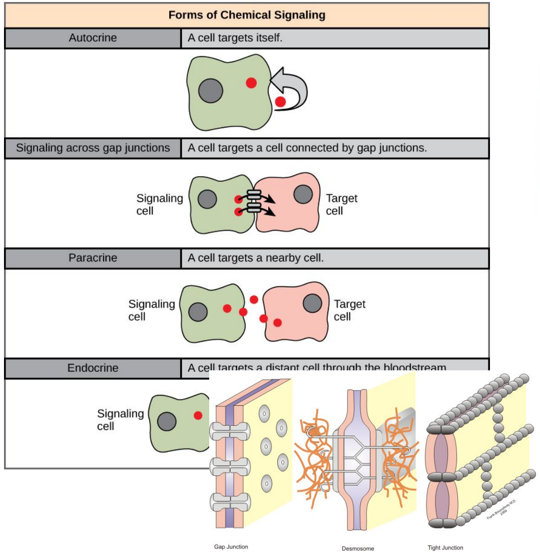


Image adapted from: National Human Genome Research Institute.

‘Omics DNA Chromatin RNA Protein Glycosylation-metabolites etc

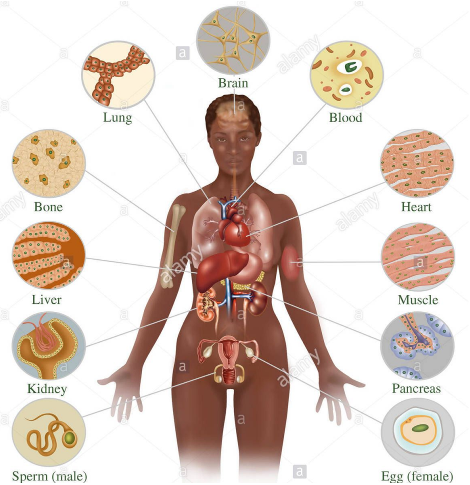


Forms of Chemical Signaling	
Autocrine	A cell targets itself.
Signaling across gap junctions	A cell targets a cell connected by gap junctions.
Paracrine	A cell targets a nearby cell.
Endocrine	A cell targets a distant cell through the bloodstream.

Connected by signnaling (paracrine, endocine Gap junctions, autocrine)

Squamous Epithelium
Columnar Epithelium
Cuboidal Epithelium
Stratified Epithelium
Pseudostratified Epithelium
Transitional Epithelium
Erythrocyte
Monocyte
Lymphocyte

Composed of organized Cells types, polarity



Human Phenotype defined by Systems, Organs that are composed of Cell Communities

Emerging Needs

Infrastructure

- Representation of each data multi-view , unified language, Cell /tissue type specific Ontologies,
- Representation/Visualization of anatomy

Benchmarking

- Methods for integration of different scales /merging later / mapping at pheno level
- Datasets to enable identification of DNA chromatin structure-> histone marks ->

Education

- As disciplines work together, Nomenclature dictionaries /common terms
- Education/Conference across discipline, especially in spatial biology - biologists learn from other fields and not reinvent GIS/weather/ecology

Products from meeting for multi-platform analysis

Open source resources

Datasets

Online- Bioc package

Code

Code for all contributed talks

Glossary/Language - Google Sheet (Data/Methods/Education) - Resource available as Awesome-sc list

White Paper

Optimistic Timeline for White Paper

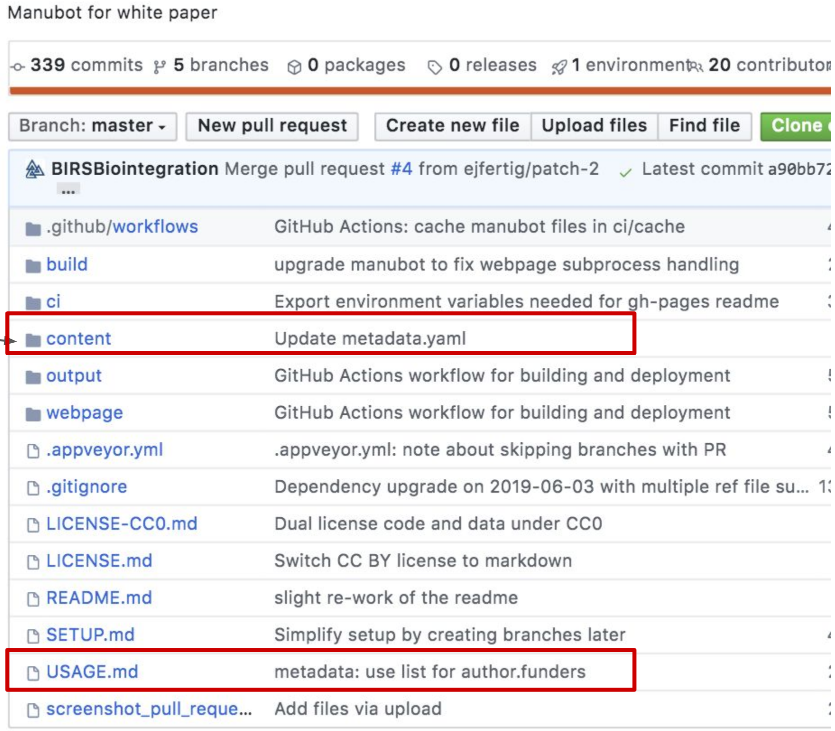
- Week 1 (June 26):
  - theme leaders push **outline** to Manubot to manage theme overlaps
  - Glossary of terms signed off
- Week 2 (July 3): **full section** written (~ 1 page + 1 Figure)
- Week 4 (July 17): **first draft** distributed to all for comments
- Week 6 (July 31): **comments back** from *all* co-authors
- Week 8 (August 14): finalise and submission

https://birsbiointegration.github.io/whitePaper/

#BIRSBioIntegration

Goal: White Paper

Manubot



Manubot for white paper

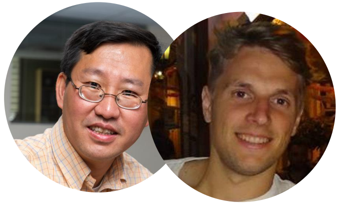
339 commits · 5 branches · 0 packages · 0 releases · 1 environment · 20 contributor

BIRSBiointegration Merge pull request #4 from ejfertig/patch-2 ✓ Latest commit a90bb72

.github/workflows	GitHub Actions: cache manubot files in ci/cache
build	upgrade manubot to fix webpage subprocess handling
ci	Export environment variables needed for gh-pages readme
content	Update metadata.yaml
output	GitHub Actions workflow for building and deployment
webpage	GitHub Actions workflow for building and deployment
.appveyor.yml	.appveyor.yml: note about skipping branches with PR
.gitignore	Dependency upgrade on 2019-06-03 with multiple ref file su…
LICENSE-CC0.md	Dual license code and data under CC0
LICENSE.md	Switch CC BY license to markdown
README.md	slight re-work of the readme
SETUP.md	Simplify setup by creating branches later
USAGE.md	metadata: use list for author.funders
screenshot_pull_reque…	Add files via upload

#manubot channel
Pull requests managed by Casey Greene, organisers and theme leaders

White Paper

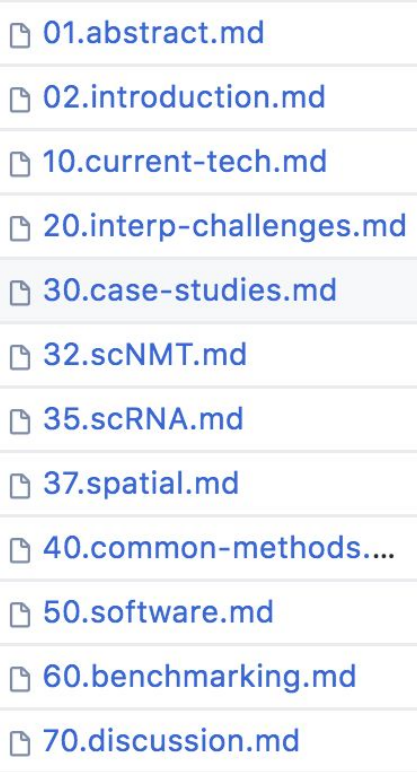


1. Spatial Transcriptomics: #seqFish_theme

2. RNA - DNA: #scNMT-seq_theme

3. Targeted Proteomics: #scTarg_Proteomics_theme

4. Summary methods: #summary_Analyses_theme



01.abstract.md

02.introduction.md

10.current-tech.md

20.interp-challenges.md

30.case-studies.md

32.scNMT.md

35.scRNA.md

37.spatial.md

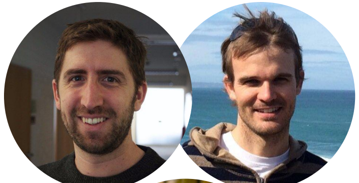
40.common-methods....

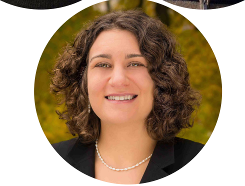
50.software.md

60.benchmarking.md

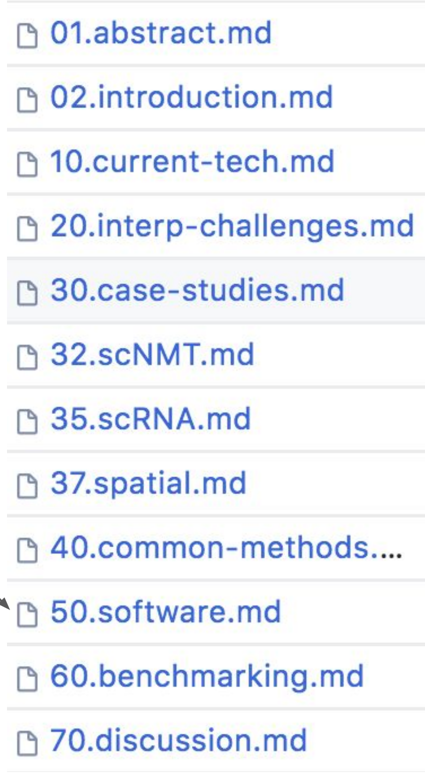
70.discussion.md

White Paper





1. Interpretation challenges: #interpretation_theme
2. Software infrastructure: #software_theme
3. Benchmarking: #benchmark_theme
4. Future Directions: #future_theme



- 01.abstract.md
- 02.introduction.md
- 10.current-tech.md
- 20.interp-challenges.md
- 30.case-studies.md
- 32.scNMT.md
- 35.scRNA.md
- 37.spatial.md
- 40.common-methods....
- 50.software.md
- 60.benchmarking.md
- 70.discussion.md

Communication will be key in the coming weeks!

Live

Zoom

Communication

Slack

BIRSBioIntegration

Datasets, code, paper

Github

https://github.com/BIRSBioIntegration

Monitor these tools and make good use of them!

Thank you for staying up late & waking up early

Interest in;

- Follow up meeting in Banff
  (deadline for application is Sep/Oct)
- Designing our own benchmarking expt and asking $$ from CZI?
- Other ideas. Please suggest.

A first poll will be distributed to state your authorship contribution.

Go to #information channel lists all important links

On behalf of the (fully zoomed) organizers - Thank You

Aedín Culhane
Dana-Farber Cancer Institute/
Harvard Chan
aedin@ds.dfci.harvard.edu

@AedinCulhane

Elana Fertig
Johns Hopkins University
ejfertig@jhmi.edu

@FertigLab

Kim-Anh Lê Cao
The University of Melbourne
kimanh.lecao@unimelb.edu.au

@mixOmics_team

Banff International Research Station
for Mathematical Innovation and Discovery

Scientific Program Coordinator: Chee Chow
Program Assistant: Dominique Vaz
Station Manager: Linda Jarigina-Sahoo
Technology Manager: Brent Kearney

@BIRS_Math