Introduction to R

Wrestling, Late Archaic (500 - 480 BCE)

Boris Steipe

UNIVERSITY OF TORONTO

Department of Biochemistry
Department of Molecular Genetics

Objectives

Be able to start up and work with **R** and R Studio;

Understand configuration files;

Be able to open files and efdit and save scripts;

Be able to work with basic **R** commands;

Be able to structure a computational task as an **R** script;

Be able to read data, select, filter, rearrange and combine;

Be able to write functions and programs;

Be able to create simple analyses;

Know where to get help.

Learning R

Most approaches to teaching **R** take you through the basics of **R** bit by bit. I had done this for the last six years ...

- constants
- vectors
- tables
- programming
- plots
- ...

... but recently I threw this all out.

Learning R

You don't want to become programmers. You want to get some biology done. And while **R** is the most important tool for that, what you are really worried about are quite different things:

- How do express my ideas in code?
- How do I even get started?
- OMG something happened! What do I do now?
- How do I keep up with things?
- How can I remember all these functions?

... and that's what I hope you'll be more comfortable with when you go home.

Learning R

You don't want to become programers. You need to get some biology done. And while **R** is the most important tool for that, what you are really worried about are quite different things:

- How do express my ideas in code?
- How do I even get started?
- OMG something happened! What do I do now?
- How do I keep up with things?
- ~~How can I remember all these functions?~~

... and that's what I hope you'll be more comfortable with when you go home.

Learning R

So we'll learn **R** by working with **R**.

Rather than learn commands in isolation, we will look at a (typical) **problem**, and develop a strategy to solve it. Part of that strategy will involve learning **R.**

But other parts are really about learning to learn.

With **R** this is particularly important.

Learning R

Here's the thing: **R** is so large, that it's virtually impossible to keep up with all of it.

Or if you tried, you wouldn't get any work done.

But the answer to "Can *x* be done with **R**?" is almost always "Yes." Someone out there has had this problem before and since **R** is so easy to extend, solutions exist.

So working with **R** realy means structuring your problem clearly. Then **you** understand it. **Then** you can ask the right questions.

Learning R

Here's what you.ll need to do to get the most out of this day:

Be active. Think ahead. We'll work on questions and you should always think: how would I approach this problem?

Take notes. Write a lot. This helps you focus.

Ask. Whenever you encounter something you don't know, or are curious about, ask. This is why you are in this room.

Play. Try things. Watch them break. Smile and fix them.

Have fun.

The Wiki

I have created a Wiki page as your main hub for scripts and resources.

Navigate to:

http://steipe.biochemistry.utoronto.ca/abc/index.php/Workshops

... follow the link to: **Saskatoon 2015 Introduction to R**

Tasks

Active participation is important. We will work through many tasks in the main script file, as the workshop proceeds. I have written a number of "Checkpoints" into the script.

You have PostIts on your desk.

Use the green PostIt to signal when you have completed a "Checkpoint".

Use the pink PostIt to signal when you need help.

Scripts

We will mostly work from scripts.
Scripts can be downloaded from the Workshop Wiki page.

Let's start by loading a script into R Studio...
... and we will continue our work in that script.

Learning R

Here is a recent (typical) paper rather randomly chosen for the type of data it uses and the type of questions it pursues...

A (TYPICAL) MODERN EXPERIMENT

Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types

Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, Ido Amit

Science (2014) 343:776-779

In multicellular organisms, biological function emerges when heterogeneous cell types form complex organs. Nevertheless, dissection of tissues into mixtures of cellular subpopulations is currently challenging. We introduce an automated massively parallel single-cell RNA sequencing (RNA-seq) approach for analyzing in vivo transcriptional states in thousands of single cells. Combined with unsupervised classification algorithms, this facilitates *ab initio* cell-type characterization of splenic tissues. Modeling single-cell transcriptional states in dendritic cells and additional hematopoietic cell types uncovers rich cell-type heterogeneity and gene-modules activity in steady state and after pathogen activation. Cellular diversity is thereby approached through inference of variable and dynamic pathway activity rather than a fixed preprogrammed cell-type hierarchy. These data demonstrate single-cell RNA-seq as an effective tool for comprehensive cellular decomposition of complex tissues.

A (TYPICAL) MODERN EXPERIMENT

Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types

Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, Ido Amit

Science (2014) 343:776-779

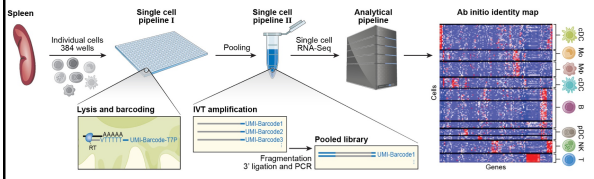


Fig. S1

A (TYPICAL) MODERN EXPERIMENT

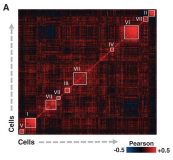
Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types

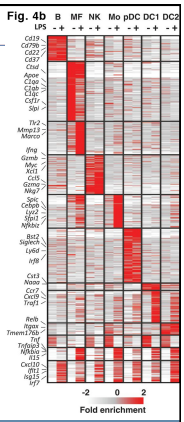
Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, Ido Amit

Science (2014) 343:776-779

In a nutshell:

- Single cell RNA-seq can be done.
- Cells can be crudely clustered into cell types.
- Cluster features can be used for classification to characterize cell types.
- Experiments can be repeated with perturbations to characterize cell-type specific responses.





A (TYPICAL) MODERN EXPERIMENT

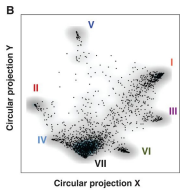
Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types

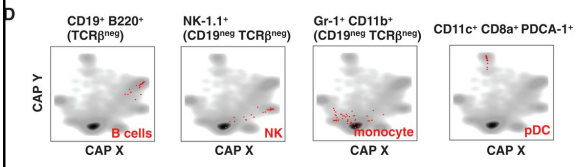
Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, Ido Amit

Science (2014) 343:776-779

So far so good but we might have questions:

- Are the "known" markers of Fig. 2 D enriched as expected in the cell types?





A (TYPICAL) MODERN EXPERIMENT

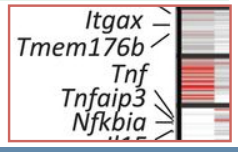
Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types

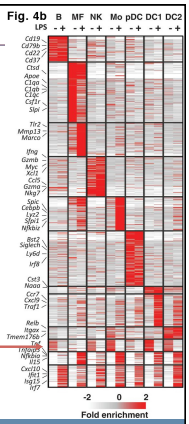
Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, Ido Amit

Science (2014) 343:776-779

So far so good but we might have questions:

- Are the "known" markers of Fig. 2 D enriched as expected in the cell types?
- What are the unlabelled genes in Figure 4?





A (TYPICAL) MODERN EXPERIMENT

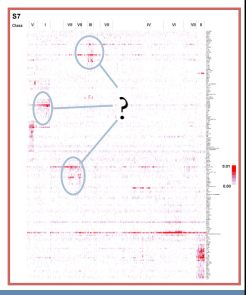
Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types

Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, Ido Amit

Science (2014) 343:776-779

So far so good
but we might have questions:

- Are the "known" markers of Fig. 2 D enriched as expected in the cell types?
- What are the unlabelled genes in Figure 4?
- Are genes that are functionally related to characteristic genes of cell types coregulated with the characteristic genes?
(I.e. are these genes functionally significant?)



A (typical) Modern Experiment

Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types

Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, Ido Amit

Science (2014) 343:776-779

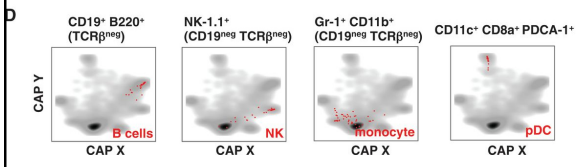
To answer our questions, we need to work with the supplementary data...

... and perhaps diverse additional data sources.

Get Practical

What do we do first?

Are the "known" markers of Fig. 2 D expressed as expected in the cell types?



Setting Up A Project

In **R**

Create a project directory

Manage your working directory definition.

Download the data files you want to work with.

Open a script template, fill in some details and save it with a meaningful name.

Enter the `setwd()` command as the first command of your script.

In **R Studio**

Use the menu.

Setup everything.

Manage your working directory definition.

Download the data files you want to work with.

Open a script template, fill in some details and save it with a meaningful name.

Software Carpentry

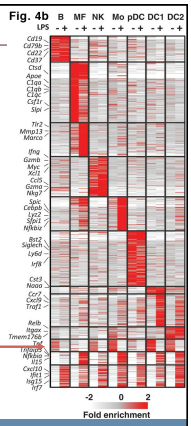
http://software-carpentry.org/

- Structure your code
- Pay attention to coding style and comments
- Do not Repeat Yourself
- Code all tasks, always use scripts
- Develop incrementally
- Use version control for everything
- Use an IDE (debugger!)
- Work with test-driven development
- Optimize later
- Collaborate

Programming

How do we express this in code?

- What are the unlabelled genes in Figure 4?

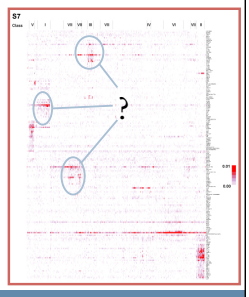


Analysis

How do we integrate various data sources;

What analyses are interesting?

- Are genes that are functionally related to characteristic genes of cell types coregulated with the characteristic genes? (I.e. are these genes functionally significant?)



boris.steipe@utoronto.ca