

STONY BROOK UNIVERSITY

RPE REPORT

Boosting the fast methods of mapping and abundance estimation of RNA-seq data

Author:
Mohsen ZAKERI

Supervisor:
Dr. Rob PATRO

*A report submitted as Research Proficiency Examination
for proceeding to candidacy of the degree of Doctor of Philosophy
in the*

COMBINE Lab
Department of Computer Science

June 26, 2017

Abstract

RNA-seq is a ubiquitous and sensitive sequencing assay to retrieve accurate short reads from whole transcriptomes. This data lets us study how transcript-level expression changes across conditions, tissues, disease states and during development, and also lets us discover new tissue-specific transcript isoforms. However, due to the sequence-level complexity of the transcriptome and the short length (hundreds of nucleotides) of high-throughput sequencing reads, accurate computational methods are required for making sense of this data and carrying out biological studies. Since the volume of sequencing data being produced is very large, and growing exponentially, fundamentally fast and scalable analysis methods have become essential. Typical processing of RNA-seq data consists of multiple computational steps such as mapping RNA-seq reads to transcripts and quantifying transcript abundance, prior to performing e.g., differential expression methods to determine condition-specific changes.

Recently methods like RapMap for mapping reads to the transcriptome and Salmon for quantifying transcript abundance, have been developed, which exhibit accuracy very similar to state-of-the-art alignment-based methods, but which are much faster. These fast methods have been demonstrated to be highly accurate, yet it is true that they discard certain information to achieve their speed. There exist rare, adversarial situations for which considering this discarded information is important to achieve the most accurate possible estimates. In this work, we propose advances that largely retain the speed of these new approaches, but which allow them to close the remaining accuracy gap in complex or adversarial situations. Specifically, we focus on approaches that expend a computational budget (i.e., in terms of time and space) that scales with the actual complexity of the underlying problem — expending less computation for easy cases, and reserving complex computations only for data where they are required. We describe two particular problems where we have adopted such a methodology. First, we developed a new mapping approach, called “selective alignment”, which achieves a similar speed to RapMap, but exhibits a higher sensitivity and specificity and which can compute the relevant edit distance and alignment information on demand. Second, we have developed a more refined, data-driven factorization of the likelihood function for the transcript quantification problem. We show that this improved factorization retains the computational efficiency of “compatibility-based” factorizations like those used in MMSeq, IsoEM, Sailfish, kallisto and Salmon, while yielding improved abundance estimates comparable to methods that adopt an un-factorized likelihood model (e.g., RSEM). Furthermore, we investigate how Bayesian methods for transcript quantification perform compared to other approaches, such as maximum likelihood, and propose how state-of-the-art inference methods might be further improved in the future.

Acknowledgements

Here I declare my deep appreciations for all the helpful guidances by my Advsiar, Dr. Rob Patro, from whom I learned a lot. I also thank lab members of Combine Lab, Fatemeh, Laraib, Hiraak and Avi who were always there for me and I owe them a lot.

Contents

Acknowledgements	v
1 Chapter Title Here	1
1.1 Welcome and Thank You	1
1.2 Learning L ^A T _E X	1
1.2.1 A (not so short) Introduction to L ^A T _E X	1
1.2.2 A Short Math Guide for L ^A T _E X	2
1.2.3 Common L ^A T _E X Math Symbols	2
1.2.4 L ^A T _E X on a Mac	2
1.3 Getting Started with this Template	2
1.3.1 About this Template	3
1.4 What this Template Includes	3
1.4.1 Folders	3
1.4.2 Files	3
1.5 Filling in Your Information in the main.tex File	4
1.6 The main.tex File Explained	5
1.7 Thesis Features and Conventions	6
1.7.1 Printing Format	6
1.7.2 Using US Letter Paper	6
1.7.3 References	6
A Note on bibtex	7
1.7.4 Tables	7
1.7.5 Figures	8
1.7.6 Typesetting mathematics	9
1.8 Sectioning and Subsectioning	9
1.9 In Closing	10
A Frequently Asked Questions	11
A.1 How do I change the colors of links?	11

List of Figures

1.1 An Electron	8
---------------------------	---

List of Tables

1.1	The effects of treatments X and Y on the four groups studied.	8
-----	---	---

List of Abbreviations

LAH List Abbreviations **Here**
WSF What (it) Stands For

Physical Constants

Speed of Light $c_0 = 2.997\,924\,58 \times 10^8 \text{ m s}^{-1}$ (exact)

List of Symbols

a	distance	m
P	power	W (J s ⁻¹)
ω	angular frequency	rad

For/Dedicated to/To my...

Chapter 1

Chapter Title Here

1.1 Welcome and Thank You

Welcome to this L^AT_EX Thesis Template, a beautiful and easy to use template for writing a thesis using the L^AT_EX typesetting system.

If you are writing a thesis (or will be in the future) and its subject is technical or mathematical (though it doesn't have to be), then creating it in L^AT_EX is highly recommended as a way to make sure you can just get down to the essential writing without having to worry over formatting or wasting time arguing with your word processor.

L^AT_EX is easily able to professionally typeset documents that run to hundreds or thousands of pages long. With simple mark-up commands, it automatically sets out the table of contents, margins, page headers and footers and keeps the formatting consistent and beautiful. One of its main strengths is the way it can easily typeset mathematics, even *heavy* mathematics. Even if those equations are the most horribly twisted and most difficult mathematical problems that can only be solved on a super-computer, you can at least count on L^AT_EX to make them look stunning.

1.2 Learning L^AT_EX

L^AT_EX is not a WYSIWYG (What You See is What You Get) program, unlike word processors such as Microsoft Word or Apple's Pages. Instead, a document written for L^AT_EX is actually a simple, plain text file that contains *no formatting*. You tell L^AT_EX how you want the formatting in the finished document by writing in simple commands amongst the text, for example, if I want to use *italic text for emphasis*, I write the `\emph{text}` command and put the text I want in italics in between the curly braces. This means that L^AT_EX is a "mark-up" language, very much like HTML.

1.2.1 A (not so short) Introduction to L^AT_EX

If you are new to L^AT_EX, there is a very good eBook – freely available online as a PDF file – called, "The Not So Short Introduction to L^AT_EX". The book's title is typically shortened to just *lshort*. You can download the latest version (as it is occasionally updated) from here: <http://www.ctan.org/tex-archive/info/lshort/english/lshort.pdf>

It is also available in several other languages. Find yours from the list on this page: <http://www.ctan.org/tex-archive/info/lshort/>

It is recommended to take a little time out to learn how to use L^AT_EX by creating several, small 'test' documents, or having a close look at several templates on:

<http://www.LaTeXTemplates.com>

Making the effort now means you're not stuck learning the system when what you *really* need to be doing is writing your thesis.

1.2.2 A Short Math Guide for L^AT_EX

If you are writing a technical or mathematical thesis, then you may want to read the document by the AMS (American Mathematical Society) called, "A Short Math Guide for L^AT_EX". It can be found online here: <http://www.ams.org/tex/amslatex.html> under the "Additional Documentation" section towards the bottom of the page.

1.2.3 Common L^AT_EX Math Symbols

There are a multitude of mathematical symbols available for L^AT_EX and it would take a great effort to learn the commands for them all. The most common ones you are likely to use are shown on this page: <http://www.sunilpatel.co.uk/latex-type/latex-math-symbols/>

You can use this page as a reference or crib sheet, the symbols are rendered as large, high quality images so you can quickly find the L^AT_EX command for the symbol you need.

1.2.4 L^AT_EX on a Mac

The L^AT_EX distribution is available for many systems including Windows, Linux and Mac OS X. The package for OS X is called MacTeX and it contains all the applications you need – bundled together and pre-customized – for a fully working L^AT_EX environment and work flow.

MacTeX includes a custom dedicated L^AT_EX editor called TeXShop for writing your '**.tex**' files and BibDesk: a program to manage your references and create your bibliography section just as easily as managing songs and creating playlists in iTunes.

1.3 Getting Started with this Template

If you are familiar with L^AT_EX, then you should explore the directory structure of the template and then proceed to place your own information into the *THESIS INFORMATION* block of the **main.tex** file. You can then modify the rest of this file to your unique specifications based on your degree/university. Section 1.5 on page 4 will help you do this. Make sure you also read section 1.7 about thesis conventions to get the most out of this template.

If you are new to L^AT_EX it is recommended that you carry on reading through the rest of the information in this document.

Before you begin using this template you should ensure that its style complies with the thesis style guidelines imposed by your institution. In most cases this template style and layout will be suitable. If it is not, it may only require a small change to bring the template in line with your institution's recommendations. These modifications will need to be done on the **MastersDoctoralThesis.cls** file.

1.3.1 About this Template

This L^AT_EX Thesis Template is originally based and created around a L^AT_EX style file created by Steve R. Gunn from the University of Southampton (UK), department of Electronics and Computer Science. You can find his original thesis style file at his site, here: <http://www.ecs.soton.ac.uk/~srg/softwaretools/document/templates/>

Steve's `ecsthesis.cls` was then taken by Sunil Patel who modified it by creating a skeleton framework and folder structure to place the thesis files in. The resulting template can be found on Sunil's site here: <http://www.sunilpatel.co.uk/thesis-template>

Sunil's template was made available through <http://www.LaTeXTemplates.com> where it was modified many times based on user requests and questions. Version 2.0 and onwards of this template represents a major modification to Sunil's template and is, in fact, hardly recognisable. The work to make version 2.0 possible was carried out by Vel and Johannes Böttcher.

1.4 What this Template Includes

1.4.1 Folders

This template comes as a single zip file that expands out to several files and folders. The folder names are mostly self-explanatory:

Appendices – this is the folder where you put the appendices. Each appendix should go into its own separate `.tex` file. An example and template are included in the directory.

Chapters – this is the folder where you put the thesis chapters. A thesis usually has about six chapters, though there is no hard rule on this. Each chapter should go in its own separate `.tex` file and they can be split as:

- Chapter 1: Introduction to the thesis topic
- Chapter 2: Background information and theory
- Chapter 3: (Laboratory) experimental setup
- Chapter 4: Details of experiment 1
- Chapter 5: Details of experiment 2
- Chapter 6: Discussion of the experimental results
- Chapter 7: Conclusion and future directions

This chapter layout is specialised for the experimental sciences, your discipline may be different.

Figures – this folder contains all figures for the thesis. These are the final images that will go into the thesis document.

1.4.2 Files

Included are also several files, most of them are plain text and you can see their contents in a text editor. After initial compilation, you will see that more auxiliary

files are created by \LaTeX or BibTeX and which you don't need to delete or worry about:

example.bib – this is an important file that contains all the bibliographic information and references that you will be citing in the thesis for use with BibTeX. You can write it manually, but there are reference manager programs available that will create and manage it for you. Bibliographies in \LaTeX are a large subject and you may need to read about BibTeX before starting with this. Many modern reference managers will allow you to export your references in BibTeX format which greatly eases the amount of work you have to do.

MastersDoctoralThesis.cls – this is an important file. It is the class file that tells \LaTeX how to format the thesis.

main.pdf – this is your beautifully typeset thesis (in the PDF file format) created by \LaTeX . It is supplied in the PDF with the template and after you compile the template you should get an identical version.

main.tex – this is an important file. This is the file that you tell \LaTeX to compile to produce your thesis as a PDF file. It contains the framework and constructs that tell \LaTeX how to layout the thesis. It is heavily commented so you can read exactly what each line of code does and why it is there. After you put your own information into the *THESIS INFORMATION* block – you have now started your thesis!

Files that are *not* included, but are created by \LaTeX as auxiliary files include:

main.aux – this is an auxiliary file generated by \LaTeX , if it is deleted \LaTeX simply regenerates it when you run the main **.tex** file.

main.bbl – this is an auxiliary file generated by BibTeX, if it is deleted, BibTeX simply regenerates it when you run the **main.aux** file. Whereas the **.bib** file contains all the references you have, this **.bbl** file contains the references you have actually cited in the thesis and is used to build the bibliography section of the thesis.

main.blg – this is an auxiliary file generated by BibTeX, if it is deleted BibTeX simply regenerates it when you run the main **.aux** file.

main.lof – this is an auxiliary file generated by \LaTeX , if it is deleted \LaTeX simply regenerates it when you run the main **.tex** file. It tells \LaTeX how to build the *List of Figures* section.

main.log – this is an auxiliary file generated by \LaTeX , if it is deleted \LaTeX simply regenerates it when you run the main **.tex** file. It contains messages from \LaTeX , if you receive errors and warnings from \LaTeX , they will be in this **.log** file.

main.lot – this is an auxiliary file generated by \LaTeX , if it is deleted \LaTeX simply regenerates it when you run the main **.tex** file. It tells \LaTeX how to build the *List of Tables* section.

main.out – this is an auxiliary file generated by \LaTeX , if it is deleted \LaTeX simply regenerates it when you run the main **.tex** file.

So from this long list, only the files with the **.bib**, **.cls** and **.tex** extensions are the most important ones. The other auxiliary files can be ignored or deleted as \LaTeX and BibTeX will regenerate them.

1.5 Filling in Your Information in the **main.tex** File

You will need to personalise the thesis template and make it your own by filling in your own information. This is done by editing the **main.tex** file in a text editor or your favourite LaTeX environment.

Open the file and scroll down to the third large block titled *THESIS INFORMATION* where you can see the entries for *University Name*, *Department Name*, etc ...

Fill out the information about yourself, your group and institution. You can also insert web links, if you do, make sure you use the full URL, including the `http://` for this. If you don't want these to be linked, simply remove the `\href{url}{name}` and only leave the name.

When you have done this, save the file and recompile `main.tex`. All the information you filled in should now be in the PDF, complete with web links. You can now begin your thesis proper!

1.6 The `main.tex` File Explained

The `main.tex` file contains the structure of the thesis. There are plenty of written comments that explain what pages, sections and formatting the \LaTeX code is creating. Each major document element is divided into commented blocks with titles in all capitals to make it obvious what the following bit of code is doing. Initially there seems to be a lot of \LaTeX code, but this is all formatting, and it has all been taken care of so you don't have to do it.

Begin by checking that your information on the title page is correct. For the thesis declaration, your institution may insist on something different than the text given. If this is the case, just replace what you see with what is required in the `DECLARATION PAGE` block.

Then comes a page which contains a funny quote. You can put your own, or quote your favourite scientist, author, person, and so on. Make sure to put the name of the person who you took the quote from.

Following this is the abstract page which summarises your work in a condensed way and can almost be used as a standalone document to describe what you have done. The text you write will cause the heading to move up so don't worry about running out of space.

Next come the acknowledgements. On this page, write about all the people who you wish to thank (not forgetting parents, partners and your advisor/supervisor).

The contents pages, list of figures and tables are all taken care of for you and do not need to be manually created or edited. The next set of pages are more likely to be optional and can be deleted since they are for a more technical thesis: insert a list of abbreviations you have used in the thesis, then a list of the physical constants and numbers you refer to and finally, a list of mathematical symbols used in any formulae. Making the effort to fill these tables means the reader has a one-stop place to refer to instead of searching the internet and references to try and find out what you meant by certain abbreviations or symbols.

The list of symbols is split into the Roman and Greek alphabets. Whereas the abbreviations and symbols ought to be listed in alphabetical order (and this is *not* done automatically for you) the list of physical constants should be grouped into similar themes.

The next page contains a one line dedication. Who will you dedicate your thesis to?

Finally, there is the block where the chapters are included. Uncomment the lines (delete the `%` character) as you write the chapters. Each chapter should be written in its own file and put into the *Chapters* folder and named **Chapter1**, **Chapter2**, etc... Similarly for the appendices, uncomment the lines as you need them. Each appendix should go into its own file and placed in the *Appendices* folder.

After the preamble, chapters and appendices finally comes the bibliography. The bibliography style (called *authoryear*) is used for the bibliography and is a fully

featured style that will even include links to where the referenced paper can be found online. Do not underestimate how grateful your reader will be to find that a reference to a paper is just a click away. Of course, this relies on you putting the URL information into the BibTeX file in the first place.

1.7 Thesis Features and Conventions

To get the best out of this template, there are a few conventions that you may want to follow.

One of the most important (and most difficult) things to keep track of in such a long document as a thesis is consistency. Using certain conventions and ways of doing things (such as using a Todo list) makes the job easier. Of course, all of these are optional and you can adopt your own method.

1.7.1 Printing Format

This thesis template is designed for double sided printing (i.e. content on the front and back of pages) as most theses are printed and bound this way. Switching to one sided printing is as simple as uncommenting the *oneside* option of the `documentclass` command at the top of the `main.tex` file. You may then wish to adjust the margins to suit specifications from your institution.

The headers for the pages contain the page number on the outer side (so it is easy to flick through to the page you want) and the chapter name on the inner side.

The text is set to 11 point by default with single line spacing, again, you can tune the text size and spacing should you want or need to using the options at the very start of `main.tex`. The spacing can be changed similarly by replacing the *singlespacing* with *onehalfspacing* or *doublespacing*.

1.7.2 Using US Letter Paper

The paper size used in the template is A4, which is the standard size in Europe. If you are using this thesis template elsewhere and particularly in the United States, then you may have to change the A4 paper size to the US Letter size. This can be done in the margins settings section in `main.tex`.

Due to the differences in the paper size, the resulting margins may be different to what you like or require (as it is common for institutions to dictate certain margin sizes). If this is the case, then the margin sizes can be tweaked by modifying the values in the same block as where you set the paper size. Now your document should be set up for US Letter paper size with suitable margins.

1.7.3 References

The `biblatex` package is used to format the bibliography and inserts references such as this one (**Reference1**). The options used in the `main.tex` file mean that the in-text citations of references are formatted with the author(s) listed with the date of the publication. Multiple references are separated by semicolons (e.g. (**Reference2**; **Reference1**)) and references with more than three authors only show the first author with *et al.* indicating there are more authors (e.g. (**Reference3**)). This is done automatically for you. To see how you use references, have a look at the `Chapter1.tex` source file. Many reference managers allow you to simply drag the reference into the document as you type.

Scientific references should come *before* the punctuation mark if there is one (such as a comma or period). The same goes for footnotes¹. You can change this but the most important thing is to keep the convention consistent throughout the thesis. Footnotes themselves should be full, descriptive sentences (beginning with a capital letter and ending with a full stop). The APA6 states: “Footnote numbers should be superscripted, [...], following any punctuation mark except a dash.” The Chicago manual of style states: “A note number should be placed at the end of a sentence or clause. The number follows any punctuation mark except the dash, which it precedes. It follows a closing parenthesis.”

The bibliography is typeset with references listed in alphabetical order by the first author’s last name. This is similar to the APA referencing style. To see how L^AT_EX typesets the bibliography, have a look at the very end of this document (or just click on the reference number links in in-text citations).

A Note on bibtex

The bibtex backend used in the template by default does not correctly handle unicode character encoding (i.e. "international" characters). You may see a warning about this in the compilation log and, if your references contain unicode characters, they may not show up correctly or at all. The solution to this is to use the biber backend instead of the outdated bibtex backend. This is done by finding this in **main.tex**: `backend=bibtex` and changing it to `backend=biber`. You will then need to delete all auxiliary BibTeX files and navigate to the template directory in your terminal (command prompt). Once there, simply type `biber main` and biber will compile your bibliography. You can then compile **main.tex** as normal and your bibliography will be updated. An alternative is to set up your LaTeX editor to compile with biber instead of bibtex, see [here](#) for how to do this for various editors.

1.7.4 Tables

Tables are an important way of displaying your results, below is an example table which was generated with this code:

```
\begin{table}
\caption{The effects of treatments X and Y on the four groups studied.}
\label{tab:treatments}
\centering
\begin{tabular}{l l l}
\toprule
\thead{Groups} & \thead{Treatment X} & \thead{Treatment Y} \\
\midrule
1 & 0.2 & 0.8 \\
2 & 0.17 & 0.7 \\
3 & 0.24 & 0.75 \\
4 & 0.68 & 0.3 \\
\bottomrule
\end{tabular}
\end{table}
```

You can reference tables with `\ref{<label>}` where the label is defined within the table environment. See **Chapter1.tex** for an example of the label and citation (e.g. Table [1.1](#)).

¹Such as this footnote, here down at the bottom of the page.

TABLE 1.1: The effects of treatments X and Y on the four groups studied.

Groups	Treatment X	Treatment Y
1	0.2	0.8
2	0.17	0.7
3	0.24	0.75
4	0.68	0.3

1.7.5 Figures

There will hopefully be many figures in your thesis (that should be placed in the *Figures* folder). The way to insert figures into your thesis is to use a code template like this:

```
\begin{figure}
\centering
\includegraphics{Figures/Electron}
\decoRule
\caption[An Electron]{An electron (artist's impression).}
\label{fig:Electron}
\end{figure}
```

Also look in the source file. Putting this code into the source file produces the picture of the electron that you can see in the figure below.



FIGURE 1.1: An electron (artist's impression).

Sometimes figures don't always appear where you write them in the source. The placement depends on how much space there is on the page for the figure. Sometimes there is not enough room to fit a figure directly where it should go (in relation to the text) and so \LaTeX puts it at the top of the next page. Positioning figures is the job of \LaTeX and so you should only worry about making them look good!

Figures usually should have captions just in case you need to refer to them (such as in Figure 1.1). The `\caption` command contains two parts, the first part, inside the square brackets is the title that will appear in the *List of Figures*, and so should be short. The second part in the curly brackets should contain the longer and more descriptive caption text.

The `\decoRule` command is optional and simply puts an aesthetic horizontal line below the image. If you do this for one image, do it for all of them.

\LaTeX is capable of using images in pdf, jpg and png format.

1.7.6 Typesetting mathematics

If your thesis is going to contain heavy mathematical content, be sure that \LaTeX will make it look beautiful, even though it won't be able to solve the equations for you.

The "Not So Short Introduction to \LaTeX " (available on CTAN) should tell you everything you need to know for most cases of typesetting mathematics. If you need more information, a much more thorough mathematical guide is available from the AMS called, "A Short Math Guide to \LaTeX " and can be downloaded from: <ftp://ftp.ams.org/pub/tex/doc/amsmath/short-math-guide.pdf>

There are many different \LaTeX symbols to remember, luckily you can find the most common symbols in [The Comprehensive \$\LaTeX\$ Symbol List](#).

You can write an equation, which is automatically given an equation number by \LaTeX like this:

```
\begin{equation}
E = mc^2
\label{eqn:Einstein}
\end{equation}
```

This will produce Einstein's famous energy-matter equivalence equation:

$$E = mc^2 \tag{1.1}$$

All equations you write (which are not in the middle of paragraph text) are automatically given equation numbers by \LaTeX . If you don't want a particular equation numbered, use the unnumbered form:

```
\[ a^2=4 \]
```

1.8 Sectioning and Subsectioning

You should break your thesis up into nice, bite-sized sections and subsections. \LaTeX automatically builds a table of Contents by looking at all the `\chapter{}`, `\section{}` and `\subsection{}` commands you write in the source.

The Table of Contents should only list the sections to three (3) levels. A `\chapter{}` is level zero (0). A `\section{}` is level one (1) and so a `\subsection{}` is level two (2). In your thesis it is likely that you will even use a `\subsubsection{}`, which is level three (3). The depth to which the Table of Contents is formatted is set

within **MastersDoctoralThesis.cls**. If you need this changed, you can do it in **main.tex**.

1.9 In Closing

You have reached the end of this mini-guide. You can now rename or overwrite this pdf file and begin writing your own **Chapter1.tex** and the rest of your thesis. The easy work of setting up the structure and framework has been taken care of for you. It's now your job to fill it out!

Good luck and have lots of fun!

Guide written by —
Sunil Patel: www.sunilpatel.co.uk
Vel: LaTeXTemplates.com

Appendix A

Frequently Asked Questions

A.1 How do I change the colors of links?

The color of links can be changed to your liking using:

```
\hypersetup{urlcolor=red}, or  
\hypersetup{citecolor=green}, or  
\hypersetup{allcolor=blue}.
```

If you want to completely hide the links, you can use:

```
\hypersetup{allcolors=.}, or even better:  
\hypersetup{hidelinks}.
```

If you want to have obvious links in the PDF but not the printed text, use:

```
\hypersetup{colorlinks=false}.
```