# Bayesian Inference for the Analysis of RNA-Seq Data

A thesis submitted by

Franziska Hoerbst

In fulfillment of the requirements for the award of the degree of

Doctor of Philosophy

John Innes Centre
Department for Computational and Systems Biology

University of East Anglia
School of Biology

Norwich, United Kingdom

2024

# Abstract

This thesis explores the application of Bayesian inference in the analysis of RNA-Sequencing data. We discuss the foundational principles of science and the mathematical articulation of belief updating via probability theory. In the Introduction we become familiar with the elegance and effectiveness of Bayesian approaches in addressing both simple examples and experimental research problems (Chapter 1). After introducing the RNA-Sequencing (RNA-Seq) technique and data (Chapter 2), we examine two use cases of Bayesian inference in the analysis of such data. First, we explore Bayes factors as a means of quantifying the evidence for gene expression changes (Chapter 3). Second, we propose Bayes factors to evaluate the evidence in RNA-Seq data for long-distance mobile messenger RNAs in plants (Chapter 6). For both applications, we perform in-depth analyses of the data (Chapters 4 and 5), test the methods and assumptions on simulated data, and compare it to currently popular published solutions (Chapters 3 and 6). We could confirm the outstanding performance of Bayes factors for the analysis of simulated and real data. Furthermore, we present and propose further inquiries, that uncover limitations and obstacles given by data collection and processing which cannot be addressed in statistical analyses (Chapters 7 and 8). The work in this thesis underscores the huge potential of Bayesian inference in enhancing scientific understanding and addresses the complexities involved in RNA-Seq data interpretation. The presented findings provide concise solutions for two statistical problems in RNA-Sequencing data analyses advancing our abilities to communicate new learnings from data in this field.

# Preface

The research presented in this thesis is a collaborative effort of scientists, coming together to improve data analysis and statistics in the field of molecular biology. There were hundreds of discussions that have shaped the work and words in this thesis. I would like to express my gratitude to all of them for their cooperation. This thesis summarises my work and contributions but would not have been possible without them. This thesis contains my thoughts and words and my code, it has been shaped, driven and written by me, but science is collaboration. Therefore, I will use the plural 'we' in the biggest parts of the thesis and only switch to singular form when the thoughts I am presenting are my own views.

Each chapter starts with a Transparency Note. They are informing about key colleagues involved in the work presented in the chapter and which publications the chapter is based on or came out of it. Even though I can mention who was involved in the work, it is literally impossible to make statements about who exactly did what. There are figures with data from one person, analysed by another, visualised by a third with the comments of a fourth. There were hours, days and weeks of parallel-coding efforts to minimise the chances for errors in our software.

Different chapters will read differently. Depending on the team of researchers shaping the work and communication, the final work changes its form. The product is a representation of all influences and philosophies. and, therefore.

For the first half of the thesis, where we introduce Bayesian inference into differential gene expression analysis, I would like to highlight, first and foremost, Richard Morris, and all members of the Morris group, in particular Melissa Tomkins, Gurpinder Singh Sidhu, and Thelonious Omori, together with our departmental colleagues Pirita Paajanen, Ander Movilla-Miangolarra, and Burkhard Steuernagel. I could sing many songs of praise about this wonderful research environment, for now, I will just say thank you for or all the time we spent discussing this topic.

The second half of the thesis contains huge efforts of a big research collaboration called PLAMORF. We bring biochemistry and structural biology together with plant physiology and computational sciences to investigate mobile mRNA in plants. We are a cooperation between the Research groups of Friedrich Kragler (Max-Planck-Institut für Molekulare Pflanzenphysiologie, Golm, Germany), Julia Kehr (Universität Hamburg, Hamburg, Germany), and Richard Morris (John Innes Centre, Norwich, United Kingdom). Here is a Thank you to all members for sharing knowledge and and learning experiences.

I enjoyed extraordinary supervision by my doctoral committee,

Richard J. Morris
John Innes Centre, Norwich, United Kingdom

Yiliang Ding
John Innes Centre, Norwich, United Kingdom

Melissa Tomkins
John Innes Centre, Norwich, United Kingdom

Fritz Kragler
Max-Planck-Institut für Molekulare Pflanzenphysiologie, Golm, Germany

Bojan Zagrovic
Max Perutz Labs & University of Vienna, Vienna, Austria

Thank you, and your accompanying teams, for mentoring my scientific process, and growth as a scientist.

All code for figures, analyses and simulations in this thesis is available on my GitHub.

This studentship as part of the PLAMORF project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 810131).

Thank you, European Union!

The thesis and myself, the student, have been examined on February 14th, 2025 by the examiners

Irene Papatheodorou
Head of Data Science at the Earlham Institute, Norwich, United Kingdom

James Locke
Sainsbury Laboratory, University of Cambridge, United Kingdom

Thank you for a very stimulating, fun, and insightful viva.

# Acknowledgements

THANK YOU,
to everyone who ever told me I can

THANK YOU,
to those who inspire me
every person
giving me a tiny piece of inspiration
at any point in my life

THANK YOU,
Familie
for raising me
for rooting me
for grounding me
for sending me off to grow
while welcoming me back any second
providing me contrast
and comfort

THANK YOU,
Tobias
and your family
und allen Gürteltieren

THANK YOU,
Austria, the internet, YouTube
for years and years of education
for global communities feeding curious brains

THANK YOU,
Science Center Netzwerk Wien
5MinutenClimateChance
UEA Biodiversity and Climate Action Network
for tons of inspiration
and the friendships that came along with it

Thank you,
Richard
for unlimited support and supervision
for your patience
for believing in me
for all the opportunities you have given me

Thank you,
Melissa
for even more of that
all the thoughts you have shared with me
and for letting me win the odd foosball game here and there to push my confidence

Thank you,
Morris dancers
for free jazz
for creating this special learning environment
for making these last years such a pleasure

Thank you,
Girls of 90P
the centre of peace and harmony
an oasis for female researchers
exploring life
expanding minds
supporting each other
morally mentally materially

Thank you,
Svenja

Thank you,
Emma

Thank you,
friends I met along the way
friends I met hiking
friends I met dancing
friends I met climbing
friends I met traveling
friends I met on trains
friends I shared food with
friends I shared coffee with
friends I had a good time with

Thank you,
friends who helped me balance
physically
friends who helped me balance
mentally
friends who keep challenging me
friends who keep supporting me

Thank you,
Fede

Thank you,
friends who care

Thank you,
for being there

Thank you,
nature, arts and the sun,
for keeping me sane

Thank you,
Josh

Thank you,
Figs and suppfigs

Thank you,
My Wensum Boys

Thank you,
Raúl

Thanks to the universe,
for all the luck that is poured over me all the time
I am very grateful

Thank you,
All of you,
I hope I can give as much as I am given

Good luck with love,
Franziska

# Prologue

## Slow down, Science, this is my love letter to you

Science, I love you. You bring joy into my life by fascinating and challenging me every day. You help me grow, you foster my inner child. You inspire me and give me hope. You amaze and remind me over and over again about the magnificence of life.

I love waking up not knowing what is going to blow my mind today, but knowing something will. I love dancing around the edges of knowledge, learning what has been explored and seeing new territory that humanity has not explored yet. I love the acceptance of failure you are teaching me, science. And that still you empower me, because only by being brave and believing in myself I can learn the arts, crafts and competences I need for heading off into the unknown by asking questions every day.

Allowing myself to ask big questions has become an important task for me during my time with you. Although it was certainly also taught and lived in my family and surroundings all those years before ... Now, science, you sometimes break my heart. I thought this was a connecting piece between those two worlds, my past and present. But how much do you allow yourself to ask big questions?

Carefully, I want to leave my criticism with you. While I am unbelievably grateful for my experience with you, I can feel a tension between us most days. I want you to know how I feel because I know we can do better. Everything I love about you is obscured by a shadow.

As students of life, we all need time and space to learn and explore. We need reminders that we can do more than we think. We need motivation to pursue visions, instead of destructive criticism that only hinders us from moving forward. Give this to humans, and they master the most difficult tasks, they are innovative and creative, and they push each other forward. Beautifully, all of what I just mentioned is not bound to resources. Sadly you showed me, as a student of science, that all of the above is a rare gift. Why is that?

Science, can't you see the structural problem underneath your feet?

Science, you can't tell me you don't understand. I know it's painful to think about it and I know it will hurt. But take my hand, I know you can do it. Be self-critical and allow things to change.

My activist nature can't look away from how you have operated the last few years. You are busy spinning the wheel, caught in the dogma of growth. And you yourself know best how dangerous a dogma is for innovation. Science, you are driving me crazy when you are driven by capitalistic forces, by growth for the sake of growth. Is this truly who you are?

Science, I love you when you are genuinely curious, when you're infecting with inspiration. I love it when you carry me off to unknown places in my imagination. Whenever the game of science gets into you, when things are done just because we can, I am losing my connection to you. Then I am disappointed and I start worrying about the state of the world. I thought you were there to help me out of that. I thought you were here to keep surprising me, and show me something new every day.

Change is hard, and change is slow. Hope is precious. Science, can't we support peaceful life on a planet full of life? With the unbelievable we constantly manage to do, can't we also do that? Let's explain our thoughts to each other. Let's empower each other. Let's get out of the fairytale of growth.

We need time. Science, you need time. Being immersed in the loops of more and more we hardly allow ourselves to ask the big question. WHAT ARE WE DOING HERE?

Science, please, slow down. Growth, the only thing that matters in life is not the only thing that matters in life.

# Contents

# Chapter 1

# An Introduction to Bayesian Statistics

## 1.1 What is science and why are we collecting data?

In science, at least the way I understand it, we want to systematically and empirically acquire *knowledge* about the natural world. This encompasses the formulation and testing of general laws through the scientific method. In biology, we want to extract knowledge about life from *data*. How can we do that?

We record data by observing and measuring the world and life around us, which transform into *information* once we assign labels or meanings to them, Figure 1.1. As a means to distill information from data, which may eventually lead to the gain of knowledge, humanity has discovered the need for statistics and philosophy. It arises from uncertainty and variability being pervasive. Statistics serves as the compass that guides us through the labyrinth of data, enabling us to derive meaningful insights and make informed decisions. In this thesis, the main contributions are efforts to organize, summarize, and interpret a deluge of data that is recently flooding biology. However, data – no matter what and how big – can always only provide a glimpse of information about a potentially complex system. Therefore, even if we are dealing with a flood, we will explore throughout this thesis what we can and cannot learn from limited bits of information and, how problems arise that statistics can or cannot solve. Just like any good tool, it can only ever help us uncover information that is actually present in the data. If information is not present, we face a problem of 'Garbage in, garbage out'.

## 1.2 Best guesses from limited information

Measuring things can be hard. This may be because of its unstable nature (How many hours of sunshine do we get in a day in Norwich?), or because measuring it is a technically challenging (How fast can flies fly?). We still want to try recording observations, of course, because we are curious, and then we can see whether there

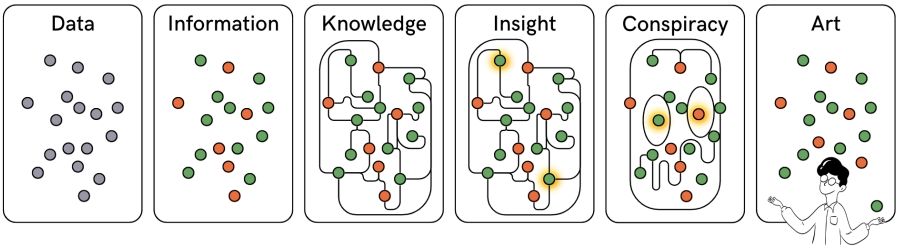



Figure 1.1: *Data* are the record of observations, which turn into *information* once we attach a label or meaning to it. There may be labeled data, however, with no information in it. In information theory, the mathematical study of information and communication, we would say it has a low entropy, meaning there is no surprise for us in the recordings. If we get surprised by information in data, it may teach us something new, and contribute to *knowledge*, once we find the connections and causalities causing the surprise in the data. Learning something new can happen at any point, maybe we record some more observations which give us new *insights* (that surprise us), from which we can deduce new knowledge that we can weave into our current beliefs. In contrast, we can also ignore all of these prior steps, methodology and philosophy and transform data, information, knowledge, and insights into a *conspiracy*. Alternatively, we can choose to set aside rationality and engage with our emotions, creating something profoundly meaningful from any of the given pieces – *art*. Artistic expression allows us to explore the emotional and subjective dimensions of our experiences, highlighting the diverse ways we can interpret and interact with the world around us.

is information in our data, that we can potentially deduce knowledge from. To achieve this, we want to measure many times to capture variability and potential measuring errors. Unfortunately, we can do our best to measure in the most excellent way possible, and still end up with limitations (in terms of information, entropy, surprise) of our data.

We can use Probability theory to express the uncertainty of a measurement. Pioneers Thomas Bayes and Pierre-Simon Laplace explored this in their work in the 18th and 19th centuries. As the story goes, their contributions were largely discredited and forgotten. Until many many decades later several researchers, among them Harold Jeffreys, Edwin Thompson Jaynes, and Richard Cox rediscovered and popularised the powers of Probability theory in statistics. They founded what we now know as Bayesian statistics.

Everything may have started with another mathematician, James Bernoulli, asking the question of how to reason in situations where one cannot argue with certainty. Bernoulli formulated a difference between deductive logic and inductive logic. Deductive reasoning moves from general principles to specific instances, and inductive reasoning moves from specific instances to general principles [1]. He was pondering over the question whether we can use the mathematics and mechanisms of probability, which can be used in games of chance, for inference problems appearing in our everyday lives ...

It is established now that we can use probability theory for inference problems and in the next pages I will try to introduce how 'Bayesian inference' works, before we will encounter new success stories later in the thesis. To the early pioneers, a probability represented a degree of belief or plausibility of an event: How much do we think something is true, given some evidence we have at hand? This is the essence of hypothesis testing and model comparison we are going to use.

## 1.3 Games of chance meet daily life

You are playing a simple card game with a full deck of cards. Your partner reveals to you the top card on the stack, now it is your turn to guess whether the next card in the stack has a higher or lower value. It's the beginning of the game, they show you a 2. Do you choose 'above' or 'below', and why?

You are rushing to a meeting in a university building you have not been in before. Suddenly, you feel your period starting. You have no female hygiene products on you. You need to find another woman to hopefully help you out. There is a junction in your path, one leading to the computer science department, one leading to the department for environmental sciences. Where do you go and why?

Chances are good, you have just intuitively done what took humanity a while to figure out consciously. Your past experiences in playing cards and navigating university buildings have informed your decision on your best choice of action. You naturally evaluated which choice is more likely to bring you success. Now, how would your decisions change, however, if you are at the end of the card game, there are 4 cards left and you have not yet seen a single 1? How would your decisions change if it is 2024 and you would assume university staff has a good representation of all genders no matter the discipline?

You can answer these questions easily because you have been confronted with them a lot of times in your life and it comes intuitive to you now. Let's pretend for a minute you are a child and you have never played a card game in your life before. What do you do? You play, and lose and win and lose until you eventually start noticing patterns. You start to get a feeling for the chances of winning with certain values. The pioneers I mentioned earlier turned this intuition into a mathematical framework: probability theory. What's the probability of winning the game by saying 'above' given your partner is showing you a 1? 2? 3? 4? 5? And so on? And furthermore, how do we learn to get better at this game by playing it lots of times?

### 1.3.1 Dissecting the games of chance

Let's say, $\theta$ is the event of you winning the game by saying 'above' and $n$ is the rank of the card that is shown to you first. We can describe our uncertainty of the events occurring with probabilities, $P(\theta)$ the probability of you winning by saying 'above' and $P(n)$, the probability of seeing a card ranked $n$, Figure 1.2.

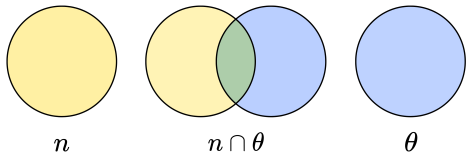



Figure 1.2: The blue event $\theta$ and the yellow event $n$ are happening with certain probabilities, $P(\theta)$ and $P(n)$. The conditional probability of $\theta$ given $n$ is found in the green overlap.

We are interested in the intersection of these events, the conditional probability of $\theta$ given $n$, $P(\theta|n)$. How likely are we to win by saying 'above' given the card we are shown. We can get this from looking at the joint probabilities of $\theta$ and $n$, the union $P(n, \theta)$, first,

$$\begin{aligned} P(n, \theta) &= P(n|\theta) \times P(\theta) \ , \\ P(\theta, n) &= P(\theta|n) \times P(n) \ . \end{aligned} \tag{1.1}$$

The probabilities $P(n, \theta)$ and $P(\theta, n)$ describe the same union of events $n \cap \theta$, hence,

$$\begin{aligned} P(n, \theta) &= P(\theta, n) \\ P(n|\theta) \times P(\theta) &= P(\theta|n) \times P(n) \ . \end{aligned} \tag{1.2}$$

From here we can get the intersection, the probability of $\theta$ given $n$. Which is the famous Bayes' theorem in action,

$$P(\theta|n) = \frac{P(n|\theta) \times P(\theta)}{P(n)} \ . \tag{1.3}$$

It provides us with an equation for a conditional probability, to calculate how likely we are to win the game by saying 'above' ($\theta$) given the card we are shown ($n$). For hypothesis testing, the card that is shown to us is nothing other than some data $D$ we would collect in an experiment, like an observation we are making. The probability that we win the game by saying 'above' is a hypothesis $H$ we are testing. Therefore, we can also write Bayes' theorem for hypothesis testing as

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} \ , \tag{1.4}$$

which we call posterior probability, $P(H|D)$, the probability that our hypothesis $H$ is true, given the data $D$ we are seeing. The posterior is determined by two factors, a prior probability $P(H)$ and a likelihood $P(D|H)$ (and a third factor $P(D)$ that is just a normalisation constant – see below).

Coming back to our card game example with our event of winning and our events of seeing certain cards $\theta$ and $n$, the posterior probability $P(\theta|D)$, determined by

the prior probability $P(\theta)$ and the likelihood of $\theta$, expressed as $P(D|\theta)$, can be written as

$$P(\theta|D) \;=\; \frac{P(D|\theta) \times P(\theta)}{P(D)} \;, \tag{1.5}$$

or more precisely,

$$P(\theta|D, H) \;=\; \frac{P(D|\theta, H) \times P(\theta|H)}{P(D|H)} \;, \tag{1.6}$$

given that everything is dependent on the hypothesis that we formulate.

### 1.3.2 Formulating hypotheses, defining models

As I stated earlier, the goal in science is to deduce knowledge from data. We want to find the general principles explaining specific observations we make. So we formulate hypotheses, that we wish to test by making observations and evaluating whether they can explain what we are seeing. A hypothesis is a specific statement or proposition about the system we are investigating and we can only test it, by also defining a model. A model in Bayesian statistics refers to a mathematical representation of the data-generating process: How did the data come about? It includes the likelihood function, which describes how the observed data is generated given certain parameters. The model incorporates prior beliefs about the parameters through the prior distribution. A hypothesis, the statement or proposition we are making, is about a parameter or a set of parameters within the context of the model. We can test hypotheses by calculating the posterior probabilities of these hypotheses given the observed data. The likelihood $P(D|\theta)$ is our means of feeding data we gathered in an experiment into the equation: We collect data ($D$) and calculate the probability of obtaining the data, given the underlying model of $\theta$, i.e. how well our model reproduces the data.

The denominator normalises the posterior probability. It is the probability of the data. This normalising factor is also called the evidence (or the marginal likelihood), and it has got this prestigious name for a reason. It will later play an important role in hypothesis testing and Bayes factors.

Probably, this is not how you have learned to play card games. You did not think about posterior probabilities, hypotheses and models. What you did is thousands of experiments, collecting experiences of failing and learning until you reached your current knowledge about card games. This learning process over the years of your life, is a knowledge updating process that is essentially the same as learning from data in our scientific experiments, follow me through Figures 1.3, 1.4, 1.5.

The posterior is a probability distribution describing our *current* knowledge. For example, the probability that we are going to win the card game, given our *current* knowledge about card games. The probability distribution over the event of


Figure 1.3: If we are playing a card game for the first time in our life we have no *prior knowledge* how to optimise our chances for winning. We have no tactics. We just start an experiment (Experiment 1), collect some experience of winning and losing, data (Data 1), and eventually know more about how to win afterwards (*posterior knowledge*). If we are playing as an adult, we have some experience in card games. This *prior knowledge* can help us to understanding the game before we have even started playing (collecting new data).

winning $\theta$ that covers all possible values (between 0 and 1), depends on the data $D$, an observations of games.

With every game we play, we learn. We do another Bayesian knowledge updating step, just as described in Equation 1.3 and 1.4. The equations formalise how to learn more about the game by updating the probability distribution over $\theta$ by including more data or collecting experience, Figure 1.4 and 1.5.

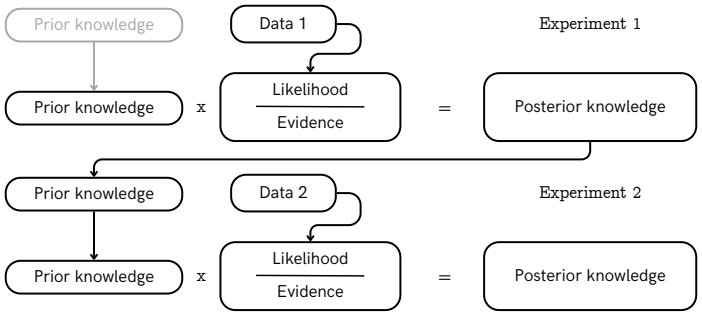



Figure 1.4: With every game we play and every experiment we conduct, we are updating our knowledge about card games, this world and how to find tampons in emergency situations. All Prior knowledge can be taken into account to update our posterior knowledge. As an adult, we have all of our past experiences behind us, making it easier to make an informed decision between 'above' and 'below'. A researcher with a lot of background knowledge will read and interpret a new paper in their field in very different light than a student immersing for the first time.

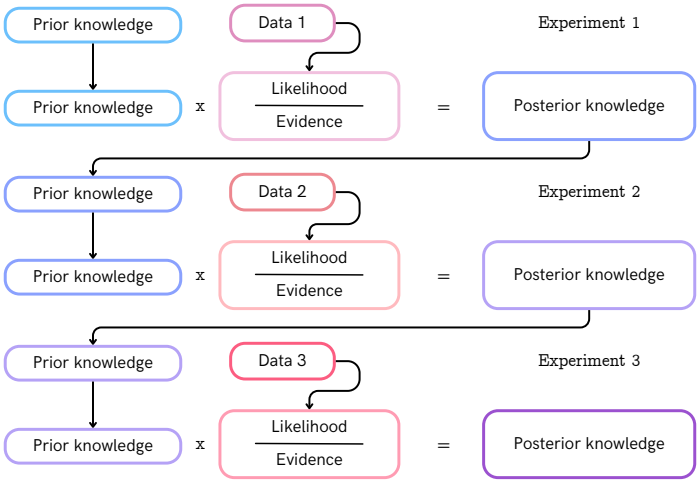



Figure 1.5: Another visualisation of the concept of Bayesian knowledge updating: This figure is exactly the same as the last ones but by applying colours we want to emphasise the learning process (starting from blue) and the mixing in of new information (shades of red) changing our current view, the posterior (purple). In this case the colours are getting more and more intense and vibrant. This is not the case in all learning processes, however. What if experiments show opposing results?

### 1.3.3 Humans interpreting probabilities

Because all of what we have talked about so far happens very fast and automatically, we do not think about it too much. Now comes the bad news: Humans tend to be bad at recognising which bits of prior knowledge are important to make good decisions. Science communicator Grant Sanderson identified the difficulty aptly, stating: 'Rationality is not about knowing facts, it's about recognising which facts are relevant' [2], [3]. For example, while we were focussing on finding a woman in this university building (especially in a stressful situation), did we consider how big the two departments are? How likely are we going to meet any person in the department? Are scientists going to be working from home? Are they going to be on fieldwork? These are the moments where theory and practice deviate far from each other and where we are losing in card games because we under- or over-estimate our chances being distracted by irrelevant facts. Prior knowledge matters.

## 1.4 The medical test paradox

A famous example demonstrating how prior information can fundamentally change the interpretation of a situation is the Medical Test Paradox. It is not a real paradox but a veridical paradox teaching us how an accurate test (for a disease, or a hypothesis) is not necessarily a predictive test (for having the disease, or a hypothesis being true).

Assume there is a breast cancer screening and 1000 women are tested. 10 of those women do suffer from breast cancer, of which 9 get a positive test result. 89 further women get a positive test result, although they do not have cancer, just like the

901 cancer-free women with negative test results, Table 1.2. You are one of the 1000 women in the test and you get your results back with a positive outcome, how likely is it, that you have cancer?

In your first shock, you might be sure you have cancer, because you consider the worst and you feel you have good evidence – a positive test result – that you suffer from the disease. But eventually you will question how accurate the test is that is turning your world upside down. Conventionally, we can calculate test accuracy statistics from the numbers given above, numbers that are important to be available for all medical tests for quality assurance. Table 1.1 will help to remind us how test statistics are usually calculated.

| Total = P+N | Test (+) | Test (-) | | |
|---|---|---|---|---|
| Condition (P) | TP | FN | TPR = Sensitivity = TP/P | FNR = FN / P |
| Condition (N) | FP | TN | FPR = FP/N | TNR = Specificity = TN/N |

Table 1.1: There are 4 possible scenarios that we can count for the outcome of a test. The result can be positive (+) and negative (-) and the condition can be given (P) or absent (N). For a person with the disease (P) receiving a positive test (+) we get a true positive hit (TP), a person with the condition (P) receiving a negative test (-) a false negative hit (FN). Subsequently we get a false positive hit (FP) for a person without the disease (N) receiving a positive test result (+), and a true negative hit (TN) for a person correctly being identified as disease-free. We can summarise the performance of our (medical) test in four values: the true positive rate (TPR) also called Sensitivity of a test, the false negative rate (FNR), the false positive rate (FPR), and the true negative rate (TNR), also known as specificity.

| Total = P+N | Test (+) | Test (-) | Σ |
|---|---|---|---|
| Condition (P) | 9 | 1 | 10 |
| Condition (N) | 89 | 901 | 990 |
| Σ | 98 | 902 | |

Table 1.2: We are using a cancer screening as an example for the medical test paradox. We are testing a group of 1000 people, of which 10 have the condition (P) and 990 do not (N). The test is showing a positive result (+) on 98 patients and a negative result (-) on 902. 89 false positive (FP) cases and 1 false negative (FN) case have occurred.

The test in our example has a sensitivity of 90%,

$$\text{sensitivity} \;=\; \frac{\text{TP}}{\text{TP}+\text{FN}} \;=\; \frac{9}{9+1} \;=\; 0.9\;, \tag{1.7}$$

and a specificity of 91%,

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{901}{901 + 89} = 0.91 \,. \tag{1.8}$$

These numbers were given to a group of gynecologists in a study in the 90s [4], where they were asked the same question as above. What are the odds that your patient actually has cancer given they received a positive test result? They were given the answers (A) 9 in 10, (B) 8 in 10, (C) 1 in 10 and (D) 1 in 100; What do you think is the right answer?

In this influential piece of research for the fields of psychology and medicine we have learned how people often misunderstand probabilities and the significance of base rates when interpreting medical test results. But *often* is actually an understatement: In this specific experiment we have learned that the group of healthcare professionals have performed worse than random in the test. Most picked the worst case (A), while (C) is the true answer. Once we see a visualisation of the test statistics, we get a better intuition for what is going on here: We underestimate what the chances to have cancer are, Figure 1.6.

We can conclude that there is the need to find better ways to communicate medical tests, if their statistics are heavily misinterpreted. But now, if medical tests are so likely to be misinterpreted, what about scientific results?

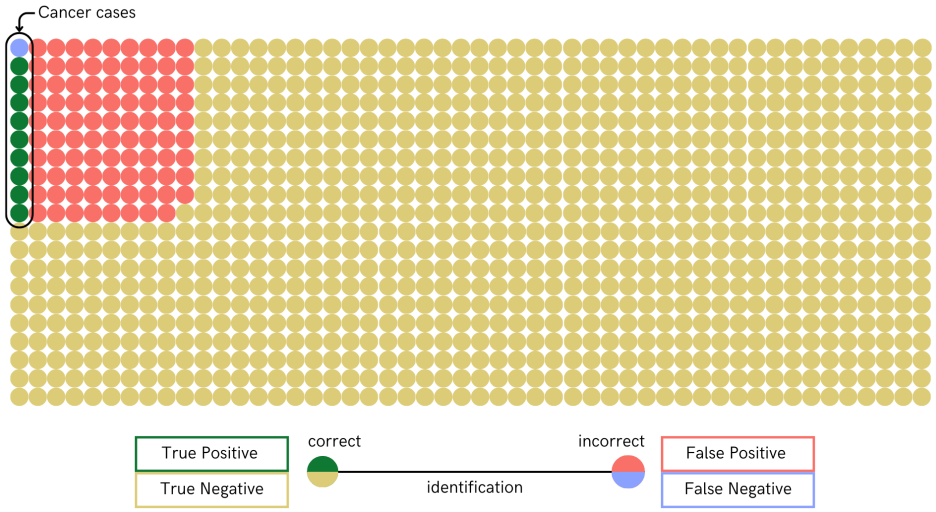



Figure 1.6: The medical test example we are using is a cancer screening with 1000 participants represented in this figure by 1000 circles. There is a 1% prevalence for the disease, the test has a sensitivity of 90%, and a specificity of 91 %. If there are 10 cancer cases among 1000 participants, receiving one of 98 positive results does update your chances of having cancer from 10 in 1000 to 1 in 10.

The implications of the medical test paradox will follow us through the rest of this thesis, out into the world and our everyday life. How do we combat the misunder-

standing of tests?

## 1.5 Introducing Bayes factors

One good way to help us correctly estimate how likely we are to suffer from a disease given a positive test result is by rephrasing the question. By what factor have the odds of carrying a disease increased, given a positive test result, as compared to before the test? This factor is called a Bayes factor. Before we do the test we might know how widespread the disease is. We can update this prior knowledge about our chances of having the disease by doing a test. It depends on the accuracy of the test how much we learn from a positive test result. To summarise the test statistics we can calculate this Bayes factor, a single value that summarises the 4 values we had been given earlier (Table 1.1) with TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative); or the 2 in form of sensitivity and specificity).

If we want to express the posterior knowledge about having the disease given a positive tests result $P(\mathrm{P}|+)$, we can use Bayes theorem,

$$P(\mathrm{P}|+) \;=\; \frac{P(+|\mathrm{P}) \times P(\mathrm{P})}{P(+)} \;=\; \frac{P(+|\mathrm{P}) \times P(\mathrm{P})}{P(+|\mathrm{P}) \times P(\mathrm{P}) \;+\; P(+|\mathrm{N}) \times P(\mathrm{N})} \;, \tag{1.9}$$

where we shall not forget to take the prior, the prevalence expressed as $P(\mathrm{P}) = (\mathrm{prior}) = \mathrm{P}/(\mathrm{P}+\mathrm{N})$, into account. We can also pronounce this with the vocabulary of test accuracy statistics where the prior is the prevalence of the disease,

$$P(\mathrm{P}|+) \quad = \quad \frac{(\mathrm{prior})(\mathrm{TP}/\mathrm{P})}{(\mathrm{prior})(\mathrm{TP}/\mathrm{P}) + (1-\mathrm{prior})(\mathrm{FP}/\mathrm{N})} \;. \tag{1.10}$$

If we express the prevalence in odds,

$$O(\mathrm{P}) \;=\; \frac{\mathrm{P}}{\mathrm{N}} \;=\; \frac{10}{990} \;, \tag{1.11}$$

we can update the prior odds of having the disease to posterior odds $O(P|+)$ of having the disease after doing a test. Given we received a positive test $(+)$, we can multiply the prior odds with a Bayes factor for the test we have done,

$$O(\mathrm{P}|+) \quad = \quad \frac{\mathrm{P}}{\mathrm{N}} \times \frac{P(+|\mathrm{P})}{P(+|\mathrm{N})} \quad = \quad O(\mathrm{P}) \times \frac{\mathrm{TP}/\mathrm{P}}{\mathrm{FP}/\mathrm{N}} \quad = \quad O(\mathrm{P}) \times \frac{\mathrm{sensitivity}}{\mathrm{FPR}} \;, \tag{1.12}$$

where the Bayes factors are tidily separated. We can remember, that if we formulate the prevalence in odds, we can simply multiply the Bayes factor with the prevalence to update our beliefs on the odds of having the disease after receiving a positive test result,

$$\text{posterior odds} \quad = \quad \text{prior odds} \;\times\; \text{Bayes factor} \;. \tag{1.13}$$

Receiving a positive test in our example changes the prior odds from 10 to 990 (1 in 100) to the correct answer (C) 1 in 10,

$$O(\mathrm{P}|+) \quad = \quad O(\mathrm{P}) \times \frac{\mathrm{TP/P}}{\mathrm{FP/N}} \quad = \quad \frac{10}{990} \times \frac{9/10}{89/990} \quad = \quad \frac{100}{990}\,, \tag{1.14}$$

with a Bayes factor of $\approx 10$.

If we tell the story about medical tests in the framework of probability theory, conveying that we are updating our prior knowledge by doing a test, would we have fewer misinterpretations about the implications of medical tests? Frankly speaking, assigning priors is already hard enough. We need to factor in the prevalence, symptoms, or contacts for contagious diseases, ... Why are we making the test interpretation another hurdle if we can learn more about probability theory and how to use it for inference problems? This will be the red thread for the rest of this thesis. We will be using this beautiful framework, and find use cases of Bayes factors for different challenges in molecular biology.

## 1.6 Bayes factor derivation for a binomial toy problem

Now that we have familiarised ourselves with the first and most important Bayesian statistics tool in this thesis, the Bayes factor, we can explore what else we can use it for. We mentioned earlier, that we can also describe learning processes in science with the posterior distribution, simply by exchanging events and their probabilities for data we are seeing and hypotheses we want to test, equation 1.4. Although we will return to the medical test paradox in the Discussion of the thesis, we can leave this exact use case of Bayes factors as a test statistic for medical tests behind us and take a look at how to use Bayes factors for other hypothesis tests. Let's return to our card game example for that.

We are watching a game of 'above or below'. As described earlier, the tactics (or lack of tactics) of a child who has never played a card game before are probably very different compared to an adult who has been confronted with similar games many times before in their life. We want to see whether the lack of tactics in this game makes a difference, or whether an adult could relax and close their eyes during that game, meaning it is all about luck anyway. We can do that using Bayes factors.

We want to find an equation for a Bayes factor that helps us evaluate after each game which hypothesis is more likely to be true. Both players, adult and kid, are shown the same cards and they give their predictions independently. We want to know whether there is a difference in their successes and how it changes over time, with more and more games the kid is exposed to. If we let our players compete in many games we might also be able to observe the learning progress of the kid (or the adult) over time. We are going to get a limited amount of data, as we are only recording a few games with our players, and so we employ Bayesian statistics here to help us make sense of our observations.

### 1.6.1 Defining the problem and stating the hypotheses

Each game consists of $N$ rounds. In each round the players will be shown a card. They have to state their guesses and we count how often the players win ($n$ successes). From this data, we can infer a success rate $q$ for each of the players. The question we are asking is essentially whether the success rates of our players are the same, or different.

We can formulate two hypotheses. Hypothesis $H_1$ states that the success rates $q_1$ and $q_2$ do not differ; understanding the probabilities behind the card game does not lead to a different outcome than making random decisions. Hypothesis $H_2$ states that the probabilities for success are different; a naive player achieves very different results than an experienced player. Therefore, the problem we are evaluating is whether,

$$q_1 \stackrel{?}{=} q_2 \ . \tag{1.15}$$

### 1.6.2 Finding a model

The simplest model to describe the process we are observing is a binomial model, as we are observing a probabilistic event with two outcomes: winning or losing. Hence, we can describe the relationships between success rate $q$, the number of successes $n$, and the number of rounds $N$ in our data with a binomial distribution,

$$P(n|N,q) = \binom{N}{n} q^n (1-q)^{N-n} \ . \tag{1.16}$$

This description implies, for example, that if we knew the success rate $q$ of a player and the number of rounds they played, $N$, we can, with the help of the binomial distribution, compute the probability of any number of successes $n$, Figure 1.7.

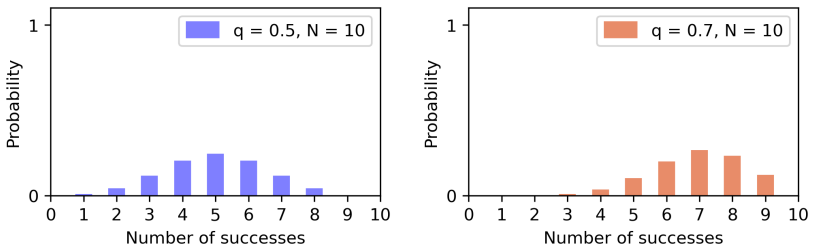



Figure 1.7: Imagine we are watching two players with different tactics represented by different success rates, 0.5 and 0.7. With the help of the binomial distribution we can compute the probability of success in $N$ rounds for all possible $n$ from 0 to 10.

### 1.6.3 Learning from data

Before the players have played the game we could guess how well they will perform, but, typically, their success rates $q_1$ and $q_2$ are unknown. However, once they have played a game (or better several) we have the outcomes or data, and we can infer the success rates using Bayes' theorem as an inverse probability problem. As we

do not know $q$, but we want to learn about $q$ from our data, we shall refer to this as $\theta$ from now on. Importantly, $\theta$ is not $q$, but the range of values that $q$ may take and over which we assign a probability distribution reflecting what we know about it.

As we want to compare our two players, child and adult, to see whether their success differs, we are, technically speaking, asking whether the successes of our players, $n_1$ for player 1 and $n_2$ for player 2,

$$q_1 \rightsquigarrow n_1 \quad \text{and} \quad q_2 \rightsquigarrow n_2 \ , \tag{1.17}$$

arose from the same success rates q (which we don't know), or not. This implies that we are comparing two models, in Hypothesis $H_1$ a single parameter $\theta$ is enough to explain all data, whereas in Hypothesis $H_2$ we need two different parameters $\theta_1$ and $\theta_2$.

Before we collect our data we don't know about the performance of our players. Our prior information only consists of the fact that we are dealing with a probabilistic event with two outcomes which we can describe with a Binomial distribution. We don't know about the outcomes.

Let's look at some results of the games, Table 1.3.

| N = 10 | Adult | Kid |
|---|---|---|
| Game 1 | 6 | 3 |
| Game 2 | 5 | 7 |
| Game 3 | 8 | 7 |

Table 1.3: Our players have played 3 games (10 rounds each) of 'above or below'. We have counted how many times they have guessed the right answer. Both players were shown the same cards and they chose their answers independently.

We can infer an expectation value $\langle\theta\rangle$ for the success rates of our players from their success counts $n$ among games $N$,

$$\langle\theta\rangle \ = \ \frac{n+1}{N+2} \ , \tag{1.18}$$

following Laplace's rule of succession to infer probabilities for events given limited observations.

How can we learn more about $\theta$ from the data, beyond an expectation value? What is the probability distribution over $\theta$, given (new) success counts of our players? Bayes theorem is back, helping us out one more time. We can write the posterior distribution $P(\theta|n)$, our newest updated knowledge, as

$$P(\theta|n) \;=\; \frac{P(n|\theta) \times P(\theta)}{P(n)} \;. \tag{1.19}$$

From the binomial distribution above we know that the number of successes $n$ depends on $\theta$ and the number of rounds $N$. We also include this bit of information in the posterior,

$$\begin{aligned} P(\theta|n,N) \;&=\; \frac{P(n|\theta,N) \times P(\theta)}{P(n,N)} \;, \\ P(\theta|D) \;&=\; \frac{P(D|\theta) \times P(\theta)}{P(D)} \;. \end{aligned} \tag{1.20}$$

To calculate the posterior we need to find all the building blocks of the equation, in other words, we need to find expressions for all of them. We remember, the posterior is determined by two factors, the prior probability $P(\theta|H)$ and the likelihood of $\theta$, $P(D|\theta,H)$, normalised by the evidence $P(D|H)$, Equation 1.5 and 1.6.

### 1.6.4 The building blocks of a posterior distribution

The likelihood is how we get data that we gathered in an experiment into the equation: We collect data ($D$) and calculate the likelihood of the parameters producing the data. We do this by employing a process model to describe the probabilistic process of how the data came about. This means we calculate the probability of observing the data given $\theta$. Hence, we use the binomial distribution we had earlier to describe the process (Equation 3.2), to define the likelihood over $\theta$,

$$P(D|\theta) \;\propto\; \binom{N}{n}\, \theta^n (1-\theta)^{N-n} \;. \tag{1.21}$$

First, we need an expression to feed our prior knowledge in a prior distribution. This distribution captures our knowledge about the system before we collect data or before we include more data and update our knowledge. We can choose whichever distribution best describes our prior knowledge of $\theta$. If we think that all values of $\theta$ are equally possible, we might assign a uniform prior.

The mathematical form of the prior influences the ease of further calculation. The Beta distribution is the conjugate distribution of the binomial distribution and therefore a convenient choice for the prior, owing to it having the same functional form (a choice that will lead to some satisfying simplifications later, extraordinarily convenient choice),

$$P(\theta|u_1,u_2) \;=\; \frac{1}{B(u_1,u_2)}\; \theta^{u_1-1}\; (1-\theta)^{u_2-1} \;=\; \mathrm{Beta}(u_1,u_2) \;, \tag{1.22}$$

where the hyper-parameters $u_1$ and $u_2$ can be used to capture existing knowledge about the success rate $\theta$. The Beta distribution is named after its normalisation factor, the Beta function $B(u_1,u_2)$,

$$B(u_1, u_2) \;=\; \int_0^1 \theta^{u_1-1}\,(1-\theta)^{u_2-1} d\theta\;. \tag{1.23}$$

Can you spot the similarities? The prior gives us the option to start with no knowledge or bias (called flat prior), where all outcomes are equally likely by setting $u_1 = u_2 = 1$. Theoretically, one could introduce a bias to favor one success rate $\theta$ over another, but as we stated earlier, there is no good reason here to do that. Figure 1.8 shows the concept of Beta priors and how we could introduce bias over $\theta$ to capture either theoretical reasoning or knowledge from previous experiments.

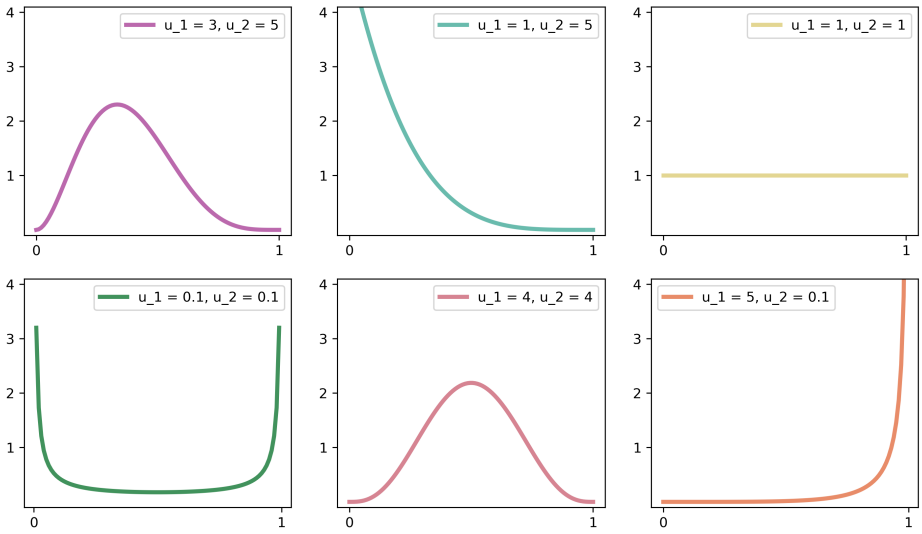



Figure 1.8: Beta distributions change their shapes according to the hyper-parameters $u_1$ and $u_2$. Choosing a 'flat prior', $u_1 = u_2 = 1$, means introducing no bias or prior information.

Finally, the evidence $P(D)$ found in the denominator of equation 3.7 is the probability that the data $D$ is produced. We are computing the probability for seeing data $D$ given all possible values of $\theta$, which means we want to sum over all probabilities for all values of $\theta$ between 0 and 1, which is essentially integrating over $P(D|\theta)$ weighed by how likely each $\theta$ is, $P(\theta)$,

$$P(D) \;=\; \int_0^1 P(D|\theta) \times P(\theta)\; d\theta\;. \tag{1.24}$$

To visualise this integration, you can picture that the evidence is the area under the curve of $P(D|\theta) \times P(\theta)$. The evidence tells us how much the possibility space given by the hypothesis collapses by looking at the data.

Note, that this is not limited to 2-dimensional spaces and can be extended to 3 or more if there are several variables you integrate over, $P(D|\theta_1, \theta_2) \times P(\theta_1, \theta_2)\;\ldots$

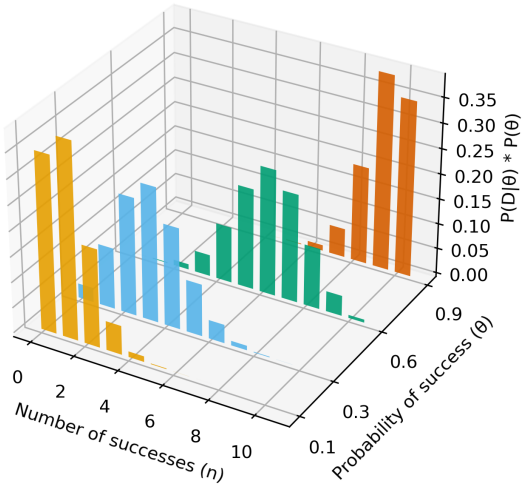



Figure 1.9: We can calculate the estimated number of successes for any probability of success $\theta$. Remember, in our calculations, we assign a probability distribution over $\theta$, instead of only considering a single value. Finding the evidence is calculating the space under under the curve of $P(D|\theta) \times P(\theta)$, given data $n$. The evidence is a measure for how much the possibility space collapses by data given the hypothesis or model.

### 1.6.5 The assembly of the posterior distribution

Now that we discussed all the building blocks of the posterior, let's put them together for our two hypotheses.

Using the binomial likelihood, we can infer the posterior distribution over the success rates of our players from the recorded games, starting from a prior for Hypothesis $H_1$,

$$P(\theta|H_1) \;=\; \text{Beta}(u_1, u_2) \;=\; \frac{1}{B(u_1, u_2)}\; \theta^{u_1-1}\; (1-\theta)^{u_2-1}\;, \tag{1.25}$$

and two priors for two parameters in Hypothesis 2,

$$P(\theta_1, \theta_2|H_2) \;=\; \text{Beta}(u_1, u_2) \times \text{Beta}(u_1, u_2)\;, \tag{1.26}$$

which we will update with the results from the games we recorded by multiplying with the likelihood. The likelihood of $\theta$ for given data $(D_1, D_2)$ and Hypotheses 1, following a process model with one parameter, can be written as

$$P(D_1, D_2|\theta, H_1) \;=\; \binom{N}{n_1}\theta^{n_1}(1-\theta)^{N-n_1} \;\times\; \binom{N}{n_2}\theta^{n_2}(1-\theta)^{N-n_2}\;. \tag{1.27}$$

For Hypothesis 2, the likelihood of the parameters, $\theta_1, \theta_2$, given $H_2$ over two models, is

$$P(D_1, D_2|\theta_1, \theta_2, H_2) \;= \binom{N}{n_1}\theta_1^{n_1}(1-\theta_1)^{N-n_1} \;\times\; \binom{N}{n_2}\theta_2^{n_2}(1-\theta_2)^{N-n_2} \;. \tag{1.28}$$

If we now put together our equation for the posterior of Hypothesis $H_1$, as we did earlier for the general case, we get,

$$P(\theta|D_1, D_2, H_1) \quad = \quad \frac{P(D_1, D_2|\theta, H_1) \times P(\theta|H_1)}{P(D_1, D_2|H_1)} \;. \tag{1.29}$$

Filled with the building blocks for the posterior for Hypothesis $H_1$ that we just described we can assemble and simplify the posterior,

$$\begin{aligned} P(\theta|D_1, D_2, H_1) \;&= \frac{\binom{N}{n_1}\frac{1}{B(u_1,u_2)}\theta_1^{n_1+u_1-1}(1-\theta_1)^{N-n_1+u_2-1}}{\binom{N}{n_1}\frac{1}{B(u_1,u_2)}\int_0^1 \theta_1^{n_1+u_1-1}(1-\theta_1)^{N-n_1+u_2-1}\,d\theta} \times \\ &\frac{\binom{N}{n_2}\frac{1}{B(u_1,u_2)}\theta_2^{n_2+u_1-1}(1-\theta_2)^{N-n_2+u_2-1}}{\binom{N}{n_2}\frac{1}{B(u_1,u_2)}\int_0^1 \theta_2^{n_2+u_1-1}(1-\theta_2)^{N-n_2+u_2-1}\,d\theta} \quad = \\ &= \frac{\theta^{n_1+n_2+u_1-1}(1-\theta)^{N+N-n_1+n_2+u_2-1}}{B(u_1+n_1+n_2, u_2+N+N-n_1-n_2)} \quad = \\ &= \text{Beta}(u_1+n_1+n_2, u_2+N+N-n_1-n_2) \;. \end{aligned} \tag{1.30}$$

Note, how the evidence (denominator) simplifies to a Beta function (magic trick of choosing a conjugate prior), an identity we recognise from Equation 3.10 and how the posterior as a whole can be expressed as a Beta distribution, compare Equation 3.9.

Analogously, for Hypothesis $H_2$ we can formulate the posterior probability distribution as

$$P(\theta_1, \theta_2|D_1, D_2, H_2) \quad = \quad \frac{P(D_1, D_2|\theta_1, \theta_2, H_2) \times P(\theta_1, \theta_2|H_2)}{P(D_1, D_2|H_2)} \;, \tag{1.31}$$

and using the same magic conjugate prior simplification tricks we end up with a posterior probability function,

$$
\begin{aligned}
P(\theta_1, \theta_2 | H_2, D_1, D_2) &= \frac{\binom{N}{n_1} \frac{1}{B(u_1,u_2)} \theta^{n_1+u_1-1}(1-\theta)^{N-n_1+u_2-1}}{\binom{N}{n_1} \frac{1}{B(u_1,u_2)} B(u_1+n_1, u_2+N-n_1)} \times \\
& \frac{\binom{N}{n_2} \frac{1}{B(u_1,u_2)} \theta^{n_2+u_1-1}(1-\theta)^{N-n_2+u_2-1}}{\binom{N}{n_2} \frac{1}{B(u_1,u_2)} B(u_1+n_2, u_2+N-n_2)} = \\
&= \text{Beta}(u_1+n_1, u_2+N-n_1) \times \\
& \quad \text{Beta}(u_1+n_2, u_2+N-n_2) \ .
\end{aligned}
\tag{1.32}
$$

We have learned now how to update our knowledge about the success rates of our players. But how can we compare the hypotheses now? We want to know which hypothesis is more likely to be true, given our data. We can find the ratio of the two probabilities to evaluate,

$$
\frac{P(H_2|D)}{P(H_1|D)} = \frac{\text{Evidence}(\text{H}_2) \ \times \ \text{Prior}(\text{H}_2)}{\text{Evidence}(\text{H}_1) \ \times \ \text{Prior}(\text{H}_1)} \tag{1.33}
$$

where $D$ is the entirety of the data, $D_1$ and $D_2$. Reading this will sound very familiar to us: The posterior odds is the Prior odds times a factor. The Bayes factor, the ratio of the evidences for our two hypotheses. We have already found a way to calculate the evidence for both of our hypotheses in the denominator of their posteriors, by integrating over likelihood and prior for all $\theta$, so we get a Bayes factor,

$$
\begin{aligned}
\text{Bayes factor} &= \frac{P(D_1, D_2|H_2)}{P(D_1, D_2|H_1)} = \\
&= \frac{B(u_1+n_1, u_2+N-n_1) \ \times \ B(u_1+n_2, u_2+N-n_2)}{B(u_1,u_2) \ \times \ B(u_1+n_1+n_2, u_2+N-n_1-n_2)} \ ,
\end{aligned}
\tag{1.34}
$$

where all except one pre-factor, $B(u_1, u_2)$, cancel and $B(u_1, u_2) = 1$ for a flat prior.

We can use this Bayes factor to update our knowledge by multiplying it with our prior knowledge about a system. This implies that for an event with a very low prior ratio, we want to see a strong Bayes factor – good evidence – to change our belief.

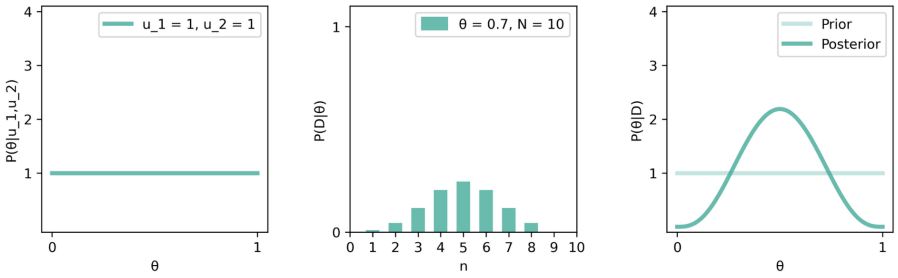



Figure 1.10: The process we have gone through in this derivation was finding (left) a likelihood function that serves as a model for the process we are watching, (middle) a prior expression to include knowledge, and (right) a posterior to express our updated belief.

### 1.6.6 Bayes factors for card games

Congratulations, you have just made it through the first Bayes factor derivation of this thesis. Two more to go! If we calculate the Bayes factors for our recorded card games we get the following table, Figure 1.11. Usually, we transform our Bayes factor to $\log_{10}$ Bayes factors. In the rest of the thesis we will always refer to $\log_{10}$ Bayes factors when talking about Bayes factors.

| N = 10 | Adult | expected success rate | Kid | expected success rate | Bayes factor |
|---|---|---|---|---|---|
| Game 1 | 6 | 0.583 | 3 | 0.333 | 0.063 |
| Game 2 | 5 | 0.500 | 7 | 0.667 | -0.141 |
| Game 3 | 8 | 0.750 | 7 | 0.667 | -0.303 |

Figure 1.11: Our two players, a kid and an adult have played 3 games with 10 rounds each. We counted the successes of them correctly guessing whether the card is 'above' or 'below'. Afterwards we inferred an expectation value $\langle\theta\rangle$ for their success rates and calculated $\log_{10}$ Bayes factors to evaluate whether we can see the differences in their tactics.

### 1.6.7 Interpreting Bayes factors

The Bayes factor is the ratio of evidences for the two hypotheses which we wish to compare. If there is is an equal probability for both hypotheses, we will get a log Bayes factor around 0, meaning we cannot conclude any explanation of the data is more likely. As soon as we can find more evidence for one of the two hypotheses in the data given our model, the $\log_{10}$ Bayes factors will grow accordingly above or below 0. If we see more evidence for hypothesis 1, the $\log_{10}$ Bayes factors will migrate into the negative numbers, in relation to how big the evidence differences are, and *vice versa*, if we see more evidence for hypothesis 2, the Bayes factors will rise, Figure 1.12.

Inspecting Figure 1.11 after you have learned that $\log_{10}$ Bayes factors around 0

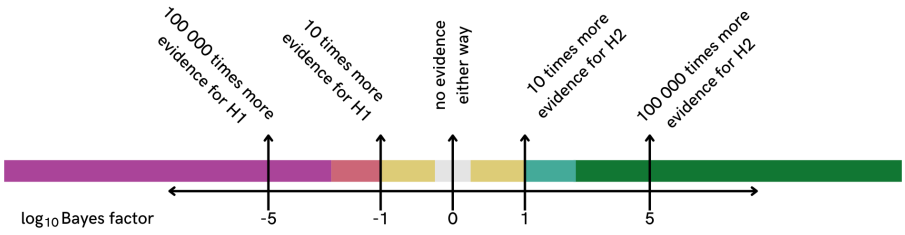



Figure 1.12: Positive $\log_{10}$ Bayes factors indicate stronger evidence for Hypothesis 2, while negative $\log_{10}$ Bayes factors represent support for Hypothesis $H_1$. Following Jaynes [5] $\log_{10}$ Bayes factors  1 are considered a sensible cutoff for binary decisions. Jeffreys [6], and similar Kass and Raftery [7] offer a verbal discription where (to summarise) absolute $\log_{10}$ Bayes factors below 0.5 barely worth mentioning, 0.5 - 1 is substantial support, above 1 is strong support and above 2 is decisive. These interpretations are built on minimal differences humans can and cannot notice in signals [6].

reflect no difference in player tactics may lead you to the conclusion that luck is everything in this card game, no inference will help you. The $\log_{10}$ Bayes factors are very small. But do you think this is the right conclusion (given the experience in your life so far)? If we pretend our players had played games with 100 rounds and we give them the same success rates, we will see how our perception changes.

| N = 100 | Adult | expected success rate | Kid | expected success rate | Bayes factor |
|---|---|---|---|---|---|
| Game 1 | 60 | 0.598 | 30 | 0.304 | 3.212 |
| Game 2 | 50 | 0.500 | 70 | 0.696 | 1.040 |
| Game 3 | 80 | 0.754 | 70 | 0.696 | -0.246 |

Figure 1.13: The importance of repetition and necessary insights become very clear when we compare the Bayes factors in Figure 1.11 and this Figure 1.13. We have gathered more data and therefore more evidence, resulting in more extreme Bayes factors. This table makes us believe that at first, kid and adult have very different success rates, and the differences decrease with the number of games they play. Interestingly, and we will encounter this in more detail in Chapter 6, it does not matter if they are playing these 300 rounds in 3 games or 5 games or 10 games, our overall conclusions will be the same. If we want to observe the learning process of the kid, however, we need to look at the process over time, of course, instead of all 300 rounds at once.

### 1.6.8 Bayes factors in the Analysis of RNA-Seq data

Bayes factors can be easier and harder to calculate, depending on how easy or hard it is to find the building blocks for the posterior distribution, Equation 1.4, and if simplifications happen like in our example. The derivation we worked through was, of course, chosen to demonstrate how simple it can be. In Chapters 3 and 6

we introduce Bayes factors in the analysis of RNA-Seq data for two different purposes. Relatively speaking, both of these examples are still simple and we could find analytical solutions for them, which is not given in many other problems out there.

## 1.7 Bayesian inference in the analysis of biological data

The central question of this thesis is: How can we extract knowledge about life from data? We will nearly exclusively look at a very popular data type in molecular biology, called RNA-Sequencing data. RNA-Sequencing is a technique to identify and count all RNA present in a biological sample. Before we go into the introduction to the technology and the content of this thesis, I want to reiterate to the beginning of the thesis and find motivation to use Bayesian inference to analyse RNA-Seq data.

We have learned in the last pages about some simple principle mechanisms and the underlying philosophy of Bayesian statistics. There are two big strengths in approaching statistical analysis like that, that we touched on so far. Firstly, we can incorporate prior knowledge into the analysis. I guess this is not only a valuable thought for the analysis but for research and life in general. By combining prior knowledge with new insights gleaned from the data, Bayesian statistics offers a more holistic and informed approach to inference, especially in situations with limited or noisy data. Secondly, the uncertainty of our measurement is quantified in the form of probability distributions, allowing for a more nuanced representation of the unknown parameters. The goal is to capture the complexity of real-world uncertainty, which brings us to biology. The Bayesian framework's ability to provide probabilistic statements about parameters and predictions offers an intuitive and comprehensive understanding of uncertainty, which is exactly what we need.

Currently, we are collecting a lot of RNA-Seq data in biology, bringing a huge amount of very interesting insights about molecular and cellular processes to us. Nevertheless, it is still a measuring technique with limitations, that is measuring probabilistic events in complex living systems. The process of the data collection is complicated, involving many steps of chemical and computational innovation we are only capable to use since a few years. This is paired with the challenge of identifying and quantifying thousands of macro-molecules that are part of dynamic processes in life. Analysing this data is unbelievably fascinating, and involves adventures from 18th-century mathematics to 21st-century innovation.

Chapter 1

# Chapter 2

# An Introduction to RNA-Seq Data

Transparency Note — Parts of this chapter are in an arXiv pre-print titled 'A Closed-Form Solution to the 2-Sample Problem for Quantifying Changes in Gene Expression using Bayes Factors' authored by Gurpinder Singh Sidhu, Melissa Tomkins, Richard J Morris and myself [8].

## 2.1 RNA-Sequencing enables quantification of RNA in cells

Every living cell of any organism on this planet follows the so-called central dogma of molecular biology – a genetic information processing pipeline and molecular production pathway. Genetic information, stored in DNA, is transcribed into RNA, of which many are translated into proteins, Figure 2.1.

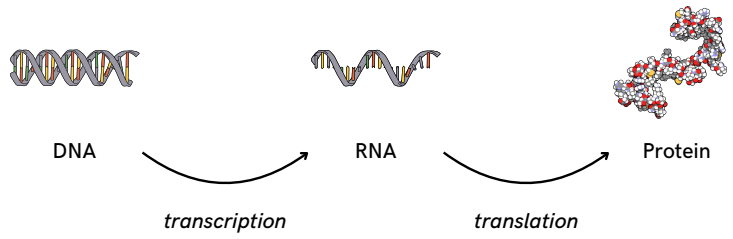



Figure 2.1: The central dogma of molecular biology.

The processes of transcription and translation are delicately regulated, and their rates of genes may change over time – orchestrating development, and life. The field of molecular biology is striving to understand and quantify all of these processes and molecules. One way to measure transcript levels in cells is by RNA-Sequencing

(RNA-Seq) [9]–[11]. In this technique, a biological sample is taken (e.g. a leaf of a plant), and by chemical and physical treatments, all RNA of all cells in the sample is extracted, while all other structures and molecules are washed away. The extracted RNA is sequenced, which results in short strings of RNA, called sequencing reads, as a digital output. Those reads can be aligned by alignment algorithms to the sequences of the genome of the organism to find their site of origin within the genetic material. For each gene in the genome, we can count the number of reads mapping to it. This number is considered proportional to the amount of RNA originating from this gene in the tissue. Hence, by sequencing all RNA found in a biological sample, we can estimate the amount of RNA we find in a tissue from the number reads assigned to a gene.

There are unlimited options for the use of RNA-Sequencing, of course. In this thesis, we are focusing on the data analysis in two different experimental setups using this technique: Differential gene expression analysis (Chapters 3 and 4), and the detection of mobile mRNA in plants (Chapters 5 and 6). For simplicity, we will only introduce the popular experiment of differential gene expression in this chapter, and go into the details of long-distance mobile mRNA detection in Chapter 5.

## 2.2 Using RNA-Seq to identify transcription regulation changes

If we carry out two (or more) RNA-Sequencing experiments under different conditions, we can compare the amounts of RNA in a sample and may learn about environmental conditions affecting the process of transcription in a tissue. In differential gene expression analysis, we seek to identify genes with a 'significant' change in their expression between the two measured conditions and call them differentially expressed genes, DEG henceforth. In Figure 2.2 we depict an overview of this experimental procedure.

## 2.3 Current statistical methods in differential gene expression analysis rely on hard cutoffs for identification criteria

Numerous review articles have summarized RNA-Seq technologies and data analysis tools for the identification of transcription changes in recent years [12]–[17]. Currently popular differential gene expression analysis tools all follow a similar workflow, depicted in Figure 2.5. Several analysis options and parameters that need to be set by researchers, combined with the challenges of technical and biological noise, may lead to difficulties in reproducing the exact results of studies, however. This issue has been discussed in several instances. Chen et al. [12], for example, illustrate in their review how using different methods for a complete analysis from raw reads to differential expression and functional enrichment analysis can lead to different results and conclusions. Rapaport et al. [17] and [18] contributed an extensive test of several packages on simulated and real data and found similar results – already 10 years ago. Shortly afterward, when the popular tool DESeq2

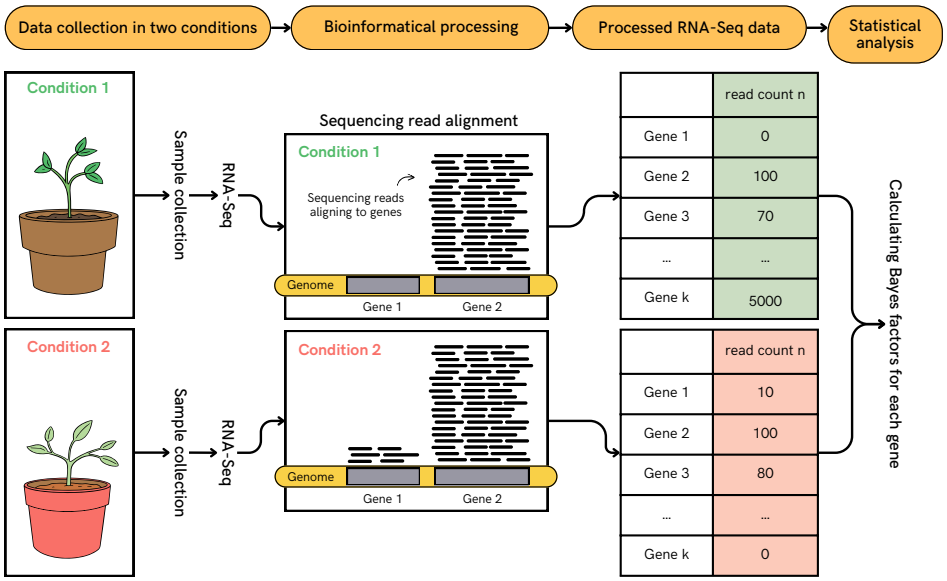



Figure 2.2: An overview of RNA-Sequencing experiments to identify differentially expressed genes. Tissue samples are collected from two conditions that are desired to be compared. The RNA in these samples is extracted separately, and sequenced (Figure 2.3). The raw sequencing data output is processed by bioinformatics pipelines (Figure 2.4). One important step thereof is the alignment of sequencing reads to the genome of the sample's organism. The processed RNA-Seq data consists of sequencing read counts for each gene. Afterwards follows the statistical analysis, which in this case is to find out which genes have changed their expression between the conditions. In Chapter 3 we will discuss how we can calculate a Bayes factor for differential gene expression for each gene, to rank genes according to the evidence in the data for changes in the amounts of RNA present in the samples.

was published, the authors also compared the performance of eight existing differential gene expression analysis approaches using simulated data [19]. Nevertheless, even though problems are reported, these analysis tools are the go-to methods for the search for differentially expressed genes. While the authors of Chen et al. [12] 'emphasize the need to both conduct comprehensive comparative analyses and justify specific study choices', we will explore a bit further *why* these differences occur later in this thesis and provide some insights on how to avoid these problems.

Differential gene expression analysis is a very popular experiment. Therefore, the process has been streamlined in the last years, resulting in a range of software tools that can carry out the full analysis: *DESeq* [20], *DESeq2* [19], *edgeR* [21]–[23], *baySeq* [24], *EBSeq* [25], to name the most popular ones, which have evolved from RNA-Seq precursor analyses of SAGE (Serial Analysis of Gene Expression) data [26], [27].

Currently, popular statistical analysis tools rely on two criteria for determining a DEG: (1) p-values below a set threshold (to determine if a 'significant' change in

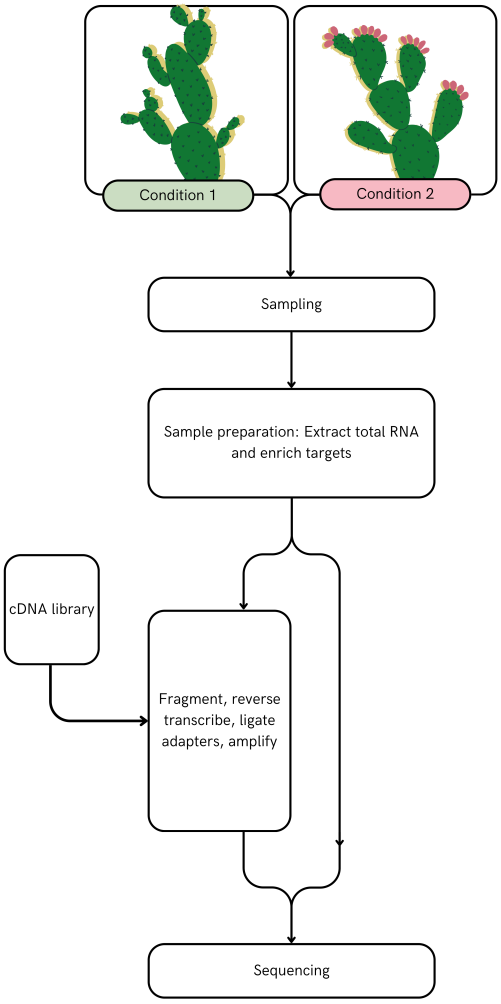



Figure 2.3: Data collection in two conditions starts with sample collection and sample preparation. It either involves converting RNA into cDNA, followed by next generation sequencing (NGS) [9], [10], or more recent methods skip converting RNA and are able to sequence RNA directly [11]. Depending on the objective of a study, scientists may choose different RNA-Sequencing technologies, varying sequencing depths as well as bioinformatics and statistical analysis pipelines [12]. After sequencing the raw data is processed in bioinformatics pipelines, continuation in Figure 2.4.

normalised read counts between the two conditions is observed) and (2) an absolute $\log_2$ fold change value above a certain threshold, where

$$\log_2(\text{fold change}) = \log_2 \frac{c_1}{c_2}, \tag{2.1}$$

with the normalised read counts $c_1$ and $c_2$ for condition 1 and condition 2, to estimate the magnitude of the change. We will refer to this $\log_2$(fold change) as fold change in the rest of this thesis.

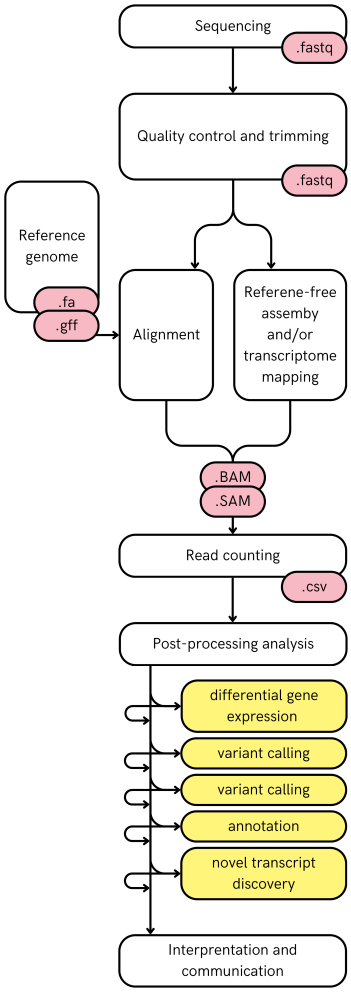



Figure 2.4: Raw sequencing data are processed in several steps (white boxes) in bioinformatics pipelines. The processing involves several data type conversions (red bubbles). To navigate the jungle of techniques, sequencing, processing and data analysis tools are reviewed constantly, e.g. in [12]–[17]. The work of this thesis is located in post-processing analyses only (yellow bubbles).

Popular cutoffs for DEGs are p-value $< 0.05$, and $\mid \log_2(\text{fold change}) \mid > 1$, sometimes $> 2$ [12], [14]–[19], [28]–[35]. This means, that current criteria are filtering for genes that have at least doubled, or even quadrupled their expression. While setting a fold change cutoff decreases the number of false positive hits, potentially interesting genes with noticeable changes that have not at least halved or doubled their expression are ignored. Considering molecular mechanisms of biology, however, there is no reason why the impact of a gene that has doubled its RNA in a cell is necessarily higher than another gene that has incurred a smaller relative change. As an alternative to deciding arbitrary fold change and p-value cut-offs, in Chapter 3 we explore the use of the Bayesian statistics toolbox in molecular biology in this popular analysis example.

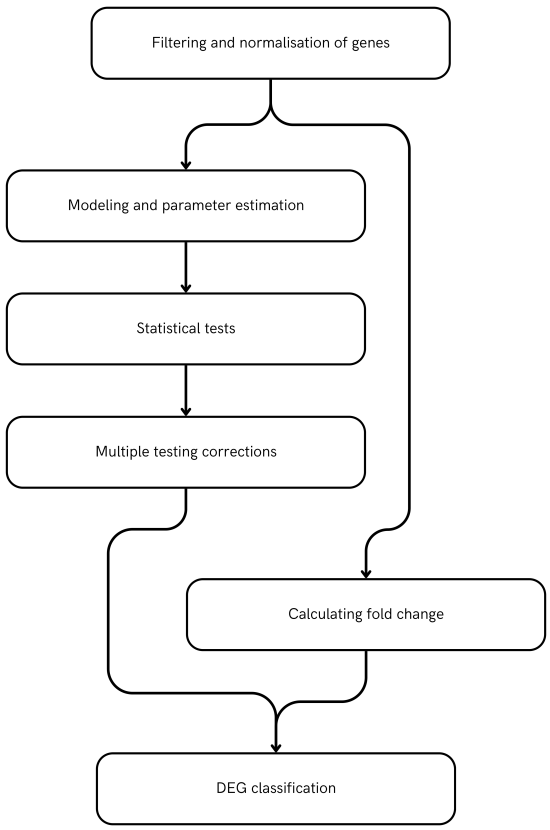



Figure 2.5: Currently used statistical software packages are mainly following the presented flowchart. Processed data, read counts for each gene, are filtered and normalised before a p-value (left side) and $\log_2$ fold change (right side) are calculated. These two values serve as classification criteria for identifying DEGs. While the overall idea of how to identify those genes is common to all packages, the exact way a p-value is calculated varies from tool to tool.

## 2.4 Thesis preview

In **Chapter 3** we present a Bayesian framework tailored to the differential gene expression analysis of processed RNA-Seq data. In contrast to current software that often involves various pre-processing stages, and applies complex statistical analyses, involving parameter estimations and multiple statistical tests, we put forward a concise mathematical equation (i.e. we provide a closed-form or analytical solution) to calculate a Bayes factor for each gene, enabling genes to be ranked according to the statistical evidence for change. In **Chapter 4** we compare our

new Bayesian framework to existing methods in the field and identify (or remind us about) limitations of RNA-Seq that can cause reproducibility issues.

In **Chapter 5** we will learn about the effort of using RNA-Seq in the detection of long-distance mobile mRNA in plants. Modern sequencing technologies paired with ancient grafting practices are state-of-the-art for investigating the mysterious long-distance movement of macromolecules, such as mRNA. We have developed a Bayesian framework to tackle sequencing noise in this detection setup, resulting in astonishing accuracy improvements, **Chapter 6**.

We carried out a re-analysis of several long-distance mobile mRNA data sets in a range of plant species which brought more complications about the use of short-read sequencing in the detection of mobile RNA to light, **Chapter 7**. As in previous chapters, we learn a lot about the limitations of RNA-Seq, ranging further than the specific use discussed in the chapter. **Chapter 8** is the final summary and discussion of the work in this thesis.

Chapter 2

# Chapter 3

# Analytical Bayesian Framework for Differential Gene Expression Analysis using RNA-Seq Data

Transparency Note — Parts of this chapter are in an arXiv pre-print titled 'A Closed-Form Solution to the 2-Sample Problem for Quantifying Changes in Gene Expression using Bayes Factors' authored by Gurpinder Singh Sidhu, Melissa Tomkins, Richard J. Morris and myself [8]. This chapter has largely profited from discussions with all co-authors of this pre-print, and Thelonious Omori. Since the completion of the first version of this thesis, the content of the chapter has been modified and in parts improved for a bioRxiv pre-print [36]. All code for figures, analyses and simulations for this chapter can be found on my GitHub.

Summary — Advances in RNA-sequencing technology have revolutionised our ability to capture the complete RNA profile in tissue samples (bulk RNA-Seq). This wealth of data allows for comparative analyses of RNA levels within cells and tissues, shedding light on transcript dynamics of developmental processes, environmental responses, and treatment effects. However, quantifying changes in gene expression still presents a challenge, given the inherent biological variability, technological limitations, and measurement errors. To address this, we introduce a Bayesian framework tailored to the differential gene expression (DGE) analysis of processed RNA-Seq data. Our framework unifies and streamlines a complex analysis, typically involving parameter estimations and multiple statistical tests, into a concise mathematical equation (i.e. we provide a closed-form or analytical solution). It can be implemented in a single line of code, enabling rapid and transparent ranking of genes according to statistical evidence for a change in gene expression. Furthermore, we introduce a method evaluating the variability between replicates. We conducted a comparison of our framework with leading differential gene expression tools, finding substantial overlaps in the results and interesting disagreements, mainly caused by pre-filtering and small differences in fold change values, which are currently driving the classification of a 'differentially expressed gene'. This motivated us to explore the possibility of ranking genes instead of classifying DEGs

moving forward.

## 3.1 Background

We have learned in Chapter 2 that RNA-Sequencing (RNA-Seq) is a powerful technique used in molecular biology to gain insights into gene expression and transcriptional activity at a tissue sample level. By sequencing all RNA molecules present in a sample, RNA-Seq can reveal which genes are expressed or which genomic regions are transcribed at a specific time during development or in response to environmental stimuli [9], [11], [28]. RNA-Seq datasets can be compared to identify statistically significant changes in the amounts of RNA between the samples (differential gene expression, DGE), or over-representation analyses to identify interesting candidate genes for further investigations [12].

Excellent review articles are available that summarise RNA-Seq technologies, associated data analysis tools, and their successes and limitations [14]–[16], [18], [28]–[34]. Motivated by studies reporting reproducibility challenges in this area [12], [17], [19], [35], we were driven to re-examine the statistical assumptions underlying differential gene expression analysis from a Bayesian perspective, Figure 3.1. This endeavor has led to the development of an exact Bayesian framework tailored to this use case, including the derivation of an analytical expression for solving this two-sample test problem in differential gene expression. Through comparisons with popular existing methods, using both simulated and read data, we found, that this analysis enhances the computational efficiency without compromising accuracy.

## 3.2 Results and Discussion

### 3.2.1 Differential gene expression analysis can be cast into the framework of Bayesian inference and model comparison

We assign to every gene $i$ an expression probability $q_i$. This probability is variable and depends on all events from transcription of gene $i$ to mapping of a corresponding read in data processing,

$$\text{DNA} \quad \xrightarrow{q_i} \quad \text{read count} \; . \tag{3.1}$$

For a known $q_i$, we can describe the probability of $n_i$ number of reads mapping to a gene $i$, out of total reads in a sample $N$ with a binomial distribution (a read maps to gene $i$, with probability $q_i$, and it does not, with probability $1 - q_i$),

$$P(n_i|N, q_i) = \binom{N}{n_i} \, q_i^n (1 - q_i)^{N - n_i} \; . \tag{3.2}$$

This implies that if we know the gene expression probability $q_i$ of a gene and the total number of all reads $N$ in an RNA-Seq experiment, we can compute the probability of any number of RNA-Seq reads $n_i$ mapping to gene $i$. Note that in total $N = \sum n_i$.

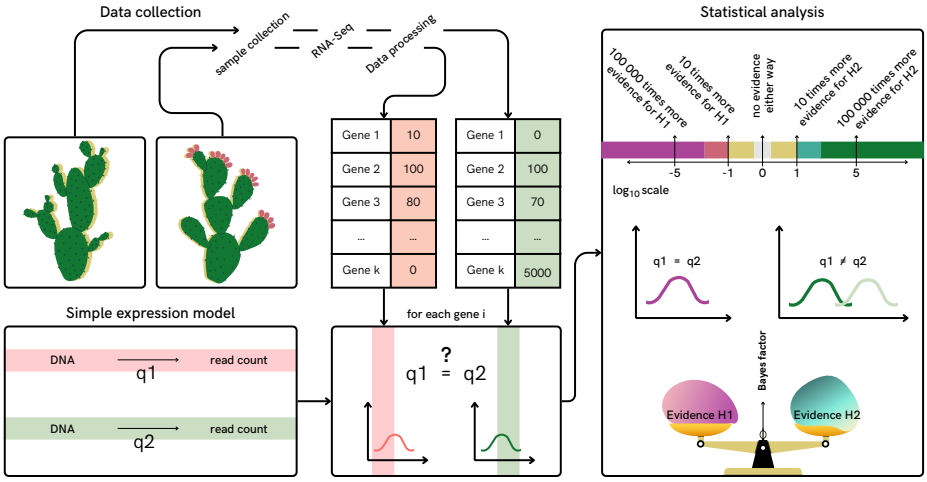



Figure 3.1: A schematic overview of our new framework for differential gene expression analysis based on Bayesian model comparison. The analysis is built on a simple gene expression model for which we infer parameters from collected, processed RNA-Seq data. We calculate the evidence (a measure for how compatible a model is with the data) for different models. The evidences for two alternative hypotheses and the corresponding models are compared by computing the ratio of them (Bayes factor). In this DGE case, these 2 hypotheses are: H1) the data can be explained by one gene expression parameter ($q_1 = q_2$), as opposed to H2) two different gene expression parameters are needed to explain the data ($q_1 \neq q_2$), which can be interpreted as a change in gene expression. Note that our knowledge of $q$ is represented by a probability distribution over the parameter and not a single value. A negative $\log_{10}$ Bayes factor for a gene lets us conclude that there is no change in expression, whereas a positive log Bayes factor marks a change. The more extreme the Bayes factor, the stronger the respective evidence. If $\log_{10}$ Bayes factors are around 0 there is no evidence either way. We can use Bayes factors to rank genes according to the confidence in an expression change, and if wanted, for binary significance decisions.

In a differential gene expression experiment (or any other two-sample test) we want to answer the question of whether two data sets $D_1$ (consisting of $N_1$ and $n_{i1}$) and $D_2$ (consisting of $N_2$ and $n_{i2}$), for an observation (gene) $i$,

$$q_{i1} \rightsquigarrow D_1 \quad \text{and} \quad q_{i2} \rightsquigarrow D_2 \tag{3.3}$$

arose from the same probability distribution $q_i$. Therefore, our problem is to decide whether

$$q_{i1} \stackrel{?}{=} q_{i2} \ , \tag{3.4}$$

which is equivalent to asking whether the expression probability $q_i$ of a gene changed between two data sets, or not.

We can calculate a Bayes factor for each gene, describing how much the RNA-Seq data supports one of two hypotheses. This will be our metric for ranking genes

according to the statistical evidence for a change in gene expression.

**Hypothesis $\mathbf{H_1}$** states that the data from both experiments can be explained by one statistical model,

$$q_i \rightsquigarrow D_1 \quad \text{and} \quad q_i \rightsquigarrow D_2 \ . \tag{3.5}$$

RNA-Seq data from the first and second experiment, i.e. mRNA levels of gene $i$, are statistically consistent, and we have no statistical support for a difference between them. The data are consistent with the biological and technical variance that might be expected between experiments.

**Hypothesis $\mathbf{H_2}$** states that the data are best explained by a different model for each experiment,

$$q_{i1} \rightsquigarrow D_1 \quad \text{and} \quad q_{i2} \rightsquigarrow D_2 \ . \tag{3.6}$$

In this case, the RNA-Seq data are unlikely to have arisen from just one $q_i$, the variance is higher and requires a separate model for each data set. Hence the data support a difference in gene expression between the samples.

### 3.2.2 Deriving Bayes factors for differential gene expression

The transcription probability $q_i$ is unknown. However, we can infer it from RNA-Seq data using Bayes' theorem as an inverse probability problem. We use $\theta_i$ to denote the range of possible values of $q_i$ (between 0 and 1) and use a probability distribution over $\theta_i$, $P(\theta_i)$, to capture our knowledge of $q_i$, with our best estimate of $q_i$ being the expectation value $\langle\theta_i\rangle$.

The posterior is a probability distribution describing *current* knowledge of $\theta_i$, given the data, $D$, i.e. the RNA-Seq read count data $(N, n_i)$,

$$P(\theta_i|D) \ = \ \frac{P(D|\theta_i) \times P(\theta_i)}{P(D)} \ . \tag{3.7}$$

The posterior is determined by three factors: the prior probability $P(\theta_i|H)$, the likelihood of $\theta_i$, either expressed as $P(\theta_i|D)$ or $\Lambda(\theta_i)$, and a normalising factor $P(D)$ in the denominator. This normalising factor is called the evidence or the marginal likelihood.

To find a likelihood function $P(D|\theta_i)$ for all possible $\theta_i$, we can use the equation 3.2 to model $P(D|\theta_i)$ as a simple binomial process

$$P(D|\theta_i) \ \propto \ \binom{N}{n_i} \theta_i^{n_i}(1-\theta_i)^{N-n_i} \ . \tag{3.8}$$

The prior distribution captures our knowledge about the system before we collect data. A convenient choice for the prior is the Beta distribution, as it is the conjugate distribution of the binomial distribution, owing to it having the same functional

form and allowing us to proceed analytically. The prior distribution can be written as

$$P(\theta_i|u_1,u_2) \;=\; \frac{1}{B(u_1,u_2)}\;\theta_i^{u_1-1}\;(1-\theta_i)^{u_2-1} \;=\; \text{Beta}(u_1,u_2)\;, \tag{3.9}$$

where the hyper-parameters $u_1, u_2 \in \mathbb{R}^+$, can be used to capture existing knowledge about the transcription probability. The Beta distribution $\text{Beta}(u_1, u_2)$ is named after its normalisation factor, the Beta function $B(u_1, u_2)$, expressed as

$$B(u_1,u_2) \;=\; \int_0^1 \theta_i^{u_1-1}\;(1-\theta_i)^{u_2-1} d\theta_i\;. \tag{3.10}$$

For instance, a reasonable estimate of $q_i$ might be the inverse of the number of genes and $u_1$ and $u_2$ could be chosen accordingly to provide a prior over $\theta_i$. Here, we choose a bias-free, flat prior ($u_1, u_2 = 1$).

Finally, the evidence $P(D)$ is the probability that the data $D$ is produced. It can be calculated by integrating over the numerator in equation 3.7

$$P(D) \;=\; \int_0^1 P(D|\theta_i)\times P(\theta_i)\;d\theta_i\;. \tag{3.11}$$

This means that if we sum up all probabilities of seeing data $D$ for all possible values of $\theta_i$, we get a value called the evidence. To visualise this, we can picture the evidence as the area under the curve of $P(D|\theta_i) \times P(\theta_i)$.

For two datasets, $D_1$ and $D_2$, we can find a posterior distribution $P(\theta_i|D)$ for Hypothesis 1 and Hypothesis 2 separately. For Hypothesis 1, the assumption is the same expression probability can explain both datasets, $\theta_i \rightsquigarrow D_1$ and $\theta_i \rightsquigarrow D_2$, resulting in

$$P(\theta_i|D_1,D_2,H_1) \quad=\quad \frac{P(D_1,D_2|\theta_i,H_1)\times P(\theta_i|H_1)}{P(D_1,D_2|H_1)}\;. \tag{3.12}$$

For Hypothesis 2, we assume that the expression probabilities are different between experiments, $\theta_{i1} \rightsquigarrow D_1$ and $\theta_{i2} \rightsquigarrow D_2$ ,

$$P(\theta_{i1},\theta_{i2}|D_1,D_2,H_2) \quad=\quad \frac{P(D_1,D_2|\theta_{i1},\theta_{i2},H_2)\times P(\theta_{i1},\theta_{i2}|H_2)}{P(D_1,D_2|H_2)}\;. \tag{3.13}$$

We define a prior for the single expression probability in Hypothesis 1,

$$P(\theta_i|H_1) \;=\; \text{Beta}(u_1,u_2) \;=\; \frac{1}{B(u_1,u_2)}\;\theta_i^{u_1-1}\;(1-\theta_i)^{u_2-1} \tag{3.14}$$

and the prior for two expression probabilities in Hypothesis 2,

$$P(\theta_{i1}, \theta_{i2}|H_2) \;=\; P(\theta_{i1}|H_2) \times P(\theta_{i2}|H_2) \;=\; \text{Beta}(u_1, u_2) \times \text{Beta}(u_1, u_2)\ . \quad (3.15)$$

Note that we have assumed the same priors over $\theta_{i1}$ and $\theta_{i2}$. The likelihood for the data $(D_1, D_2)$ given Hypothesis 1 can be written as

$$P(D_1, D_2|\theta_i, H_1) \;=\; \binom{N_1}{n_{i1}} \theta_i^{n_{i1}} (1-\theta_i)^{N_1 - n_{i1}} \;\times\; \binom{N_2}{n_{i2}} \theta_i^{n_{i2}} (1-\theta_i)^{N_2 - n_{i2}}\ . \quad (3.16)$$

For Hypothesis 2, the likelihood of the data given $H_2$ depends on the two expression parameters of the model, $\theta_{i1}, \theta_{i2}$,

$$P(D_1, D_2|\theta_{i1}, \theta_{i2}, H_2) \;=\; \binom{N_1}{n_{i1}} \theta_{i1}^{n_{i1}} (1-\theta_{i1})^{N_1 - n_{i1}} \;\times\; \binom{N_2}{n_{i2}} \theta_{i2}^{n_{i2}} (1-\theta_{i2})^{N_2 - n_{i2}}\ . \quad (3.17)$$

Thus, the posterior of Hypothesis 1 simplifies, thanks to conjugate prior, to the following equation,

$$\begin{aligned}
P(\theta_i|D_1, D_2, H_1) \;&=\; \frac{P(D_1, D_2|\theta_i, H_1) \times P(\theta_i|H_1)}{P(D_1, D_2|H_1)} \;= \\
&=\; \frac{\binom{N_1}{n_{i1}} \theta_i^{n_{i1}} (1-\theta_i)^{N_1 - n_{i1}}}{\binom{N_1}{n_{i1}} \int_0^1 \theta_i^{n_{i1}} (1-\theta_i)^{N_1 - n_{i1}} d\theta_i} \times \\
&\quad \frac{\binom{N_2}{n_{i2}} \theta_i^{n_{i2}} (1-\theta_i)^{N_2 - n_{i2}}}{\binom{N_2}{n_{i2}} \int_0^1 \theta_i^{n_{i2}} (1-\theta_i)^{N_2 - n_{i2}} d\theta_i} \times \\
&\quad \frac{\frac{1}{B(u_1, u_2)}\ \theta_i^{u_1 - 1} (1-\theta_i)^{u_2 - 1}}{\frac{1}{B(u_1, u_2)} \int_0^1 \theta_i^{u_1 - 1} (1-\theta_i)^{u_2 - 1} d\theta_i} \;= \\
&=\; \frac{\theta_i^{n_{i1} + n_{i2} + u_1 - 1} (1-\theta_i)^{N_1 + N_2 - n_{i1} - n_{i2} + u_2 - 1}}{B(u_1 + n_{i1} + n_{i2}, u_2 + N_1 + N_2 - n_{i1} - n_{i2})} \;= \\
&=\; \text{Beta}(u_1 + n_{i1} + n_{i2}, u_2 + N_1 + N_2 - n_{i1} - n_{i2})\ .
\end{aligned} \quad (3.18)$$

Note how the evidence (denominator) simplifies to a Beta function; an identity we recognise from Equation 3.10 and how the posterior as a whole can be expressed as a Beta distribution, compare Equation 3.9.

Analogously, for Hypothesis 2 we can formulate and simplify the posterior probability distribution to

$$
\begin{aligned}
P(\theta_{i1}, \theta_{i2} | D_1, D_2, H_2) &= \frac{P(D_1, D_2 | \theta_{i1}, \theta_{i2}, H_2) \times P(\theta_{i1}, \theta_{i2} | H_2)}{P(D_1, D_2 | H_2)} = \\
&= \frac{\binom{N_1}{n_{i1}} \theta_{i1}^{n_{i1}} (1-\theta_{i1})^{N_1 - n_{i1}}}{\binom{N_1}{n_{i1}} \int_0^1 \theta_{i1}^{n_{i1}} (1-\theta_{i1})^{N_1 - n_{i1}} d\theta_{i1}} \times \\
&\quad \frac{\binom{N_2}{n_{i2}} \theta_{i2}^{n_{i2}} (1-\theta_{i2})^{N_2 - n_{i2}}}{\binom{N_2}{n_{i2}} \int_0^1 \theta_{i2}^{n_{i2}} (1-\theta_{i2})^{N_2 - n_{i2}} d\theta_{i2}} \times \\
&\quad \frac{\frac{1}{B(u_1, u_2)} \theta_{i1}^{u_1 - 1} (1-\theta_{i1})^{u_2 - 1}}{\frac{1}{B(u_1, u_2)} \int_0^1 \theta_{i1}^{u_1 - 1} (1-\theta_{i1})^{u_2 - 1} d\theta_{i1}} \times \\
&\quad \frac{\frac{1}{B(u_1, u_2)} \theta_{i2}^{u_1 - 1} (1-\theta_{i2})^{u_2 - 1}}{\frac{1}{B(u_1, u_2)} \int_0^1 \theta_{i2}^{u_1 - 1} (1-\theta_{i2})^{u_2 - 1} d\theta_{i2}} = \\
&= \text{Beta}(u_1 + n_{i1}, u_2 + N_1 - n_{i1}) \times \\
&\quad \text{Beta}(u_1 + n_{i2}, u_2 + N_2 - n_{i2}) \ .
\end{aligned}
\tag{3.19}
$$

The Bayes factor (BF) is a ratio of the statistical evidences $P(D|H_1)$ and $P(D|H_2)$ supporting the two hypotheses. We already found the evidences for both of our hypotheses in differential gene expression analysis in the denominator of their posteriors.

$$
\begin{aligned}
BF_{21} &= \frac{P(D_1, D_2|H_2)}{P(D_1, D_2|H_1)} = \\
&= \frac{\binom{N_1}{n_{i1}} \frac{1}{B(u_1,u_2)} B(u_1+n_{i1}, u_2+N_1-n_{i1})}{B(u_1+n_{i1}+n_{i2}, u_2+N_1+N_2-n_{i1}-n_{i2})} \times \\
&\quad \frac{\binom{N_2}{n_{i2}} \frac{1}{B(u_1,u_2)} B(u_1+n_{i2}, u_2+N_2-n_{i2})}{\binom{N_1}{n_{i1}} \binom{N_2}{n_{i2}} \frac{1}{B(u_1,u_2)}} = \\
&= \frac{B(u_1+n_{i1}, u_2+N_1-n_{i1}) \times B(u_1+n_{i2}, u_2+N_2-n_{i2})}{B(u_1,u_2) \times B(u_1+n_{i1}+n_{i2}, u_2+N_1+N_2-n_{i1}-n_{i2})}
\end{aligned}
\tag{3.20}
$$

Where all except one pre-factor, $B(u_1, u_2)$, cancel. For a flat prior, $B(u_1, u_2) = 1$.

As is common practice in differential gene expression analysis, we proceed to calculate an inferred $\log_2$ fold change,

$$
\text{inferred } \log_2 \text{ fold change} = \log_2 \left\{ \left( \frac{u_1+n_2}{u_2+N_2-n_2} \right) / \left( \frac{u_1+n_1}{u_2+N_1-n_1} \right) \right\} . \tag{3.21}
$$

In our framework, under the assumption the replicates $r_j$ for conditions $j$ are consistent with each other, all $N_{jr}$ of all replicates can be summed to $N_j$ for each condition $j$ and all $n_{ijr}$ to $n_{ij}$ to calculate $BF$ and inferred $\log_2$ fold change for each gene $i$.

### 3.2.3 Ranking genes according to statistical evidence for expression change

Bayes factors are a measure of confidence in one hypothesis over another after seeing the data and assuming that both hypotheses were equally probable *a priori*: for each gene the data are consistent with one expression probability ($H_1$) vs. the data support there being two underlying expression probabilities ($H_2$), i.e. a change in gene expression has occurred. We $\log_{10}$ transform the Bayes factors. A log Bayes factor of 0 means there is an equal probability for both hypotheses, whereas a $\log BF_{21} > 0$ favours Hypothesis 2 (change in gene expression) and $\log BF_{21} < 0$ favours Hypothesis 1 (no change in gene expression), see Figure 3.1. Assigning each gene a $\log_{10}$ Bayes factor, $\log BF_{21}$, enables ranking them according to the evidence supporting gene expression change given the RNA-Seq data.

Should a simple 'change' or 'no change' outcome be preferred, Bayes factors can be used as criteria for a binary classification of DEGs. Following published literature [5], [7], we choose a cut-off of $\log BF_{21} > 1 \rightarrow$ differentially expressed gene (DEG).

Furthermore, Bayes factors reflect the number of mapping reads between the RNA-Seq samples, see Figure 3.2: A high Bayes factor means high confidence in a change in gene expression and high relative numbers of change. This means, if a gene has a lot of reads mapping to it that support a change relative to the same gene under different conditions, then the magnitude of the Bayes factors will reflect this. This is something to keep in mind, as we are comparing evidences to get a measure for confidence and confidence will rise as we have more points of measurement (number of reads).

### 3.2.4 Ranking genes according to their variability across replicates

In our framework, under the assumption the replicates $r$ are consistent with each other, all $N_{jr}$ of all replicates can be summed to $N_j$ for each condition $j$ and all $n_{ijr}$ to $n_{ij}$ to carry out the analysis proposed above. For monitoring the variability in the data, we extended the Bayesian framework to quantify the consistency of a gene's expression across replicates. Again, we calculate a Bayes factor, this time to rank by variability.

Using the same equations as derived above, we can ask whether two replicates have consistent expression for each gene. This framework can be extended to ask whether any number of replicates, $k$, have consistent expression.

We define two hypotheses. **Hypothesis $\mathbf{H_1}$** states that the data from all replicates can be explained by statistical model with only one expression probability for each gene,

$$q_i \rightsquigarrow D_1 \quad \ldots, \quad q_i \rightsquigarrow D_k, \tag{3.22}$$

i.e the data are consistent with the biological and technical variance that might be expected between replicates. **Hypothesis $\mathbf{H_2}$** states the data are best explained by a separate expression probability for each replicate,

$$q_{i1} \rightsquigarrow D_1, q_{i2} \rightsquigarrow D_2, \quad \ldots, \quad q_{ik} \rightsquigarrow D_k. \tag{3.23}$$

In principle, the number of models need not be equal to the number of replicates, and any number greater than 1 and less or equal to $k$ could be explored, i.e. we would be asking whether various subsets of replicates are consistent. Due to the combinatorics for large numbers of replicates (see below), we limit ourselves here to the extreme case of every replicate being different. If the replicates are consistent, we can simply sum all the $n_{ir}$ for replicate $r$ for each gene $i$, and likewise for the $N_r$ with the above framework.

For each gene we calculate a Bayes factor, describing how much the RNA-Seq data supports Hypothesis 2 over Hypothesis 1, i.e. how consistent the expression is between replicates. We use this Bayes factor for identifying genes that vary strongly between replicates.

For Hypothesis 1 the evidence is given by

$$P(D_1, D_2, \ldots, D_k|H_1) \quad = \quad \frac{B(n_i + u_1, N - n_i + u_2)}{B(u_1, u_2)} \times \prod_{r=1}^{k} \binom{N_r}{n_{ir}}, \qquad (3.24)$$

with $N = \sum N_r$ and $n_i = \sum n_{ir}$ for all replicates $r$. The evidence for Hypothesis 2 is given by

$$P(D_1, D_2, \ldots, D_k|H_2) = \frac{1}{B^k(u_1, u_2)} \prod_{r=1}^{k} \binom{N_r}{n_{ir}} B(n_{ir} + u_1, N_r - n_{ir} + u_2), \quad (3.25)$$

with $N = \sum N_r$ and $n_i = \sum n_{ir}$. By computing the ratio of the evidence for Hypothesis 2 over Hypothesis 1, we have a general way of testing $k$ models over 1 model,

$$\begin{aligned} BF_{k1} \quad &= \quad \frac{P(D_1, D_2, \ldots, D_k|H_2)}{P(D_1, D_2, \ldots, D_k|H_1)} \quad = \\ &= \quad \frac{\frac{1}{B^k(u_1, u_2)} \prod_{j=1}^{k} \binom{N_j}{n_{ij}} B(n_{ij} + u_1, N_j - n_{ij} + u_2)}{\frac{B(n_i + u_1, N - n_i + u_2)}{B(u_1, u_2)} \times \prod_{j=1}^{k} \binom{N_j}{n_{ij}}} \quad = \qquad (3.26) \\ &= \quad \frac{\prod_{j=1}^{k} B(n_{ij} + u_1, N_j - nij + u_2)}{B^{k-1}(u_1, u_2) \times B(n_i + u_1, N - n_i + u_2)}. \end{aligned}$$

### 3.2.5 The more data, the stronger the evidence

Before we started using the framework on biological data, we investigated the general behavior of Bayes factors and inferred $\log_2$ fold change and their relationships for different total read depths and number of reads mapping to single genes, see Figure 3.2. Furthermore, we have investigated what happens if the total read depth between conditions or treatments varies, see Figures 3.3 and 3.4. As expected, the more data, the stronger the evidence, and the more pronounced the Bayes factors. Note that due to the chosen statistical approaches and normalisation steps other software packages cannot account for effects of differing total read depths.

### 3.2.6 Bayesian differential gene expression inference results are in agreement with other methods

Many tools are available for differential gene expression analysis [22]–[27]. Currently, *DESeq2* [19], [20] and *edgeR* [21], [22] are two of the most popular software packages for differential gene expression analysis. Therefore, we compared the performance and results using Bayes factors for differential gene expression with these

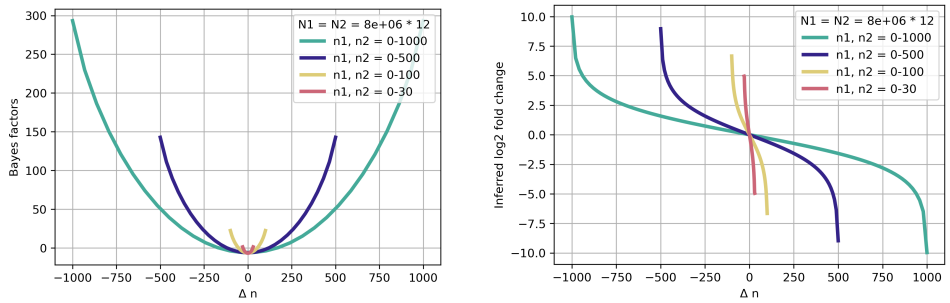



Figure 3.2: Bayes factors and absolute inferred fold change values rise with more data (increased number of reads in RNA-Seq experiments). Bayes factors and inferred $\log_2$ fold change have been calculated following the equations for Bayes factors described above in the Results. The total number of reads in both *in silico* experiments is set to $8 * 10^6 * 12$. This number follows the average read depth in the RNA-Seq study of Schurch et al. [35] and their recommended number of 12 biological replicates. In colors, we see the functions depending on the difference between $n_1$ and $n_2$ the number of reads mapping to a gene in two different conditions, ($\Delta\ n = n_1 - n_2$). We document what happens if the total read depth between conditions or treatments varies in Figures 3.3 and 3.4.

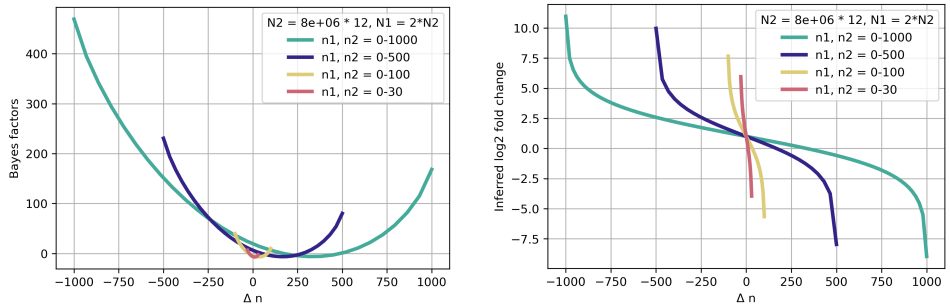



Figure 3.3: Differences in overall read depths between experiments (while $n_1, n_2$ stay in similar magnitudes), distort the symmetry. Again, we can see how Bayes factors and absolute inferred fold change values rise with more data (increased number of reads in RNA-Seq experiments). Bayes factors and inferred $\log_2$ fold change have been calculated following the equations in the Method section (derivation in Section 3.2.2). Here, one experiment has double total number of reads compared to the other, $n_1, n_2$ keep the same magnitudes. The total number of reads in the in silico experiments is set and we see variation in the read depths per genes. In colours we plotted the functions dependent on the difference between $n_1$ and $n_2$, $\Delta n = n_1 - n_2$, the number of reads mapping to a gene in the two different conditions.

two packages.

We conducted differential gene expression analysis on a yeast data set collected by Schurch et al. which has been designed for comparative differential gene expression analysis studies [35]. We took the set of 42 wild-type (WT) and 44 mutant replicates, and performed analyses using *DESeq2*, *edgeR*, and our analytical Bayesian framework, here referred to as *bayexpress*. The overlaps of the identified genes

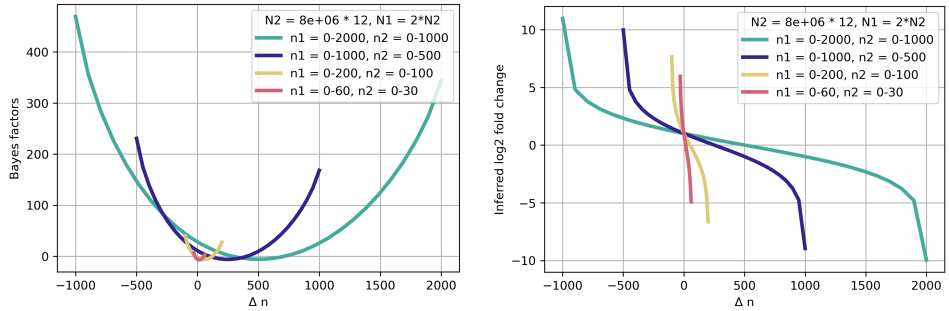



Figure 3.4: Differences in overall read depths in combination with $n_1, n_2$ rising in proportion, distort the symmetry accordingly. Once more we see how Bayes factors and absolute inferred fold change values rise with more data (increased number of reads in RNA-Seq experiments). Bayes factors and inferred $\log_2$ fold change have been calculated following the equations in the Method section (derivation in Section 3.2.2). Here, one experiment has double total number of reads compared to the other, in contrast to Figure 3.3, here $n_1, n_2$ are in relation to the raised read depth in one experiment. The total number of reads in the in silico experiments is set and we see variation in the read depths per genes. In colours we plotted the functions dependent on the difference between $n_1$ and $n_2$, the number of reads mapping to a gene in the two different conditions.

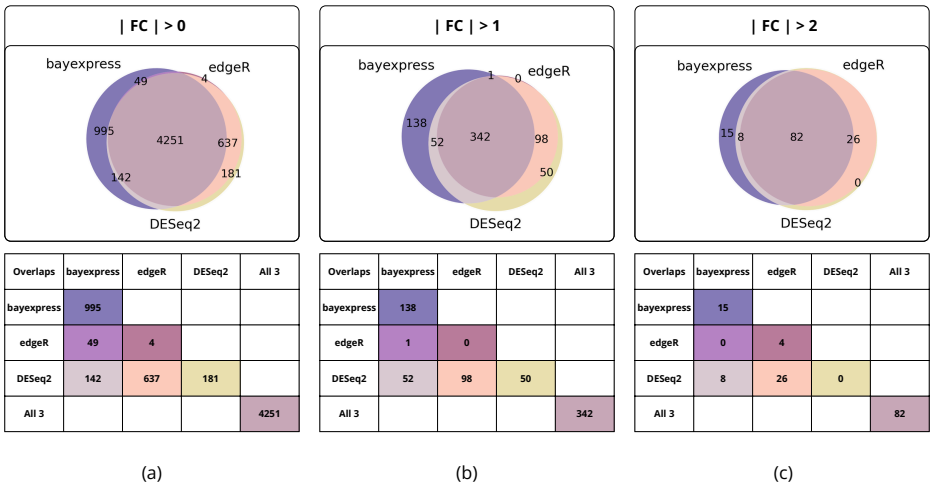
based on different labelling criteria are shown in Figure 3.5. Overall, the agreement between the three methods is high, especially *DESeq2* and *edgeR* seem to agree.

### 3.2.7 Differences in results arise from fold change cut-offs and pre-filtering

Whilst the agreement between *DESeq2*, *edgeR* and *bayexpress* overall is good, there are discrepancies – especially for our Bayesian framework *bayexpress*. We explored the disagreements and found that they arise because of differences in $\log_2$ fold change values between packages, or because of pre-filtering data (*edgeR* was set to filter genes with 0 values in any of the replicates), Figure 3.6. In Figure 3.7 and 3.8 we show a selection of example genes illustrating the differences of the classification results we found in Figure 3.5 – all of which can be explained by pre-filtering or small differences in fold change values.

### 3.2.8 Bayes factors do not require a correction for the length of genes

In Section 3.2.5 we stated that Bayes factors grow with the evidence given in data to calculate them. The more reads map to a gene, for example, the stronger the evidence we can find to support the hypotheses. Therefore, we investigated the relationships between gene lengths, q-values and Bayes factors, because longer genes have a higher probability for reads to map to them. We found, however, that in practice the analysis is not sensitive to gene lengths and no further normalisation is required, Figure 3.9. We document an extreme example of what happens if the total read depth between conditions or treatments varies and how this is not an




| Overlaps | bayexpress | edgeR | DESeq2 | All 3 |
|---|---|---|---|---|
| bayexpress | 995 | | | |
| edgeR | 49 | 4 | | |
| DESeq2 | 142 | 637 | 181 | |
| All 3 | | | | 4251 |

(a)

| Overlaps | bayexpress | edgeR | DESeq2 | All 3 |
|---|---|---|---|---|
| bayexpress | 138 | | | |
| edgeR | 1 | 0 | | |
| DESeq2 | 52 | 98 | 50 | |
| All 3 | | | | 342 |

(b)

| Overlaps | bayexpress | edgeR | DESeq2 | All 3 |
|---|---|---|---|---|
| bayexpress | 15 | | | |
| edgeR | 0 | 4 | | |
| DESeq2 | 8 | 26 | 0 | |
| All 3 | | | | 82 |

(c)

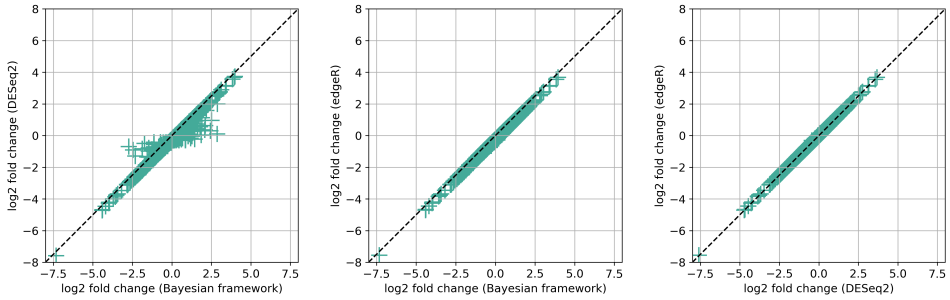
Figure 3.5: Currently popular statistical tools *DESeq2*, *edgeR* and our new Bayesian framework *bayexpress* show decent overlaps in the identification of DEGs between WT and a Snf2-mutant in yeast. Especially *DESeq2* and *edgeR* show very similar results. At the top we are showing Venn-diagrams of classified DEGs, at the bottom the same data in table format. The three diagrams show three different sets of criteria for DEG: (a) Significant change (p-value $< 0.05$, or Bayes factor $> 1$) and $|\log_2$ fold change $| > 0$, (b) Significant change (p-value $< 0.05$, or Bayes factor $> 1$) and $|\log_2$ fold change $| > 1$, (c) Significant change (p-value $< 0.05$, or Bayes factor $> 1$) and $|\log_2$ fold change $| > 2$. In total there are 7126 yeast genes, the experiment had 42 WT replicates, and 44 Snf2-mutant replicates. More information on the data can be found in the original publication [35], the code to run the analysis can be found on Github.

issue for the framework in Figures 3.3 and 3.4.

## 3.3 Conclusions

### 3.3.1 A Bayesian framework to rank genes based on the evidence for a change in gene expression between experiments

Our proposed Bayesian framework offers an exact solution for the two-sample-test problem in differential gene expression analysis on RNA-Seq data. It has been tested on both simulated (Figure 3.2) and real RNA-Seq data and compared to existing packages (Figure 3.5). A single equation can provide a rank for genes based on the statistical support for change in gene expression. Our model uses the binomial distribution, which is, with no further knowledge, an optimal (maximum entropy, least-biased) probability assignment [5] and delivers sensible results for both simulated and real data, as well as comparisons to popular published methods. We noticed differences between our inferred fold change and the fold change values calculated by *DESeq2* and *edgeR* (Figure 3.6) which explained a lot of variation between classification results of packages. A second cause for differences we identified is a pre-filtering step excluding zero-read genes which was activated in




(a) (b) (c)

Figure 3.6: The $\log_2$ fold change values of different statistical packages are not calculated in the exact same way and therefore show differences, resulting in DGE classification differences. This is the reason why we distinguish our value from others by calling it 'inferred' $\log_2$ fold change. Here we are plotting the $\log_2$ fold change of *DESeq2*, *edgeR* and the inferred $\log_2$ fold change of the Bayesian framework in relation to each other. The lines show density histograms for the scatter plots. Data: 7126 yeast genes (42 WT replicates, 44 Snf2-mutant replicates) [35]. (a) Overall, the fold change values are shifted and do not match perfectly. We checked the genes where *DESeq2* and the Bayesian framework do not agree (off-diagonal) and found that those are genes with no reads in at least one replicate. These are the genes where the effect of Laplace's rule of succession (which results from the choice of uniform prior) is visible. (b) Again, with the Bayesian framework and *edgeR* comparison we can see an overall shift, which leads to classification differences if we choose cutoffs according to fold change. Note that for positive fold changes the Bayesian framework will identify more genes as DEG, meanwhile for negative fold changes fewer compared to the other two packages. Furthermore, *edgeR* had a filter on in our analysis for genes with $n_1|n_2 = 0$, hence we do not see any genes here that do not follow the diagonal line. (c) *DESeq2* and *edgeR* match nearly perfectly.

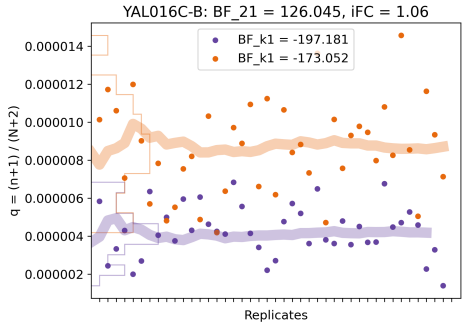
*edgeR*.

Analogous to bulk RNA-Seq data handling, there are numerous articles discussing single-cell RNA-Seq data [37]–[40]. While our framework is not specifically tailored to single-cell sequencing data, we posit that with appropriate modifications, it holds potential for adaptation to such data.

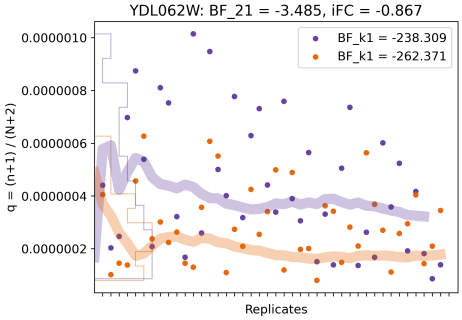
Furthermore, despite us not having made use of the prior in this framework yet, there is huge potential in doing so in the future. Information about the system (e.g. known expression changes between tissues or organisms) and experimental design (e.g. batch-effects or alike) may be fed into the analysis.

### 3.3.2 A Bayesian framework to rank genes according to variability between replicates

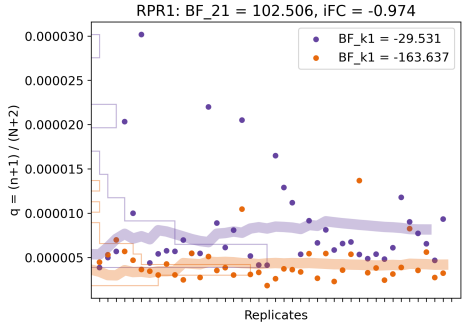
Variability in RNA-Seq experiments has been raised as an issue many times before [12], [14]–[20], [22], [24], [25], and we are only at the beginning of grasping how much of it is technical or biological nature [41]–[44]. We propose to use Bayes



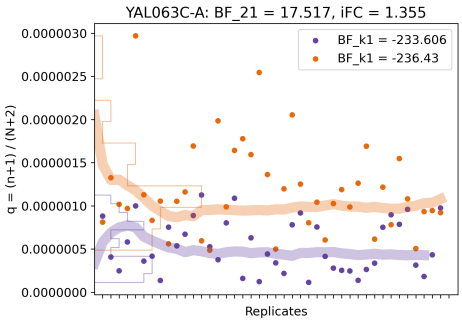

(a) YAL016C-B is one of 138 genes which are marked as DEGs in *bayexpresss* but not the other two packages.




(b) YDL062W is one of 50 genes that are DEGs according to *DESeq2* but not the other two.

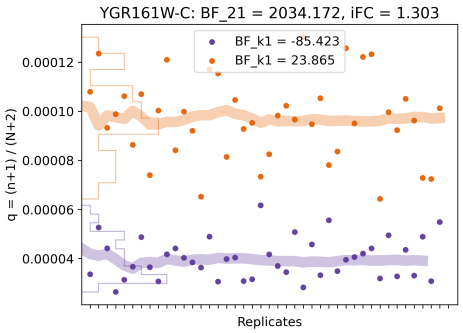



(c) RPR1 is one of 98 genes that are DEGs according to *DESeq2*, *edgeR* but not *bayexpress*.

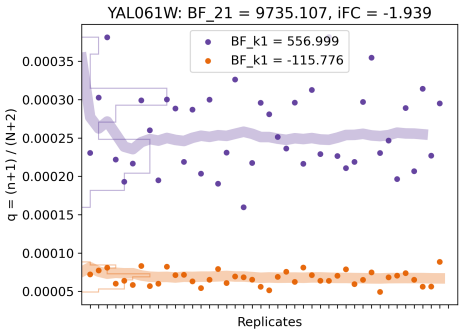



(d) YAL063C-A is one of 52 genes that are DEGs according to *DESeq2*, *bayexpress* but not *edgeR*.

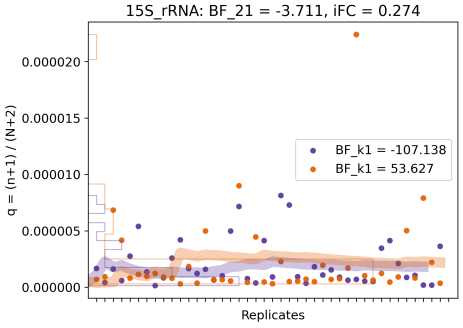
Figure 3.7: Disagreements in the classification of DEGs arise because of pre-filtering and small differences in fold change values, which are one of two defining criteria. Here, and continued in Figure 3.8 we plot q-values for 7 example genes across all replicates (42 WT, 44 Snf2-mutant) in a yeast experiment [35]. The WT is seen in purple, and the Snf2-mutant in orange. Note the variability of the y-axis ranges. The fine lines are density histograms along the y-axis, and the thick lines are estimated means of q, updated with each new replicate. For each gene we can find a Bayes factor for differential expression ($BF_{21}$) and an inferred $\log_2$ fold change at the top, both are calculated taking all 42/44 replicates into account. Bayes factors for consistency of replicates ($BF_{k1}$) can be found for each genotype in the boxes. The example genes have been selected to cover all sets in the Venn-diagram in Figure 3.5 (b). Continued in Figure 3.8.




(a) YGR161W-C is the only gene that is a DEG according to *edgeR*, *bayexpress* but not *DESeq2*.




(b) YAL061W is one of 342 genes that are DEGs according to all 3 packages.




(c) Finally, 15S rRNA is one of 6445 genes that none of the packages have identified as DEG.

| Package | bayexpress | | edgeR | | DESeq2 | |
|---|---|---|---|---|---|---|
| | iFC | BF_21 | FC | p-value | FC | p-value |
| YAL016C-B | 1.060 | 126.045 | 0.782 | 1.090e-13 | 0.775 | 1.267e-18 |
| YCR108C | 1.514 | -4.288 | NaN | NaN | 1.142 | 1.242e-02 |
| RPR1 | -0.974 | 102.506 | -1.275 | 1.489e-12 | -1.301 | 4.319e-17 |
| YAL063C-A | 1.355 | 17.517 | NaN | NaN | 1.097 | 2.824e-10 |
| YGR161W-C | 1.303 | 2034.172 | 1.011 | 1.928e-26 | 0.998 | 1.641e-50 |
| YAL061W | -1.939 | 9735.107 | -2.224 | 4.721e-51 | -2.235 | 1.953e-241 |
| 15S_rRNA | 0.274 | -3.711 | NaN | NaN | -0.137 | 6.884e-01 |

(d) *DESeq2*, *bayexpress* and *edgeR* results for the shown example genes in Figure 3.7 (a)-(d) and Figure 3.8 (a)-(c).

Figure 3.8: Disagreements in the classification of DEGs arise because of pre-filtering and small differences in fold change values, which are one of two defining criteria. Here, as a continuation of Figure 3.7, we are plotting q-values for 7 example genes across all replicates (42 WT, 44 Snf2-mutant) in a yeast experiment [35]. The WT is seen in purple, and the Snf2-mutant in orange. Note the variability of the y-axis ranges. The fine lines are density histograms along the y-axis, and the thick lines are estimated means of q, updated with each new replicate. For each gene we can find a Bayes factor for differential expression ($BF_{21}$) and an inferred $\log_2$ fold change at the top, both are calculated taking all 42/44 replicates into account. Bayes factors for consistency of replicates ($BF_{k1}$) can be found for each genotype in the boxes. The example genes have been selected to cover all sets in the Venn-diagram in Figure 3.5 (b).

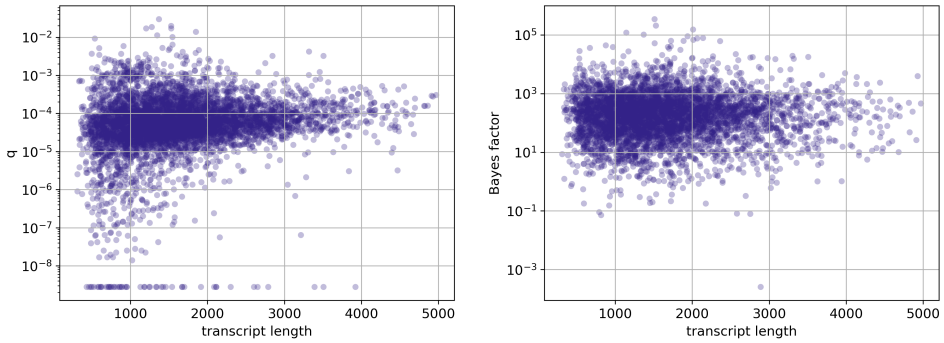



Figure 3.9: Bayes factors do not require gene length corrections. We cannot see a correlation between transcript lengths and q-values of genes (log-scale, across all replicates in the Yeast data set [35]) and neither can we see a correlation between transcript lengths and Bayes factors of genes (log-scale, across all replicates). Shown are 5219 of the 7126 yeast genes, data and code to make these plots are available on Github. We conclude that Bayes factors are not gene length dependent and that no additional normalisation is necessary.

factors as a metric for the consistency of replicates to rank genes according to their variability to monitor this issue in the interpretation of the DGE results. Of course, the quality of this ranking is dependent on available information about the consistency of a gene, i.e. the number of replicates. Moreover, we speculate that this part of the framework can be extended to identify DEGs in multiple comparison studies.

### 3.3.3 An Analytical Bayesian Framework in DGE speeds up the data analysis

Although other packages use Bayesian approaches [24], [25], [45], [46], the exact framework and model we are using here, which has been generally described by [47], is novel in differential gene expression analysis.

Whilst this may not be a decisive factor for many applications, analytical frameworks can drastically speed up data analysis. For a single run on 42/44 replicates in yeast we can already see the differences: 5s for *edgeR*, 14s for *DESeq2*, and 0.1s for bayexpress. Whenever we want to implement DGE analysis into bigger processes, like bootstrapping experiments e.g. [35], the differences start to sum up. We conducted a bootstrapping experiment which took 333s using *edgeR*, 1176s using *DESeq2*, 6s for bayexpress, which is a good improvement in light of optimising computational processes.

### 3.3.4 Ranking by Bayes factors instead of classifying by fold-change cutoffs provides a new way to communicate DGE results

One source for differences in the classification of DEGs by different packages is prefiltering of data (e.g. excluding genes with 0 reads). A second source is differences in the calculation of fold change values resulting in different values (Figure 3.6), big enough to make differences in classification via cutoffs. Both of these issues are

reflected in the examples in Figure 3.7 and 3.8. Having fold change value cutoffs as decisive factors gives those values big importance, although there is no reason to discard genes that have not doubled or quadrupled their expression. By setting arbitrary fold change cutoffs we are loosing interesting candidates with good statistical evidence for change. Only before RNA-Seq results were trusted without qRT-PCR validation, there have been discussions about the accuracy of both techniques below certain magnitudes [48], [49], hence why fold change cutoffs were set. Biologically speaking we cannot generalize that a certain fold change in expression is more significant than another. In that sense, these cutoffs are only serving as false positive filters, as a trade-off for accepting higher false negative hits. For increasing accuracy and reproducibility in general, however, we may simply need to increase the number of replicates. Ranking genes by Bayes factors (according to statistical support) can help to identify follow-up candidates without filtering out interesting genes due to fold change cutoffs. If Volcano plots [33], [50] are needed, Bayes factors for differential gene expression can be plotted against inferred $\log_2$ fold change values. Ranked list of genes could be compared using rank comparison algorithms [51], [52].

## 3.4 Materials and methods

All real data we used and all simulated data we created with all of our scripts to repeat the study and reproduce all figures can be found on our Github.

### 3.4.1 How to use Bayes factors for differential gene expression analysis

All presented equations can be coded in just a few lines of code and can be used to calculate Bayes factors for differential gene expression ($BF_{21}$), Bayes factors for consistency of replicates ($BF_{k1}$), and inferred $\log_2$ fold change values, for all genes from processed RNA-Seq data (read counts). We provide a Python and R implementation on our Github.

### 3.4.2 Yeast data

Our package comparison has been inspired by a yeast study conducted by Schurch et al. [35], [53] in 2016. They performed RNA-Seq on 42 wild type (WT) and 44 Snf2-mutant replicates and carried out a bootstrapping study to find out how many biological replicates RNA-Seq studies need and which packages to use for the analysis. Thanks to their detailed documentation and data availability, we could base a lot of our work on their data.

We used their fully processed data downloaded from their Github. Originally they had 48 replicates but only worked with 42/44, as documented in [35], [53], which we followed.

# Chapter 4

# What is a differentially expressed gene?

Transparency Note — This chapter is an extension to Chapter 3. It has largely profited from discussions about differential gene expression with Gurpinder Singh Sidhu, Thelonious Omori, Melissa Tomkins and Richard J. Morris. Since the completion of the first version of this thesis, the content of the chapter has been modified and, in parts, improved for a bioRxiv pre-print [54]. All code for figures, analyses and simulations for this chapter can be found on my GitHub.

Summary — In the last chapter we learned about a new Bayesian framework we developed for ranking genes according to their statistical evidence for gene expression change in RNA-Seq data using Bayes factors. To compare our method to existing popular analyses we had to define what exactly a 'differentially expressed genes' was, and started challenging the current definitions. Here we present a series of bootstrapping experiments on a Yeast data set (1) exploring how different researchers and statistical packages define 'differentially expressed genes', (2) demonstrating which trade-offs appear with setting arbitrary cutoffs, and (3) building motivation for moving away from binary classification towards ranking methods. We will learn about the importance of the number of replicates in a study, how variability can easily be misinterpreted as differential expression, and in which ways we can or cannot avoid these problems.

## 4.1 Introduction

### 4.1.1 The insights promised by RNA-Seq data

In Molecular Biology we strive to understand the global orchestration of cellular processes. Identifying all components and their concentrations in cells and tissues over time, development, environmental stimuli and treatments is a very compelling vision. It is therefore not surprising that techniques in this field are in high demand and, hence, rapidly evolving. RNA-Sequencing is the best snapshot of all RNA in a tissue sample (bulk RNA-Seq) or isolated cell (single-cell RNA-Seq) we can get at the moment. Numerous advances are published and reviewed regularly, helping researchers to navigate the best-practice jungle, documenting strengths and weaknesses of the techniques [14]–[16], [18], [28]–[34]. As these methodologies

are high-throughput, it becomes increasingly difficult to validate their effectiveness and reliability in real-world applications. This situation highlights a critical gap in the field, where the innovation potential must be matched by rigorous testing and validation processes to ensure that the results obtained are both accurate and meaningful. The ongoing evolution of these techniques presents an exciting opportunity for researchers, but it also necessitates a careful approach to ensure that the tools we develop can be trusted. As we strive to enhance our understanding of cellular dynamics, it is essential to establish robust frameworks for assessing the accuracy of these new methods, thereby ensuring that technological advancement and scientific rigor develop hand in hand. At the moment, e.g. we cannot tell biological (changes in gene expression or RNA accessibility) and technical fluctuations (everything under the term 'sequencing error', and bioinformatic limitations) apart, and both of them have probabilistic behaviours and components. Even if we try to validate RNA-Seq, there are challenges: The accuracy of qRT-PCRs or similar methods is under investigation itself [55], [56] and we don't know if we are able to sequence all RNA or only an accessible fraction.

### 4.1.2 The reliability of DGE identification using RNA-Seq is under discussion

Motivated by studies reporting reproducibility challenges in RNA-Seq analysis [12], [17], [19], [35], we re-examined the statistical assumptions underlying differential gene expression analysis from a Bayesian statistics perspective. The newly proposed framework in Chapter 3 enables us to rank genes by Bayes factors for differential gene expression. Here in Chapter 4 we investigated where reproducibility issues come from in three RNA-Seq data bootstrapping experiments. The results teach us about (1) how differences in the results of statistical analyses between analysis tools arise due to arbitrary cutoffs and the variability of the data; (2) how easily variable data can produce what we currently define a 'differentially expressed gene', and (3) how we can take the variability of genes into account when interpreting DGE results. All of these results support our proposal to rethink binary classifications in RNA-Seq results, given the limitations of the technique and often low numbers of repetitions of experiments.

## 4.2 Results

### 4.2.1 Bootstrapping experiments underline explanations we found for disagreements between methods

In the previous chapter we conducted a comparison between methods – following similar studies by others [12], [17], [19], [35]. We used current statistics packages to carry out differential expression analysis on a comprehensive yeast data set and presented and discussed the results (Chapter 3). Comparing our Bayesian framework, Bayes factors for differential expression, *DESeq2* [19], [20], and *edgeR* [21], [22], we found substantial overlaps in the results but also disagreements, which we would like to discuss further, after gaining new insights thanks to the following bootstrapping study.

Here, we aim to illustrate instances of genes uncovered during our exploration of the method comparison, where statistical software packages to identify DEGs did not

agree. Some of these instances have prompted a reevaluation of our understanding of what truly constitutes a DEG. We are using the yeast data we introduced in Chapter 3, an RNA-Seq experiment with 42 wild type and 44 mutant replicates collected by Schurch et al. [35].

We present a series of plots for single genes. First, we look at normalised read counts ($q = (n + 1)/(N + 2)$ with $n$ reads mapping to the gene of interest and $N$ total reads in the experiment) of 42/44 replicates of the yeast study (wild-type vs. mutant), plot at the top. Second, the plots below, are the results of a bootstrapping experiment comparing three statistical methods: Bayes factors for differential expression (see Chapter 3), *DESeq2* [19], [20], and *edgeR* [21], [22]).

We conducted a bootstrapping experiment to evaluate whether the identification of DEGs depended on the number of replicates. The results for this are presented under the q-value plots. We generated 100 data sets by sampling 3, 6, 12, or 20 replicates from the pool of 42/44 replicates and performed differential gene expression analysis using the different methods on those 100 data sets. The numbers (and colors) are the counts out of 100, how often the gene was labelled as DEG. We applied four different criteria: (1) Bayes factor $> 1$ and inferred $|\log_2$ fold change $| > 1$; (2) *edgeR*: p-value $< 0.05$ and $|\log_2$ fold change $| > 1$; (3) *DESeq2*: p-value $< 0.05$ and $|\log_2$ fold change $| > 1$; (4) Bayes factor $> 1$ and no inferred $\log_2$ fold change cutoff.

The first set of examples (Figure 4.1: 'YAL016C-B', 'YAL031W-A', 'YAR035W', 'YAR068W') has been taken of a pool of genes that were identified as DEG by Bayes factors (compare criteria (1) and (4)), but not the other two packages, see Figure 3.5 (b) in Chapter 3. These genes are not highly expressed but show clear expression changes between WT and mutant, still, they are not reliably identified by *edgeR* (2) and *DESeq2* (3), even if 20 replicates are given. The reason for them not being picked up is due to their only slight variations in fold change values. We can also see that for the Bayesian framework to identify some of them requires a high number of replicates.

The second set of examples (Figure 4.2: 'YGL228W', 'YBR078W', 'YIL094C', 'YGL253W') has been chosen to continue asking questions about fold change cutoffs. None of these genes show up as DEG, even though some of them show noticeable changes. Furthermore, YIL094C is a beautiful example to talk about the group of genes that are identified as DEG by *DESeq2* and *edgeR* but not the Bayesian framework: The reason is a (seemingly almost preposterous) difference in how exactly $\log_2$ fold change values are calculated. We compared the 3 fold changes and found that our inferred value is consistently slightly lower, see Appendix Chapter 3. In summary, while certain genes may not be desired as DEGs due to their lack of biological significance or the influence of confounding factors, it is equally important to be aware of the potential for overlooking significant genes in our analyses. For some of these example genes here we were wondering why we would not want them to show up as DEGs and how many DEGs we have overlooked in analyses so far.

The third set of genes (Figure 4.3: 'YGR192C', 'YOR383C', 'YHR174W', 'YDR077W') shows some of the highest Bayes factors in the analysis across all replicates, show-

ing clear separations between WT and mutant. Note, that the Bayes factors grow to astronomical numbers when summing up the evidence over 42/44 replicates, in line with increasing amounts of data supporting a consistent inference. Still, some of these genes are not reliably picked up when fold change cutoffs are introduced. Whether a smaller fold-change is biologically less relevant remains to be explored – currently there seems no apparent reason to think that. There are discussions that some genes may show changes in expression that are statistically significant but biologically irrelevant. For instance, small changes in genes that are consistently highly expressed in all conditions might not contribute meaningfully to the biological processes under investigation. Therefore, some people conclude that including such genes as DEGs could lead to misleading interpretations of the underlying biology. Excluding a lot of genes, however, is actively ignoring information. From our point of view, this might make us blind to identifying potential mechanisms of importance. Another line of argument might be to relate concentration changes to binding constants for a specific binding event. If a small fold change is sufficient to take a concentration from below the binding constant to above the binding constant then a gene will have an impact, even though the changes are slow.

The fourth set of examples (Figure 4.4: 'YNL232W', 'YLR329W', 'YDR291W', 'YPR164W') is a negative control showing that analyses also successfully identify genes where the change is not extreme enough. On average over 42/44 replicates we can see distinct tendencies of the genes in WT and mutant, hence why they obtained a positive Bayes factor. However, the changes are not big enough to cross fold change thresholds and also for a lower number of replicates Bayes factors would often not support any expression change.

The fifth set of examples (Figure 4.5: 'snR32', 'YPL032C', 'YNL034W', 'YPL030W') presents genes with negative Bayes factors (taking all replicates into account). Note, that in one of them the average lines are separated but we do not get a positive Bayes factor because there are not many reads mapping (see y-axis scale).

To summarise the findings of our statistical software package comparison, we see both beautiful agreements and concerning differences. Most disagreements can be explained by arbitrary cutoffs, which support our previous suggestion to renounce binary classifications and introduce ranking-based methods in the future (Chapter 3). Bayes factors enable this ranking and can also replace $\log_2$ fold change rankings, for which we have no judicious biological backup regardless. We can clearly see how with a higher number of replicates, we get better-defined data and, therefore, better accuracy of the analysis.

### 4.2.2 RNA-Sequencing delivers a lot of numbers, but do we need more?

Despite these challenges in classification, RNA-Seq is a huge success, because it delivers numbers – lots of them. The quality of those numbers is dependent on many factors: the experimental setup and technique, of course, but also earlier research like the quality of genome assemblies, and the number of biological replicates. Unfortunately, the latter is a good practice that is still often not implemented, even though it is a way to increase the resolution of RNA-Seq. In the following section we want to demonstrate how easily we can get significant results from variability in

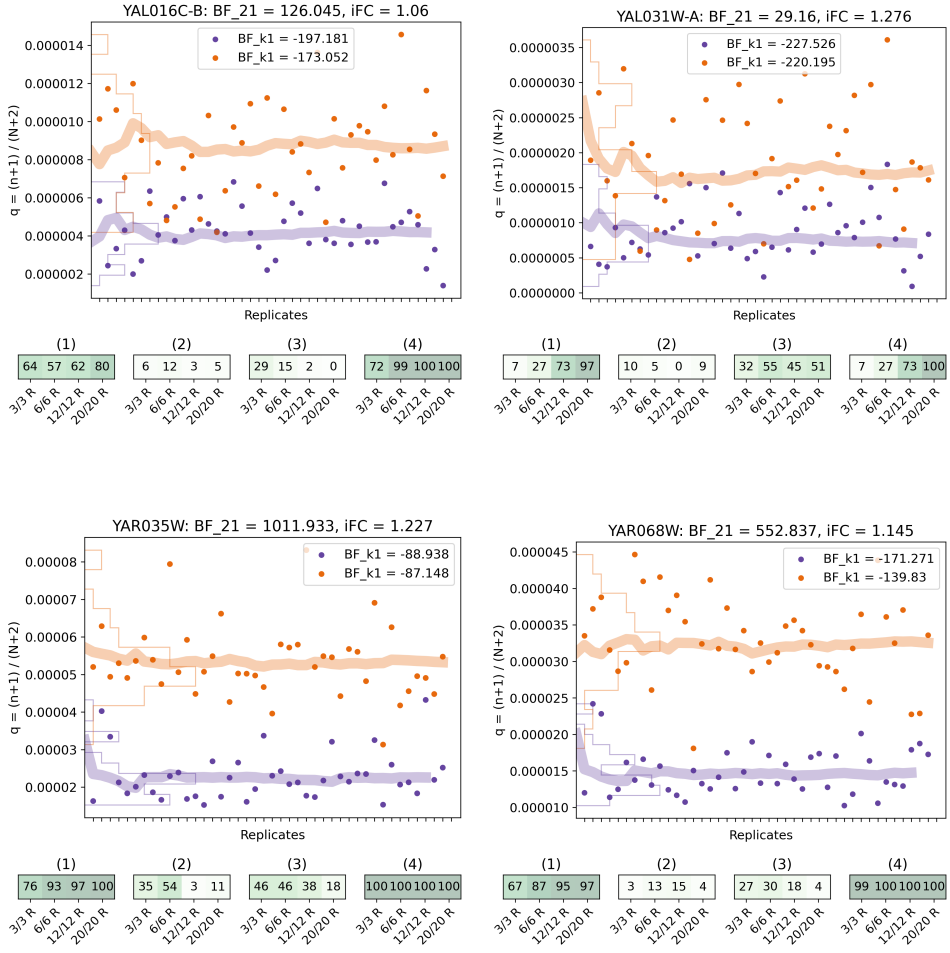



Figure 4.1: Example genes in yeast that move between DEG and no DEG classification depending on arbitrary cutoffs. The four examples presented tend to be identified by Bayes factors (if sufficient replicates are given) but not *edgeR* and *DESeq2*. The plots at the top show inferred $q$-values (normalised read counts, $q = (n+1)/(N+2)$ with $n$ reads mapping to the gene of interest and $N$ total reads in the experiment) for one gene per plot across 42/44 RNA-Seq replicates [35]. The WT is seen in purple, and the Snf2-mutant in yellow. Note the variability of the y-axis ranges. The fine lines are density histograms along the y-axis, and the thick lines are estimated means of $q$, updated with each replicate. For each gene we can find a Bayes factor and inferred $\log_2$ fold change at the top, those are calculated taking all replicates into account. The numbers at the bottom of each plot are the results of the bootstrapping experiments: The number of times (x/100) the gene has been identified as DEG, for data sets with 3, 6, 12 and 20 replicates ((1) Bayes factor $> 1$, and $|\log_2$ fold change$| > 1$; (2) *edgeR*: p-value $< 0.05$ and $|\log_2$ fold change$| > 1$; (3) *DESeq2*: p-value $< 0.05$ and $|\log_2$ fold change$| > 1$; (4) Bayes factor $> 1$, no $\log_2$ fold change cutoff). *edgeR* had a filter for 0 read genes on, which is causing some differences. Bayes factors in boxes are Bayes factors for testing the consistency of the number of mapping reads across replicates (see Chapter 3). If the gene is marked* it is part of the list of highly variable genes identified in another bootstrapping experiment, see Figure 4.6.

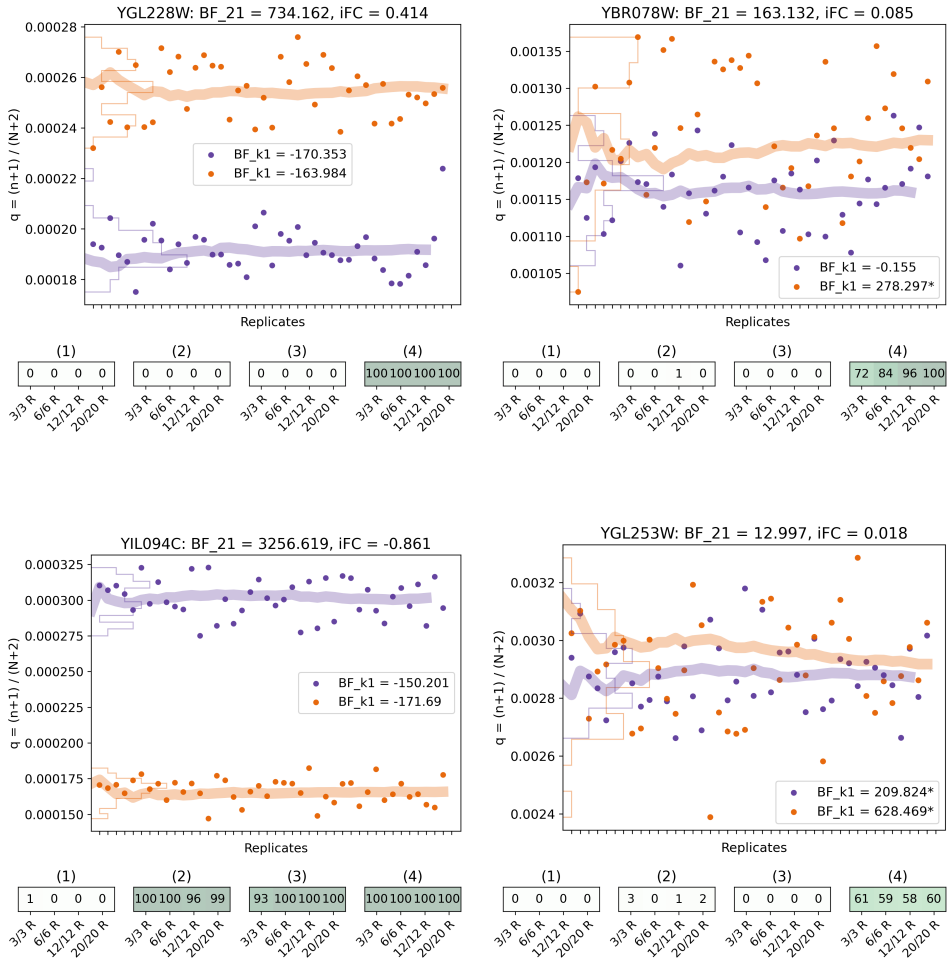



Figure 4.2: Example genes that move between classes depending on arbitrary $\log_2$ fold change cutoffs. Note, that the inferred fold-change in the Bayesian framework is consistently lower than the fold-change calculated by *DESeq2* and *edgeR*. The plots at top show inferred $q$-values (normalised read counts, $q = (n+1)/(N+2)$ with $n$ reads mapping to the gene of interest and $N$ total reads in the experiment) for one gene per plot across 42/44 RNA-Seq replicates [35]. The WT is shown in purple, the Snf2-mutant in yellow. Note the variability of the y-axis ranges. The fine lines are density histograms along the y-axis, and the thick lines are estimated means of $q$, updated with each replicate. For each gene a Bayes factor ($BF_{21}$) and inferred $\log_2$ fold change is given at the top of the plot, those are calculated taking all replicates into account. The numbers at the bottom of each plot are the results of the bootstrapping experiments: The number of times (x/100) the gene has been identified as DEG, for data sets with 3, 6, 12 and 20 replicates ((1) $BF_{21} > 1$, and $|\log_2$ fold change$| > 1$; (2) *edgeR*: p-value $< 0.05$ and $|\log_2$ fold change$| > 1$; (3) *DESeq2*: p-value $< 0.05$ and $|\log_2$ fold change$| > 1$; (4) $BF_{21} > 1$, no $\log_2$ fold change cutoff). *edgeR* had a filter for 0 read genes on, which is causing some differences. Bayes factors in boxes are Bayes factors $BF_{k1}$ for testing the consistency of the number of mapping reads across replicates (Chapter 3). If the gene is marked* it is part of the list of highly variable genes identified in another bootstrapping experiment, see Figure 4.6.

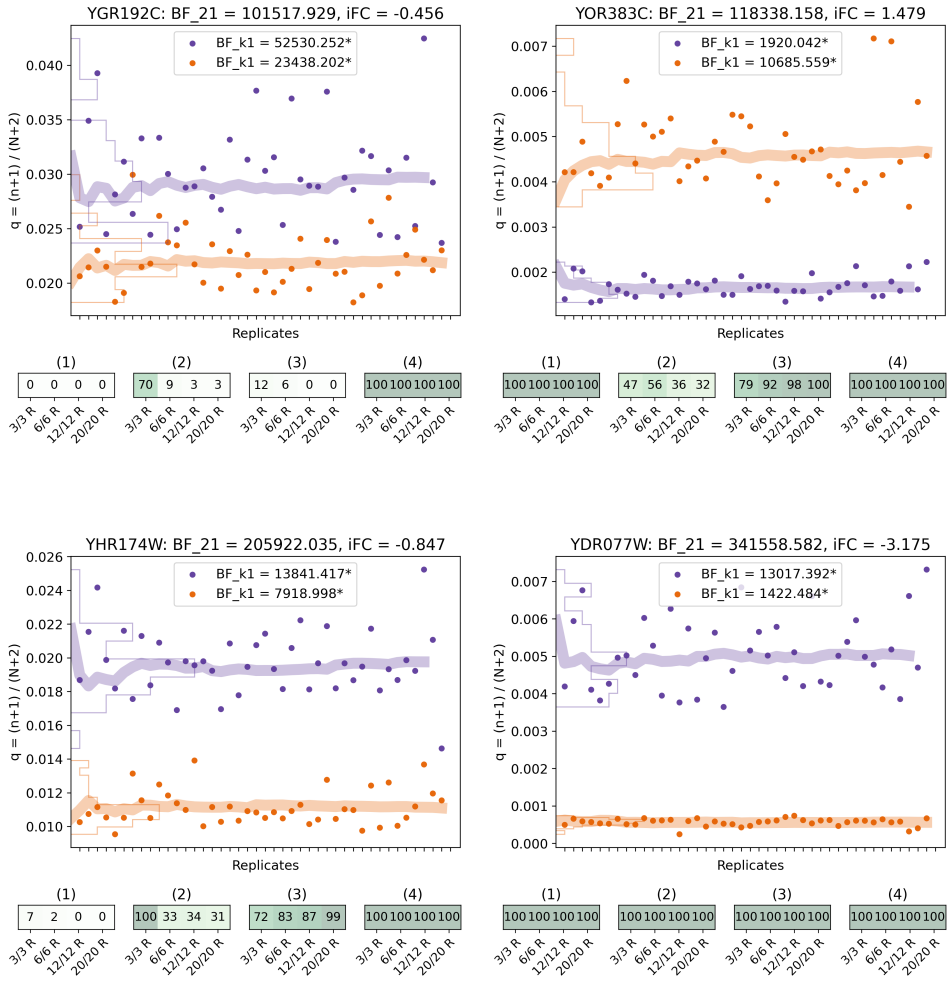



Figure 4.3: Example genes with some of the highest Bayes factors in the analysis across all replicates are not always identified as DEGs. The plots at the top show inferred $q$-values (normalised read counts, $q = (n+1)/(N+2)$ with $n$ reads mapping to the gene of interest and $N$ total reads in the experiment) for one gene per plot across 42/44 RNA-Seq replicates [35]. The WT is seen in purple, the Snf2-mutant in orange. Note the variability of the y-axis ranges. The fine lines are density histograms along the y-axis, and the thick lines are estimated means of $q$, updated with each replicate. For each gene a Bayes factor ($BF_{21}$) and inferred $\log_2$ fold change is given at the top of the plot, those are calculated taking all replicates into account. The numbers at the bottom of each plot are the results of the bootstrapping experiments: The number of times (x/100) the gene has been identified as DEG, for data sets with 3, 6, 12 and 20 replicates ((1) $BF_{21} > 1$, and $|\log_2$ fold change$| > 1$; (2) *edgeR*: p-value $< 0.05$ and $|\log_2$ fold change$| > 1$; (3) *DESeq2*: p-value $< 0.05$ and $|\log_2$ fold change$| > 1$; (4) $BF_{21} > 1$, no $\log_2$ fold change cutoff). *edgeR* had a filter for 0 read genes on, which is causing some differences. Bayes factors in boxes are Bayes factors $BF_{k1}$ for testing the consistency of the number of mapping reads across replicates (Chapter 3). If the gene is marked* it is part of the list of highly variable genes identified in another bootstrapping experiment, see Figure 4.6.

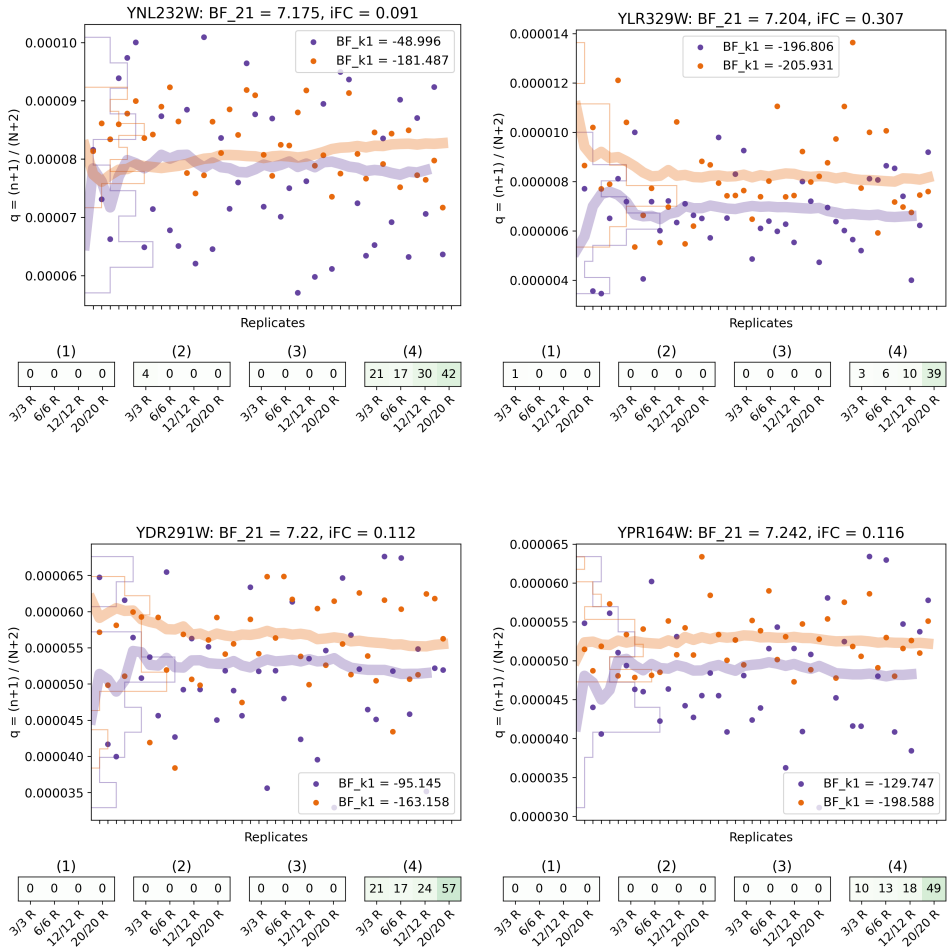



Figure 4.4: Different analyses do agree in the identification of no change, or change that can only be identified with a high number of replicates. The plots at the top show inferred $q$-values (normalised read counts, $q = (n+1)/(N+2)$ with $n$ reads mapping to the gene of interest and $N$ total reads in the experiment) for one gene per plot across 42/44 RNA-Seq replicates [35]. The WT is seen in purple, the Snf2-mutant in orange. Note the variability of the y-axis ranges. The fine lines are density histograms along the y-axis, and the thick lines are estimated means of $q$, updated with each replicate. For each gene a Bayes factor ($BF_{21}$) and inferred $\log_2$ fold change is given at the top of the plot, those are calculated taking all replicates into account. The numbers at the bottom of each plot are the results of the bootstrapping experiments: The number of times (x/100) the gene has been identified as DEG, for data sets with 3, 6, 12 and 20 replicates ((1) $BF_{21} > 1$, and $|\log_2$ fold change$| > 1$; (2) *edgeR*: p-value $< 0.05$ and $|\log_2$ fold change$| > 1$; (3) *DESeq2*: p-value $< 0.05$ and $|\log_2$ fold change$| > 1$; (4) $BF_{21} > 1$, no $\log_2$ fold change cutoff). *edgeR* had a filter for 0 read genes on, which is causing some differences. Bayes factors in boxes are Bayes factors $BF_{k1}$ for testing the consistency of the number of mapping reads across replicates (Chapter 3). If the gene is marked* it is part of the list of highly variable genes identified in another bootstrapping experiment in Section 4.2.4, or Figure 4.6.

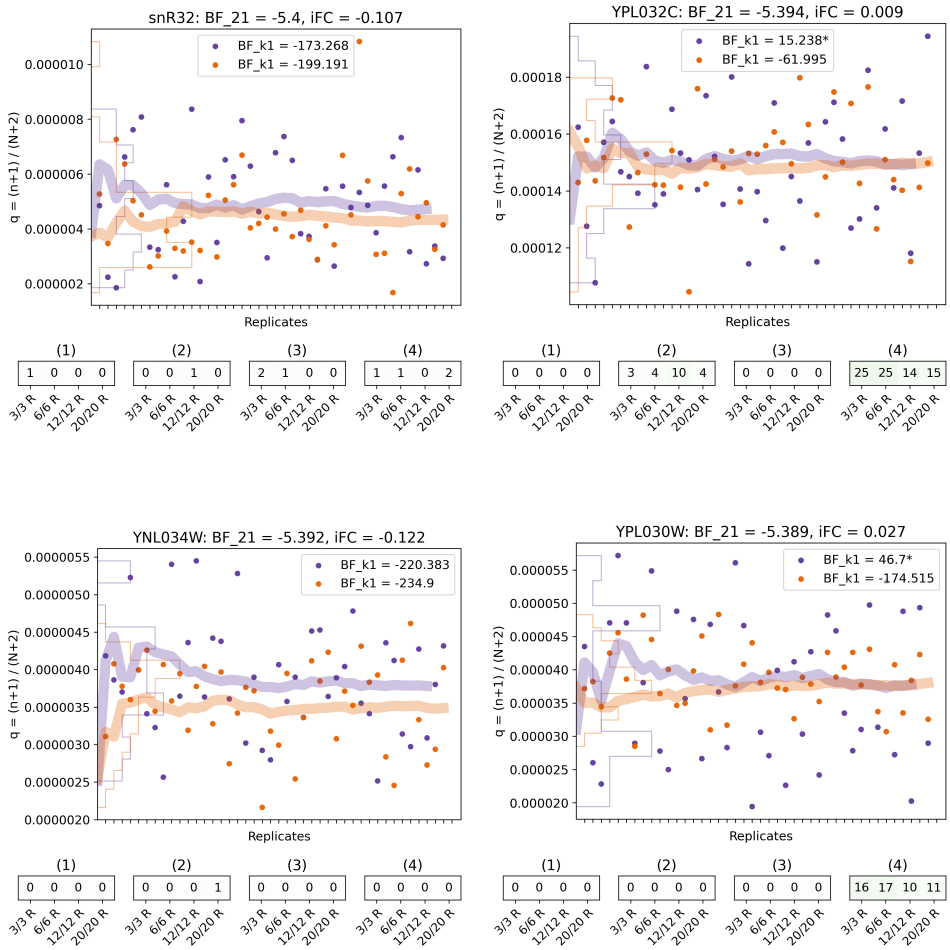



Figure 4.5: A set of genes resulting in negative Bayes factors, if all replicates are taken into account. Here we are looking at 4 example genes with negative Bayes factors, in other words, no statistical support for gene expression change (taking all replicates into account). The plots at the top show inferred $q$-values (normalised read counts, $q = (n+1)/(N+2)$ with $n$ reads mapping to the gene of interest and $N$ total reads in the experiment) for one gene per plot across 42/44 RNA-Seq replicates [35]. The WT is seen in purple, the Snf2-mutant in orange. Note the variability of the y-axis ranges. The fine lines are density histograms along the y-axis, and the thick lines are estimated means of $q$, updated with each replicate. For each gene a Bayes factor ($BF_{21}$) and inferred $\log_2$ fold change is given at the top of the plot, those are calculated taking all replicates into account. The numbers at the bottom of each plot are the results of the bootstrapping experiments: The number of times (x/100) the gene has been identified as DEG, for data sets with 3, 6, 12 and 20 replicates ((1) $BF_{21} > 1$, and $|\log_2$ fold change$| > 1$; (2) *edgeR*: p-value $< 0.05$ and $|\log_2$ fold change$| > 1$; (3) *DESeq2*: p-value $< 0.05$ and $|\log_2$ fold change$| > 1$; (4) $BF_{21} > 1$, no $\log_2$ fold change cutoff). *edgeR* had a filter for 0 read genes on, which is causing some differences. Bayes factors in boxes are Bayes factors $BF_{k1}$ for testing the consistency of the number of mapping reads across replicates (Chapter 3). If the gene is marked* it is part of the list of highly variable genes identified in another bootstrapping experiment in Section 4.2.4, or Figure 4.6.

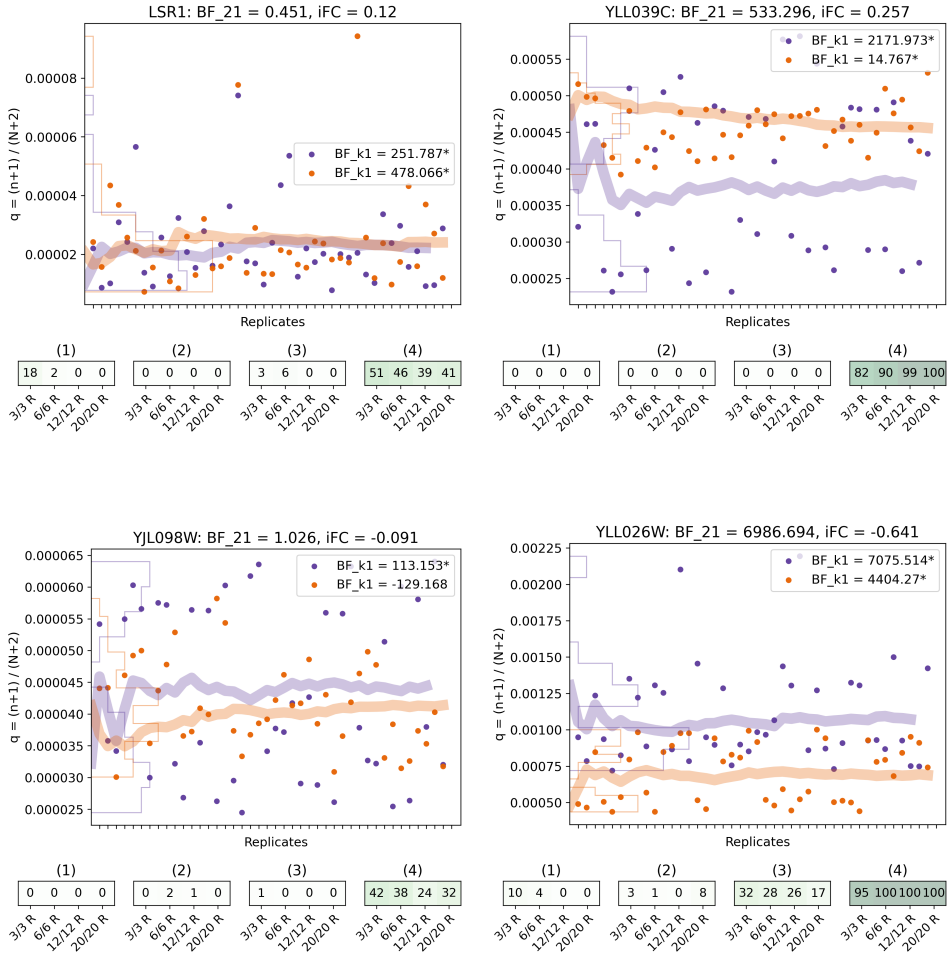



Figure 4.6: Bayes factors cannot inform about high variability and inconsistency of genes but can still reflect expression changes for those potentially problematic cases. The plots at the top show inferred $q$-values (normalised read counts, $q = (n+1)/(N+2)$ with $n$ reads mapping to the gene of interest and $N$ total reads in the experiment) for one gene per plot across 42/44 RNA-Seq replicates [35]. The WT is seen in purple, the Snf2-mutant in orange. Note the variability of the y-axis ranges. The fine lines are density histograms along the y-axis, and the thick lines are estimated means of $q$, updated with each replicate. For each gene a Bayes factor ($BF_{21}$) and inferred $\log_2$ fold change is given at the top of the plot, those are calculated taking all replicates into account. The numbers at the bottom of each plot are the results of the bootstrapping experiments: The number of times (x/100) the gene has been identified as DEG, for data sets with 3, 6, 12 and 20 replicates ((1) $BF_{21} > 1$, and $|\log_2$ fold change$| > 1$; (2) *edgeR*: p-value $< 0.05$ and $|\log_2$ fold change$| > 1$; (3) *DESeq2*: p-value $< 0.05$ and $|\log_2$ fold change$| > 1$; (4) $BF_{21} > 1$, no $\log_2$ fold change cutoff). *edgeR* had a filter for 0 read genes on, which is causing some differences. Bayes factors in boxes are Bayes factors $BF_{k1}$ for testing the consistency of the number of mapping reads across replicates (Chapter 3). Starting top left, 'LSR1' shows beautifully how this gene would often be wrongly identified as differentially expressed, by chance, due to its high variability. 'YLL039C' seems to be expressed in a bimodal fashion in WT, switching to one state in Snf2, and also 'YJL098W' and 'YLL026W' show interesting bimodal patterns. Just like before, all of those are examples of interesting genes we miss out on when applying fold change cutoffs.

data, especially with low repetition numbers. Is a 'statistically significant' result, a result we can communicate without uncertainty?

### 4.2.3 How easily can variable data masquerade as differential expression?

In this experiment, we sought to illustrate the potential impact of variable data on the identification of DEGs. To emphasize this point, we sampled an additional set of 100 bootstrapped data sets with 3 and 10 replicates exclusively using wild-type data. We calculated Bayes factors for differential expression ($BF_{21}$) for all genes in the data set, and found many 'significantly differentially expressed' genes. The objective was to highlight the challenge arising with data variability and the ease at which a variable gene, of which there are many (more on that in the third experiment), can be misidentified as a DEG. Our analysis revealed that while most DEGs exhibited expected moderate $\log_2$ fold change values, a surprising number of genes, particularly in data sets with only 3 replicates, displayed noteworthy deviations, see Figure 4.7. In Figures 4.8, 4.9 and 4.10 we show how these numbers change for higher Bayes factor cutoffs and demonstrate that we can find strong evidence for 'differential expression', especially in experiments with only 3 replicates.

This insightful experiment mirrors the concept of using Bayes factors to quantify the variability of genes, as previously described (Chapter 3). Calculating this factor for each gene can caution researchers about genes with inconsistent replicates. However, the metric is dependent on the resolution of the data, which can only be raised by increasing the number of replicates. The exploration underscores once more the importance of accounting for data variability and exercising prudence in the interpretation of differential gene expression results, particularly in scenarios with limited replicates. The findings presented here serve as a compelling reminder of the complexities inherent to differential gene expression analysis and the need for robust methodologies to navigate the challenges posed by variable data.

### 4.2.4 Identification of consistently inconsistent genes

It is evident from the last two sections that there are genes with high variability in the data set. The issue has been raised many times before [12], [14]–[20], [22], [24], [25], and we are only at the beginning of grasping how much of it is of technical or biological nature [41]–[44]. In this experiment, we wanted to find how many genes with inconsistent expression there are in the data set. Previously, we defined how to calculate Bayes factors to test the consistency of replicates and rank genes according to their variability between replicates (Chapter 3). In order to identify highly variable genes in the yeast data set [35] we calculated Bayes factors for consistency. However, we also went further and identified consistently inconsistent genes. We carried out a third bootstrapping experiment, where we shuffled a pool of 42/44 replicates 100 times (see 100 colourful lines in Figure 4.11) and counted how many genes of the 7126 genes in yeast are identified as inconsistent using Bayes factors as previously defined (Chapter 3), while increasing the number of replicates. We can learn from this experiment that it takes quite a few more than 12 replicates for our results to converge. By chance, the luxurious situation of around 40 replicates is enough to conclude there are around 1600 consistently inconsistent genes in the WT and around 900 in the Snf2-mutant. Taking all replicates into account we find

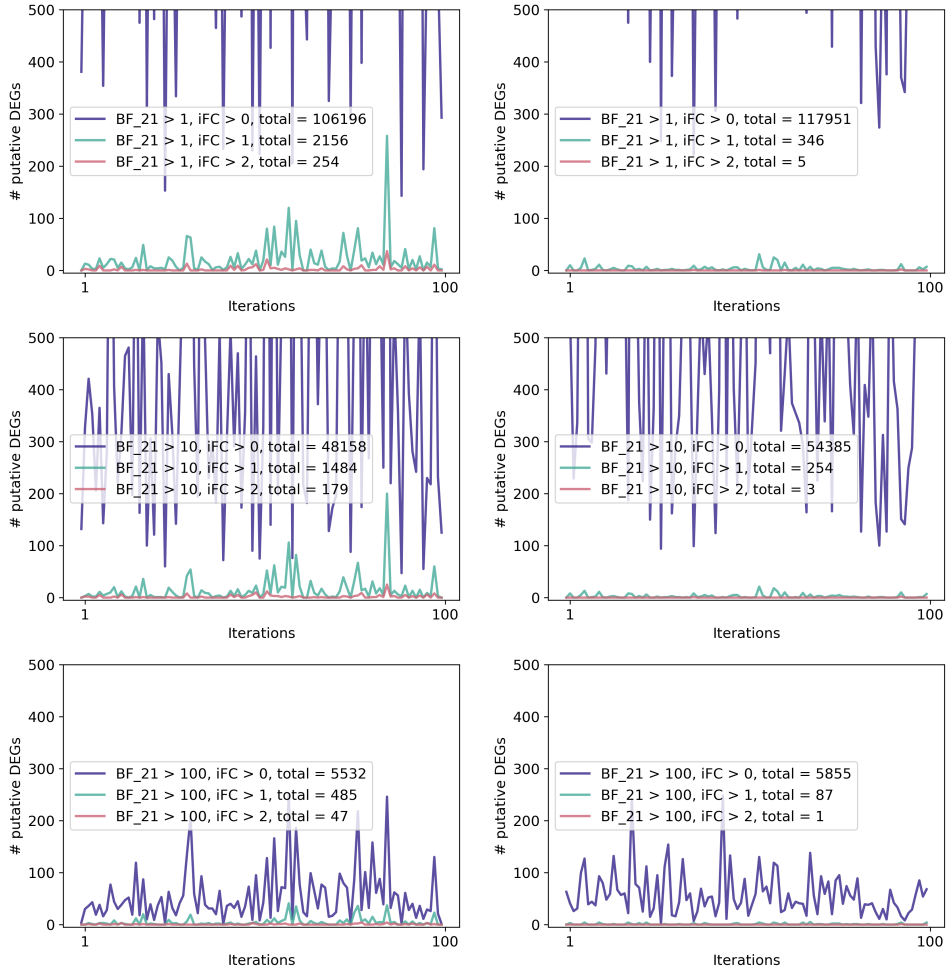



Figure 4.7: A control experiment (wild type vs. wild type comparison) demonstrating how variability in data is leading to false identification of differentially expressed genes. The number of putative DEGs (y-axis) identified in 100 bootstrapped data sets (x-axis), given 3 different criteria (purple, green and red). The total number of genes is 7126, 44 WT replicates (wild type only) of a yeast study have been used to sample the individual data sets (Data: [35]). On the left the number of replicates in each sampled data set is 3, on the right hand side it is 10. The first row is using a Bayes factor cutoff of 1, the second row uses a Bayes factor cutoff of 10 and the third row uses a Bayes factor cutoff of 100. We only show counts up to 500 here, for full data visit our Github.

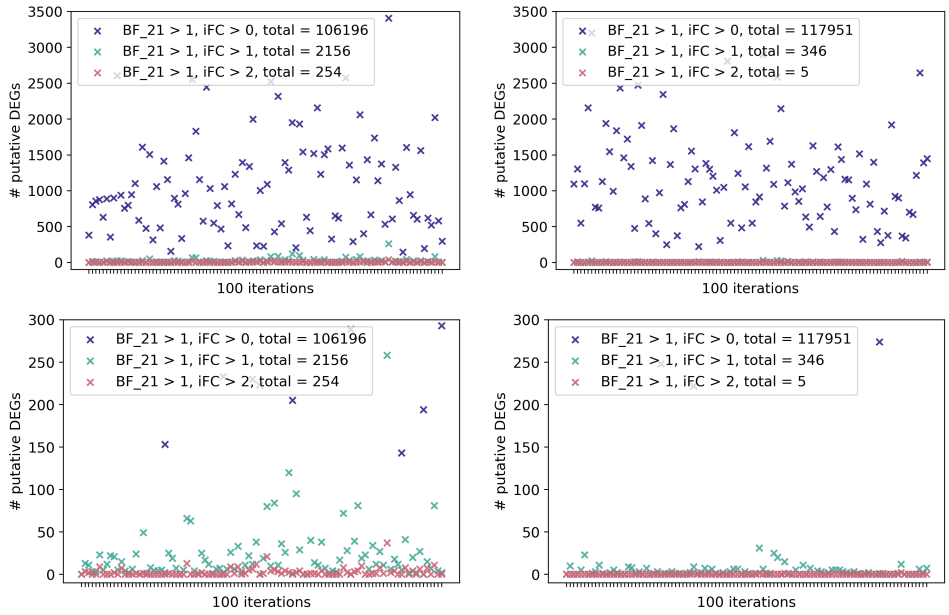



Figure 4.8: A control experiment comparing wild type vs. wild type in a differential gene expression analysis demonstrating how variability in data leads to false positive identification of differentially expressed genes. The number of putative DEGs (y-axis) identified over 100 bootstrapping iterations (x-axis), given 3 different criteria (in colours). The total number of genes is 7126, 44 WT replicates (wild type only) of a yeast study have been used to sample the individual data sets (Data: [35]). On the left the number of replicates in each sample is 3, on the right hand side it is 10. Top: all data points, bottom: zoom into top plot. For this Figure we used a Bayes factor cutoff of 1, in Figure 4.9 we increase the cutoff to 10, and in 4.10 to 100 to demonstrate that we can find strong evidence for differentially expressed genes in the control, with decreasing numbers of putative DEGs for higher numbers of replicates.

an overlap of 1426 and a union of 1633 genes for WT and an overlap of 765 and a union of 922 in the mutant across 100 bootstrapping iterations. The union numbers drastically increase, however, when reducing the number of replicates. This is in line with the results we discovered in Figure 4.7. The genes found in the union list are marked* in Figure 4.1, 4.2, 4.3, 4.4, 4.5. To highlight this group we discuss examples of variable genes we identified ('LSR1', 'YLL039C', 'YJL098W', 'YLL026W') in Figure 4.6. We did take a look at those examples and realized, how the assumption for the Bayesian framework (all replicates are consistent with each other) is not met and nevertheless we get sensible results – but only if given 42/44 replicates.

For big numbers of replicates, we can use Bayes factors to evaluate the variability of replicates, and bootstrapping, as a warning system for genes where we need to be cautious with interpreting results. This could enable us to communicate variability in data sets. We considered filtering highly variable genes but decided not to, when we saw the interesting example genes in Figure 4.6 we would miss.

Based on our observations of the described data, we can proceed with interpreting differential gene expression analysis on these genes only if enough replicates are

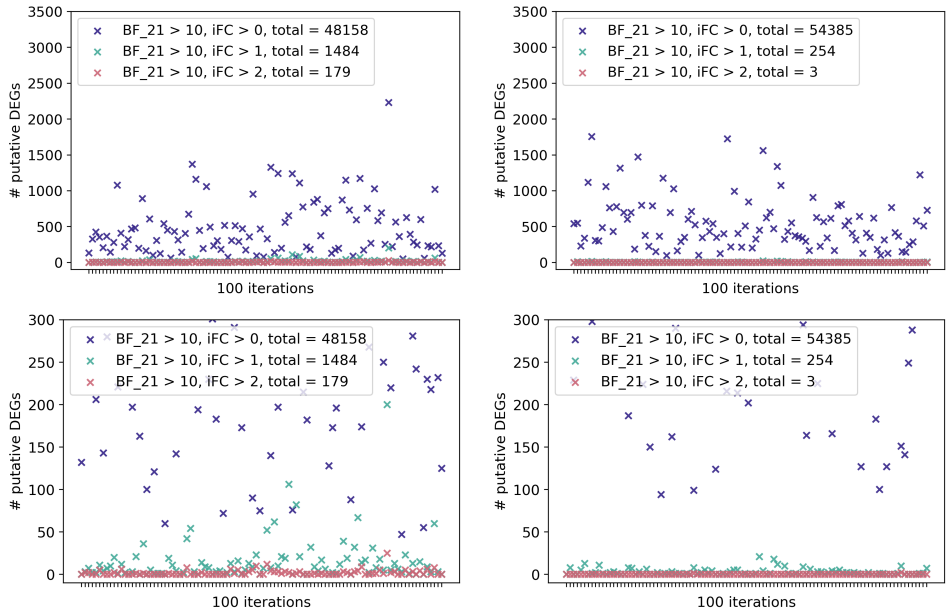



Figure 4.9: A control experiment comparing wild type vs. wild type in a differential gene expression analysis demonstrating how variability in data leads to false positive identification of differentially expressed genes. The number of putative DEGs (y-axis) identified in 100 bootstrapped data sets (x-axis), given 3 different criteria (purple, green and red). The total number of genes is 7126, 44 WT replicates (wild type only) of a yeast study have been used to sample the individual data sets (Data: [35]). On the left the number of replicates in each sample is 3, on the right hand side it is 10. Top: all data points, bottom: zoom into top plot. For this Figure we used a Bayes factor cutoff of 10, in Figure 4.8 we set the cutoff to 1, and in 4.10 to 100 to demonstrate that we can find strong evidence for differentially expressed genes in the control, with decreasing numbers of putative DEGs for higher numbers of replicates. For more information and code, visit our Github.

given. We advise exercising caution in interpreting Bayes factors for these genes, as illustrated by several examples in Figure 4.6. Limited numbers of replicates may lead to wrong impressions for highly variable genes, see genes we identified in a bootstrapping experiment in Figure 4.11.

## 4.3 Discussion

Three different bootstrapping experiments of a Yeast experiment show the limitations of differential gene expression analysis using RNA-Seq. In the first experiment, we compared different statistical methods for differential gene expression analysis and learned how insufficient numbers of replicates produce false positive classifications in differential gene expression analysis and how fold change cutoffs drive potential false negative classifications of DEGs. Of course, these statements are made in a binary-thinking world, which is at the heart of our criticism. Statistical methods have a decent agreement in their results, increasing consensus comes with better-defined data (more replicates). In the second experiment, the WT-WT

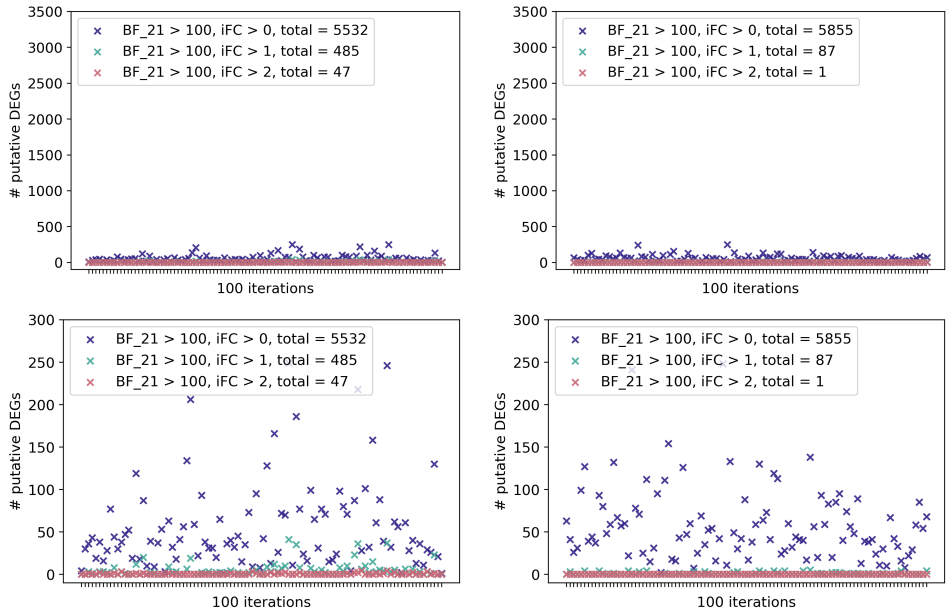



Figure 4.10: A control experiment comparing wild type vs. wild type in a differential gene expression analysis demonstrating how variability in data leads to false positive identification of differentially expressed genes. The number of putative DEGs (y-axis) identified in 100 bootstrapped data sets (x-axis), given 3 different criteria (purple, green and red). The total number of genes is 7126, 44 WT replicates (wild type only) of a yeast study have been used to sample the individual data sets (Data: [35]). On the left the number of replicates in each sample is 3, on the right hand side it is 10. Top: all data points, bottom: zoom into top plot. For this Figure we used a Bayes factor cutoff of 1, in Figure 4.8 we set the cutoff to 1 and in 4.9 to 10, to demonstrate that we can find strong evidence for differentially expressed genes in the control, with decreasing numbers of putative DEGs for higher numbers of replicates. For more information and code, visit our Github.

control experiment, we have shown how high variability of a gene across replicates in RNA-Seq, due to biological and technical fluctuations, may lead to false positive DEGs. In the third experiment, we found that there are around 1600 consistently inconsistent genes (of 7126) in the WT and around 900 consistently inconsistent genes in the Snf2-mutant. Especially those genes are likely to produce 'false positive DEGs' if not sufficient replicates are given in studies. Considering the high number of problematic genes, especially because we are handling a population of yeast cells in this case, raises questions about the value and insights gained by large-scale techniques like this.

### 4.3.1 What can we expect from RNA-Seq?

Of course, there is a huge potential and vision behind transcriptomics. All of those challenges together result in reproducibility issues, however, which have been documented many times before. The large data output of RNA-Seq builds up expectations for deep insights from differential gene expression studies. However, we do not have information on the resolution of RNA-Seq nor the variability of genes

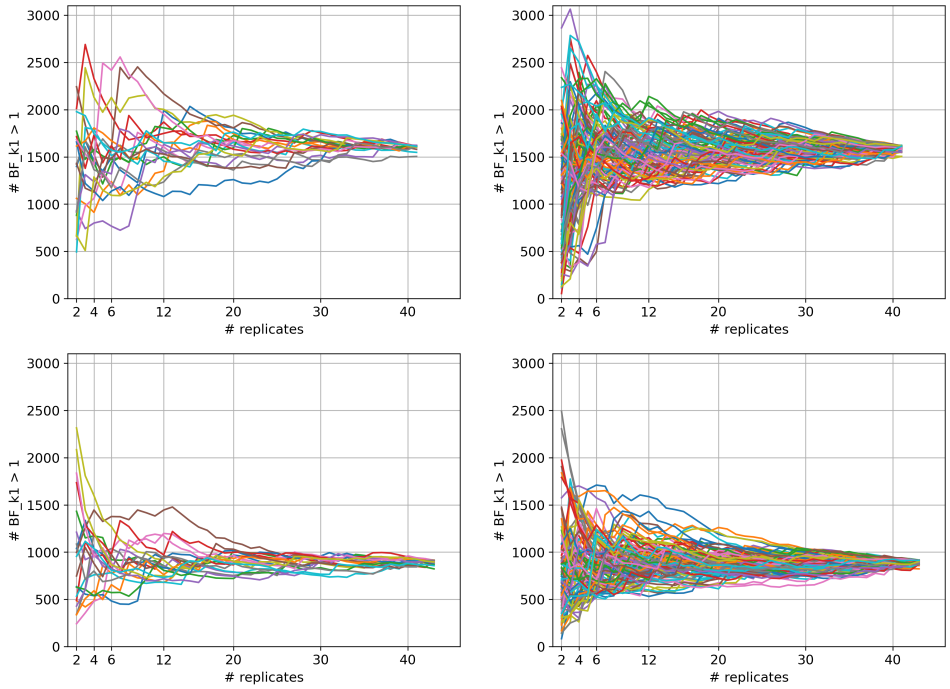



Figure 4.11: We identified around 1600 consistently inconsistent genes in the WT (top) and around 900 consistently inconsistent genes in the Snf2-mutant (bottom). We calculated Bayes factors for testing the consistency of gene expression ($BF_{k1}$) for increasing numbers of replicates (2-42/44) in 20 (left) and 100 (right) bootstrapping iterations (each one colourful line in the plots). On the y-axis we count the number of genes that have been identified as inconsistent across replicates. The criterion is a Bayes factor for consistency, $BF_{k1} > 1$. On the x-axis we increase the number of replicates we take into account. We can see how overall there is a different number of inconsistent genes in WT and mutant. Fascinatingly, this test also beautifully shows how in Bayesian statistics the order in which data is added (in form of more replicates) does not matter, and so when all information (all replicates) are taken into account, we always reach the same conclusion.

without many repetitions of experiments.

The vision of a fully described cell is a huge motivator. But does RNA-Seq deliver what it promises to get us closer to this goal? How much information is there for us in RNA-Seq data and how much knowledge can we deduce from it? (This a the moment to go back to Figure 1.1 at the start of the thesis.) Indeed, we do get a lot of numbers that seem very attractive to us in the information age. These numbers feel like a big insight, despite the lack of information on their resolution. We need to be careful not to be tricked by our brains to read, interpret and communicate them as solid truth. Especially with a scientific culture of publishing 'positive' results only we solely expose ourselves to RNA-Seq success stories. How many RNA-Seq data sets have (for sometimes good reasons) vanished under desks somewhere? We don't know.

Nevertheless, identifying limitations always serves as a guide for improving techniques. As we gain a deeper understanding of cellular processes at the mechanistic level, we can develop solutions to enhance RNA accessibility for RNA-Seq. Advances in sequencing techniques, such as the emergence of long-read sequencing, will address parts of alignment issues and improve genome assemblies. Furthermore, the importance of replication becomes evident as we continue to learn and refine our experimental and technical tools. In parallel, we have the opportunity to enhance algorithms in bioinformatics and improve statistical methods by incorporating the Bayesian statistics toolbox, which offers elegant and efficient solutions.

### 4.3.2 Where do we go from here?

We do want to give a motivational outlook after these cleanup efforts. How can we update our RNA-Seq workflows to contribute more knowledge to the puzzle of life?

1. We can increase the number of replicates to increase the resolution of the data. Schurch et al. answered the question: 'How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?' in [35] and conclude 6-12 replicates are needed, dependent on the resolution at which we wish to detect differences. Liu et al. investigated the tactics of raising sequencing depth vs. increasing biological replication [57]. Gierlinski et al. have developed a method to identify outliers among replicates [53]. While we acknowledge the trade-off between time and resources, our findings, as well as previous studies, underscore the substantial decrease in insight gained from an RNA-Seq experiment if an insufficient number of replicates are employed. This, in turn, increases the likelihood of drawing erroneous conclusions.

2. We can learn a lot from well-documented RNA-Seq data bases and previous studies. The big variances we get for genes across replicates have been discussed and challenged statistical method development since early days of RNA-Seq [35], [53]. We can learn about the variability of our data and whether it is of biological (be it individual cells [44], [58], [59] or organisms [41], [42]) or technical nature [43] from the wealth of existing data. In addition to that, we can investigate how e.g. pooled samples [60]–[62] vs. individuals in multicellular and unicellular organisms change the picture.

3. Documenting and publishing the results of any study, no matter the outcome, will feed the investigations mentioned in the previous point.

4. Updating our statistical frameworks to communicate outcomes of studies in a more sophisticated way may help us to correctly evaluate the outcomes of experiments. The way we can avoid fold change cutoffs, and therefore the accuracy trade-offs coming with them, by using Bayes factors for differential gene expression, serves as a motivator for exploring how Bayesian inference and further causal inference could shape research in the future.

## 4.4 Methods

All real data we used and all simulated data we created is found, with all of our scripts to repeat our study and reproduce all figures, on our Github. The introduction of Bayes factors for differential gene expression was given in the previous

chapter and we have applied it exactly how documented and derived in our former work.

### 4.4.1 Yeast data

Our package comparison has been heavily inspired by a yeast study conducted by Schurch et al. [35], [53] in 2016. They performed RNA-Seq on 42 wild type (WT) and 44 Snf2-mutant replicates and carried out a bootstrapping study to find out how many biological replicates RNA-Seq studies need and which packages to use for the analysis. Thanks to their detailed documentation and data availability, we could base a lot of our work on their data. We used their fully processed data downloaded from their Github. Originally they had 48 replicates but only worked with 42/44, as documented in [35], [53], which we followed.

### 4.4.2 Bootstrapping DGE analysis

We created 100 data sets randomly sampling 3, 6, 12, or 20 replicates from the pool of 42/44 yeast replicates without replacement (meaning, an individual replicate can appear only once within each data set). Afterwards, we set up a pipeline to analyse the data using *edgeR*, *DESeq2* and our Bayesian framework which was used throughout our work, from package comparison (Figures 4.1, 4.2, 4.3, 4.4, 4.5 and 4.6) to the control experiment on wild-type only data (Figures 4.7, 4.8, 4.9, and 4.10) and the identification of consistently inconsistent genes (Figure 4.11). All data sets and results have been stored and published, and so is all code to reproduce the full procedure.

# Chapter 5

# RNA-Seq in the detection of long-distance mobile mRNA in plants

Transparency Note — Parts of this chapter have been uploaded to bioRxiv in a manuscript titled 'Re-analysis of mobile mRNA datasets highlights challenges in the detection of mobile transcripts from short-read RNA-Seq data' authored by Pirita Paajanen, Melissa Tomkins, Ruth Veevers, Michelle Heeney, Hannah Rae Thomas, Federico Apelt, Eleftheria Saplaoura, Saurabh Gupta, Margaret Frank, Dirk Walther, Christine Faulkner, Julia Kehr, Friedrich Kragler, Richard J. Morris and myself [63]. Other parts may be recognised from two further of our articles emerging from the PLAMORF project, which are central to Chapter 6, that are authored by subgroups of the former author list [64], [65].

Summary — The long-distance transport of messenger RNAs (mRNAs) has been shown to be important for several developmental processes in plants. A popular method for identifying travelling mRNAs is to perform RNA-Seq (RNA-Sequencing) on grafted plants. This approach depends on the ability to correctly assign sequenced mRNAs to the genetic background from which they originated. The assignment is often based on the identification of single-nucleotide polymorphisms (SNPs) between otherwise identical sequences. A major challenge is therefore to distinguish SNPs from sequencing errors. In the following 3 chapters we are presenting our efforts to investigate and improve the accuracy of mobile mRNA detection using RNA-Seq data. Here, we give an introduction to the experimental setup and explain our motivation that led to the work presented in Chapter 6.

## 5.1 Introduction

### 5.1.1 mRNAs exit cells and move long distances in plants

At the beginning of my doctoral studies, it was widely acknowledged that a multitude of messenger RNAs (mRNAs) molecules are transported over significant distances within plants [66], [67] and between parasitic plants and their hosts [68]–[70]. It seems, for some mysterious reason, that mRNAs exit the cells where they

are transcribed to travel in vascular tissues from one part of a plant to another. Non-cell-autonomous functions of these transported RNAs have been described in several instances [71], such as acting as signals and influencing development in cells distinct from their origin [72], [73]. These processes include leaf development in tomato [74], meristem maintenance and root growth in *Arabidopsis thaliana* [75], [76] or tuberisation in potato [77]. Reports of mRNAs travelling over long-distances coupled with evidence of mRNAs acting non-cell-autonomously have given rise to the proposal that mobile mRNAs may act as signaling agents in a novel long-distance communication system [66], [71], [78]–[80].

The broad range of possible purposes and tasks conducted by travelling biomolecules, in general, are under investigation. The consequences of the movement can be seen in a variety of observable effects. While for some molecules signaling functions have been documented (proteins, siRNAs, miRNAs) [81]–[84], for mRNAs – the main interest in this investigation – different functions are debated [80], [85]. For some specific examples of mobile mRNAs we have descriptions of their impact [74], [86]–[88], for many more, we remain puzzled what the function and relevance of mobility is [67], [89], [90]. The identification of transported RNAs, their origins, destinations, and the conditions under which they are transported is crucial for understanding their potential roles in signaling processes [85].

### 5.1.2 How do mRNA move long distances?

The exact mechanisms for the movement of huge molecules like mRNAs are still missing. For mRNAs to move from cell-to-cell, they most likely need to cross the cell walls through pores that link the cytoplasm of neighbouring cells with each other, called plasmodesmata [75], [91]. How exactly this movement occurs is yet to be discovered. Evidence for plasmodesmal trafficking, despite the limitations of size, is puzzling. A full mRNA, potentially folded into a 3-dimensional structure, fitting through these passages, suggests that there may be active and specific translocation mechanisms that either unfold the mRNA and shuttle a molecule through the plasmodesmata or enlarge the cytoplasmic sleeve. There is also evidence for RNAs and proteins travelling together [92], [93], again lacking exact mechanistic explanations. So far, all of those ideas have only been described, but not shown to actually happen, leaving the riddle of how such large molecules move between cells effectively unsolved [91].

Movement inside cells may be via diffusion, local movement from cells to neighboring cells may happen by some kind of active transport to increase efficiency. However, as soon as we consider larger distances, these modes of transport become less effective. We can do a simple back-of-the-envelope calculation based on measurements in mammals [94], [95]: If an mRNA is travelling a distance of 1 cm by diffusion, it would take 1000 days, approximately 20 days by non-selective active transport, and approximately 2 days, or less, by selective active transport. For these numbers, however, we have not yet taken movement through plasmodesmata into account, which would significantly increase the times [96], [97]. In order for long-distance transport to happen within biologically reasonable time scales, it has been considered, that long-distance transport of macromolecules occurs in the vascular system of plants [69], [98], [99].

The vasculature of plants consists of two counterparts. In the xylem, a non-living pipe-like structure, water and nutrients taken up by the roots are transported to the shoot, or wherever they are needed. In the living tube-structure called phloem, water and photosynthates, forming the phloem sap, are transported. This sap has been proposed to be a good transport medium for macromolecules [69], [98]. The idea is inspired by parasitic plants, such as *Cuscuta pentagona*, a plant feeding on the phloem sap of *Arabidopsis thaliana*. It has been shown that approximately 30% of the Arabidopsis transcriptome can be detected in the phloem of Cuscuta [69], [98], supporting the hypotheses that (a) there are mRNAs in the phloem and (b) it is the phloem where the translocation of mobile mRNAs takes place [99]. The flow of the sap may reduce the transport time of mRNAs from several days to minutes, which would comfortably lie within the expected range of half-lives of most mRNAs [100], [101].

### 5.1.3 Which mRNA move long distances?

There were several efforts to catalog macromolecules in vascular tissues. The appearance of RNA and protein was so abundant, that researchers started talking about phloem transcriptomes and proteomes. These detection efforts led to publications of phloem transcriptomes in Arabidopsis, Melon, and Legumes [89], [102]–[104], reporting the presence of several hundreds of mobile mRNAs.

The amount of mRNA molecules that re-enter cells at a destination is very low. Reports place the imported numbers of mobile mRNAs in cells at less than 0.1% of the endogenous levels [105]–[107], i.e. only a fraction of the observed transcriptional noise [108]. If transcription of an mRNA resulted in 1000 transcripts then the expected stochastic fluctuations would be greater than $\pm$ 30 transcripts (typically over-dispersed compared to a Poisson distribution [108], [109]), whereas the expected number of imported non-endogenous transcripts of the same mRNA would be only 1. Should mobile mRNA be part of a communication system, this raises questions about how such signals can be read and decoded. Ideas from information theory have been put forward as a means of analysing and testing mRNA signaling as a communication system [85], [90]. These inquiries are likely to be important for unravelling the potential signaling role and biological function of mobile mRNAs, and would be greatly aided by knowing which mRNAs move, when they travel, from where to where and under which conditions. Furthermore, determining which mRNAs are translocated might allow for features to be identified, such as physico-chemical properties or sequence and structural patterns (motifs, zip-codes), that relate to transport or functional characteristics. For instance, a previously identified mobility motif that was found to be enriched in the low abundance subset that turned out to relate to mRNA stability [110], inline with recent findings which suggest that mRNA localisation can be explained by mRNA stability [95]. There is therefore significant interest in the detection of mobile mRNAs and many reports and datasets are meanwhile available [111]. Nevertheless, extensive data mining efforts, including our own, have failed to find predictive features for mobility in the sequences of currently annotated transcripts [112].

### 5.1.4 The detection of mobile RNAs often exploits the ancient technique of grafting

Combining root stocks and vegetative material (scions) of different plants is an ancient practice called grafting, that has been and is still used in traditional horticulture to confer abiotic and biotic resistance to crops [113]. Also in the detection of mRNAs, the property of plants to re-establish vascular and cellular connections after cutting proves to be useful. It has become the method of choice to combine different plants and find RNA molecules that have crossed the graft-junction, while having some control over contamination [67].

Early breakthroughs identified long-distance mobile signalling agents, among them flowering regulators in plants, called florigen [114], of which Flowering Locus T is now known to be a contributing factor [115], [116]. Later on, with the establishment of modern sequencing technologies and computational analyses, the field moved on to high-throughput detection methods, enabling the transcriptome-wide search for mobile mRNAs. From what we have learned about the likelihood of mobile mRNAs travelling through the phloem, it seems probable that as soon as a graft-junction heals and the vascular system reconnects, hundreds of phloem-mobile mRNAs can cross the graft junction. Indeed, using grafting followed by high-throughput transcriptomics (RNA-Seq), studies over the last decade have identified hundreds to potentially thousands of graft-mobile mRNAs in a variety of plant species [69], [89], [98], [106], [107], [117]–[122].

### 5.1.5 High-throughput detection of mobile mRNAs in plants using grafting followed by RNA-Seq

Using RNA-Seq to identify all mRNAs extracted from plant tissue from grafts between different genetic backgrounds, such as different species, accessions, ecotypes, or cultivars, has been widely employed for identifying transported transcripts over graft junctions and potentially over long distances within a plant [69], [89], [98], [106], [107], [117]–[122]. However, the identification of transported transcripts keeps challenging the field, as existing graft-mobile mRNA classifications show poor reproducibility [63], [67]. These inconsistencies may be attributed to experimental conditions and technical differences that hinder direct comparability, or the nature of mobile mRNA transport [67].

Identifying which transcripts belong to which genotype in grafted plants (between two different genotypes) requires a means of distinguishing them. Choosing closely related genotypes that differ in known positions between their genomes, single nucleotide polymorphisms (SNPs), allows for transcripts to be identified based on these allele-specific nucleotides in RNA-Seq data, Figure 5.1. For grafts between different cultivars or ecotypes of the same species, such as the *Arabidopsis thaliana* Col-0 and Ped ecotypes [98], or between closely related species, such as grapevines species *Vitis palmata, Vitis girdiana*, Vitis hybrid C3309, *Vitis vinifera* cultivar Riesling [106], transcripts from different genotypes can be identified based on the SNPs between them. Typically, a requirement is made for a defined number of RNA-Seq reads to have a SNP that corresponds to the alternative allele for a transcript to be assigned to a foreign genotype. These values are chosen carefully to reduce the risk of sequencing noise biasing the identification. Previously published criteria include at least one read if it is covering a minimum of two SNPs [106], or

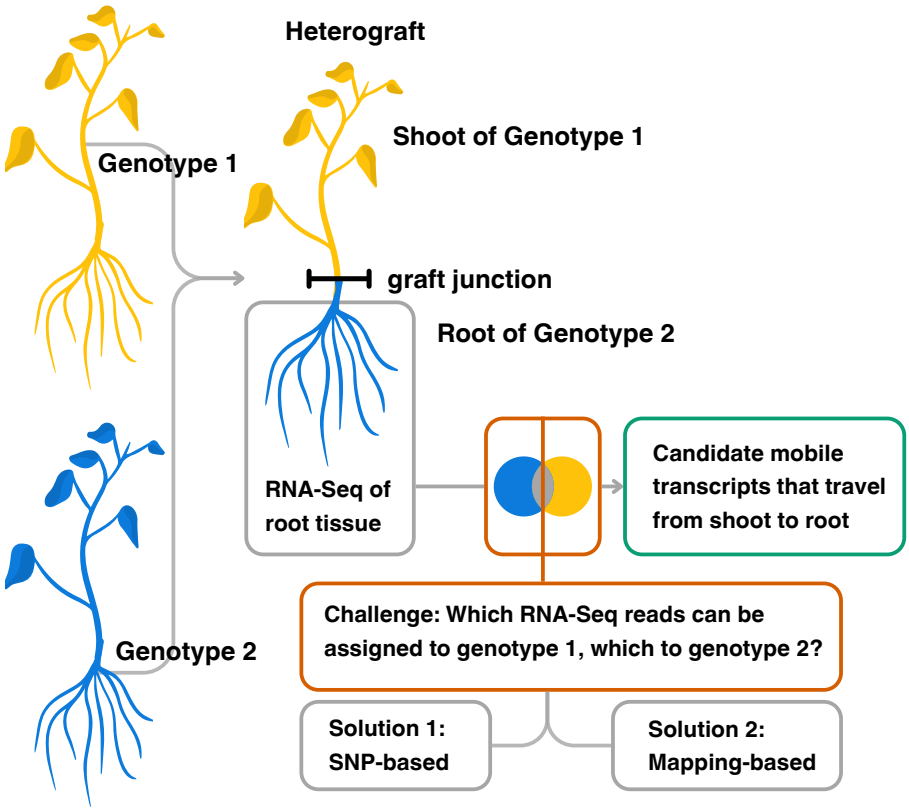



Figure 5.1: Grafting coupled with RNA-Seq has been employed to identify transcripts that move from tissue of one genotype/species/ecotype/cultivar into tissue of another genotype/species/ecotype/cultivar across the graft junction. Shown here is a grafting-based strategy for identification of mRNAs that move from shoot (scion) to root (stock), from genotype 2 to genotype 1, using a scion:stock=genotype 2:genotype 1 heterograft. The same strategy can be used to identify transcripts that move from shoot to root from genotype 1 to genotype 2 using a genotype 1:genotype 2 graft. Transcripts that move root to shoot can be identified by analysing mRNAs in shoot tissue. Natural grafts, such as those established between the parasitic dodder plant and its host plants, can be used in place of artificial grafts. A key challenge in all such approaches is how to assign transcripts to each genotype; methods for doing so are based on (1) SNP (single nucleotide polymorphism) identification or on (2) the alignment to different reference genomes. For grafts from the same species, or similar genotypes, SNPs can be used to distinguish between genotypes and thus identify the source genotype of each transcript (1). For grafts between different species, mapping (2) each RNA-Seq read to the genome assemblies can be an effective method for determining which transcripts are specific to one species. In this thesis we will only focus on SNP-based approaches.

two [106], three [123] or four [98] reads covering the same SNP. When these criteria are met, the transcript is defined as mobile.

## 5.2 Motivation and Aim for use of Bayesian statistics in the detection of mobile mRNA

### 5.2.1 Needles in haystacks and mobile RNAs in sequencing data

We have learned earlier, that the imported numbers of mobile mRNAs is less than 0.1% of the endogenous levels [105]–[107]. Sequencing has known technological limitations which lead to every nucleotide having a small probability of being assigned to a different nucleotide (base-calling error). Illumina sequencing machines produce base-calling errors at a rate of approximately 0.1 to 1% per base on average [124], [125]. The detection of mobile mRNAs using SNP differences between two grafted plants therefore faces a challenging signal-to-noise problem. We asked how well the above mentioned absolute read count criteria take this into account, and came up with a simple estimate on how many currently mobile annotated mRNAs can be explained by sequencing noise, instead of evidence for mobility.

Sequencing providers often give a quality assurance, for instance, that 85% of the reads have at least Q30 (i.e. a base-calling error of less than $10^{-3}$). Thus, assuming an RNA-Seq experiment delivers 20 million reads using paired-end sequencing of 2 x 100 bp, we will have a total of 4 billion base calls of which we can expect up to approximately 4 million to be incorrect. Usually, further Phred score filtering is applied to enhance the quality, and reduce the number of errors [98]. Base-calling errors are not the only potential source of change in the expected nucleotide identity. Reverse transcriptases can introduce base changes with an error rate of approximately $10^{-5}$ to $10^{-4}$; the reverse transcription reaction error can be different for different nucleotides, for instance a 'G' to 'A' bias [126], [127], and may exhibit a range of artifacts [128]. In the above example, this could result in a further 400,000 potential nucleotide substitutions. Multiple effects give rise to differences between the sequenced fragments and the corresponding genome sequence. Some of these differences may appear in SNP positions and be indistinguishable from a SNP supporting the alternate allele.

The alignment of RNA-Seq reads to the genomes of the two grafted plants is a process with two outcomes (endogenous genotype, foreign genotype), for which, with no further knowledge, the binomial distribution represents an optimal (maximum entropy, least-biased) probability assignment [5]. We denote the probability of a SNP matching the alternate allele by $q$. Using a (cumulative) binomial distribution, we assessed existing mobile mRNA criteria by computing the probability of transcripts being classified as being mobile based on sequencing noise, Figure 5.2. For low read-depths the probability of the criteria for mobile mRNA being met by chance is extremely low, however, this increases with higher read-depths, meaning that the false positives are to be expected for high-coverage experiments. The high coverages employed in mobile mRNA studies result in high read-depths and hence more absolute numbers of errors.

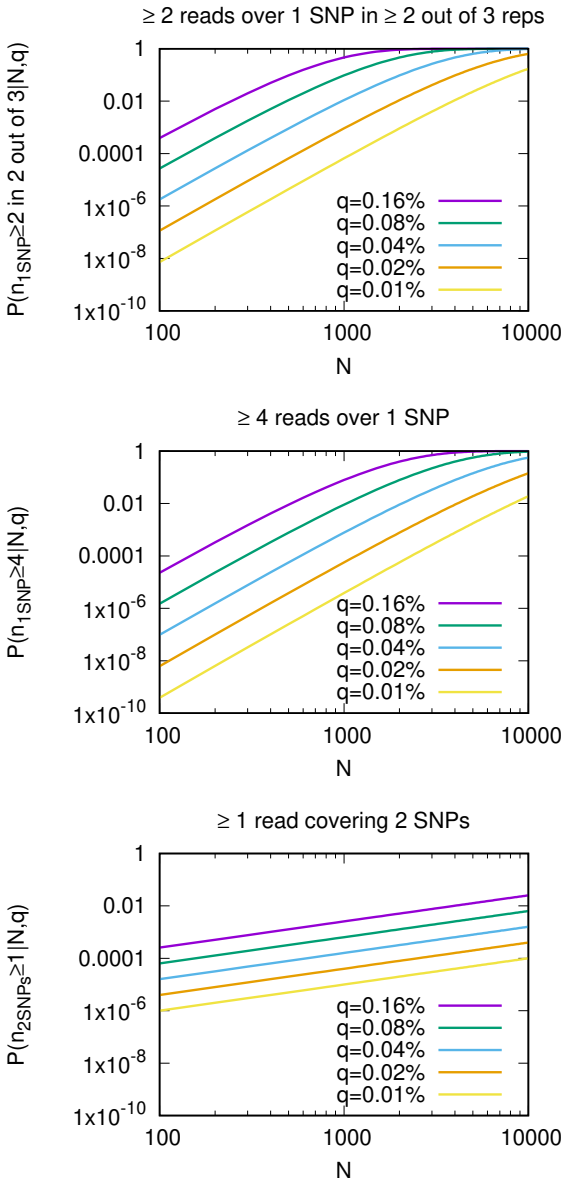



Figure 5.2: Published criteria for defining mobile mRNAs based on absolute read counts suffer from read-depth dependencies. Shown are three different mobile mRNA definitions (top, middle, bottom) and their dependence on read-depth ($N$) and on the rate of a SNP matching to the alternate allele ($q$). The number of read counts over one SNP that correspond to the alternate allele is denoted by $\mathrm{n_{1SNP}}$, over two SNPs by $\mathrm{n_{2SNPs}}$. Note that both axes are on a log-scale. The requirement for co-occurring SNPs on one read is more stringent and less likely to occur by chance at higher read-depths. If the rate of finding the alternate allele was a consequence of base-calling errors only, then we would expect $q \approx 10^{-Q/10}/3$, where $Q$ is the Phred score cut-off. For a Phred score of Q30 ($q < 0.0\dot{3}\%$), these criteria are robust up to moderately high (several hundred) read-depths and would be unlikely to occur by chance.

| | # of reported mobile mRNAs | # consistent with q = 0.1% | # consistent with q = 0.01% |
|---|---|---|---|
| *A. thaliana* | 2006 | 1455 | 1086 |
| *V. girdiana* | 1130 | 945 | 384 |

Table 5.1: Total numbers of reported mobile mRNAs across all experiments in *A. thaliana* [98] and *V. girdiana* [106] that can be explained by expected sequencing noise. Assuming a probability of 0.1% and 0.01% for finding the alternate allele, this table lists how many transcripts have been reported to be mobile and of these how many have occurrences of the alternate allele (supporting mobility) that are consistent with this probability, i.e. no statistical evidence for mobility. The results are based on a defined background error rate that is the same for all SNPs. The hypotheses for whether the data can be explained by background errors or whether the experimental observations go beyond what would be expected from background errors alone (and are therefore potential candidates for mobile transcripts) were evaluated following established procedures [5], [64].

### 5.2.2 Many previously identified mobile mRNA candidates show signals consistent with sequencing noise

Criteria based on absolute read counts exhibit a read-depth dependency (Figure 5.2), resulting in a bias towards highly abundant transcripts being assigned as mobile. This bias has been reported and discussed in a previous study in our lab [110]. We therefore investigated how many previously published mobile transcripts identified by RNA-Seq are consistent with technological noise and read-depth. Table 5.1 shows how many mobile mRNAs would be consistent with different assumed error rates, $q$, for the alternate allele. The consistency of the data with these error rates was performed using Bayes factors [64], as we will describe in the next chapter. If the probability for the alternate allele, $q$, was purely a consequence of base-calling errors, we would expect $q$ to be approximately a third of the base-calling error; for a Phred quality score cut-off of Q30 this would result in a base-calling error rate of $10^{-3}$ and a $q$ of approximately $10^{-3}/3$ or $q \approx 0.0\dot{3}\%$. This q-value falls within a range for which we find that a substantial number of previously identified mobile mRNAs [98], [106], [107] were classified based on numbers of RNA-Seq reads over SNP positions that are in line with what would be expected from sequencing noise, Table 5.1. We point out that we used the same value of $q$ for each base in this analysis which does not take position-specific error rates into account. Nevertheless, this quick analysis suggests that a large proportion of the annotated mobile mRNAs, based on available RNA-Seq data, may lack statistical support.

### 5.2.3 Identifying mobility signals by employing Bayesian inference

By combining grafting with RNA-Seq we wish to detect mobile RNAs in plants to learn which mRNAs are travelling from which parts of the plant to which organs. Previously published studies using this detection method have reported hundreds

or thousands of mRNAs. Often, we do not have many replicates for those experiments. If so, there were different ways to summarise the evidence from several replicates or experiments. In many cases, the unions of identified long-distance mobile mRNAs were reported as final numbers. In Figure 5.3 we have visualised the overlaps between several samples in a published mobile mRNA experiment. We want to highlight that the majority of mobile mRNAs in this experiment, with 5 pooled samples taken from 5 organs, have been reported in the root and rosette samples. There are 19 mobile mRNAs that have been detected in all 5 samples, among the union of 1606 reported mobile mRNAs. We have described the challenge earlier, that using grafting combined with RNA-Seq depends on the ability to correctly assign sequenced mRNAs to the genetic background from which they originated. The assignment is often based on the identification of SNPs between otherwise identical sequences. A major challenge is therefore to distinguish these SNPs from sequencing errors. The read-depth dependency, which is introduced by criteria based on absolute read counts, is one possible explanation for a previous observation that mRNA abundance and mobility are correlated [110]. The definition of mobile mRNA based on absolute read counts gives rise to what appears to be non-selective, abundance-dependent transport of hundreds or thousands of mRNAs. The 19 long-distance mobile mRNAs identified in all 5 samples in the experiment show high read counts in all samples, with 4 interesting exceptions, Table 5.2.

Mobile mRNA detection faces a signal-to-noise problem in the detection of long-distance mobile mRNAs in plants. For successful identification of mobile transcripts, we need to demonstrate which data can be explained by sequencing errors only, and which data require an explanation beyond that. To address this problem, we developed a Bayesian framework for distinguishing sequencing errors from signals for mobile RNAs. We include prior information about sequencing errors in the analysis, and identify transcripts where the data cannot be explained by errors only. For each gene (with SNPs between grafted genotypes) we can calculate a Bayes factor, informing about the evidence for mobility across all SNPs and all replicates in an experiment. In the following chapter, we derive and demonstrate how these Bayes factors can be computed analytically using RNA-Seq data over all the SNPs in an mRNA. We present simulations to evaluate the performance of the proposed framework and show how Bayes factors accurately identify graft-mobile transcripts. The comparison to other mobility criteria using simulated data shows how not taking the variability in read-depth, error rates, and multiple SNPs per transcript into account can lead to incorrect classification. Furthermore, we elaborate on the pitfalls of filtering for sequencing errors or focussing on single SNPs within an mRNA. Moreover, we provide experimental design suggestions for successful graft-mobile mRNA detection that emerged from our results.

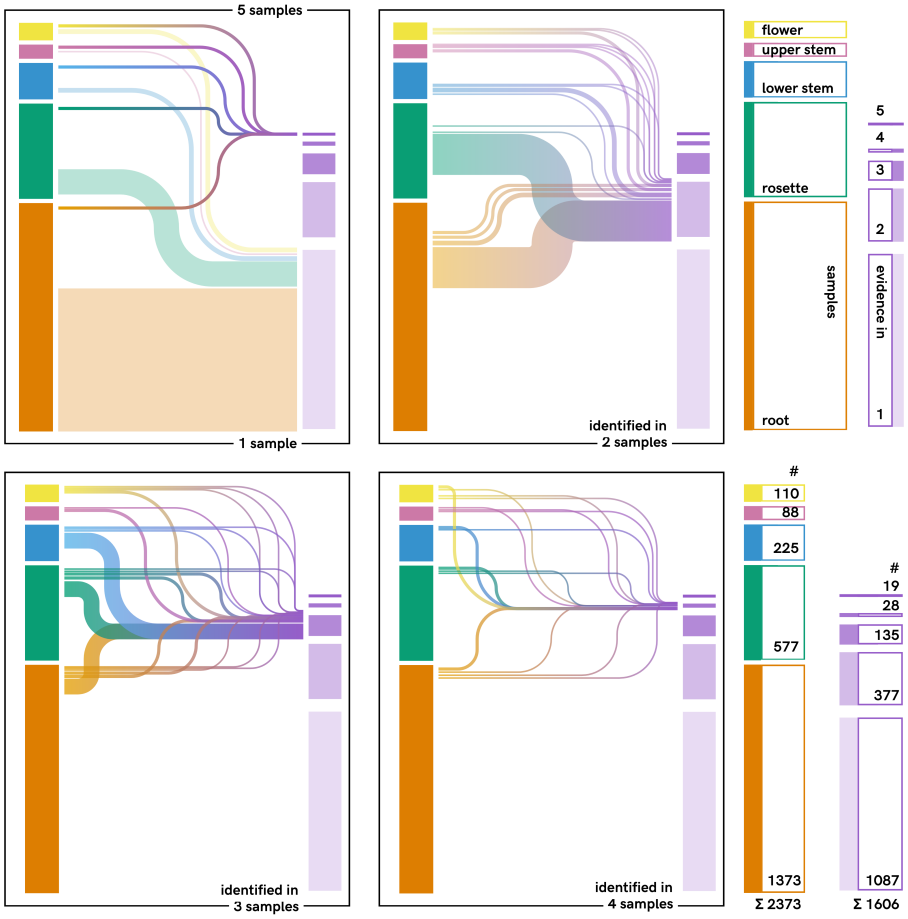



Figure 5.3: Many annotated long-distance mobile mRNAs are based on evidence from a single sample. Thieme et al. [98] have reported 1606 mobile mRNAs from an experiment on one plant with 5 samples from 5 organs (flower, upper stem, lower stem, rosette, root). In this figure we want to draw attention to the differences in evidences for mobility we have for those 1606. While there are 19 reported long-distance mobile mRNAs that have been found in all 5 samples, there are also 1373 reported cases among the 1606 that only have evidence from 1 sample. Furthermore, we notice large differences in the number of mobile mRNAs identified in different organs. This may have biological reasons, of course, but we cannot draw conclusions without more replicates. Top left we see the 19 long-distance mobile mRNAs that have been identified in 5 samples and the 1087 mobile mRNAs where there is only one sample providing evidence. Note that a large proportion of them are found in the root and rosette samples. Top right we see the 377 long-distance mobile mRNAs with evidence found in two samples. Again, a large proportion is coming from root and rosette. Bottom left and bottom right we see 135 and 28 mobile mRNAs detected in three and four organs, respectively.

| | | root | | rosette | | lower stem | | upper stem | | flower | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | # SNPs | # reads | # SNPs *** | # reads | # SNPs *** | # reads | # SNPs *** | # reads | # SNPs *** | # reads in flower | # SNPs *** |
| AT1G6593 | 16 | 63941 | 12 | 69729 | 11 | 38925 | 6 | 36809 | 1 | 34298 | 1 |
| AT1G7808 | 6 | 16392 | 4 | 12097 | 4 | 12974 | 2 | 18598 | 1 | 18848 | 1 |
| AT1G3571 | 78 | 1626 | 52 | 228413 | 6 | 94857 | 2 | 51279 | 1 | 42720 | 2 |
| AT1G0848 | 12 | 14937 | 7 | 11315 | 1 | 8215 | 3 | 11692 | 1 | 14378 | 1 |
| AT1G7447 | 14 | 7823 | 12 | 68819 | 3 | 67888 | 2 | 59250 | 1 | 56665 | 1 |
| AT1G3292 | 7 | 8925 | 3 | 40127 | 3 | 21694 | 1 | 31143 | 1 | 20700 | 1 |
| AT2G3797 | 13 | 85523 | 11 | 79669 | 11 | 44220 | 11 | 59078 | 3 | 45160 | 2 |
| AT3G1624 | 17 | 31525 | 16 | 23177 | 4 | 83987 | 3 | 62408 | 1 | 88527 | 1 |
| AT3G1320 | 8 | 18608 | 5 | 9943 | 3 | 8017 | 2 | 9786 | 1 | 11959 | 1 |
| AT3G1086 | 3 | 10223 | 1 | 9294 | 2 | 7931 | 1 | 11417 | 1 | 13280 | 1 |
| AT3G5580 | 18 | 95 | 18 | 120324 | 5 | 90575 | 1 | 80546 | 4 | 68522 | 1 |
| AT4G3096 | 10 | 22504 | 10 | 31149 | 5 | 24243 | 5 | 47026 | 1 | 53755 | 2 |
| AT4G0932 | 5 | 35932 | 5 | 45543 | 5 | 47391 | 5 | 46526 | 2 | 47397 | 1 |
| AT4G1619 | 16 | 45752 | 15 | 139373 | 10 | 114015 | 7 | 135871 | 2 | 96703 | 2 |
| AT4G1034 | 8 | 221 | 8 | 78537 | 1 | 78207 | 3 | 78764 | 1 | 77224 | 3 |
| AT5G5477 | 15 | 14358 | 12 | 113914 | 3 | 112624 | 3 | 113051 | 1 | 111471 | 1 |
| AT5G0153 | 13 | 2636 | 12 | 118139 | 3 | 118629 | 3 | 109167 | 3 | 93215 | 2 |
| AT5G2785 | 13 | 81103 | 12 | 71397 | 13 | 70685 | 4 | 68310 | 1 | 98339 | 1 |
| AT5G0981 | 16 | 65197 | 16 | 86884 | 15 | 95512 | 8 | 97355 | 2 | 111496 | 3 |

Table 5.2: A set of 19 consistently identified transcripts, that show high read numbers and could therefore be in line with sequencing errors. In Figure 5.3 we have seen that in one experiment, 1606 mobile mRNAs were reported across pooled samples from 5 organs (root, rosette, lower stem, upper stem, and flower). Of those 1606 reported long-distance mobile mRNAs, 19 have been detected in all 5 samples. In this table we take a closer look at this set of 19 genes and the number of reads mapping to them (all 4 alleles included) in different samples (root, rosette, lower stem, upper stem, flower). We show the numbers of SNPs for each transcript (# SNPs) that have been taken into account to evaluate evidence for mobility in the data and how many of those SNP-positions show evidence for mobility of the transcript, the presence of the alternate allele (# SNPs ***). Note, that all of the genes in the set have at least 7000 reads mapping to them, except 4 examples in the root sample. These 4 genes are not known to be expressed in the root, puzzlingly [129]–[132]. Referring back to Figure 5.2, we were wondering how many of those could be explained by sequencing errors?

Chapter 5

# Chapter 6

# Bayesian inference in the Detection of mobile mRNA in Plants

Transparency Note — The work presented in this chapter is a collaborative effort of Melissa Tomkins, Saurabh Gupta, Federico Apelt, Julia Kehr, Friedrich Kragler, Richard J. Morris and myself. The described framework – how to use Bayes factors in the detection of mobile mRNAs – has been published in 2022 [64]. A software package to carry out this analysis has been published in 2023 [65].

Summary — In Chapter 5 we learned about the long-distance mobility of mRNAs and that one way to investigate this phenomenon is combining grafting of plants with RNA-Sequencing. Because the translocation of mRNAs is thought to happen in very low numbers, sequencing errors present a major issue in distinguishing signals (SNPs from different genotypes) from noise in the analysis. To tackle this problem we show in this chapter how we can employ Bayesian inference to significantly improve detection accuracy by incorporating prior information about expected sequencing error rates into the analysis. We demonstrate how Bayes factors can be computed analytically using sequencing read counts (from processed RNA-Seq data) over all SNPs in an mRNA. We create simulated data to evaluate the performance of the proposed framework, and to compare Bayes factors with other analyses. We find excellent accuracy improvements over previously applied mobility criteria. Our results uncover experimental design suggestions (mainly concerning read-depth) for improved graft-mobile mRNA detection and show the pitfalls of filtering for sequencing errors or focusing on single SNPs within an mRNA.

## 6.1 Method description

### 6.1.1 Introducing Bayes factors in the detection of mobile mRNA using grafting followed by RNA-Seq

In a grafting experiment, two genotypes of a species (ecotypes, accessions, or cultivars) are combined to create a heterograft, see Figure 6.1 for a reminder of the setup

and an overview of the new analysis workflow. We will refer to the vegetative material that is grafted onto another plant as the scion, and the plant onto which the scion is grafted as the stock. If a transcript that originates from the stock is found in scion tissue or a transcript from the scion is found in stock tissue, then a plausible interpretation is that it has moved over the graft junction. This inference depends on the ability to distinguish between mRNA molecules from the two grafted plants.

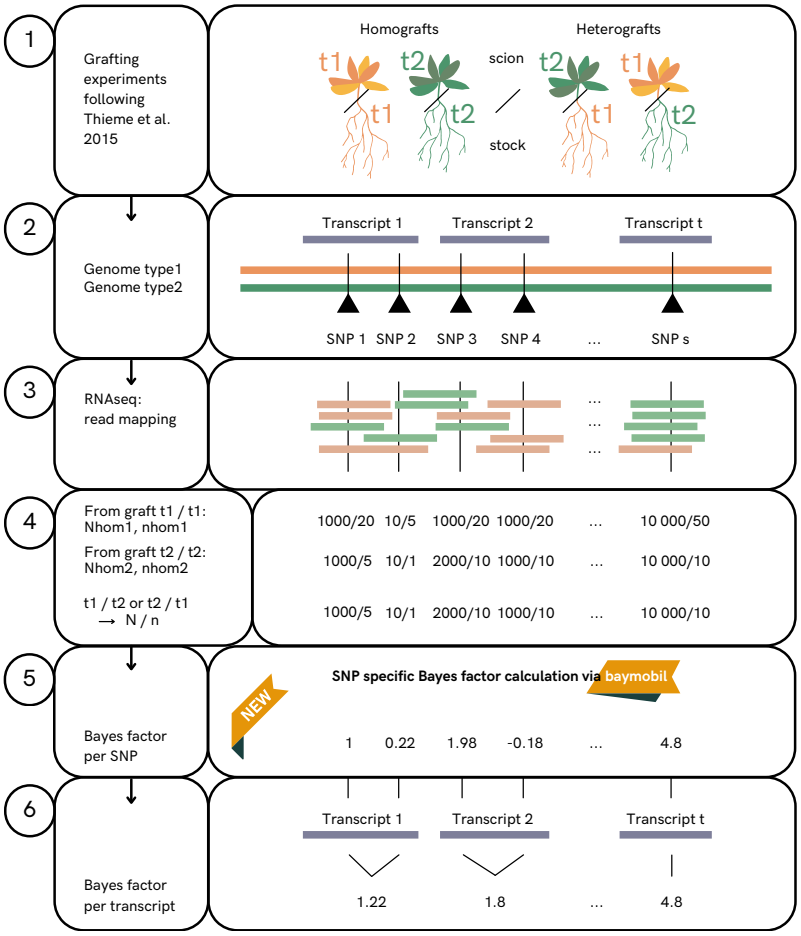



Figure 6.1: Workflow overview for the detection of graft-mobile mRNAs: Homo- and hetero-grafting experiments are conducted between 2 types (t1/t1, t2/t2, t1/t2, t2/t1) and RNA-Seq is performed on tissue samples. RNA-Seq reads are mapped to the genomes of type 1 and type 2. At all SNP positions $s$ between the types t1 and t2, the aligned reads are counted (for homografts $Nhom$, for heterografts $N$) and according to the type specific nucleotide at the SNP position assigned to one type (resulting in some $nhom$ in $Nhom$ and likewise $n$ in $N$). These numbers are input for the probabilistic framework and can be used for calculating a Bayes factor for each SNP. Afterwards, we can combine the SNP-level Bayes factors to a transcript-level Bayes factor (which gives us better accuracy).

Homologous genes from closely related organisms often have highly similar sequences. Sequences that are identical cannot be distinguished. We assume that

distinguishable homologous mRNAs from the two grafted plants with different genetic backgrounds (genotypes) differ in a number of SNPs. Other differences, such as insertions or deletions, are possible but not considered here.

Our objective is to establish a framework for determining the genotype from which transcripts have originated. When sampling tissue from one of the grafted genotypes (local), we may find SNPs associated with the other genotype (distal). The challenge lies in assessing whether the RNA-Seq reads for these positions, which could indicate transcripts that have crossed the graft junction, can be explained by expected biological and technical variation in the local genotype.

Here, we construct a probabilistic description of how to distinguish sequencing errors from graft-mobile transcripts using a Bayesian formalism for which we derive exact solutions. We leverage Bayesian inference to evaluate the evidence for a transcript being graft-mobile over the evidence for the data being consistent with sequencing or processing errors. The analysis takes into account read depths, SNP-specific error rates, all replicates, and multiple SNPs per transcript. It assumes that the SNP positions have been identified from the comparison of high-quality genomes or transcriptome data, and it focuses on sequencing errors at those positions without accounting for uncertainty in those positions.

The proposed Bayesian approach is evaluated using simulated data, demonstrating its ability to accurately identify graft-mobile transcripts and outperform other methods for classifying transcripts based on RNA-Seq reads. Based on true and false positives rates, we find that filtering SNPs with measurable errors, and not taking error rates or read depths adequately into account have significant detrimental consequences on the classification performance. The proposed Bayesian approach overcomes these issues and with controlled simulated data we can accurately identify graft-mobile transcripts. Additionally, the analysis using simulated data shows that combining information from multiple SNP positions increases classification accuracy, achieving near-perfect classification accuracy over a wide range of parameters.

### 6.1.2 Problem definition

In the following, a genotype can be either a species, an accession, an ecotype or a cultivar of a plant. We denote the two types used for grafting as t1 and t2. Different graft combinations may result in homografts, i.e. stock:scion is t1:t1 or t2:t2, and heterografts, i.e. stock:scion is t1:t2 or t2:t1, see Figure 6.1.

Our objective is to distinguish homologous transcripts from the grafted plants that exhibit variations in nucleotide positions (SNPs). By conducting RNA-Seq analysis on samples from both the scion and stock of grafted plants and aligning the reads to reference genomes from genotype 1 and genotype 2, we anticipate finding read counts for the presence of the corresponding alleles of t1 and/or t2 for each SNP position. Thieme *et al.* [98] have pioneered this experimental setup and offer details on original grafting and RNA-Seq protocols.

For each SNP, we represent the total number of reads mapping as $N$ and the total number of reads containing SNPs that correspond to a genotype different from the sampled tissue's genotype as $n$ (for brevity, we will refer to $n$ as the reads mapping to the *other* type). As an example, if we conduct RNA-Seq on tissue obtained from a scion of genotype t1, the reads that align best to genotype t2 at a specific SNP will be represented by $n$. In the case of a sample taken from a homograft t1:t1, if $n$ reads align to genotype t2, these occurrences must be considered as 'errors' that occurred somewhere during the process.

Potential sources of error encompass biological variation, such as mutations or the presence of different splice variants, as well as technical issues during library preparation or PCR amplification steps, mapping errors to the reference genome and errors therein, and the limitations of bioinformatic tools and associated parameters. In Figure 6.2, we delineate the relationships between the numbers from processed RNA-Seq data and the actual reads from each genetic background. The pivotal question revolves around whether the data substantiate the presence of transcripts from the other genotype (e.g., from t2) in the tissue of the sampled genotype (e.g., t1), and the quantification of reads that can consequently be attributed as graft-mobile.

For the sake of simplifying the explanation, we denote the genetic background of the sampled tissue, the local genotype, always as '1' (where '1' could actually be either genotype t1 or t2). We assume that a transcript has a set of SNPs, $S = s$, for which the RNA-Seq analysis results in data in the form $D = D_s = N_s, n_s$, where $N_s$ is the total number of reads for SNP $s$, of which $n_s$ map to the other, non-sampled tissue of another genetic background, the distal genotype (denoted by '2').

We want to compare the evidence for two opposing hypotheses which we define as follows.

**Hypothesis $\mathbf{H_1}$** states the data can be explained by a statistical model with only one genetic background, i.e. RNA-Seq reads that appear to be from a second genetic background are consistent with sequencing errors, mapping errors, frequencies of somatic mutations, presence of splice variants and any other process that can be expected to introduce mis-assignments in the sampled tissue of genotype 1 $\rightarrow$ **the data are consistent with the expected biological and technical variance from only one genotype**.

**Hypothesis $\mathbf{H_2}$** states the data are best explained by a statistical model that includes transcripts from distal grafted tissue, i.e. there are RNA-Seq reads from transcripts from a second genetic background that are unlikely to have arisen from the expected variance of genotype 1 $\rightarrow$ **the data support the presence of RNA-Seq reads from two genotypes and the transcript potentially being graft-mobile**.

Hypothesis 1 thus posits that RNA-Seq reads from only one genetic background are sufficient to explain the data, whereas hypothesis 2 requires that two genetic backgrounds are present. If statistical evidence for two genetic backgrounds is found in the data then a plausible inference is that transcripts have moved across

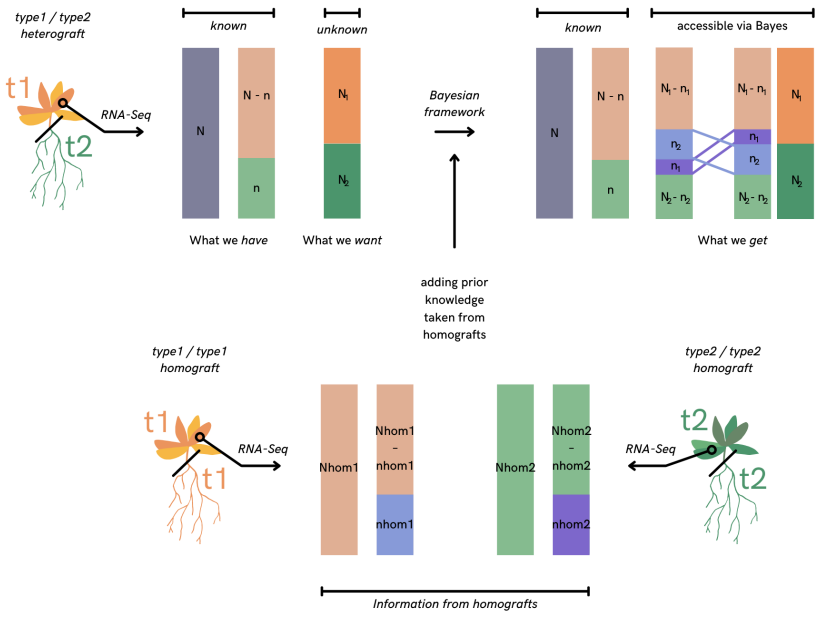



Figure 6.2: Relationships between measured and inferred RNA-Seq read numbers: From bioinformatics pipelines we can retrieve numbers of reads that map to certain positions: $N$ and $n$. $N$ reads map at the position of SNP $s$, $n$ of these carry the allele of the other type and seem to be from graft-mobile transcripts (See Figure 6.1). However, there is always an error of some kind. Only $N_1$ actually originate from type 1 and $N_2$ actually originate from type 2. Nevertheless, we do not know $N_1$, $N_2$. How can we get numbers for the actually translocated transcripts? We can build a probabilistic framework for occurring errors by looking at prior knowledge [64] we have: the error rates in the homograft data (Why we can do this is discussed in [64]). In this way we can infer numbers for the errors $n_1$ and $n_2$ (from $Nhom$ and $nhom$ values) and, therefore, also the reads from the graft-mobile transcripts $N_1$ and $N_2$.

the graft junction.

We can compare hypotheses using the posterior odds ratio [1], [5],

$$\frac{P(H_2|D)}{P(H_1|D)} \ . \tag{6.1}$$

When both hypotheses, $H_1$ and $H_2$, are equally likely *a priori*, $P(H_1) = P(H_2)$, the posterior odds ratio becomes equal to the marginal likelihood ratio (Bayes factor),

$$BF \ = \ \frac{P(D|H_2)}{P(D|H_1)} \ , \tag{6.2}$$

where $P(D|H_1)$ and $P(D|H_2)$ are the marginal likelihoods, also known as evidences.

If the ratio of graft-mobile transcripts to non-graft-mobile transcripts were known then this information could be used as the prior ratio for $P(H_2)$ over $P(H_1)$. Here,

we will assume a 1:1 ratio and use Bayes factors. Following Jaynes [5], we use the logarithm (of base 10) of the Bayes factor. A Bayes factor $\log BF$ of 0 means there is an equal probability for both hypotheses, whereas $\log BF > 0$ favours hypothesis 2 (graft-mobile) and $\log BF < 0$ favours hypothesis 1 (errors), remember Figure 1.12 in Chapter 1 about how to interpret Bayes factors. The higher the $\log_{10}$ Bayes factors in the detection of mobile mRNAs, the more evidence we find for the mobility of mRNA. The Bayes factor is the ratio between the evidences supporting two different hypotheses: the data are consistent with expected sequencing and mapping errors vs. the data support the transcript being graft-mobile. Assigning each mRNA a $\log_{10}$ Bayes factor enables ranking them according to the confidence of mobility given the RNA-Seq data. Furthermore, the Bayes factors give quantitative information about the ratio between translocated and local transcripts in the RNA-Seq sample taken from the heterograft. A high Bayes factor, therefore, means high confidence in the ability for this mRNA to be mobile plus high relative numbers of translocated individual transcripts. If wanted or necessary, Bayes factors can be used as criteria for a binary classification of mRNAs in mobile vs. non-mobile. We recommend following published literature [5], [7] and choosing a cut-off of $\pm 1$ ($\log_{10}$ Bayes factor $\leq -1$ non-mobile, $\log_{10} BF \geq 1$: mobile), rather than data-specific ranges.

### 6.1.3 The statistical evidence for a transcript being graft-mobile can be computed from the contributions from its SNPs

To compute the evidence, $P(D|H_1)$, we need to integrate over all parameters associated with $H_1$, $p'$. We assume these parameters can be specific to each SNP $s$ in a transcript,

$$p' = (p'_1, \ldots, p'_{|S|}), \tag{6.3}$$

where $|S|$ is the cardinal number of $S$, i.e. the number of SNPs in the transcript. Assuming the parameters are independent between SNPs, the likelihood can be expressed as the product of the likelihoods for each SNP,

$$P(D|H_1, p') = \prod_{s \in S} P(D_s|H_1, p'_s). \tag{6.4}$$

The evidence for $H_1$ over all data related to the transcript can thus also be expressed as a product,

$$P(D|H_1) = \prod_{s \in S} P(D_s|H_1), \tag{6.5}$$

where

$$P(D_s|H_1) = \int P(D_s|H_1, p'_{s,1}) \times P(p_{s,1}|H_1)\ dp'_{s,1} \tag{6.6}$$

is the evidence for $H_1$ for data $D_s$ associated with SNP $s$.

Analogously we can derive equations for $H_2$, whereby there may be a different number of parameters associated with each SNP for $H_2$,

$$p'' = (p''_1, \ldots, p''_{|S|})\ , \tag{6.7}$$

compared to $H_1$.

The parameters are explained in the following sections and the full derivation. Following the above procedure, we can express the evidence $P(D|H_2)$ as a product of SNP-based evidences,

$$P(D|H_2) = \prod_{s \in S} P(D_s|H_2). \tag{6.8}$$

The log Bayes factor of hypothesis $H_2$ (graft-mobile, i.e. there are reads from two genotypes) over $H_1$ ('errors', i.e. reads from only one genotype) can thus be written as the sum over log Bayes factors for each SNP $s$ within a transcript (as set of SNPs, $S$),

$$\log BF \;=\; \sum_{s \in S} \log \frac{P(D_s|H_2)}{P(D_s|H_1)} \;=\; \sum_{s \in S} \log BF_s \;. \tag{6.9}$$

This factorisation into contributions from each SNP allows us to focus on deriving equations for a single SNP. We then sum these contributions to obtain an overall log Bayes factor for a transcript being graft-mobile ($H_2$) or not ($H_1$).

### 6.1.4 SNP-specific evidence for $H_1$ and $H_2$ can be computed from the posterior distribution over $N_2$

RNA-Seq analysis of grafted plant tissue may result in some reads that align best to type 1 and others to type 2 at SNP positions. Just as above, we denote the total number of reads over a SNP by $N$ and the total number of reads associated with genotype 2 by $n$. If the data came from a homograft or a non-grafted plant, we can assume that $N - n$ reads were correct and $n$ reads have sequencing errors or were mismapped, i.e. there is a SNP-specific error rate that we can infer from the data. A suitable likelihood $\Lambda(\theta)$, for a two-outcome event is the binomial distribution, where $\theta$ is the error rate we wish to infer,

$$\Lambda(\theta) \;=\; \binom{N}{n} \, \theta^n \, (1-\theta)^{N-n} \;. \tag{6.10}$$

The conjugate prior of the binomial distribution is the Beta distribution, $\mathrm{Beta}(u_1, u_2)$,

$$P(\theta|u_1, u_2) \;=\; \frac{1}{B(u_1, u_2)} \, \theta^{u_1 - 1} \, (1-\theta)^{u_2 - 1} \;=\; \mathrm{Beta}(u_1, u_2) \;, \tag{6.11}$$

where $u_1$ and $u_2$ are hyper-parameters and the normalising factor, $B(u_1, u_2)$, is the Beta function,

$$\mathrm{B}(z_1, z_2) = \int_0^1 t^{z_1 - 1} (1-t)^{z_2 - 1} \, dt \;, \tag{6.12}$$

which results in the posterior having the same functional form, i.e. a Beta distribution,

$$P(\theta|D) = \mathrm{Beta}(u_1 + n, u_2 + N - n) = \mathrm{Beta}(\alpha, \beta) \;, \tag{6.13}$$

for which $u_1$ and $u_2$ are hyper-parameters of the prior and $\alpha$ and $\beta$ are the updated Beta distribution parameters of the posterior, see derivation in Section 6.1.7. The probability distribution over SNP-specific error rates can thus be described by a Beta distribution, Beta$(\alpha, \beta)$.

For a grafting experiment that involves plants with genotypes t1 and t2, we can infer the following Beta distribution parameters from the homograft data: ($\alpha_s^{\text{scion } t1}$, $\beta_s^{\text{scion t1}}$) and ($\alpha_s^{\text{stock } t1}$, $\beta_s^{\text{stock t1}}$) from $t1$:$t1$, and ($\alpha_s^{\text{scion } t2}$, $\beta_s^{\text{scion t2}}$) and ($\alpha_s^{\text{stock } t2}$, $\beta_s^{\text{stock t2}}$) from $t2$:$t2$ for each SNP position $s$.

Examples of inferred posterior distributions for two different priors, error rates and number of reads are shown in Figure 6.3. As expected, the choice of prior loses relevance with increasing read depth.

Replicates can be used to update the posterior distributions. The information content of a dataset remains the same regardless how it is subdivided or the order in which it is processed, and this is reflected in the mechanics of Bayesian updates [5]. Figure 6.4 depicts how information is combined (always updating to the latest state of knowledge) within the Bayesian framework described here. As expected this leads to the same results no matter how the data are split (e.g. into replicates).

### 6.1.5 The number of reads from transported transcripts can be inferred from heterograft data

As described above, we can focus on each SNP position in a transcript individually and then combine the contributions from all SNPs. Given $N$ total RNA-Seq reads from sampled tissue of genotype 1 of which $n$ contain SNPs associated with genotype 2, we want to know how many reads actually came from genotype 2, $N_2$, see Figure 6.2. Assuming that the biological and technical variation associated with RNA-Seq analysis will be similar between homografts and heterografts, we can use homograft data as a reference against which to evaluate the heterograft data. As described in the derivation in Section 6.1.7, we can derive the posterior distribution over $N_2$, $P(N_2|D)$, which can be expressed analytically as a summation of Beta functions that include $N$ and $n$ as parameters. From this posterior distribution, the expected ratio of reads from transported transcripts can be computed,

$$\langle N_2 \rangle = \sum_{N_2=0}^{N} N_2 \times P(N_2|D) \ , \tag{6.14}$$

and the ratio as $r_2 = \langle N_2 \rangle / N$ for each SNP $s$. For a transcript, the expected ratio is given by averaging over all SNPs. Figure 6.5 shows how Bayes factors change as a function of $n$ and how well $N_2$ can be estimated for different homograft read depths.

### 6.1.6 SNP-specific evidence for $H_1$ and $H_2$ can be computed from the posterior distribution over $N_2$

We now show how both $P(H_1|D)$ and $P(H_2|D)$ can be computed from the above described distribution $P(N_2|D)$. Under hypothesis $H_1$, any reads with SNPs as-

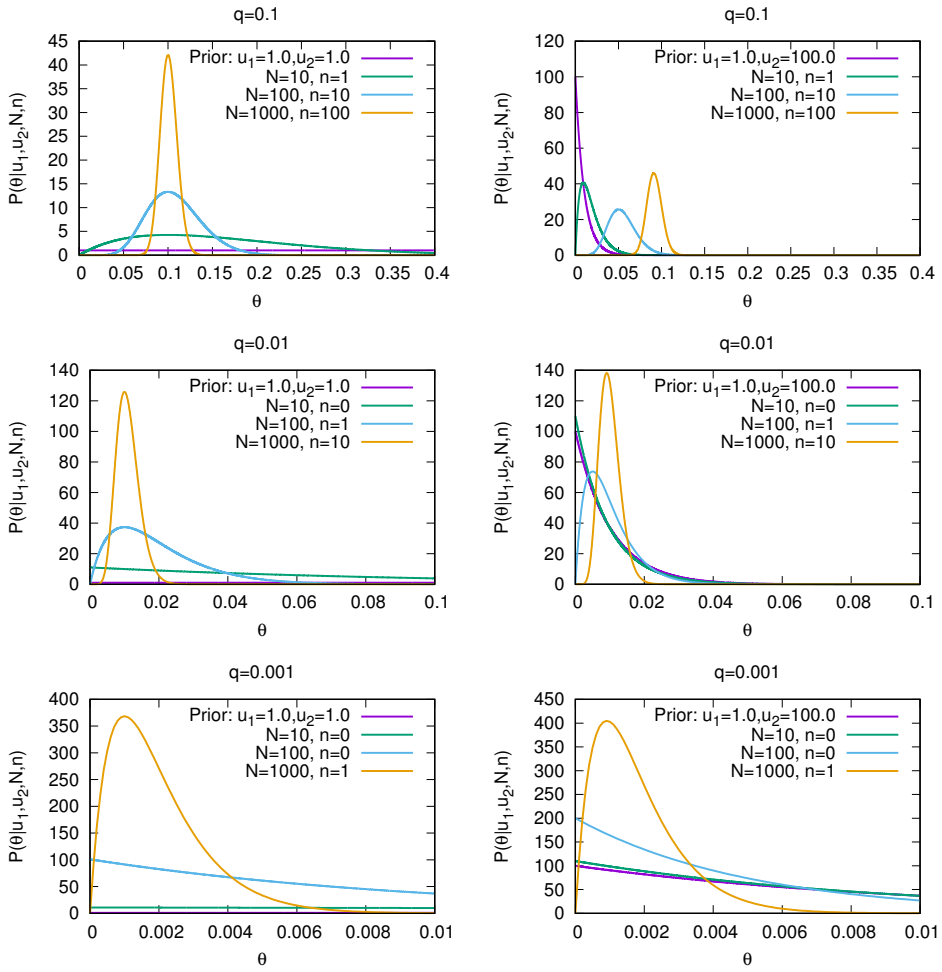



Figure 6.3: The choice of prior distribution becomes less relevant as the read depth increases. Each subplot shows the inferred posterior distribution for a different true error rate, $q$, and an example of number of reads generated with this error rate. The plots on the left start from a uniform prior ($u_1 = 1, u_2 = 1$), the plots on the right from a more extreme prior ($u_1 = 1, u_2 = 100$). With increasing read depths, the posterior distributions become more and more similar (yellow curves) for both priors. Note that the scales of $\theta$ vary between rows in order to show the interesting differences, likewise the scales of the y-axis are different.

sociated with the distal genotype must be considered errors, i.e. there are no transcripts from genotype 2 in the data, $N_2 = 0$. Under hypothesis $H_2$, there are reads from transcripts from genotype 2 and $N_2$ is a natural number. For $N_2 > 0$ there are $N$ possible values for $N_2$ and we can compare the evidence for each of them against $H_1$. The uniform prior distribution over $N_2$ (from 0 to $N$, see derivation in Section 6.1.7) ensures that each instance of $N_2 > 0$, which corresponds to $H_2$, will have the same prior probability as $N_2 = 0$ that corresponds to $H_1$, thus resulting in an equal prior for $H_1$ and $H_2$ and the posterior ratio being equal to the Bayes factor. The maximum log posterior odds ratio over all possible $N_2$ values for a transcript with $|S|$ SNPs can now be computed from $P(N_2|D)$,

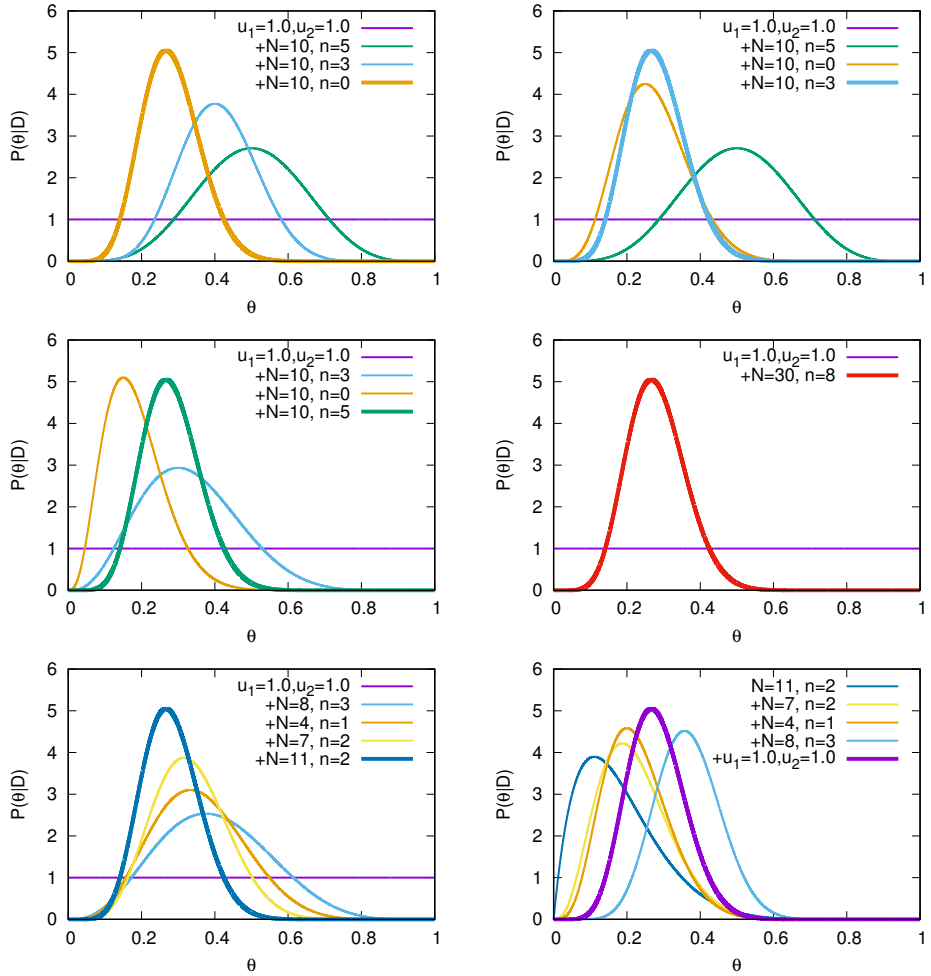



Figure 6.4: The posterior distribution over $\theta$ for the full dataset does not depend on how the data is grouped or the order of data. The curves correspond to the prior and the updated posterior after seeing different sets of data. The final posterior is shown as a thick line and is the same regardless of the order in which the data are processed. Furthermore, we can subgroup the data in whatever way we like or even treat one experiment as the prior and the prior as an experiment and we get the same result.

$$
\begin{aligned}
\log \frac{P(H_2|D)}{P(H_1|D)} &= \sum_{s\in S} \log \frac{\max\{P(N_2|D_s)\}_{N_2>0}}{P(N_2=0|D_s)} = \sum_{s\in S} \log BF_{s,21} = \\
&= \sum_{s\in S} \frac{\max\{P(D_s|N_2)\}_{N_2>0}}{P(D_s|N_2=0)} \times \log \frac{P(D_s|H_2)P(H_2)}{P(D_s|H_1)P(H_1)} = \\
&= \frac{\prod_{s\in S} P(D_s|H_2)}{\prod_{s\in S} P(D_s|H_1)} = \prod_{s\in S} \frac{P(D_s|H_2)}{P(D_s|H_1)} ,
\end{aligned}
\tag{6.15}
$$

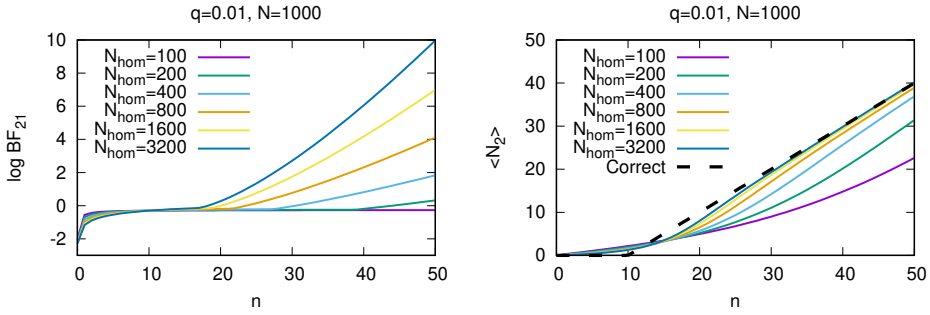



Figure 6.5: The read depth of the homograft data influences the Bayes factors and the numbers of inferred graft-mobile transcripts. The left plot shows how the Bayes factor changes as a function of the number of reads, $n$, that map to the other genotype for different homograft read depths. For an assumed error rate, $q$, of 0.01 and a total number of reads, $N$, of 1000, the expected number of mismatches would be 10. As the plot shows, the Bayes factor is negative for $n < qN = 10$ and moves closer to 0 as $n$ approaches the expected number of mismatches. The Bayes factor remains negative somewhat beyond the expected number of reads as these low numbers are still consistent with the inferred error distribution. As $n$ increases further, the Bayes factor becomes positive, favouring the hypothesis that reads may have originated from genotype 2. This change in Bayes factor is shown for different read depths from the homograft data. The right plot shows the behaviour of the inferred number of reads from genotype 2, $N_2$, as a function of $n$ and of the read depth from the homograft data, $N_{\text{hom}}$. With higher read depths from the homograft data and therefore more accurate inferences of the error probability distribution, the estimated number of reads from the genotype 2, $\langle N_2 \rangle$ approaches the correct value (black dashed line).

where max is over all positive $N_2$. Given that this approach has $P(H_1) = P(H_2)$, the above equation is equal to the sum of SNP-specific Bayes factors, $\sum_{s \in S} \log BF_{s,21}$, for each transcript. The expression for $P(N_2|D_s)$ is given in the derication in Section 6.1.7. The above equation can be used to assess whether a transcript is graft-mobile from RNA-Seq data and the results interpreted like we discussed in the Introduction, Figure 1.12.

### 6.1.7 Derivation of Bayes factors in the identification of mobile mRNAs

We assume that every nucleotide position, and thus every SNP, may have RNA-Seq errors associated with it, which we describe by an error rate $q$. This error rate may depend on the tissue that is being sampled, the genetic background, and may be different at different genomic locations. If we knew the error rate at each SNP, $q$, we could compute the probability of any number of reads being assigned to either of the two grafted genotypes, following the binomial distribution for a total number of reads at that SNP of $N$,

$$P(n|N,q) \quad = \quad \binom{N}{n} \, q^n \, (1-q)^{N-n} \, . \tag{6.16}$$

For instance, data from tissue of genotype A may consist of a total number of reads of $N$ of which $N-n$ best align to the reference genome of genotype A and $n$ align to reference genome of genotype B, i.e. the transcripts have SNPs that are associated with the grafted ecoptypes. The error rate, $q$, can be inferred from RNA-Seq data from various homograft combinations or non-grafted plants using Bayes' theorem,

$$P(\theta|D) \quad = \quad \frac{\Lambda(\theta)\ P(\theta)}{P(D)}\ , \tag{6.17}$$

where $\theta$ covers all possible values of the error rate $q$, and $D$ is the RNA-Seq read count data $(N, n)$ for the SNP under investigation in whichever tissue was sampled (genotype 1). Index 1 refers to the sampled genotype and index 2 refers to the other grafted genotype. To compute the posterior distribution over $\theta$, we need the likelihood, $\Lambda(\theta)$, and the prior, $P(\theta)$.

We can use the above binomial distribution to define the likelihood over $\theta$,

$$\Lambda(\theta) = P(D|\theta) = \binom{N}{n}\theta^n(1-\theta)^{N-n}\ . \tag{6.18}$$

A convenient analytical choice of prior for a binomial likelihood is its conjugate prior, the Beta distribution, due to it having the same functional form,

$$P(\theta_1|u_1, u_2) \ = \ \frac{1}{B(u_1, u_2)}\ \theta^{u_1-1}\ (1-\theta)^{u_2-1} \ = \ \text{Beta}(u_1, u_2)\ , \tag{6.19}$$

where $u_1$ and $u_2$ are hyper-parameters. The normalisation factor of the Beta distribution is known as the Beta function,

$$B(u_1, u_2) = \int_0^1 \theta^{u_1-1}(1-\theta)^{u_2-1}\ d\theta = \frac{\Gamma(u_1)\Gamma(u_2)}{\Gamma(u_1+u_2)}\ , \tag{6.20}$$

where $\Gamma(x)$ is the Gamma function.

The expectation value $\langle\theta\rangle$ for $\theta$ over $\text{Beta}(u_1, u_2)$ is

$$\langle\theta\rangle \quad = \quad \frac{u_1}{u_1+u_2}\ , \tag{6.21}$$

and the variance is given by

$$V(\theta) \quad = \quad \frac{u_1\ u_2}{(u_1+u_2)^2(u_1+u_2+1)}\ . \tag{6.22}$$

Using the binomial likelihood, we can infer the posterior distribution over the error rate from given RNA-Seq read data for each SNP, starting from a Beta prior $P(\theta) = \text{Beta}(u_1, u_2)$,

$$
\begin{aligned}
P(\theta|N,n) &= \frac{\Lambda(\theta)\ P(\theta)}{P(D)} = \\
&= \frac{\binom{N}{n}\ \theta^{n+u_1-1}\ (1-\theta)^{N-n+u_2-1}}{\binom{N}{n}\int_0^1 \theta^{n+u_1-1}\ (1-\theta)^{N-n+u_2-1}\ d\theta} = \\
&= \mathrm{Beta}(u_1+n, u_2+N-n)\ .
\end{aligned}
\tag{6.23}
$$

The mathematics of inference thus reduces to updating the Beta distribution with new count data. The expectation value for $\theta$ is updated from the prior expectation $\langle\theta\rangle$ 6.21 to the posterior expectation

$$
\langle\theta\rangle\ =\ \frac{u_1+n}{u_1+u_2+N}\ . \tag{6.24}
$$

For a non-informative, uniform prior ($u_1 = 1$, $u_2 = 1$), the best estimate of the error rate $q$, becomes

$$
\langle\theta\rangle\ =\ \frac{n+1}{N+2}\ , \tag{6.25}
$$

which is Laplace's rule of succession for estimating probabilities from a limited number of observations.

For heterograft data, we have $N$ total reads of which $N_1$ could be from genotype 1 and $N_2$ from genotype 2, with $N = N_1 + N_2$, Figure 6.2. We assume in the following that the RNA-Seq associated errors will be comparable between homografts and heterografts and between replicates.

The probability of observing $n_1$ errors from $N_1$ reads from genotype 1, given an error rate $q_1$ for genotype 1 is,

$$
P(N_1,n_1|q_1)\ =\ \binom{N_1}{n_1}\ q_1^{n_1}\ (1-q_1)^{N_1-n_1}\ , \tag{6.26}
$$

and the probability of observing $n_2$ errors from $N_2$ reads, given an error rate $q_2$ is,

$$
P(N_2,n_2|q_2)\ =\ \binom{N_2}{n_2}\ q_2^{n_2}\ (1-q_2)^{N_2-n_2}\ . \tag{6.27}
$$

The joint probability of observing $N_1, n_1, N_2, n_2$ is thus

$$
P(N_1,n_1,N_2,n_2|q_1,q_2)\ =\ \binom{N_1}{n_1}q_1^{n_1}(1-q_1)^{N_1-n_1} \times \binom{N_2}{n_2}q_2^{n_2}(1-q_2)^{N_2-n_2}\ . \tag{6.28}
$$

If $N_2$ and $n_2$ were known, we could compute the probability of observing $n$ being associated with genotype 2 out of $N$ total reads. We can use this to define a

likelihood function (replacing the known $q_1$ and $q_2$ by $\theta_1$ and $\theta_2$ that we wish to infer),

$$\begin{aligned} P(D|N_2, n_2, \theta_1, \theta_2) &= \binom{N - N_2}{n - N_2 + n_2} \theta_1^{n+N_2-n_2} \, (1-\theta_1)^{N-n+n_2} \times \\ & \qquad \binom{N_2}{n_2} \theta_2^{n_2} \, (1-\theta_2)^{N_2-n_2} \, . \end{aligned} \tag{6.29}$$

To compute $P(D|N_2)$ we also need to consider the dependencies on $n_2, \theta_1, \theta_2$. If we knew the posterior distribution $P(N_2, n_2, \theta_1, \theta_2|D)$, we could compute $P(N_2|D)$ by summation over $n_2$ (from 0 to $N_2$) and integration over $\theta_1$ and $\theta_2$ (from 0 to 1),

$$P(N_2|D) = \sum_{n_2=0}^{N_2} \int_0^1 \int_0^1 P(N_2, n_2, \theta_1, \theta_2|D) \, d\theta_1 \, d\theta_2 \, . \tag{6.30}$$

From Bayes's theorem we have

$$P(N_2, n_2, \theta_1, \theta_2|D) = \frac{P(D|N_2, n_2, \theta_1, \theta_2) \times P(N_2, n_2, \theta_1, \theta_2)}{P(D)} \, , \tag{6.31}$$

which leads to

$$\begin{aligned} P(N_2|D) &= \sum_{n_2=0}^{N_2} \int_0^1 \int_0^1 P(N_2, n_2, \theta_1, \theta_2|D) \, d\theta_1 \, d\theta_2 \; = \\ &= \sum_{n_2=0}^{N_2} \int_0^1 \int_0^1 \frac{P(D|N_2, n_2, \theta_1, \theta_2) \times P(N_2, n_2, \theta_1, \theta_2)}{P(D)} \, d\theta_1 \, d\theta_2 \, . \end{aligned} \tag{6.32}$$

We know that $N_2$ can vary between 0 and $N$ and $n_2$ can be at most $N_2$, so as a prior for $N_2$ we choose a uniform distribution over $(0, N)$ and for $n_2$ a uniform distribution over $(0, N_2)$. For $\theta_1$ and $\theta_2$ we take the distributions inferred from the homograft data, $P(\theta_1) = \text{Beta}(\alpha_1, \beta_1)$ and $P(\theta_2) = \text{Beta}(\alpha_2, \beta_2)$ as priors. This results in a joint prior of

$$P(N_2, n_2, \theta_1, \theta_2) = \frac{1}{N+1} \times \frac{1}{N_2+1} \times \text{Beta}(\alpha_1, \beta_1) \times \text{Beta}(\alpha_2, \beta_2) \, . \tag{6.33}$$

With this prior, the posterior distribution over $N_2$ becomes

$$\begin{aligned} P(N_2|D) &= \frac{1}{P(D)} \sum_{n_2=0}^{N_2} \int_0^1 \int_0^1 P(D|N_2, n_2, \theta_1, \theta_2) \times P(N_2, n_2, \theta_1, \theta_2)\ d\theta_1\ d\theta_2\ = \\ &= \frac{1}{P(D)(N+1)} \sum_{n_2=0}^{N_2} \frac{1}{N_2+1} \binom{N-N_2}{n-N_2+n_2} \binom{N_2}{n_2} \times \\ &\quad \int_0^1 \int_0^1 \theta_1^{n-N_2+n_2} \times (1-\theta_1)^{N-n+n_2} \theta_2^{n_2} (1-\theta_2)^{N_2-n_2} \times \\ &\quad \frac{\theta_1^{\alpha_1-1}(1-\theta_1)^{\beta_1-1}}{B(\alpha_1,\beta_1)} \times \frac{\theta_2^{\alpha_2-1}(1-\theta_2)^{\beta_2-1}}{B(\alpha_2,\beta_2)}\ d\theta_1\ d\theta_2\ = \\ &= \frac{1}{P(D)(N+1)B(\alpha_1,\beta_1)B(\alpha_2,\beta_2)} \sum_{n_2=0}^{N_2} \frac{1}{N_2+1} \binom{N-N_2}{n-N_2+n_2} \binom{N_2}{n_2} \times \\ &\quad B(n-N_2+n_2+\alpha_1, N-n+n_2+\beta_1)\ B(n_2+\alpha_2, N_2-n_2+\beta_2)\ . \end{aligned} \tag{6.34}$$

The normalisation constant $P(D)$, can be obtained by summing the above equation over $N_2$ and setting the result to 1. This leads to an expression for $P(D)$,

$$\begin{aligned} P(D) &= \frac{1}{(N+1)B(\alpha_1,\beta_1)B(\alpha_2,\beta_2)} \sum_{N_2=0}^{N} \sum_{n_2=0}^{N_2} \frac{1}{N_2+1} \binom{N-N_2}{n-N_2+n_2} \binom{N_2}{n_2} \times \\ &\quad B(n-N_2+n_2+\alpha_1, N-n+n_2+\beta_1)\ B(n_2+\alpha_2, N_2-n_2+\beta_2)\ . \end{aligned} \tag{6.35}$$

We now have an analytical expression for the posterior over $N_2$ with which the expectation value for $N_2$ can be computed,

$$\langle N_2 \rangle = \sum_{N_2=0}^{N} N_2 \times P(N_2|D)\ . \tag{6.36}$$

The ratio of the inferred number of reads from genotype 2 over the total number of reads, $r_2 = \langle N_2 \rangle / N$, for each SNP can be averaged over all SNPs to give the expectation value of the percentage of transported reads averaged over each transcript,

$$\langle r_2 \rangle = \sum_{s \in S} \frac{r_2}{|S|}\ . \tag{6.37}$$

The posterior probability $P(H_1|D)$ for Hypothesis $H_1$ and can be expressed as

$$P(H_1|D) = \frac{P(D|H_1)\ P(H_1)}{P(D)}\ , \tag{6.38}$$

and consequently we can find the posterior probability $P(H_2|D)$ for Hypothesis $H_2$ as

$$P(H_2|D) = \frac{P(D|H_2)\ P(H_2)}{/P(D)}\ . \tag{6.39}$$

The normalisation constant, or evidence $P(D)$, can be written as

$$P(D) = P(D|H_1)\ P(H_1)\ +\ P(D|H_2)\ P(H_2)\ . \tag{6.40}$$

Hypothesis 1 corresponds to $N_2 = 0$ and the probability of $H_1$ can be obtained by setting $N_2 = 0$ in $P(N_2|D)$ and renormalising accordingly,

$$P(H_1|D) \quad \propto \quad P(N_2 = 0|D)\ . \tag{6.41}$$

Hypothesis 2 corresponds to a value of $N_2 \neq 0$. Above, we assigned to every $N_2$ an equal prior probability of $1/(N+1)$, which also results in an equal prior for $H_1$ and $H_2$ for a defined $N_2$. Choosing the $N_2$ with the maximum posterior, $N_2^*$, will lead to highest posterior for $H_2$ amongst the available options,

$$P(H_2|D) \propto P(N_2 = N_2^*|D)\ , \tag{6.42}$$

and represents the most favourable choice for $N_2$ suggested by the available data. The proportionality factor is the same in both cases and results from the renormalisation due the change from $(N+1)$ hypotheses over $N_2$ to 2 hypotheses, $H_1$ and $H_2$. The log posterior odds ratio can now be expressed as

$$\log \frac{P(H_2|D)}{P(H_1|D)} \quad = \quad \log \frac{P(N_2 = N_2^*|D)}{P(N_2 = 0|D)} \quad = \quad \log \text{Bayes factor}\ , \tag{6.43}$$

which, as the priors are the same, is equal to the Bayes factor of hypothesis 2 over hypothesis 1 at each SNP $s$. The key terms are obtained from previous expressions,

$$P(N_2 = 0|D) \quad = \quad \frac{1}{P(D)(N+1)B(\alpha_1, \beta_1)} \binom{N}{n} B(n + \alpha_1, N - n + \beta_1) \tag{6.44}$$

and

$$\begin{aligned} P(N_2 = N_2^*|D) \quad = \quad & \frac{1}{P(D)(N+1)B(\alpha_1, \beta_1)B(\alpha_2, \beta_2)} \times \\ & \sum_{n_2=0}^{N_2^*} \frac{1}{N_2^* + 1} \binom{N - N_2^*}{n - N_2^* + n_2} \binom{N_2^*}{n_2} \times \\ & B(n - N_2^* + n_2 + \alpha_1, N - n + n_2 + \beta_1) \times \\ & B(n_2 + \alpha_2, N_2^* - n_2 + \beta_2)\ . \end{aligned} \tag{6.45}$$

## 6.2 Validation against labeled data

As most predicted graft-mobile mRNAs from RNA-Seq data have not been validated, there is uncertainty regarding their labelling, making an evaluation of the classification performance problematic. We circumvent this problem by creating datasets for which we know exactly which transcripts are assigned to be graft-mobile and which not (see Methods). The use of simulated data gives us full control over important parameters such as error rates and read depths while testing the accuracy of our method.

### 6.2.1 Error rates can be accurately inferred from read counts per SNP in homografts

A first question was how well we can infer the true error rate from simulated data (see Methods). Figure 6.6 shows the inferred error rate, $\langle\theta\rangle$, plotted against the true error rate $q$ for a range of $q$ from 0 to 1 and for read depth $N$ from 10 to 10,000. For low read depths ($N = 10$) the possible number of outcomes is small, leading to a visibly discretised set of inferred error rates with significant variation. For $N = 100$, the inferred error rates match already well to the true error rates. For $N = 1000$ and above, the estimates are accurately defined with high precision. This suggests that with suitably high RNA-Seq read depths (relative to the error rate), we can expect the inferred error rate per SNP to be a fair reflection of the true error rate.

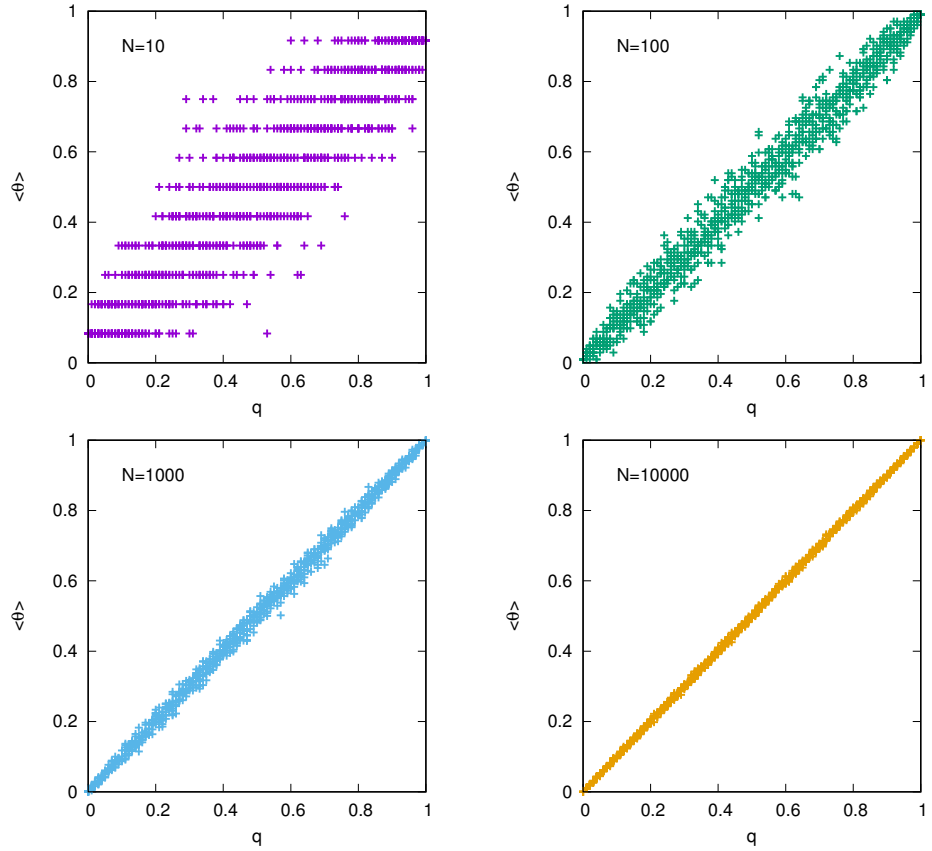



Figure 6.6: Error rate estimates from the inferred Beta distribution improve with higher read depth and accurately capture the known error rate in simulated data. Data were generated by randomly assigning reads to the other genotype, $n$, based on a defined error rate, $q$, for a total number of reads, $N$. The hyper-parameters were set to $u_1 = 1$ and $u_2 = 1$ (uniform distribution). The estimated error rate was computed from the expectation of $\theta$ over the posterior distribution, $\langle\theta\rangle = (u_1 + n)/(u_1 + u_2 + N)$. Ten datasets were generated for every value of $q$ in steps of 0.01 between 0 and 1 for different values of $N$.

### 6.2.2 Negative and positive controls are captured well by Bayes factors for individual SNPs

We compare hypotheses by evaluating the evidence for one hypothesis over another. Here, we use logarithm base 10 of the Bayes factor for hypothesis 2 over hypothesis 1, meaning that a value of zero arises when the data give no evidence either way, a negative value when the data is consistent with expected error rates and a positive value if graft-mobile transcripts are present, as discussed above and back in the Introduction (Figure 1.12).

To investigate whether the Bayes factor calculation would correctly assign a negative value for cases where there are only errors (no graft-mobile transcripts), we generated datasets (see Methods) for different error rates and different read depths. Data were generated that represent both homograft combinations to infer their posterior distributions for the error rates.

Separate data were generated from the same process with the same error rates for each SNP and treated as heterograft data. We found that the Bayes factors from most simulated datasets correctly favour hypothesis 1 that the observed data arose from a process with consistent error rates between homo- and heterograft data. However, as expected, there are several exceptions due to the stochastic nature of the simulations, Figure 6.6. The variability in Bayes factors depends on how well the read depth captures the underlying error rate, for instance an error rate of $q = 0.01$ will not be well represented by a read depth of 100 or less.

After confirming that the Bayes factors perform satisfactorily on negative controls, we next validated the approach on positive controls. We used simulated data for which additional reads from genotype 2 were added (graft-mobile transcripts), thus making them less consistent with the inferred homograft error rates. As anticipated, the closer the number of reads from genotype 2 is to the expected number of errors, the more challenging it is to distinguish graft-mobile transcripts from errors, Figure 6.7. If the error rate for a SNP is $q$, then we would expect to have on average $qN$ reads that are errors. The standard deviation of this value is $\sqrt{q(1-q)N}$. If we infer the error rate from the homograft data, we obtain a distribution over the inferred error rate, $\theta$, the expectation of which, $\langle\theta\rangle$, is our best estimate for $q$. The detection of reads from graft-mobile transcripts is therefore limited by the available data through the variation (precision) in the inferred error rate. We conclude that the Bayes factors perform well at the individual SNP level but that false assignments can be expected, in particular for low read depths that fail to represent the underlying error rates.

### 6.2.3 Combining the evidence across SNPs increases the accuracy of classification

A major advantage of the proposed framework is its ability to combine the evidence across multiple SNPs within a transcript. If the data from several SNPs of a transcript are incompatible with expected errors, then this enhances the evidence of the transcript being graft-mobile. Conversely, if the data from only one out of several SNPs within a transcript are found to deviate from expected errors, and data from the other SNPs are likely errors, then this could result in evidence

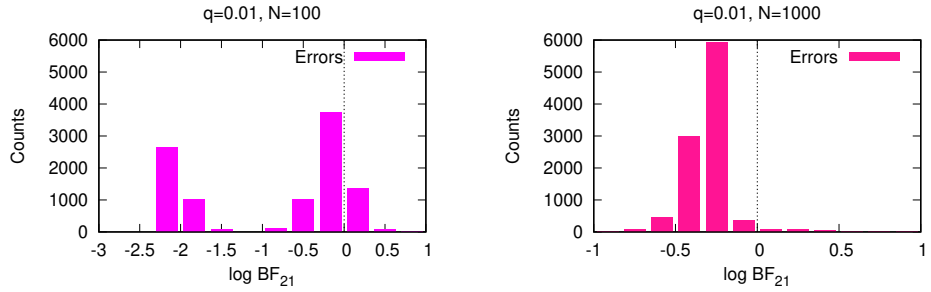



Figure 6.7: The Bayes factors for simulated data with a defined error rate become more tightly distributed around negative values when the read depth is sufficient to accurately estimate the error rate in the homograft data. Simulated homograft data following a binomial distribution for a given error rate, $q$, and read depth, $N$, were generated and compared against other such datasets that were treated as heterograft data (but with no graft-mobile transcripts). Here, $N$ is the same for homograft and heterograft datasets. The plots show the results from 10000 simulated datasets for two values of $N$. For $q = 0.01$ and $N = 100$ (left), the read depth in the homograft data is insufficient to accurately estimate the error rate, leading to erratic inferences for individual SNPs in the heterograft datasets. All SNPs are consistent with the underlying error rate (as they were generated stochastically from the same model), yet we find 1439 instances out of 10000 with $\log_1 0$ Bayes factor $> 0$ due to the broad inferred homograft error distributions. For $q = 0.01$ and $N = 1000$ (right), we get better estimates of the underlying error rates and a tighter distribution of Bayes factors. In this case, we get 204 instances out of 10000 with $\log_1 0$ Bayes factor $> 0$, i.e. (marginally) favouring the wrong hypothesis.

against the transcript being graft-mobile. We can demonstrate this effect by showing how the distribution of Bayes factors for mobile and non-mobile populations changes as we sum across all of the SNPs in a transcript, Figure 6.8 and Figure 6.9. Interestingly, for anything other than borderline cases of low read depth, this improvement saturates in the simulated data and little is to be gained beyond a SNP number of approximately 3, Figure 6.10.

## 6.3 Comparison to other methods

One key advantage of using Bayes factors is that we are not classifying transcripts *per se* but instead evaluating the evidence for them being graft-mobile, or not. The Bayes factors are thus used to rank our confidence in a transcript being graft-mobile given the data. The strength of the evidence can be assessed using well-established ranges [1], [5], [7], Figure 1.12 in the Introduction.

To compare the presented Bayesian approach with alternative criteria for defining mobile mRNA, we defined and implemented two approaches, Method A and Method B, inspired by previous publications [98], [117]. Method A determines a transcript as graft-mobile based on the number of reads mapping to genotype 2 being above a predefined threshold of 3, in two out of three replicates. Method B filters out SNPs with reads mapping to genotype 2 in data from genotype 1 homo-

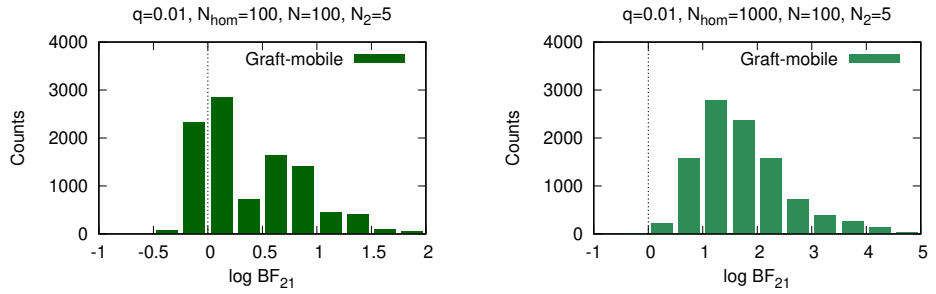



Figure 6.8: The Bayes factors for simulated data with added reads, $N_2$, from the other genotype (graft-mobile) are mostly positive but depend on how well the homograft error rates are defined. Simulated data from homograft data following a binomial distribution for a given error rate, $q$, and read depth, $N$, were generated and compared against other such datasets with added reads from the other genotype to represent heterograft data. Both plots show the results from 10000 stochastic simulations. For $q = 0.01$, $N_{\text{hom}} = 100$ and 5 added reads from the other genotype (left), there were several instances that marginally support the incorrect hypothesis ($\log_1 0$ Bayes factor $< 0$) but these vanish for higher read depths and therefore better defined error rate distributions (right). As shown in Figure 6.5, the more reads from the other genotype we have, the higher the Bayes factors.

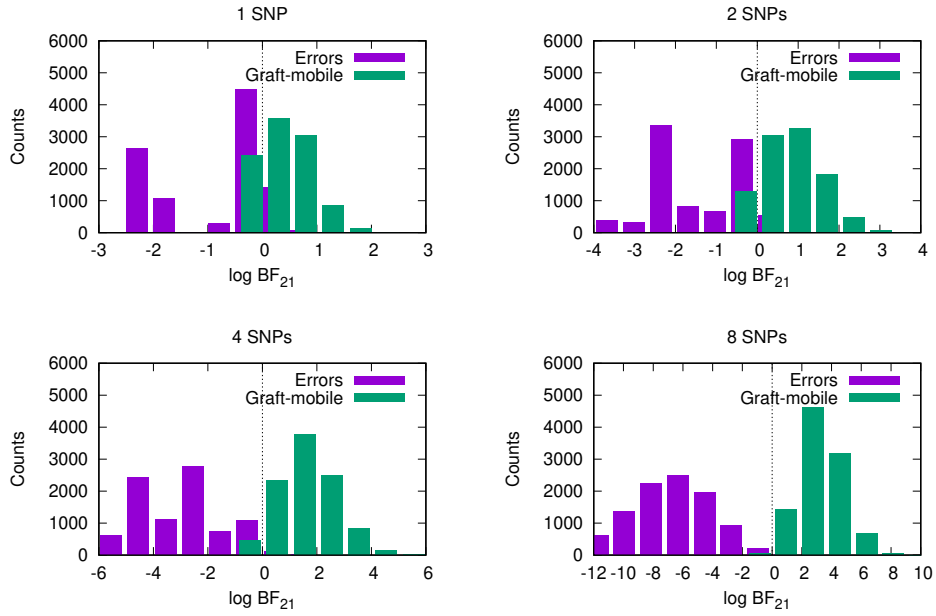



Figure 6.9: The inclusion of data from multiple SNPs per transcript enhances the distinction between likely errors and graft-mobile transcripts. Each plot shows the result of 10000 stochastic simulations with a different number of SNPs each with $q = 0.01$, $N_{\text{hom}} = 100$, $N = 100$, $N_2 = 5$.

grafts and then assigns SNPs as being graft-mobile when reads in the heterograft data are above 3 in two replicates [117]. So, Method B corresponds to Method A

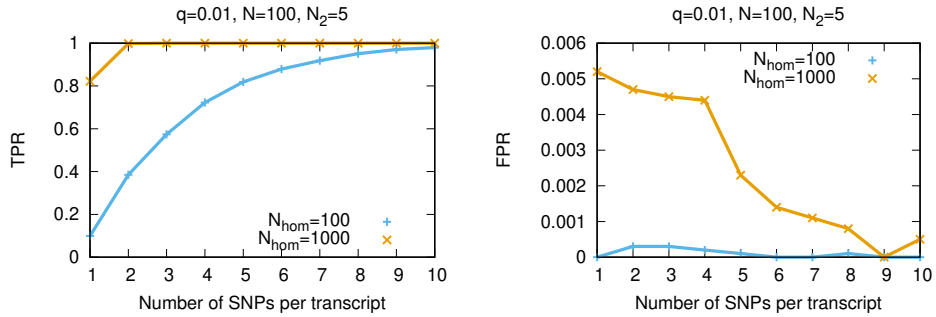



Figure 6.10: Combining data covering multiple SNPs per transcript enhances the classification accuracy. Each plot shows the result of 10000 stochastic simulations with a different number of SNPs each with $q = 0.01$, $N_{\text{hom}} = 100$, $N = 100$, $N_2 = 5$. We assign a transcript as being graft-mobile if $\log_{10}$ Bayes factor $\geq 1$ and say the data are consistent with errors otherwise. TPR is the true positive rate, FPR is the false positive rate (see Methods).

using only pre-filtered SNPs. Figure 6.11 show a comparison of the different methods. We find that as the read depth increases, more and more false positives occur using Method A, which leads to a drop in accuracy, Figure 6.11. This is because if every nucleotide has an error rate, then higher read depths will result in higher absolute numbers of errors. The conservative nature of Method B means that it has a low rate of false positives, but it is unable to detect graft-mobile transcripts when sequencing errors are present, as they always are, in the homograft data. Method B is thus unable to detect graft-mobile transcripts, unless the error rates are very low, Figure 6.11.

Confidently inferring a low error rate per SNP from RNA-Seq data requires a high read depth. Consequently, SNPs with both low read depths and low error rates were not detected in real RNA-Seq data [98], Figure 6.14, 6.15. The classification found in the simulated data is also observed using artificial data (see Methods) generated from published RNA-Seq data from Thieme *et al.* [98], Figure 6.12. Using a $\log_{10}$ Bayes factor $\geq 1$ (see Methods) to classify transcripts as graft-mobile and analysing the same dataset with the Bayesian method delivers both high TPRs and low FPRs, reflected in high accuracy (Figures 6.11, 6.12, and 6.13). The approximately equal read depths between the homograft data and the blended heterograft data reduces the difference in performance between the Bayesian approach and Methods A and B (Methods A and B don't take read-depths into account which leads to a higher mis-classification rate when differences are present). Consistent with the observations made above, the Bayesian approach shows an excellent TPR (and accuracy) for simulations with reads from genotype 2 that exceed the expected number of errors in genotype 1, Figure 6.13. The analysis of existing data in terms of sequencing depth and error rate per SNP, Figures 6.14 and 6.15, shows that experimental data falls into a parameter regime where mis-classifications of Method A and B can be expected. We conclude that the Bayes factors perform well and significantly better than our implementation of Method A and Method B.

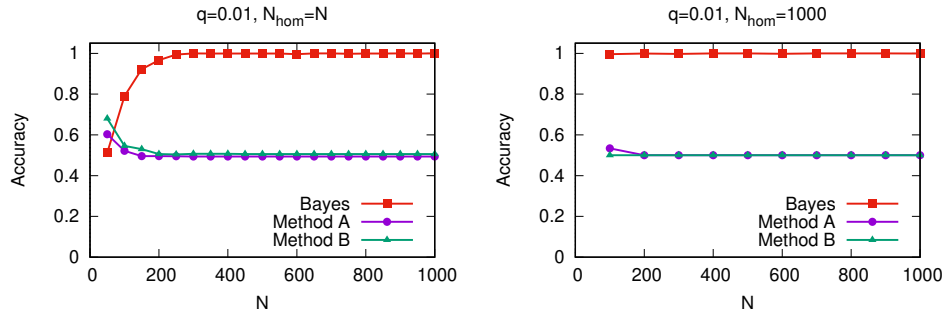



Figure 6.11: Filtering SNPs based on observed errors or determining mobile transcripts based on absolute read counts leads to poor classification accuracy for simulated data. Using our Bayesian approach (Bayes, red curves) we assign a transcript as being graft-mobile if $\log_{10}$ Bayes factor $\geq 1$. Methods A and B are explained in the main text. Here we have three replicates per transcript. The Bayesian approach sums the evidence over replicates whereas Methods A (magenta) and B (green) require that two out of three replicates show evidence for mobility. Both plots shows the accuracy, (TP + TN) / (TP + TN + FP + FN), of each method over 1000 simulated datasets for different read depths, $N$, for an error rate of $q = 0.01$. The left plot shows how the accuracy varies for different values of $N$ for homograft read depths equal to $N$. The right plot shows how the accuracy varies for different values of $N$ for fixed homograft read depths of 1000. The convergence of Method A and B towards ≈0.5 is a consequence of the balanced dataset with a 1:1 ratio of transcripts from one genotype (i.e. not graft-mobile) and two genotypes (i.e. graft-mobile), meaning that their performance is essentially little better than random for data similar to the simulated datasets. For an unbalanced dataset the accuracy could drop below 0.5 for Method A and B. An analogous analysis for true and false positive rates using data from published RNA-Seq studies is shown in Figure 6.12.

## 6.4 Methods

### 6.4.1 Dataset generation

We used two approaches for generating test data, simulation and blending. The first uses purely simulated data based on the underlying statistical model described below. The second mixes RNA-Seq data from existing homograft experiments to generate an artificial heterograft dataset. Neither approach is likely to capture the full variation and noise inherent in real datasets but have the advantage of having known labels that allow us to evaluate our method in a controlled manner.

### 6.4.2 Simulation of RNA-Seq data

Simulated datasets are generated based on a binomial distribution with an error rate $q$ for each SNP. A random number generator is used to provide stochasticity in line with the expected variance of a binomial distribution. For the homograft datasets, each SNP is assigned a number of reads ($N$) and a value for $q$. A range of values for $N$ and $q$ were used and are given in the individual figures. For each read, from 1 to $N$, a uniform random number is drawn and if this number is greater than $q$ then the read is assigned to genotype 1, otherwise to genotype 2. The generated reads per SNP thus represent a discrete realization of a stochastic process with a defined error rate, $q$, with $N$ reads assigned to genotype 1 and

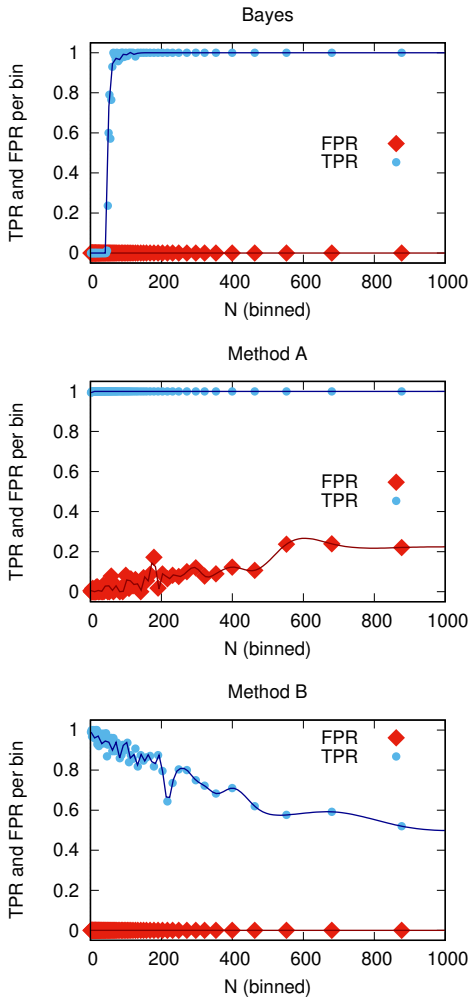



Figure 6.12: A comparison of methods on artificially generated heterograft data from real RNA-Seq data shows a similar trend in performance to the simulated datasets. The Bayesian approach (top), Method A (middle) and Method B (bottom) are evaluated in terms of their true (TPR) and false positive rates (FPR) for blended real data using a blend factor of $p = 0.1$ (see Methods). Only SNPs of comparable read depths between datasets were used and the data were binned for $N$ (see Methods). TPR and FPR are shown per SNP. Note that for similar read depths at each SNP location between homografts and heterografts, Method A and B make less mis-classifications than they would for different read depths. Without prior knowledge of the error rates, the Bayesian approach requires sufficient sequencing read depth to build an accurate error model.

$n$ reads assigned to genotype 2. The heterograft data are generated using the same process but with the addition of further reads, $N_2$, from genotype 2 that represent mobile transcripts. Different numbers of added mobile transcripts, $N_2$, are used to evaluate the sensitivity of the method. For measuring the classification accuracy (see below) we use balanced datasets throughout to not distort any of the performance metrics.

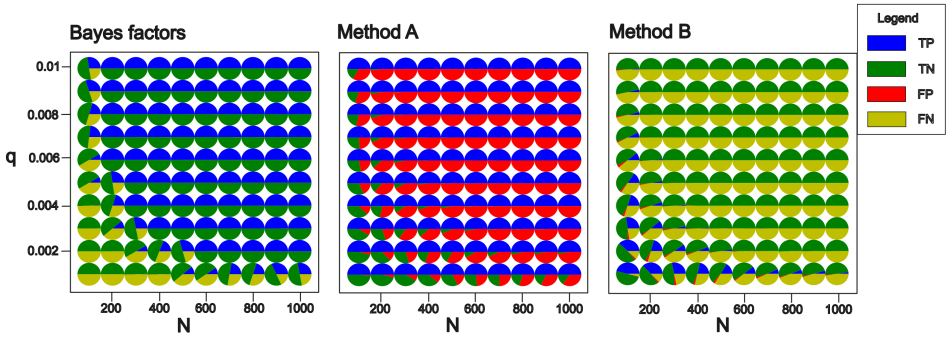



Figure 6.13: The classification performance of graft-mobile mRNAs depends on the error rates ($q$) and the read depths ($N$) of the RNA-Seq data. The simulation conditions are the same as in Figure 6.11. $N$ is the same homograft and heterograft data. The numbers of transcripts correctly labelled (TP and TN) based on Bayes factors (BF) increases with read depth (left). Higher read depths capture the error rates from the homograft data, resulting in clearer separation based on Bayes factors between hypothesis 2 (graft-mobile) and hypothesis 1 (errors). However, as sequencing errors in heterograft data are more likely to arise with higher read depths, we see a rise in false positives (transcripts with errors being incorrectly labelled as being graft-mobile, FP) for Method A (middle). Conversely, the increase in errors in the homograft data with read depth, leads to an increase in false negatives (graft-mobile transcripts being incorrectly labelled as non-graft-mobile, FN) in Method B (right). Therefore, both Methods A and B display decreasing classification performance with increasing read depth. Graft-mobile transcripts have been shown to be present in low numbers [106], therefore requiring high sequencing depths in order to detect them and necessitating methods able to distinguish between errors and graft-mobile transcripts.

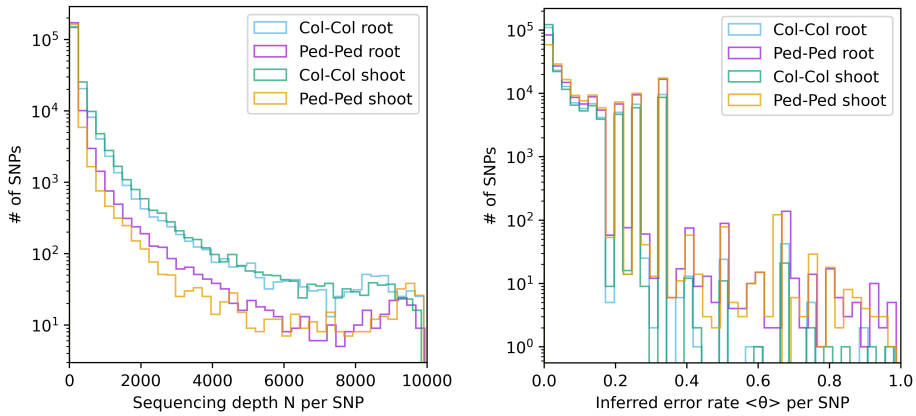



Figure 6.14: Sequencing read depth $N$ and inferred error rate $\langle\theta\rangle$ per SNP can vary significantly over a broad range. The distribution of the sequencing depths per SNP (left) and inferred error rates per SNP (right) from homograft RNA-Seq data from Thieme *et al.* [98]. The genotypes used in this experiment were Arabidopsis Col-0 and Ped-0. The best estimate of the error rate is $\langle\theta\rangle = (n+1)/(N+2)$, with $n$ the number of reads carrying the SNP of genotype of the non-sampled tissue and $N$ the sequencing depth, i.e. number of RNA-Seq reads mapping to a SNP position.

### 6.4.3 Blending real RNA-Seq data

Whilst the simulated data are inherently noisy, they follow an underlying statistical model that may not capture the variability inherent in experimental data. We therefore generated labelled data based on existing RNA-Seq datasets. We took RNA-Seq homograft data from the Arabidopsis Col-0 and Ped-0 accessions [98]. We then 'titrated' one dataset into the other to create a labelled heterograft dataset, following

$$(1-p) \times \text{Col} - 0 + p \times \text{Ped} - 0 \ , \tag{6.46}$$

where $p$ is the blending proportion. For instance p=0.1 would include 10% of the RNA-Seq reads for each SNP from Ped-0 and add them to the reduced $(1-p)$ reads in Col-0. If a SNP in the Ped-0 dataset had 100 reads assigned to the Ped-0 genotype then for p=0.1 we would add 10 of these reads to create a graft-mobile transcript. We selected SNPs that had a comparable read depth in Col-0 and Ped-0. Note that this approach creates data homo- and heterograft data of similar read depths, providing favourable conditions for Method A and B. Given the spread and limited data compared to the simulations we used adaptive binning to provide approximately equal-sized sets to evaluate the performance of the method as a function of read depth $N$.

### 6.4.4 Bayesian classification criterion

The Bayes factor denotes the evidence provided by the data that supports one hypothesis over another [1], [5]. We use these Bayes factors to rank our confidence in a transcript being graft-mobile, however, to evaluate our approach against existing methods that classify into mobile and non-mobile, we need a binary output. Following standard practice [5], [7], we use a value $\log_{10}$ Bayes factor $\geq 1$ to select hypothesis 2 over hypothesis 1. It is worth noting that we could also determine transcripts for which there is strong evidence for them not being graft-mobile (e.g. those with $\log_{10}$ Bayes factor $\leq -1$). This latter point is important for curating high-confidence datasets for training using positive and negative examples.

### 6.4.5 True and false positive rates

The classification performance is evaluated using accuracy, and true and false positive rates. Each mRNA is given one of four different labels: an mRNA that is correctly classified as graft-mobile is a true positive (TP); an mRNA incorrectly classified as graft-mobile is a false positive (FP); an mRNA correctly classified as non-graft-mobile is a true negative (TN); an mRNA that is incorrectly determined to be non-graft-mobile is a false negative (FN). From this, we can compute the true positive rate (TPR) as the proportion of the graft-mobile mRNAs that are classified correctly from all the graft-mobile mRNAs, (TP / (TP + FN)), and the false positive rate (FPR) is the proportion of the non-graft-mobile mRNAs that are incorrectly assigned out of all of the non-graft-mobile mRNAs (FP / (FP + TN)). The accuracy is calculated as (TP + TN) / (TP + TN + FP + FN). The accuracy is 0 if no classes are identified correctly (TP=0, TN=0) and 1.0 for error-free classification (FP=0, FN=0).

### 6.4.6 Availability of code, data and materials

A detailed tutorial with test data and results for the statistics and simulation package *baymobil* can be found on our Github. It is licensed under a MIT software license.

## 6.5 Discussion

Long-distance transport of mRNAs has been shown to be important for plant development, and recently also in mammalian systems [133]. A popular method for detecting transported mRNAs in plants involves grafting plants with different genetic backgrounds and using RNA-sequencing techniques to determine whether transcripts have traversed the graft junction. To do so robustly requires consideration of the associated sequencing and alignment errors that are known to occur.

Here, we have presented an approach for distinguishing between expected variation (e.g. biological variation, RNA-Seq errors, processing errors) in RNA-Seq pipelines and putative graft-mobile transcripts. We set up the problem as an hypothesis-testing framework founded in Bayesian inference for which we derived an exact analytical solution. A key, yet unverified, assumption inherent to the method and other approaches that compare to homograft data is that the biological and technical variance of the RNA-Seq analysis in homografts and heterografts will be similar.

For equal prior probabilities for hypotheses 1 and 2, the posterior probabilities ratios are equivalent to Bayes factors. Bayes factors are computed for each SNP position based on the associated read counts and then combined into a Bayes factor per transcript. Replicates can be handled analogously. Bayes factors allow transcripts to be ranked based on data supporting their graft-mobility. Higher Bayes factors are indicative of more translocated individual transcripts (the expectation value of which can be computed separately).

We show that RNA-Seq error rates can be accurately estimated using the presented methodology (Figure 6.6) and how the read depth influences the precision (Figure 5.2, Table 5.1, and Figures 6.5, 6.3, 6.6) and the inferred number of graft-mobile transcripts (Figure 6.5). Multiple SNPs and replicates can be readily accounted for by summing their contributions (Equation 6.9, Figures 6.10, 6.4, 6.8). We validated our approach extensively with simulations (Figures 6.9, 6.7, 6.8, 6.9). We used simulated RNA-Seq data to allow for accurate quantification of the performance of the method, avoiding the uncertainties inherent in previously published graft-mobile mRNA labels. An additional advantage is that the use of simulated RNA-Seq data allows us to evaluate the performance over a range of parameters. Further validation was carried out using datasets derived from published homograft data [98], Figure 6.12.

The evaluation of different mobile mRNA detection methods showed that approaches using absolute number thresholds of reads per SNP on selected positions lead to high rates of mis-classified transcripts, Figures 6.11, 6.12 and 6.13. Classifying graft-mobile transcripts based on absolute read numbers without consideration of error rates and read depths (Method A) leads to the assignment of mobility to many transcripts for which statistical support is not present (high false positive

rate).

On the other hand, filtering SNPs with errors in the homograft data (Method B), results in a reduction in the number of detected mobile transcripts (low true positive rate), Figures 6.11, 6.12 and 6.13. The Bayesian method relies on reference data for the estimation of error rates and the quality and quantity of this data influences its performance. Here, we used a non-informative uniform prior to evaluate the method but a more suitable choice of prior for the error rates based on existing experimental data and RNA-Seq error analyses [134] would help overcome the limitations arising from low read depths (Figure 6.3).

The approach could be extended in several ways. We have simplified the outcome at each SNP position to map to the two genetic backgrounds used for grafting, resulting in a binomial likelihood. Including errors for all four nucleotides could provide a better description of the overall nucleotide variability, leading to a replacement of the bionomial likelihood by a multinomial distribution. The conjugate prior of the multinomial distribution is the Dirichlet distribution, however, we have not explored how tractable this approach would be to solve analytically in full. A key assumption of the presented approach is that the error rates per SNP are comparable between experiments. This assumption will need to be checked for real data. If necessary the inferred error rate distributions may need to be adjusted to take large deviations into account (for instance, by increasing the spread by reducing the Beta function parameters). Another addition would be to include sequencing quality scores within the framework. These extensions will be the subject of future developments after the careful analysis of existing datasets to assess shortcomings in the presented developments.

To summarise, this contribution presents a Bayesian framework that takes account of read depths, error rates, replicates, and multiple SNPs per transcript, providing a powerful means for distinguishing graft-mobile mRNAs from RNA-Seq errors. As graft-mobile mRNAs are often rare compared to endogeneous mRNAs [106], RNA-Seq read depths needs to be sufficiently high [117] and chosen with care and every effort should be made to reduce the risk of contamination [89]. Detecting rare events can be statistically challenging. Error rates from either non-grafted plants or homografts provide useful reference values for the number of sequencing errors to expect for different genomic locations. The higher the read depth of the reference dataset, the better the error rates can be inferred.

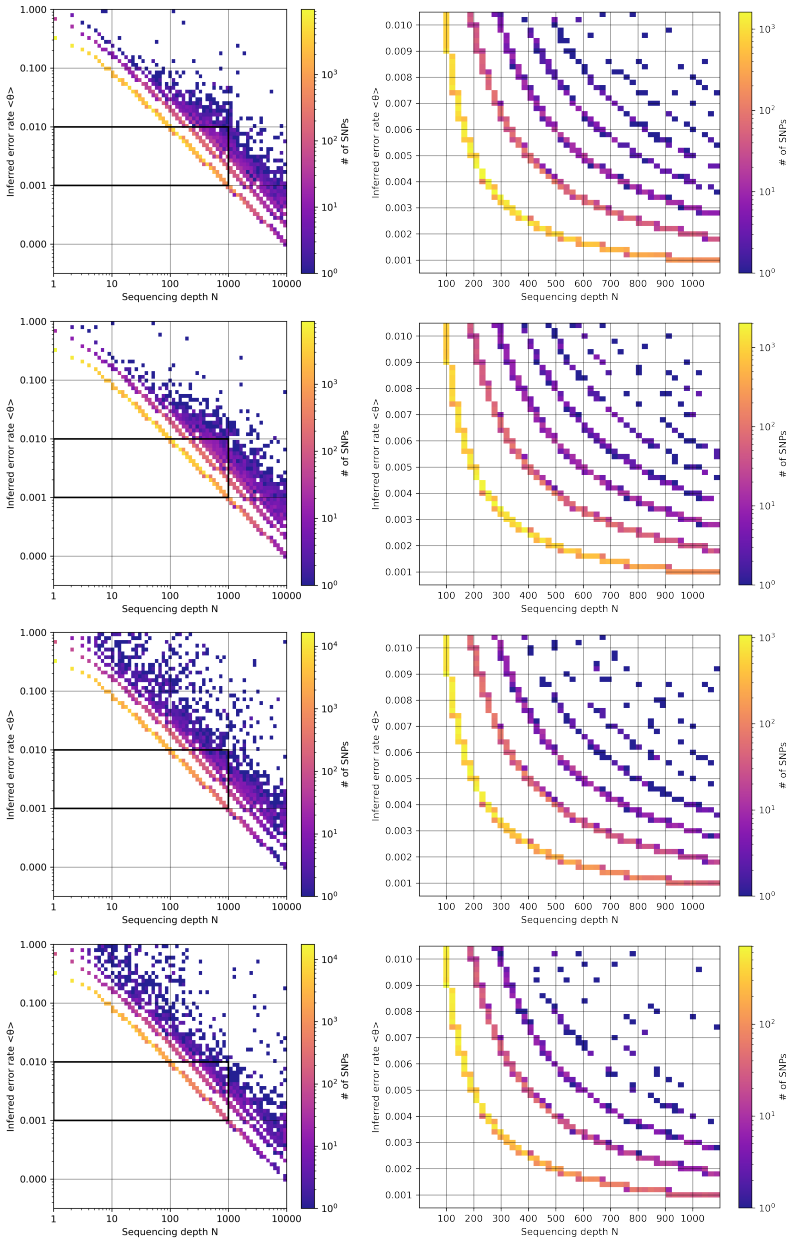



Figure 6.15: Real RNA-Seq data does not fall into parameter regimes that are favourable for Method A and B. Sequencing depths per SNP vs. inferred error rates per SNP from homograft data (4 samples from top to bottom: Col-Col root, Col-Col shoot, Ped-Ped root, and Ped-Ped shoot) from Thieme *et al.* [98]. The left hand side plots show the relationship across the whole dataset on log-log scales. In the right hand side plots we zoom into the marked window of the left plot and lose the log-log scale (compare to Figure 6.13). Each intersection in the grid marks one of the pie plots in Figure 6.13. $N$ is the number of reads mapping to a SNP position. The best estimate of the error rate $q$ is $\langle\theta\rangle = (n+1)/(N+2)$, with $n$ the number of reads carrying the SNP of the genotype of the non-sampled tissue, and $N$ number of all reads mapping to the SNP position. The colour indicates the number of SNPs for each pair of $\langle\theta\rangle$ and $N$.

# Chapter 7

# The adventurous endeavors of applying Bayes factors in the detection of mobile mRNAs on real RNA-Seq data

Transparency Note — The learnings uncovered during our efforts to apply Bayes factors on real RNA-Seq data to detect mobile mRNAs are documented in a re-analysis study carried out by Pirita Paajanen, Melissa Tomkins, Ruth Veevers, Michelle Heeney, Hannah Rae Thomas, Federico Apelt, Eleftheria Saplaoura, Saurabh Gupta, Margaret Frank, Dirk Walther, Christine Faulkner, Julia Kehr, Friedrich Kragler, Richard J. Morris and myself [63].

Summary — After demonstrating how we can improve the detection accuracy for mobile mRNA in RNA-Seq data using Bayesian inference we couldn't wait to see the improvements when applying it to real data. We started re-analysing several published data sets with the aim to establish a high-confidence dataset of mobile mRNAs. Our results suggest that previously reported variability between experiments can also be explained by a lack of detection accuracy, which has implications for existing mobile mRNA annotations and challenges current views on long-distance mRNA communication. We could confirm our concern about sequencing errors and detection criteria, but found further analysis problems that we couldn't resolve yet. Therefore, this chapter does not present the obvious results that we and the community had so eagerly awaited (Bayes factors for all genes, high-confidence mobile lists, ...), but we discuss why cannot present them. We hope that the uncovered issues are not only appreciated by the mobile mRNA community but also travel further into other areas in genomics. In particular studies using single nucleotide polymorphisms (SNPs) to detect variants may be affected by similar problems.

## 7.1 Proud and oblivious, we set off on our journey ...

In 2023, equipped with our new statistical analysis, we sought to establish a high-confidence dataset of mobile mRNAs. We had the goal of using this dataset for learning the determinants of mobility, and to develop and train predictive models. We planned on using 7 data sets for the meta-study: A study on the parasitic plant *Cuscuta pentagona* feeding on Arabidopsis [69], a study using *Arabidopsis thaliana* (grafting the ecotypes Col and Ped) [98], a grapevine study using *Vitis palmata, Vitis girdiana*, Vitis hybrid C3309, and *Vitis vinifera* cultivar Riesling [106], and studies using watermelon (*Citrullus lanatus*) [117], *Nicotiana bethamiana* [105], *Solanum lycopersicum* and *Nicotiana benthamiana* [120], and cucumber [107]. In the following sections of this thesis, we will summarise the relevant results for mobile mRNA detection using SNP-based approaches. In the paper we also inspected a second detection strategy, grafting two different species followed by RNA-Sequencing, so that the genome differences are bigger than SNPs only.

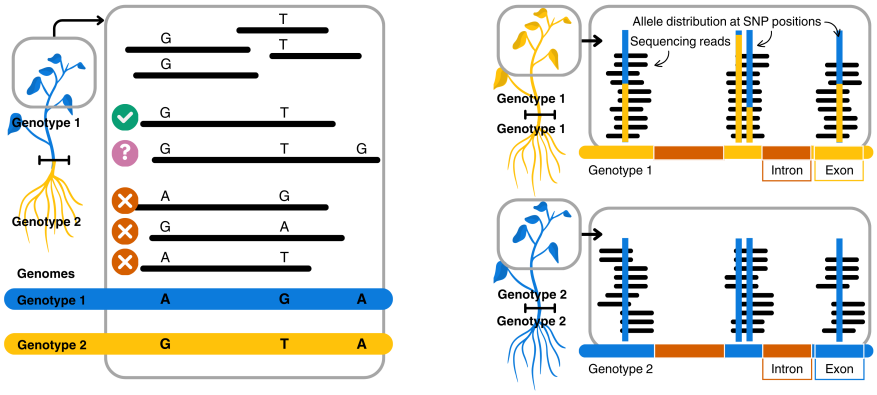
Our re-analysis efforts have implications for existing mobile mRNA annotations and challenge current views on the extent of long-distance mRNA communication. To this date, we have not yet found a way to robustly detect mobile mRNAs from RNA-Seq data, which makes us question whether this popular experimental setup is suitable for high-throughput identification of mobile mRNAs at this point in time.

Looking in detail into the data supporting the annotation revealed several confounding problems in current methods for detecting mobile mRNAs that have not been considered in prior analyses to our knowledge. The re-analysis results suggest that the reported variability between experiments [67] is partly a consequence of the criteria that have been employed to classify mRNAs as being mobile, and partly a bioinformatical data processing problem, of which parts cannot be solved with a noise filter only. We have shown earlier in Chapter 5 and 6 that criteria based on absolute read counts lead to a dependence on read-depth for simulated data [64]. In this later work, we could confirm this dependency for real data, but found more issues. We show that criteria based on mapping to genomes depend on genome assembly completeness and quality. Furthermore, we identify other sources of variation, such as gene copies, that can bias the analysis.

## 7.2 ... and all we found were rocks along the way

### 7.2.1 Issue 1: Can we find co-occurring SNPs acting as a positive control?

In the last chapter, we investigated the impact of sequencing noise at the individual SNP level. During the re-analysis we had the idea to search for the strongest pieces of evidence we can find for mRNA mobility. If SNPs are located closely together, a single RNA-Seq read may cover more than one SNP (we will refer to these as co-occurring SNPs). Requiring RNA-Seq reads to cover co-occurring SNPs that support the alternate allele results in a less pronounced read-depth dependence than single SNP criteria. We therefore investigated RNA-Seq reads that cover multiple SNPs, Figure 7.1 (a). We did find such reads in the data, and they were particularly easy to find in two samples from one, single replicate, experiment.




(a) (b)

Figure 7.1: Allelic differenes in multiple SNPs per read and the appearance of heterozygosity (in homozygous species) can be used to check the viability of SNPs and exclude potentially problematic transcripts from the analysis. (a) SNPs can be in close proximity, and therefore it can happen that several SNPs are recorded in the same RNA-Seq read. In this example, genotype 1 has three SNPs very close to each other: A, G and A (yellow bar). In genotype 2, we find G, T, A (magenta bar) in those positions. In this schematic example, reads from the shoot of Genotype 1 are mapped to Genotype 2. If all covered loci carry the allele of Genotype 2, we are observing evidence for the read being from Genotype 2 and the associated transcript being potentially mobile. On the other hand, if only one locus carries the allele of Genotype 2, the outcome is inconclusive. (b) *Arabidopsis thaliana* is a selfing species, so we expect homozygocity at all positions for all reads mapping to the genome at all positions. However, for duplicated genes (magenta) in Genotype 1, which may be single copy genes in Genotype 2 (yellow), short read sequencing and mapping to Genotype 2, can give rise to what appears to be heterozygocity. When there are two alleles present in the homograft data (magenta and yellow), we may be observing pseudo-heterozygocity. As the identification of such pseudo-SNPs depends on read-depth, it is important to exclude the full gene, rather than only the seemingly heterozygous SNPs. For deeper explanations and examples in the data, see full manuscript [63].

Currently, we are looking into quantifying the occurrences.

### 7.2.2 Issue 2: Are Pseudo-SNPs and Pseudo-heterozygosity affecting the data processing?

The examination of reads over multiple SNPs revealed another confounding factor in the identification of mobile mRNA; several loci showed apparent heterozygosity in the homograft data, Figure 7.1 (b). While the *Arabidopsis thaliana* ecotype *Columbia* (Col-0) assembly is of excellent quality, the genome assemblies of other Arabidopsis ecotypes are not all of the same quality. For instance, a recent study of the *Landsberg erecta* (Ler-1) ecotype [135] of *Arabidopsis thaliana* identified 105 single copy genes present only in Ler-1 but not in Col-0, and 334 single copy orthologs that had an additional copy in only one of the genome assemblies. It has

been estimated that 10% of the annotated genes in Arabidopsis have copy-number variation [135], [136]. Differences in gene copy numbers can lead to reads not mapping correctly, which gives rise to pseudo-SNPs and pseudo-heterozygosity [135], [136]. Whilst challenging to quantify, a high fraction of reads with variation at specific positions within a single genome can be indicative of pseudo-heterozygosity [135], [136]. Meaning, in addition to technological noise, there are also biological causes that could be interpreted as SNPs of an alternate allele. This is an issue we haven't found a solution for yet.

### 7.2.3 Issue 3: How do we handle outlier samples?

We mentioned above how one experiment (flowering Ped/Col) contained a significant number of reads covering multiple SNPs from the other genotype. Two samples from this experiment accounted for 1373 and 577 of the annotated mobile mRNAs from the root and rosette, respectively, recall Figure 5.3 and Table 5.2 [98]. This prompted questions about when a sample should be considered an outlier with technical origin.

### 7.2.4 Issue 4: How informative are the SNP-positions we are looking at?

To investigate the differences between SNPs and non-SNPs, we computed the nucleotide distributions at different positions as a negative control. Given the expected low numbers of mobile mRNAs relative to local transcripts, most SNPs in mobile mRNAs will not have sufficient coverage to detect foreign alleles, Figure 7.3. Those SNPs that have reads matching the alternate allele may provide evidence in support of the associated mRNA being from another genotype and potentially being trafficked into the sampled tissue. Figure 7.2 shows the distribution of reads that match to the alternate allele, $n$, over the total number of reads, $N$, for each SNP in the mobile population of two example datasets [98], [106]. We can see that several SNPs have support for the alternate allele. We would expect these distributions to be different from another, non-SNP, positions in the sequence. However, looking at all neighbouring positions of SNPs in the mobile population and computing the number of reads with second most frequent nucleotide, $m$, over the most frequent and second most frequent nucleotides, $M$, we find little difference between SNPs and non-SNPs. More reads with nucleotides from the alternate allele at a SNP position should lead to a shift in the distribution towards higher $n/N$ values, i.e. $n/N > m/M$, but this is not observed. This is in line with prior results that suggest that many reads that have been interpreted as supporting the alternate allele are actually consistent with sequencing noise.

## 7.3 Conclusion

In plants, local cell-to-cell transport of mRNA occurs via plasmodesmata [75], [91], and long-distance translocation is thought to be facilitated by the phloem [99]. The transport of several mRNAs, both cell-to-cell and over long distances, has been experimentally validated by methods that include the detection of the mobile mRNA via qRT-PCR, a translated protein, or a phenotypic change distal to the site of transcription of either an endogenous or transgene-derived sequence. The evidence for transport and translation of selected mRNAs is compelling [137], [138] and several

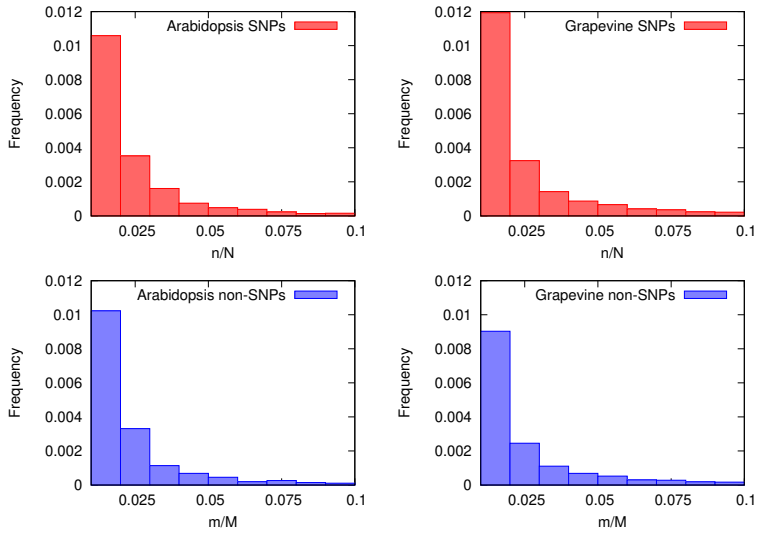



Figure 7.2: The distributions of nucleotides at SNP and non-SNP positions can be informative for evaluating the evidence for the alternate allele. Top (red): Histograms of the ratio of the number reads that match to the alternate allele, $n$, over the number of reads of local and foreign reads, $N$, for each SNP position in the mobile population on examples from Arabidopsis [98] and grapevine [106] on the left and right, respectively. Several SNPs have reads that match to the alternate allele. The full distributions are depicted in Figure 7.3. Bottom (blue): Histograms of the ratio of the number reads that match to the second most frequent nucleotide, $m$, over the sum of the number of reads over the most frequent and second most frequent nucleotide, $M$, for neighbouring positions to SNPs. An overlay of the distributions is given in Figure 7.4. In these examples, the SNP distributions are similar to non-SNP distributions, suggesting that both may be a consequence of the same technological noise.

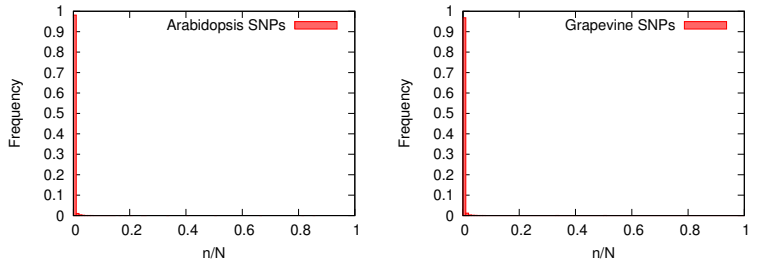



Figure 7.3: Most SNPs in mobile mRNAs do not have evidence for the alternate allele. Histograms of the ratio of the number of reads that match to the alternate allele, $n$, over the number of reads of local and foreign reads, $N$, for each SNP position in the mobile population on examples from Arabidopsis [98] and grapevine [106].

mobile mRNAs have been experimentally confirmed. Based on former RNA-Seq studies, several thousand transcripts have been reported to be transported over long-distances in plants [69], [98], [105]–[107], [117], [120]. The potential for signaling and the possibility of finding mobility motifs that would greatly enhance our ability to engineer designed constructs for transgene-free genome editing [138], make the study of mobile mRNAs particularly important.

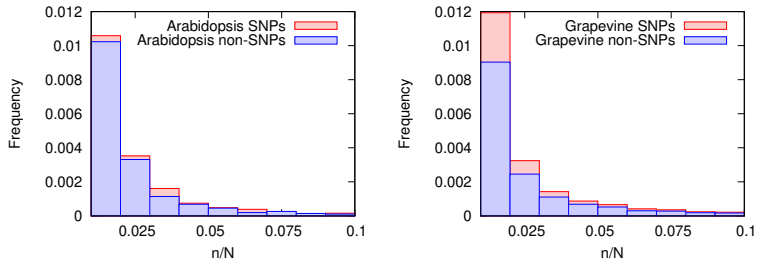



Figure 7.4: Overlay of the ratios of nucleotides at SNP and non-SNP position. Histograms of the ratio of the number reads that match to the alternate allele, $n$, over the number of reads of local and foreign reads, $N$, for each SNP position in the mobile population on examples from Arabidopsis [98] and grapevine [106] on the left and right, respectively. For non-SNP positions the ratio of the number reads that match to the second most frequent nucleotide, $m$, over the sum of the number of reads over the most frequent and second most frequent nucleotide, $M$, is depicted. The frequencies are somewhat different but the actual values of $n/N$ are similar for SNP and non-SNP positions.

Despite significant recent progress in the elucidation of mechanisms underlying long-distance mRNA movement, many key questions remain. Why are so many transcripts transported? What is their function? How are they transported? In particular, bi-directional transport and movement against the presumed direction of phloem flow seems puzzling. Why are the numbers and identity of mobile mRNAs so variable between experiments?

In this recent work, we re-analysed RNA-Seq data and critically reviewed associated bioinformatic pipelines from several key publications on mobile mRNAs. Identifying putative mobile mRNAs relies on distinguishing from which of the two genomes of the grafted plants an RNA-Seq read arose. As shown, this step is far from trivial. Once a transcript has been identified as foreign to the sampled tissue and likely to be from the other grafted genotype, then a common hypothesis is that it may have been transported from the grafted tissue into the sampled tissue, i.e. the transcript is a mobile mRNA. However, even after stringent controls, contamination can rarely be excluded. Strictly speaking, the current criteria used to define a transcript as being mobile should rather, therefore, be for a transcript to be from a foreign genotype. Furthermore, the detection accuracy can be debated after taking into account newly uncovered issues.

Our re-analysis of grafting data suggests that the evidence from RNA-seq data supporting many candidate mobile mRNAs cannot be discerned from technological noise, genome assembly issues, and associated subtleties in data processing. In fact, our negative controls concern us: In those cases where homograft data was available, we found little difference between homograft and heterograft or between SNPs and neighbouring positions. If this is the case and the current mobile mRNA detection methods based on RNA-Seq are susceptible to technical and biological noise, then we can make several predictions.

For transcripts from the alternate allele, we would expect RNA-Seq reads that cover multiple SNPs to have all SNPs corresponding to a specific genotype, whereas for technological sequencing noise, such as base-calling errors or reverse transcriptase errors, this co-occurrence is statistically highly improbable and would not be expected. We point out that both non-selective mRNA transport and contamination would be expected to give rise to high numbers of transcripts from the alternate allele that correlate with the read depth in the source tissue and are therefore challenging to distinguish.

In the extreme case of technological noise inadvertently being a main contributor to the SNP-based classification of mobile mRNAs (by defining a transcript as being mobile based on absolute read counts), we would expect that mainly endogenously-expressed transcripts are found to be mobile. We would therefore not expect to find an mRNA which was over-expressed in a source tissue but which is not expressed in the sampled tissue to be identified as mobile. Furthermore, we would expect a pronounced read-depth dependence and considerable variation between experiments due to the nature of sequencing associated errors. This read-depth dependence (arising from previously used criteria to define mobile mRNAs and potentially also from contamination) is likely to be the underlying cause for the observed dependence on abundance [110], i.e. abundance is causative for mobility but not in the way we had envisioned and does not have a mechanistic explanation but is rather a consequence of how mobile mRNAs had been defined. Interestingly, in mammalian systems, mRNA transport was also found to be pervasive across the genome, in a non-selective and expression dependent manner [139].

Base-calling errors, reverse transcriptase errors, differences in coverage, batch effects, potential contamination, copy number variation and genome reference quality all represent alternate hypotheses that also explain data for current mobile mRNA assignments from RNA-Seq experiments. Overall, it is not clear whether short-read RNA-Seq is a viable technique for identifying mobile mRNAs when the transported numbers are so low and below the typical noise levels of this technology. It thus becomes important to evaluate different hypotheses rather than to define transcripts as mobile based on chosen criteria. Although several mRNAs that were identified as being mobile based on RNA-Seq data have been experimentally validated, we cannot rule out that others might be false positives.

We have a lot to learn about the phenomenon of long-distance mRNA transport and its function. This re-analysis of available RNA-seq datasets highlights the complexity of identifying mobile mRNAs and raises questions about the extent of mRNA communication in plants. We hope that the presented considerations and analyses will contribute to driving this exciting field forward.

## 7.4 Methods

The study summarised in this chapter was led by Pirita Paajanen, and is documented in further detail in a recent preprint [63]. Briefly, we used the following published datasets, including the archived reads from NCBI.

*Cuscuta pentagona* [69]: PRJNA257158 (incomplete and partly corrupt);
*Vitis vinifera* [106]: SRP058158 and SRP058157;
*Solanum lycopersicum, Nicotiana benthamiana* [120]: SRP111187;
*Arabidopsis thaliana* [98]: PRJNA271927;
*Citrullus lanatus* L. [117]: PRJNA553072.

We also used the deposited Supplementary datasets to obtain the numbers of identified mRNAs. For each of the grafting studies, we downloaded the reference genome sequence that matched the one that was used in the original paper with the same annotations; these were all publicly available on Ensembl plants. The raw reads were mapped to the references using hisat2 (v.2.1.0) [140], and processed by samtools (v1.9) [141], the expression levels were quantified with stringtie (v1.3.5) [142]. The variants were called with bcftools (v1.10.2) [141]. The NCBI nucleotide database was downloaded on (21/Oct/2022) and blast+ v2.9.0 [143] was used for alignments. The error-rate corrected evaluation of evidence for mobility was performed using the *baymobil* python package, the method introduced in Chapter 6 [64], [65].

# Chapter 8

# Discussion

## 8.1 What we have learned in this thesis

In **Chapter 1**, we gently revisited the fundamental reasons behind our pursuit of science and how we derive knowledge from data. We brought to our awareness how we learn and update our beliefs, and that we can mathematically articulate this process using probability theory. As science is all about learning and updating our knowledge through observations and experiments, we elucidated the rationale for employing probability theory in and on science throughout the thesis. We explored the human intuition of this mathematical and philosophical framework which offers concise and beautiful solutions to the example problems in the Introduction and the real-world applications discussed in subsequent chapters. Already in the explanatory problems we uncovered what solutions Bayesian statistics can provide, and where limitations of extrinsic nature lie. Examples of upper boundaries may be set by available data which can be only giving limited information and, therefore, restricting inference we can draw from them. This exploration sets the stage for the discussions that follow. We hope to have conveyed the utility of Bayesian approaches in scientific inquiry and generated appetite to learn more by pointing out its elegance.

In **Chapter 2**, we acquainted ourselves with the prominent data type of this thesis: RNA-Sequencing data. We found out how it helps us to learn about RNA populations in tissues, how experiments are set up to investigate changes in RNA populations, and how the data is traditionally processed and analysed. We have prepared ourselves for diving into the work presented in this thesis. We are merging what we have learned in Chapter 1 and Chapter 2 and bring Bayesian inference into the analysis of RNA-Seq data.

In **Chapter 3**, we introduced the application of Bayes factors in the analysis of RNA-Seq data for identifying gene expression changes. In the Introduction we were observing the advantages of using a single Bayes factor to describe medical test statistics, instead of relying on multiple conventional accuracy indices. In Chapter 3 we calculate Bayes factor which allow us to quantify the evidence for changes in gene expression between two samples. Similar to the example in the Introduction we observe the beauty of the method and its ability to condense the

communication of results into a single value. This enhances clarity and allows us to rank (instead of classify) genes according to evidence for expression change. With this, we are opening new avenues to communicate our confidence in hypotheses and avoid arbitrary cutoffs that are often applied in traditional analyses using p-values, $\log_2$ fold change, and such.

In **Chapter 4**, we examined the variability inherent to RNA-Seq data and discussed how this variability, when combined with data sparsity, can lead to pitfalls in the interpretation. While RNA-Seq data holds substantial potential for uncovering biological insights, its effectiveness is contingent upon a thorough understanding of both biological and technical fluctuations that may influence the results. Properly accounting for these variations is essential to harness the full power of RNA-Seq analyses and avoid common traps that can misguide conclusions.

In **Chapter 5**, we learned about the fascinating phenomenon of mobile mRNAs in plants and the inventive efforts to develop a high-throughput detection technique that merges ancient grafting methods with cutting-edge sequencing technologies. In **Chapter 6**, we employed Bayes factors to assess the evidence for mRNA mobility using RNA-Seq data from grafted plants. While we were once again impressed by the performance of the Bayesian approach, we have recently encountered additional issues that have raised concerns about the reliability of this detection method. In **Chapter 7** we summarised the recently uncovered structural problems underlying the usage of RNA-Seq data from grafted plants that must be addressed to achieve successful detection of mobile mRNA.

## 8.2 What more is there to say?

I am absolutely stunned by what we can do in science in 2024. Advances in sequencing technologies allow us to get insights into molecular processes within tissues and cells, offering tremendous opportunities to unravel how gene regulation orchestrates development and responses to the environment [9]–[11]. Furthermore, we have found countless creative applications of the technology in research and medicine (e.g. detecting mobile mRNA in plants). The impact of RNA-Sequencing is, nevertheless, tightly linked to our ability to handle the data. And I am surprised by how much more there is for us to do to improve this.

RNA-Sequencing data comes with measuring uncertainties that have been challenging the statistics community already for a while [12], [14]–[19], [28]–[35]. Given variability is the big challenge this data exposes us to, it does surprise me, that not more efforts have been put into finding solutions in the Bayesian universe for these problems – so far. In this thesis, I have presented my work as part of a team of curious, eager and fun scientists, full of determination to do this. We started our journey with the challenge of detecting mobile mRNA in plants, from where we moved to the popular search for differential gene expression. We have found beautiful analytical solutions to calculate Bayes factors in both cases, showing convincing performance and providing a more intuitive solution. Despite discussions for, against or how to exactly use Bayes factors [144]–[146] and hypothesis testing in general [147]–[150], we found the quantification and ranking solutions a compelling first step, aside from the analysis being fast (concomitant of the exact solutions

given) and the circumvention of arbitrary cutoffs. It is somewhat delusive to compare the methods presented in this thesis to existing methods in the field, seeing the fundamental differences in mathematics and philosophy. Bayesian statistics provide opportunities to communicate and interpret research results in a different way – framed in a world view of ever-updating knowledge. Of course, we still did do our best to show with simulated and real data how our methods provide sensible results which are in some cases more and in some cases less compatible with the results from other analyses. Comparing methods without a ground-truth will always end up in circular arguments, as the absence of an objective standard makes it impossible to definitively assess which method yields more accurate or reliable results.

Overall and nevertheless, we have also learned about problems where Bayes factors and Bayesian inference can not solve the issue.

## 8.3 Why Bayes factors cannot solve all our problems

Critical voices (like certain parts of my own brain) will push this discussion in the direction of whether or how to exactly use Bayes factors [144]–[146] and hypothesis tests [147]–[151]. I agree that this is a fair point to raise. Using Bayes factors in the analysis the way we did it is still an employment of a hypothesis test, and we have not put forward a sophisticated process model for the high-throughput analysis we do. We are using Bayes factors for asking a simple question, far from the search for mechanisms of the phenomena we are investigating. Of course, this would ultimately be the goal and the most effective way to learn from our data. I am curious and excited about the efforts that are made to do exactly that, and in what other ways we can make use of the huge, growing amounts of RNA-Seq data around us. For the moment, however, we see our coherent and clear approach as an educational chance for the field, introducing knowledge updating and ranking methods. Calculating Bayes factors and ranking our results may in some places be the best we can do with existing data right now. It should be straightforward since the medical test paradox story in the Introduction, that there are certainly use cases of Bayes factor that are worthwhile. The methods we presented can be seen as a first step towards establishing Bayesian methodology that can grow far beyond this initial work, now that we have shown it to be worth further exploration.

In the future we can build up on that, moving towards more complex Bayesian and Causal inference problems. Furthermore, the data accumulation around us is also not happening without our involvement. For future experiments, we can upgrade our strategy. We can do fewer experiments but make sure we have enough information in our data per experiment to produce reproducible results. This is (as we have explored in Chapter 4) in RNA-Seq often bound to a higher number of replicates. We can explore the use of Causal inference for our experiments and make efforts to find mechanistic models while learning about the powers and limitations of all statistics.

### 8.3.1 Improvements ready to be explored

In the Introduction I bravely stated, that statistics can help us to make informed decisions from limited data. Of course, there is still a lower limit to that. If we remind ourselves of the card game between the kid and the adult, we recall that for data from only a limited number of rounds, we did not get very strong Bayes factors (Figure 1.11) and we could only get a clearer picture after our players played 100 games (Figure 1.13). Now in the analysis of RNA-Seq data: Do we have limited data or limited information in our data?

An open discussion for me is, whether the numbers we get in processed RNA-Seq data are deceptive or not. We are producing many, and huge numbers in RNA-Seq (compare Figures 3.2, 3.3, 3.4 and 6.15) which lead to very sharply defined likelihood functions in our Bayes factor calculations. The process to get these high numbers may, however, not fully be reflected in our simple binomial model. This leads to very extreme Bayes factors that are arguably justifiable, but given the variability of some genes (deeply explored in Chapter 4) also a risk for over-interpretation. Surely, this is only a problem when considering absolute Bayes factors – as opposed to ranking. Furthermore, often we can learn about specific genes of interest and their expression patterns *a priori* from existing data (bases). In the same way, that a medical test updates our prior odds of having a disease, the Bayes factor for a gene updates our belief about a gene's expression changes upon a stimulus.

Needless to say, we did consider choosing other process models, trying to encounter this issue by taking amplification steps into account, for example. There is potential for expanding to a multinomial, instead of a bionomial framework, considering that we may want to look at whole gene regulatory networks instead of single genes. Finally, we decided to stick with the simplest model as a first step, as we get sensible results. For educational purposes simplicity is a well-advised start, and from here: may it grow, expand, and thrive.

## 8.4 Why Bayesian statistics cannot solve all our problems

It's often said, but one of the most noteworthy lessons I (re-)learned during my PhD is the importance of allowing yourself to ask big questions. WHAT ARE WE SEARCHING FOR? — is potentially the most frequent sentence in my notebook. In my feeling, embracing these fundamental inquiries is what truly propels us forward in our research and understanding. In retrospect, asking this question in the light of mobile mRNA detection using RNA-Seq reveals a potential flaw in the whole approach. If we iterate back to our entrance example of a cancer screening in the Introduction and how much we learn from doing a certain medical test, it may occur to us, that calculating a Bayes factor for mobile mRNA detection is a task that can hardly be accomplished. I do not mean calculating a Bayes factor for transcript mobility, but for the test itself. The difficulty is, that no no ground truth exists to date, due to limited validation. A second issue is, that, as far as we now, the prevalence of mobile mRNAs is very low. This means, we are applying a test of unknown accuracy for a low-prevalence event to a big population. It could be, that the underlying structure of our problem makes it practically impossible to

find mobile mRNAs in RNA-Seq data without a lot of false positives. To use an analogy, it is like we are fishing for a rare, small fish with a fishing net; if we want to catch any of them we need a fine net, resulting in a lot of bycatch, Figure 8.1.

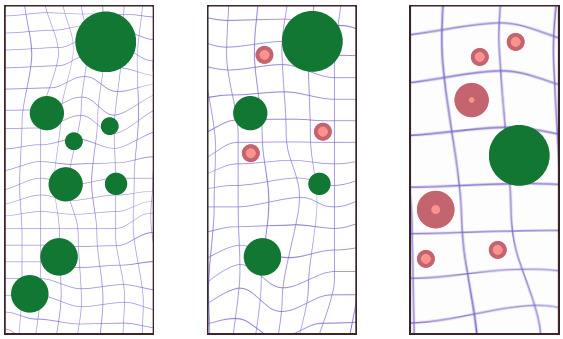


Figure 8.1: Catching mobile mRNAs, differentially expressed genes, or tuna requires making a decision about the coarseness of a fishing net, or a detection criterion. Tweaking the coarseness changes the outcome, the ratio of target hits to bycatch, or shuffling around true positives, true negatives, false positives, and false negatives. For RNA-Sequencing in molecular biology we do not have information on the accuracy of the test. Neither do we know about the coarseness of the net. In our work (Chapter 6) we introduced Bayes factors into the analysis. We enabled the incorporation of prior knowledge in the analysis, with the goal to make the fishing net adjustable and avoid bycatch. However, the cues we intended to use for adjusting coarseness turned out to be unreliable, Figures 7.2, 7.3, 7.4.

What probability theory can help us with, are any problems that require inference. For example, we can estimate how much noise there is in the data (as we did in Chapter 5). What probability theory cannot do, of course, is distill information and further knowledge from data, that is not given in the data (Chapter 4 and 7, or the card game example in the introduction with 10 vs. 100 rounds, Figures 1.11 and 1.13). Given the insights we learned about in Chapter 5 about how rare we think the phenomenon of macromolecular long-distance traveling is, if we do a test (let's for now ignore the fact that we are screening with unknown accuracy), our posterior knowledge will not grow a lot by employing a test with big bycatch. Just like in the medical test paradox study mentioned in the Introduction [4], the common cognitive bias known as the base rate fallacy strikes again, Figure 1.6. Theoretically, as we have seen for simulated data, introducing Bayes factors and prior knowledge drastically improves the detection. In practice, just like in medical tests, it is hard to find reliable prior information. Only a second source of unrelated evidence may help us.

### 8.4.1 Forever trapped in the accuracy-error dilemma

Research methodologist and transparency activist Vera Wilde works with signal detection problems or mass screenings for low-prevalence problems across diverse

domains [152], [153]. She has made the effort to collect problematic cases from a broad range of disciplines to raise awareness for the limitations of mass screenings, which can cause negative side effects even if the screenings are done with good intentions, Table 8.1.

| Application | Target | Bycatch |
|---|---|---|
| Trawling | Tuna | Dolphin |
| Polygraph | Spies, Terrorists | Non-spies, non-terrorists |
| iBorderCtrl | Bad crossings | Innocent crossings |
| ChatControl | CSAM | Innocent coms |
| Asymptomatic cancer screenings | Deaths | Healthy people |
| Lifestyle diseases | Big problems | Mild cases |
| Advanced medical imaging | See problems | Harmless anomalies |
| Educational ethics | Plagiarism and AI use in writing | Innocent students |
| Misinformation | Provably wrong | Ambiguity, dissent |
| Disinformation | Hostile propaganda | Ambiguity, dissent |
| Long-distance travelling molecules | mobile mRNA in plants | Sequencing and bioinformatics issues |

Table 8.1: A list of mass-screenings for low-prevalence problems that all run into the same issues with bycatch. The list was collected by Vera Wilde [152], [153], we have added the detection of mobile mRNA. No matter what exactly we are searching for, maximizing true-positive rates and minimizing false-positive rates are in tension. A reliable second stream of evidence may drastically change this picture by providing solid prior knowledge to build upon.

In the detection of mobile mRNAs it is on us to decide now how we want to confront the puzzle. One approach is to keep pushing technology and analysis forward and, hence, find more accurate detectors for this mass detection. Given that mRNA mobility is considered a rare event, however, we might want to put our efforts into investigating the phenomenon from a different angle, the search for mechanistic details. An important message in Vera Wilde's work is the criticism of tech-solutionism, which is often employed when limitations of a technique are visible, while the low probabilities of the events keep being neglected. In tech-solutionism, the focus of solving a problem (from health care, and security to researching scientific phenomena) lies in finding a technical solution for it. Huge efforts are put into improving the detection methods of the mass screenings in Table 8.1. Visions of how these improvements are happening often involve the newest innovations, like artificial intelligence, or in our case potentially single-cell sequencing. A similar experience has occurred to our research project during the last few years. Of course, we are investing all of our energies to learn how we can use RNA-Sequencing data to detect mobile mRNAs. I was myself certainly drawn towards this birds-eye perspective on the phenomenon, and very curious about these big data sets when I started. The biological research question is incredibly fascinating and the data has been collected by our collaborators in hours and hours of tedious lab work, after refining protocols for years. We didn't want to give up on finding

evidence for mobile mRNA in the sequencing data. You do not want to give up too early; What is science without believing in finding solutions for all problems? Given the problems about mass screenings I have touched on now, however, do we still think the potential lies in high-throughput detection?

## 8.5 How Bayesian statistics can help us accept our problems

Bayesian statistics may not be able to provide us with a technical solution for improving the mobile mRNA detection accuracy with current data and methodology (what we initially thought when we started introducing Bayes factors in the analysis). But probabilistic thinking provides us with a philosophy and mathematical framework to go far beyond that. We can identify the underlying limitations of high-throughput detection given by the phenomenon we are investigating.

We can argue we always should think about a net harm/benefit analysis to know which tests or screenings help us in science and society, and which ones hurt us. Thankfully, false positive hits in mobile mRNA detection are certainly not as harmful as some other problematic mass screening examples mentioned in Table 8.1. Still, there are vast amounts of human time and resources put into hypotheses and further research, formulated on potential false positive hits or based on thin evidence. Initially, investigation of all avenues is equally valuable in science. After everything we have learned in our research project in the last years, we started to ask ourselves whether we want to keep focusing on finding technological solutions for our mass detection problem. Building knowledge from a strong evidence base would be what probabilistic thinking advises us.

In science, at least the way I understand it, we want to systematically and empirically acquire knowledge about the natural world. So we are measuring things, even though this can be hard. We still want to try recording observations, because we are curious. We can see afterwards, whether there is information in our data that we can deduce knowledge from. This process is where probabilistic thinking comes in with a huge potential. The questions Bernoulli was pondering over, the use of the mathematics of probability for inference problems that are appearing in life and science, are just as relevant as ever. We are dealing with inherently complex data, because life is complex and the world is complex. Probability theory, however, can help us to articulate this complexity and communicate the uncertainty of our understanding of life and the world. Our observations require this kind of description because we have probabilistic associations of imperfectly recorded cues that we wish to understand. Technological solutions often only provide us with ways to categorise our seemingly probabilistic world. Probably.

# Bibliography


[1] D. Sivia and J. Skilling, *Data Analysis: A Bayesian Tutorial*. Oxford University Press, Jun. 2006, 259 pp., ISBN: 978-0-19-856831-5. Google Books: lYMSDAAAQBAJ.

[2] G. Sanderson. "Bayes theorem, the geometry of changing beliefs - YouTube." (2019), [Online]. Available: https://www.youtube.com/watch?v=HZGCoVF3YvM (visited on 09/16/2024).

[3] G. Sanderson. "The medical test paradox, and redesigning Bayes' rule - YouTube." (2020), [Online]. Available: https://www.youtube.com/watch?v=lG4VkPoG3ko (visited on 09/16/2024).

[4] G. Gigerenzer, *Calculated Risks: How to Know When Numbers Deceive You*. Simon and Schuster, 2002, 328 pp., ISBN: 978-0-7432-5423-6. Google Books: KJ7nrlJqcRYC.

[5] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, Apr. 10, 2003, 764 pp., ISBN: 978-0-521-59271-0. Google Books: tTN4HuUNXjgC.

[6] H. Jeffreys, *The Theory of Probability*. OUP Oxford, Aug. 6, 1998, 474 pp., ISBN: 978-0-19-158967-6. Google Books: vh9Act9rtzQC.

[7] R. E. Kass and A. E. Raftery, "Bayes Factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, Jun. 1, 1995, ISSN: 0162-1459. DOI: 10.1080/01621459.1995.10476572.

[8] F. Hoerbst, G. S. Sidhu, M. Tomkins, and R. J. Morris. "A Closed-Form Solution to the 2-Sample Problem for Quantifying Changes in Gene Expression using Bayes Factors." arXiv: 2406.19989 [q-bio, stat]. (Jun. 28, 2024), [Online]. Available: http://arxiv.org/abs/2406.19989 (visited on 07/31/2024), pre-published.

[9] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: A revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 1 Jan. 2009, ISSN: 1471-0064. DOI: 10.1038/nrg2484.

[10] S. Behjati and P. S. Tarpey, "What is next generation sequencing?" *Archives of Disease in Childhood. Education and Practice Edition*, vol. 98, no. 6, pp. 236–238, Dec. 2013, ISSN: 1743-0593. DOI: 10.1136/archdischild-2013-304340. PMID: 23986538.

[11] M. Furlan, I. Tanaka, T. Leonardi, S. de Pretis, and M. Pelizzola, "Direct RNA Sequencing for the Study of Synthesis, Processing, and Degradation of Modified Transcripts," *Frontiers in Genetics*, vol. 11, p. 394, 2020, ISSN: 1664-8021. DOI: 10.3389/fgene.2020.00394. PMID: 32425981.

[12] J.-W. Chen, L. Shrestha, G. Green, A. Leier, and T. T. Marquez-Lago, "The hitchhikers' guide to RNA sequencing and functional analysis," *Briefings in Bioinformatics*, vol. 24, no. 1, bbac529, Jan. 1, 2023, ISSN: 1477-4054. DOI: 10.1093/bib/bbac529.

[13] K. R. Kukurba and S. B. Montgomery, "RNA Sequencing and Analysis," *Cold Spring Harbor Protocols*, vol. 2015, no. 11, pp. 951–969, Apr. 13, 2015, ISSN: 1559-6095. DOI: 10.1101/pdb.top084970. PMID: 25870306.

[14] S. Anders, D. J. McCarthy, Y. Chen, *et al.*, "Count-based differential expression analysis of RNA sequencing data using R and Bioconductor," *Nature Protocols*, vol. 8, no. 9, pp. 1765–1786, 9 Sep. 2013, ISSN: 1750-2799. DOI: 10.1038/nprot.2013.099.

[15] A. Conesa, P. Madrigal, S. Tarazona, *et al.*, "A survey of best practices for RNA-seq data analysis," *Genome Biology*, vol. 17, p. 13, Jan. 26, 2016, ISSN: 1474-760X. DOI: 10.1186/s13059-016-0881-8. PMID: 26813401.

[16] K. Van den Berge, K. M. Hembach, C. Soneson, *et al.* "RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis." (Jul. 1, 2019), prepublished.

[17] F. Rapaport, R. Khanin, Y. Liang, *et al.*, "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data," *Genome Biology*, vol. 14, no. 9, p. 3158, Sep. 10, 2013, ISSN: 1474-760X. DOI: 10.1186/gb-2013-14-9-r95.

[18] X. Zhou, H. Lindsay, and M. D. Robinson, "Robustly detecting differential expression in RNA sequencing data using observation weights," *Nucleic Acids Research*, vol. 42, no. 11, e91, Jun. 17, 2014, ISSN: 0305-1048. DOI: 10.1093/nar/gku310.

[19] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, no. 12, p. 550, Dec. 5, 2014, ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8.

[20] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, R106, Oct. 27, 2010, ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-10-r106.

[21] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: A Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 1, 2010, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp616.

[22] Y. Chen, A. T. L. Lun, and G. K. Smyth, "From reads to genes to pathways: Differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline," no. 5:1438, Aug. 2, 2016. DOI: 10.12688/f1000research.8987.2.

[23] D. J. McCarthy, Y. Chen, and G. K. Smyth, "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation," *Nucleic Acids Research*, vol. 40, no. 10, pp. 4288–4297, May 1, 2012, ISSN: 0305-1048. DOI: 10.1093/nar/gks042.

[24] T. J. Hardcastle and K. A. Kelly, "baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data," *BMC Bioinformatics*, vol. 11, no. 1, p. 422, Aug. 10, 2010, ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-422.

[25] N. Leng, J. A. Dawson, J. A. Thomson, *et al.*, "EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments," *Bioinformatics*, vol. 29, no. 8, pp. 1035–1043, Apr. 15, 2013, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt087.

[26] J. Lu, J. K. Tomfohr, and T. B. Kepler, "Identifying differential expression in multiple SAGE libraries: An overdispersed log-linear model approach," *BMC Bioinformatics*, vol. 6, no. 1, p. 165, Jun. 29, 2005, ISSN: 1471-2105. DOI: 10.1186/1471-2105-6-165.

[27] M. D. Robinson and G. K. Smyth, "Small-sample estimation of negative binomial dispersion, with applications to SAGE data," *Biostatistics*, vol. 9, no. 2, pp. 321–332, Apr. 1, 2008, ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxm030.

[28] Z. Su, P. P. Łabaj, S. Li, *et al.*, "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium," *Nature Biotechnology*, vol. 32, no. 9, pp. 903–914, 9 Sep. 2014, ISSN: 1546-1696. DOI: 10.1038/nbt.2957.

[29] L. A. Corchete, E. A. Rojas, D. Alonso-López, J. De Las Rivas, N. C. Gutiérrez, and F. J. Burguillo, "Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis," *Scientific Reports*, vol. 10, no. 1, p. 19 737, Nov. 12, 2020, ISSN: 2045-2322. DOI: 10.1038/s41598-020-76881-x.

[30] C. M. Koch, S. F. Chiu, M. Akbarpour, *et al.*, "A Beginner's Guide to Analysis of RNA Sequencing Data," *American Journal of Respiratory Cell and Molecular Biology*, vol. 59, no. 2, pp. 145–157, Aug. 2018, ISSN: 1044-1549. DOI: 10.1165/rcmb.2017-0430TR. PMID: 29624415.

[31] J. Costa-Silva, D. Domingues, and F. M. Lopes, "RNA-Seq differential expression analysis: An extended review and a software tool," *PLOS ONE*, vol. 12, no. 12, e0190152, Dec. 21, 2017, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0190152.

[32] J. Costa-Silva, D. S. Domingues, D. Menotti, M. Hungria, and F. M. Lopes, "Temporal progress of gene expression analysis with RNA-Seq data: A review on the relationship between computational methods," *Computational and Structural Biotechnology Journal*, vol. 21, pp. 86–98, Jan. 1, 2023, ISSN: 2001-0370. DOI: 10.1016/j.csbj.2022.11.051.

[33] A. McDermaid, B. Monier, J. Zhao, B. Liu, and Q. Ma, "Interpretation of differential gene expression results of RNA-seq data: Review and integration," *Briefings in Bioinformatics*, vol. 20, no. 6, pp. 2044–2054, Nov. 27, 2019, ISSN: 1477-4054. DOI: 10.1093/bib/bby067.

[34] R. Stark, M. Grzelak, and J. Hadfield, "RNA sequencing: The teenage years," *Nature Reviews Genetics*, vol. 20, no. 11, pp. 631–656, 11 Nov. 2019, ISSN: 1471-0064. DOI: 10.1038/s41576-019-0150-2.

[35] N. J. Schurch, P. Schofield, M. Gierliński, *et al.*, "How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?" *RNA*, vol. 22, no. 6, pp. 839–851, Jun. 2016, ISSN: 1355-8382. DOI: 10.1261/rna.053959.115. PMID: 27022035.

[36] F. Hoerbst, G. S. Sidhu, T. Omori, M. Tomkins, and R. J. Morris, "A Bayesian framework for ranking genes based on their statistical evidence for differential expression," 2025. DOI: 10.1101/2025.01.20.633909.

[37] T. Chari and L. Pachter. "The Specious Art of Single-Cell Genomics." (Dec. 22, 2022), pre-published.

[38] C. Ahlmann-Eltze and W. Huber, "Comparison of transformations for single-cell RNA-seq data," *Nature Methods*, pp. 1–8, Apr. 10, 2023, ISSN: 1548-7105. DOI: 10.1038/s41592-023-01814-1.

[39] X. Li and C.-Y. Wang, "From bulk, single-cell to spatial RNA sequencing," *International Journal of Oral Science*, vol. 13, no. 1, pp. 1–6, 1 Nov. 15, 2021, ISSN: 2049-3169. DOI: 10.1038/s41368-021-00146-0.

[40] G. H. Putri, S. Anders, P. T. Pyl, J. E. Pimanda, and F. Zanini, "Analysing high-throughput sequencing data in Python with HTSeq 2.0," *Bioinformatics*, vol. 38, no. 10, pp. 2943–2945, May 15, 2022, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac166.

[41] S. Cortijo and J. C. W. Locke, "Does Gene Expression Noise Play a Functional Role in Plants?" *Trends in Plant Science*, vol. 25, no. 10, pp. 1041–1051, Oct. 1, 2020, ISSN: 1360-1385. DOI: 10.1016/j.tplants.2020.04.017. PMID: 32467064.

[42] S. Cortijo, Z. Aydin, S. Ahnert, and J. C. Locke, "Widespread inter-individual gene expression variability in Arabidopsis thaliana," *Molecular Systems Biology*, vol. 15, no. 1, e8591, Jan. 2019, ISSN: 1744-4292. DOI: 10.15252/msb.20188591.

[43] A. Varabyou, S. L. Salzberg, and M. Pertea, "Effects of transcriptional noise on estimates of gene and transcript expression in RNA sequencing experiments," *Genome Research*, vol. 31, no. 2, pp. 301–308, Feb. 2021, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.266213.120.

[44] I. S. Araújo, J. M. Pietsch, E. M. Keizer, *et al.*, "Stochastic gene expression in Arabidopsis thaliana," *Nature Communications*, vol. 8, no. 1, p. 2132, 1 Dec. 14, 2017, ISSN: 2041-1723. DOI: 10.1038/s41467-017-02285-7.

[45] R. Kelter, "Bayesian and frequentist testing for differences between two groups with parametric and nonparametric two-sample tests," *WIREs Computational Statistics*, vol. 13, no. 6, e1523, 2021, ISSN: 1939-0068. DOI: 10.1002/wics.1523.

[46] H. Chen, M. Centola, S. F. Altschul, and H. Metzger, "Characterization of Gene Expression in Resting and Activated Mast Cells," *The Journal of Experimental Medicine*, vol. 188, no. 9, pp. 1657–1668, Nov. 2, 1998, ISSN: 0022-1007, 1540-9538. DOI: 10.1084/jem.188.9.1657.

[47] K. M. Borgwardt and Z. Ghahramani. "Bayesian two-sample tests." version 1. arXiv: 0906.4032 [cs]. (Jun. 22, 2009), pre-published.

[48] C. Everaert, M. Luypaert, J. L. V. Maag, *et al.*, "Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data," *Scientific Reports*, vol. 7, no. 1, p. 1559, May 8, 2017, ISSN: 2045-2322. DOI: 10.1038/s41598-017-01617-3.

[49] T. Coenye, "Do results obtained with RNA-sequencing require independent verification?" *Biofilm*, vol. 3, p. 100 043, Jan. 13, 2021, ISSN: 2590-2075. DOI: 10.1016/j.bioflm.2021.100043. PMID: 33665610.

[50] K. Blighe, S. Rana, and M. Lewis. "EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling." (2023), [Online]. Available: https://bioconductor.org/packages/release/bioc/vignettes/EnhancedVolcano/inst/doc/EnhancedVolcano.html#download-the-package-from-bioconductor (visited on 04/17/2024).

[51] A. Tarsitano. "Nonlinear Rank Correlations." (Sep. 9, 2008), [Online]. Available: http://www.ecostat.unical.it/Tarsitano/Publications/Pagepubs/pub02a4.htm (visited on 04/17/2024).

[52] W. Webber, A. Moffat, and J. Zobel, "A similarity measure for indefinite rankings," *ACM Transactions on Information Systems*, vol. 28, no. 4, 20:1–20:38, Nov. 23, 2010, ISSN: 1046-8188. DOI: 10.1145/1852102.1852106.

[53] M. Gierliński, C. Cole, P. Schofield, *et al.*, "Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment," *Bioinformatics (Oxford, England)*, vol. 31, no. 22, pp. 3625–3630, Nov. 15, 2015, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btv425. PMID: 26206307.

[54] F. Hoerbst, G. S. Sidhu, M. Tomkins, and R. J. Morris. "What is a Differentially Expressed Gene?" (Feb. 1, 2025), pre-published.

[55] H. Zhang, L. Cao, J. Brodsky, *et al.*, "Quantitative or digital PCR? A comparative analysis for choosing the optimal one for biosensing applications," *TrAC Trends in Analytical Chemistry*, vol. 174, p. 117 676, May 1, 2024, ISSN: 0165-9936. DOI: 10.1016/j.trac.2024.117676.

[56] J. M. Ruijter, R. J. Barnewall, I. B. Marsh, *et al.*, "Efficiency Correction Is Required for Accurate Quantitative PCR Analysis and Reporting," *Clinical Chemistry*, vol. 67, no. 6, pp. 829–842, Jun. 1, 2021, ISSN: 0009-9147. DOI: 10.1093/clinchem/hvab052.

[57] Y. Liu, J. Zhou, and K. P. White, "RNA-seq differential expression studies: More sequence or more replication?" *Bioinformatics*, vol. 30, no. 3, pp. 301–304, Feb. 1, 2014, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt688. PMID: 24319002.

[58] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, "Stochastic Gene Expression in a Single Cell," *Science*, vol. 297, no. 5584, pp. 1183–1186, Aug. 16, 2002. DOI: 10.1126/science.1070919.

[59] P. S. Swain, M. B. Elowitz, and E. D. Siggia, "Intrinsic and extrinsic contributions to stochasticity in gene expression," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12 795–12 800, Oct. 2002. DOI: 10.1073/pnas.162041399.

[60] M. Konczal, P. Koteja, M. T. Stuglik, J. Radwan, and W. Babik, "Accuracy of allele frequency estimation using pooled RNA-Seq," *Molecular Ecology Resources*, vol. 14, no. 2, pp. 381–392, 2014, ISSN: 1755-0998. DOI: 10.1111/1755-0998.12186.

[61] A. P. Rajkumar, P. Qvist, R. Lazarus, *et al.*, "Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq," *BMC Genomics*, vol. 16, no. 1, p. 548, Jul. 25, 2015, ISSN: 1471-2164. DOI: 10.1186/s12864-015-1767-y.

[62] A. Takele Assefa, J. Vandesompele, and O. Thas, "On the utility of RNA sample pooling to optimize cost and statistical power in RNA sequencing experiments," *BMC Genomics*, vol. 21, no. 1, p. 312, Apr. 19, 2020, ISSN: 1471-2164. DOI: 10.1186/s12864-020-6721-y.

[63] P. Paajanen, M. Tomkins, F. Hoerbst, *et al.*, "Re-analysis of mobile mRNA datasets highlights challenges in the detection of mobile transcripts from short-read RNA-Seq data," *bioRxiv*, pp. 2024–07, 2024. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.07.25.604588.abstract (visited on 07/31/2024).

[64] M. Tomkins, F. Hoerbst, S. Gupta, *et al.*, "Exact Bayesian inference for the detection of graft-mobile transcripts from sequencing data," *Journal of The Royal Society Interface*, vol. 19, no. 197, p. 20 220 644, Dec. 14, 2022. DOI: 10.1098/rsif.2022.0644.

[65] F. Hoerbst, R. J. Morris, and M. Tomkins, "Baymobil: A Python package for detection of graft-mobile mRNA using exact Bayesian inference on RNA-Seq data," In Review, preprint, Feb. 14, 2023. DOI: 10.21203/rs.3.rs-2520491/v1.

[66] B.-K. Ham and W. J. Lucas, "Phloem-Mobile RNAs as Systemic Signaling Agents," in *Annual Review of Plant Biology, Vol 68*, S. S. Merchant, Ed., vol. 68, Palo Alto: Annual Reviews, 2017, pp. 173–195, ISBN: 978-0-8243-0668-7. DOI: 10.1146/annurev-arplant-042916-041139.

[67] J. Kehr, R. J. Morris, and F. Kragler, "Long-Distance Transported RNAs: From Identity to Function," *Annual Review of Plant Biology*, vol. 73, no. 1, pp. 457–474, May 20, 2022, ISSN: 1543-5008, 1545-2123. DOI: 10.1146/annurev-arplant-070121-033601.

[68] R. David-Schwartz, S. M. Runo, B. T. Townsley, J. S. Machuka, and N. R. Sinha, "Long-distance transport of mRNA via parenchyma cells and phloem across the host-parasite junction in Cuscuta.," *The New phytologist*, vol. 179 4, pp. 1133–41, 2008.

[69] G. Kim, M. L. LeBlanc, E. K. Wafula, C. W. dePamphilis, and J. H. Westwood, "Genomic-scale exchange of mRNA between a parasitic plant and its hosts," *Science*, vol. 345, no. 6198, pp. 808–811, Aug. 15, 2014, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1253122. PMID: 25124438.

[70] J. K. Roney, P. A. Khatibi, and J. H. Westwood, "Cross-species translocation of mRNA from host plants into the parasitic plant dodder," *Plant Physiology*, vol. 143, no. 2, pp. 1037–1043, Dec. 2006, ISSN: 0032-0889. DOI: 10.1104/pp.106.088369. eprint: https://academic.oup.com/plphys/article-pdf/143/2/1037/35942462/plphys_v143_2_1037.pdf.

[71] W. J. Lucas, B.-C. Yoo, and F. Kragler, "RNA as a long-distance information macromolecule in plants," *Nature Reviews Molecular Cell Biology*, vol. 2, no. 11, pp. 849–857, 2001. DOI: 10.1038/35099096.

[72] A. K. Banerjee, M. Chatterjee, Y. Yu, S.-G. Suh, W. A. Miller, and D. J. Hannapel, "Dynamics of a mobile RNA of potato involved in a long-distance signaling pathway," *The Plant Cell*, vol. 18, no. 12, pp. 3443–3457, Dec. 2006, ISSN: 1040-4651. DOI: 10.1105/tpc.106.042473. eprint: https://academic.oup.com/plcell/article-pdf/18/12/3443/36906497/plcell_v18_12_3443.pdf.

[73] E. Saplaoura and F. Kragler, "Chapter One - Mobile Transcripts and Intercellular Communication in Plants," in *The Enzymes*, ser. Developmental Signaling in Plants, C. Lin and S. Luan, Eds., vol. 40, Academic Press, Jan. 1, 2016, pp. 1–29. DOI: 10.1016/bs.enz.2016.07.001.

[74] M. Kim, W. Canio, S. Kessler, and N. Sinha, "Developmental changes due to long-distance movement of a homeobox fusion transcript in tomato," *Science*, vol. 293, no. 5528, pp. 287–289, 2001. DOI: 10.1126/science.1059805.

[75] M. Kitagawa, P. Wu, R. Balkunde, P. Cunniff, and D. Jackson, "An RNA exosome subunit mediates cell-to-cell trafficking of a homeobox mRNA via plasmodesmata," *Science*, vol. 375, no. 6577, pp. 177–182, 2022. DOI: 10.1126/science.abm0840. eprint: https://www.science.org/doi/pdf/10.1126/science.abm0840.

[76] L. Yang, V. Perrera, E. Saplaoura, *et al.*, "m5C Methylation Guides Systemic Transport of Messenger RNA over Graft Junctions in Plants," *Current Biology*, vol. 29, no. 15, 2465–2476.e5, Aug. 2019, ISSN: 09609822. DOI: 10.1016/j.cub.2019.06.042.

[77] D. J. Hannapel and A. K. Banerjee, "Multiple Mobile mRNA Signals Regulate Tuber Development in Potato," *Plants (Basel, Switzerland)*, vol. 6, no. 1, p. 8, Feb. 10, 2017, ISSN: 2223-7747. DOI: 10.3390/plants6010008. PMID: 28208608.

[78] R. A. Jorgensen, R. G. Atkinson, R. L. S. Forster, and W. J. Lucas, "An RNA-based information superhighway in plants," *Science*, vol. 279, no. 5356, pp. 1486–1487, 1998. DOI: 10.1126/science.279.5356.1486. eprint: https://www.science.org/doi/pdf/10.1126/science.279.5356.1486.

[79] Z. Spiegelman, G. Golan, and S. Wolf, "Don't kill the messenger: Long-distance trafficking of mRNA molecules," *Plant Science*, vol. 213, pp. 1–8, 2013, ISSN: 0168-9452. DOI: 10.1016/j.plantsci.2013.08.011.

[80] N. Winter and F. Kragler, "Conceptual and Methodological Considerations on mRNA and Proteins as Intercellular and Long-Distance Signals," *Plant and Cell Physiology*, vol. 59, no. 9, pp. 1700–1713, Sep. 1, 2018, ISSN: 0032-0781. DOI: 10.1093/pcp/pcy140.

[81] D. Walther and F. Kragler, "Limited Phosphate: Mobile RNAs convey the message," *Nature Plants*, vol. 2, p. 16040, Apr. 5, 2016, ISSN: 2055-0278. DOI: 10.1038/nplants.2016.40. PMID: 27249568.

[82] S. Jin, Z. Nasim, H. Susila, and J. H. Ahn, "Evolution and functional diversification of FLOWERING LOCUS T/TERMINAL FLOWER 1 family genes in plants," *Seminars in Cell & Developmental Biology*, 1. Hormonal Signalling in Plant Development by Tom Bennett2. RIP Kinases in Cell Regulation by James Vince, vol. 109, pp. 20–30, Jan. 1, 2021, ISSN: 1084-9521. DOI: 10.1016/j.semcdb.2020.05.007.

[83] B. C. Yoo, F. Kragler, E. Varkonyi-Gasic, *et al.*, "A systemic small RNA signaling system in plants," *Plant Cell*, vol. 16, no. 8, pp. 1979–2000, Aug. 2004, ISSN: 1040-4651. DOI: 10.1105/tpc.104.023614.

[84] S.-C. Yoo, C. Chen, M. Rojas, *et al.*, "Phloem long-distance delivery of FLOWERING LOCUS T (FT) to the apex," *Plant Journal*, vol. 75, no. 3, pp. 456–468, Aug. 2013, ISSN: 0960-7412. DOI: 10.1111/tpj.12213.

[85] R. J. Morris, "On the selectivity, specificity and signalling potential of the long-distance movement of messenger RNA," *Current Opinion in Plant Biology*, 43 Physiology and Metabolism 2018, vol. 43, pp. 1–7, Jun. 1, 2018, ISSN: 1369-5266. DOI: 10.1016/j.pbi.2017.11.001.

[86] D. Hannapel, P. Sharma, and T. Lin, "Phloem-mobile messenger RNAs and root development," *Frontiers in Plant Science*, vol. 4, 2013, ISSN: 1664-462X. DOI: 10.3389/fpls.2013.00257.

[87] V. Haywood, T.-S. Yu, N.-C. Huang, and W. J. Lucas, "Phloem long-distance trafficking of GIBBERELLIC ACID-INSENSITIVE RNA regulates leaf development," *The Plant Journal*, vol. 42, no. 1, pp. 49–68, 2005. DOI: 10.1111/j.1365-313X.2005.02351.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-313X.2005.02351.x.

[88] P. Hao, X. Lv, M. Fu, *et al.*, "Long-distance mobile mRNA CAX3 modulates iron uptake and zinc compartmentalization," *EMBO reports*, vol. 23, no. 5, e53698, 2022. DOI: 10.15252/embr.202153698. eprint: https://www.embopress.org/doi/pdf/10.15252/embr.202153698.

[89] M. Notaguchi, "Identification of phloem-mobile mRNA," *Journal of Plant Research*, vol. 128, no. 1, pp. 27–35, 2015. DOI: 10.1007/s10265-014-0675-6.

[90] M. Heeney and M. H. Frank, "The mRNA mobileome: Challenges and opportunities for deciphering signals from the noise," *The Plant Cell*, M. Axtell, Ed., koad063, Mar. 7, 2023, ISSN: 1040-4651, 1532-298X. DOI: 10.1093/plcell/koad063.

[91] E. E. Tee and C. Faulkner, "Plasmodesmata and intercellular molecular traffic control," *New Phytologist*, vol. n/a, no. n/a, 2024. DOI: 10.1111/nph.19666. eprint: https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/nph.19666.

[92] J. Kehr and F. Kragler, "Long distance RNA movement," *New Phytologist*, vol. 218, no. 1, pp. 29–40, 2018, ISSN: 1469-8137. DOI: 10.1111/nph.15025.

[93] A. Ostendorp, S. Ostendorp, Y. Zhou, *et al.*, "Intrinsically disordered plant protein PARCL colocalizes with RNA in phase-separated condensates whose formation can be regulated by mutating the PLD," *Journal of Biological Chemistry*, vol. 298, no. 12, p. 102 631, 2022, ISSN: 0021-9258. DOI: 10.1016/j.jbc.2022.102631.

[94] H. Y. Park, H. Lim, Y. J. Yoon, *et al.*, "Visualization of dynamics of single endogenous mRNA labeled in live mouse," *Science*, vol. 343, no. 6169, pp. 422–424, 2014. DOI: 10.1126/science.1239200. eprint: https://www.science.org/doi/pdf/10.1126/science.1239200.

[95] I. Loedige, A. Baranovskii, S. Mendonsa, *et al.*, "mRNA stability and m6A are major determinants of subcellular mRNA localization in neurons," *Molecular Cell*, vol. 83, no. 15, 2709–2725.e10, Aug. 2023.

[96] E. E. Deinum, B. M. Mulder, and Y. Benitez-Alfonso, "From plasmodesma geometry to effective symplasmic permeability through biophysical modelling," *Elife*, p. 40, 2019. [Online]. Available: https://elifesciences.org/articles/49000.

[97] A. Hughes, C. Faulkner, R. J. Morris, and M. Tomkins, "Intercellular communication as a series of narrow escape problems," *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, vol. 7, no. 2, pp. 89–93, 2021. DOI: 10.1109/TMBMC.2021.3083719.

[98] C. J. Thieme, M. Rojas-Triana, E. Stecyk, *et al.*, "Endogenous Arabidopsis messenger RNAs transported to distant tissues," *Nature Plants*, vol. 1, no. 4, p. 15 025, Apr. 2015, ISSN: 2055-0278. DOI: 10.1038/nplants.2015.25.

[99] E. A. Tolstyko, A. A. Lezzhov, S. Y. Morozov, and A. G. Solovyev, "Phloem transport of structured RNAs: A widening repertoire of trafficking signals and protein factors," *Plant Science*, vol. 299, p. 110 602, 2020, ISSN: 0168-9452. DOI: 10.1016/j.plantsci.2020.110602.

[100] R. Narsai, K. A. Howell, A. H. Millar, N. O'Toole, I. Small, and J. Whelan, "Genome-Wide Analysis of mRNA Decay Rates and Their Determinants in Arabidopsis thaliana," *The Plant Cell*, vol. 19, no. 11, pp. 3418–3436, Nov. 2007, ISSN: 1040-4651. DOI: 10.1105/tpc.107.055046. PMID: 18024567.

[101] R. S. Sorenson, M. J. Deshotel, K. Johnson, F. R. Adler, and L. E. Sieburth, "¡i¿Arabidopsis¡/i¿ mRNA decay landscape arises from specialized RNA decay substrates, decapping-mediated feedback, and redundancy," *Proceedings of the National Academy of Sciences*, vol. 115, no. 7, E1485–E1494, 2018. DOI: 10.1073/pnas.1712312115. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1712312115.

[102] R. Deeken, P. Ache, I. Kajahn, J. Klinkenberg, G. Bringmann, and R. Hedrich, "Identification of *Arabidopsis thaliana* phloem RNAs provides a search criterion for phloem-based transcripts hidden in complex datasets of microarray experiments," *The Plant Journal*, vol. 55, no. 5, pp. 746–759, Sep. 2008, ISSN: 09607412, 1365313X. DOI: 10.1111/j.1365-313X.2008.03555.x.

[103] A. Omid, T. Keilin, A. Glass, D. Leshkowitz, and S. Wolf, "Characterization of phloem-sap transcription profile in melon plants," *Journal of Experimental Botany*, vol. 58, no. 13, pp. 3645–3656, Oct. 2007, ISSN: 0022-0957. DOI: 10.1093/jxb/erm214.

[104] C. Rodriguez-Medina, C. A. Atkins, A. J. Mann, M. E. Jordan, and P. M. Smith, "Macromolecular composition of phloem exudate from white lupin (Lupinus albusL.)," *BMC Plant Biology*, vol. 11, no. 1, p. 36, 2011. DOI: 10.1186/1471-2229-11-36.

[105] A. Bombarely, H. G. Rosli, J. Vrebalov, P. Moffett, L. A. Mueller, and G. B. Martin, "A draft genome sequence of nicotiana benthamiana to enhance molecular plant-microbe biology research," *Molecular Plant-Microbe Interactions®*, vol. 25, no. 12, pp. 1523–1530, 2012. DOI: 10.1094/MPMI-06-12-0148-TA. eprint: https://doi.org/10.1094/MPMI-06-12-0148-TA.

[106] Y. Yang, L. Mao, Y. Jittayasothorn, *et al.*, "Messenger RNA exchange between scions and rootstocks in grafted grapevines," *BMC Plant Biology*, vol. 15, no. 1, p. 251, Oct. 19, 2015, ISSN: 1471-2229. DOI: 10.1186/s12870-015-0626-y.

[107] Z. Zhang, Y. Zheng, B.-K. Ham, *et al.*, "Vascular-mediated signalling involved in early phosphate stress response in plants," *Nature Plants*, vol. 2, p. 16 033, Apr. 4, 2016, ISSN: 2055-0278. DOI: 10.1038/nplants.2016.33. PMID: 27249565.

[108] B. Munsky, G. Neuert, and A. van Oudenaarden, "Using gene expression noise to understand gene regulation," *Science*, vol. 336, no. 6078, pp. 183–187, 2012. DOI: 10.1126/science.1216379. eprint: https://www.science.org/doi/pdf/10.1126/science.1216379.

[109] G. Gorin, J. J. Vastola, M. Fang, and L. Pachter, "Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments," *Nature Communications*, vol. 13, no. 1, p. 7620, Dec. 9, 2022, ISSN: 2041-1723. DOI: 10.1038/s41467-022-34857-7.

[110] A. Calderwood, S. Kopriva, and R. J. Morris, "Transcript Abundance Explains mRNA Mobility Data in Arabidopsis thaliana," *The Plant Cell*, vol. 28, no. 3, pp. 610–615, Mar. 1, 2016, ISSN: 1040-4651, 1532-298X. DOI: 10.1105/tpc.15.00956. PMID: 26952566.

[111] D. Guan, B. Yan, C. Thieme, *et al.*, "PlaMoM: A comprehensive database compiles plant mobile macromolecules," *Nucleic Acids Research*, vol. 45, no. D1, pp. D1021–D1028, Oct. 2016, ISSN: 0305-1048. DOI: 10.1093/nar/gkw988. eprint: https://academic.oup.com/nar/article-pdf/45/D1/D1021/8847352/gkw988.pdf.

[112] M. Fu, Z. Xu, H. Ma, *et al.*, "Characteristics of long-distance mobile mRNAs from shoot to root in grafted plant species," *Horticultural Plant Journal*, May 26, 2023, ISSN: 2468-0141. DOI: 10.1016/j.hpj.2023.05.009.

[113] K. Mudge, J. Janick, S. Scofield, and E. E. Goldschmidt, "A history of grafting," in *Horticultural Reviews*, John Wiley and Sons, Ltd, 2009, pp. 437–493, ISBN: 978-0-470-59377-6. DOI: 10.1002/9780470593776.ch9. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470593776.ch9.

[114] C. M. K., "Concerning the hormonal nature of plant development processes," *Doklady Akademii Nauk SSSR*, vol. 16, pp. 227–230, 1937. [Online]. Available: https://cir.nii.ac.jp/crid/1570572701266709632.

[115] P. A. Wigge, M. C. Kim, K. E. Jaeger, *et al.*, "Integration of spatial and temporal information during floral induction in ¡i¿arabidopsis¡/i¿," *Science*, vol. 309, no. 5737, pp. 1056–1059, 2005. DOI: 10.1126/science.1114358. eprint: https://www.science.org/doi/pdf/10.1126/science.1114358.

[116] L. Corbesier, C. Vincent, S. Jang, *et al.*, "FT protein movement contributes to long-distance signaling in floral induction of ¡i¿arabidopsis¡/i¿," *Science*, vol. 316, no. 5827, pp. 1030–1033, 2007. DOI: 10.1126/science.1141752. eprint: https://www.science.org/doi/pdf/10.1126/science.1141752.

[117] Y. Wang, L. Wang, N. Xing, *et al.*, "A universal pipeline for mobile mRNA detection and insights into heterografting advantages under chilling stress," *Horticulture Research*, vol. 7, p. 13, Jan. 1, 2020, ISSN: 2052-7276. DOI: 10.1038/s41438-019-0236-1.

[118] M. Notaguchi, T. Higashiyama, and T. Suzuki, "Identification of mRNAs that Move Over Long Distances Using an RNA-Seq Analysis of Arabidopsis/Nicotiana benthamiana Heterografts," *Plant and Cell Physiology*, vol. 56, no. 2, pp. 311–321, Feb. 1, 2015, ISSN: 0032-0781. DOI: 10.1093/pcp/pcu210.

[119] J. Xu, M. Zhang, G. Liu, X. Yang, and X. Hou, "Comparative transcriptome profiling of chilling stress responsiveness in grafted watermelon seedlings," *Plant Physiology and Biochemistry*, vol. 109, pp. 561–570, 2016, ISSN: 0981-9428. DOI: 10.1016/j.plaphy.2016.11.002.

[120] C. Xia, Y. Zheng, J. Huang, *et al.*, "Elucidation of the Mechanisms of Long-Distance mRNA Movement in a Nicotiana benthamiana/Tomato Heterograft System," *Plant Physiology*, vol. 177, no. 2, pp. 745–758, Jun. 1, 2018, ISSN: 0032-0889. DOI: 10.1104/pp.17.01836.

[121] N. Liu, J. Yang, X. Fu, *et al.*, "Genome-wide identification and comparative analysis of grafting-responsive mRNA in watermelon grafted onto bottle gourd and squash rootstocks by high-throughput sequencing," *Molecular Genetics and Genomics*, vol. 291, no. 2, pp. 621–633, 2016. DOI: 10.1007/s00438-015-1132-5.

[122] W. Li, S. Chen, Y. Liu, *et al.*, "Long-distance transport RNAs between rootstocks and scions and graft hybridization," *Planta*, vol. 255, no. 5, p. 96, Mar. 29, 2022, ISSN: 1432-2048. DOI: 10.1007/s00425-022-03863-w.

[123] T. Wang, Y. Zheng, C. Xu, *et al.*, "Movement of ACC oxidase 3 mRNA from seeds to flesh promotes fruit ripening in apple," *Molecular Plant*, 2024, ISSN: 1674-2052. DOI: 10.1016/j.molp.2024.06.008.

[124] F. Pfeiffer, C. Gröber, M. Blank, *et al.*, "Systematic evaluation of error rates and causes in short samples in next-generation sequencing," *Scientific Reports*, vol. 8, no. 1, p. 10 950, 2018. DOI: 10.1038/s41598-018-29325-6.

[125] N. J. Loman, R. V. Misra, T. J. Dallman, *et al.*, "Performance comparison of benchtop high-throughput sequencing platforms," *Nature Biotechnology*, vol. 30, no. 5, pp. 434–439, 2012. DOI: 10.1038/nbt.2198.

[126] A. Fungtammasan, M. Tomaszkiewicz, R. Campos-Sánchez, K. A. Eckert, M. DeGiorgio, and K. D. Makova, "Reverse transcription errors and RNA–DNA differences at short tandem repeats," *Molecular Biology and Evolution*, vol. 33, no. 10, pp. 2744–2758, Jul. 2016, ISSN: 0737-4038. DOI: 10.1093/molbev/msw139. eprint: https://academic.oup.com/mbe/article-pdf/33/10/2744/17473292/msw139.pdf.

[127] W. Li and M. Lynch, "Universally high transcript error rates in bacteria," *eLife*, vol. 9, C. R. Landry, P. J. Wittkopp, and J. Masel, Eds., e54898, May 2020, ISSN: 2050-084X. DOI: 10.7554/eLife.54898.

[128] J. Verwilt, P. Mestdagh, and J. Vandesompele, "Artifacts and biases of the reverse transcription reaction in RNA sequencing," *RNA (New York, N.Y.)*, vol. 29, no. 7, pp. 889–897, Jul. 2023.

[129] "Beddie: Natural root grafts in New Zealand trees - Google Scholar." (), [Online]. Available: https://scholar.google.com/scholar_lookup?author=A.+D.+Beddie&title=Natural+root+grafts+in+New+Zealand+trees&publication_year=1942&journal=Trans.+Proc.+R.+Soc.&volume=71 (visited on 02/17/2021).

[130] "Central Dogma of Molecular Biology — Nature." (), [Online]. Available: https://www.nature.com/articles/227561a0 (visited on 02/24/2021).

[131] "Mathematical Modelling in Plant Biology — Bookshare." (), [Online]. Available: https://www.bookshare.org/browse/book/2299688 (visited on 03/04/2021).

[132] "Seidel: Ueber Verwachsungen von Stammen und Zweigen... - Google Scholar." (), [Online]. Available: https://scholar.google.com/scholar_lookup?author=C.+F.+Seidel&title=Ueber+Verwachsungen+von+Stammen+und+Zweigen+von+Holzgew%C3%A4chsen+und+ihren+Einfluss+auf+das+Dickenwachsthum+der+betreffenden+Theile&publication_year=1879&journal=Naturwiss.+Ges.+Isis+Dresden+Sitzber.&volume=1879 (visited on 02/17/2021).

[133] M. Segel, B. Lash, J. Song, *et al.*, "Mammalian retrovirus-like protein PEG10 packages its own mRNA and can be pseudotyped for mRNA delivery," *Science*, vol. 373, no. 6557, pp. 882–889, Aug. 20, 2021, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.abg6155.

[134] N. Stoler and A. Nekrutenko, "Sequencing error profiles of Illumina sequencing instruments," *NAR Genomics and Bioinformatics*, vol. 3, no. 1, lqab019, Mar. 1, 2021, ISSN: 2631-9268. DOI: 10.1093/nargab/lqab019.

[135] L. Zapata, J. Ding, E.-M. Willing, *et al.*, "Chromosome-level assembly of ¡i¿Arabidopsis thaliana¡/i¿ L¡i¿er¡/i¿ reveals the extent of translocation and inversion polymorphisms," *Proceedings of the National Academy of Sciences*, vol. 113, no. 28, E4052–E4060, 2016. DOI: 10.1073/pnas.1607532113. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1607532113.

[136] B. Jaegle, R. Pisupati, L. M. Soto-Jiménez, R. Burns, F. A. Rabanal, and M. Nordborg, "Extensive sequence duplication in Arabidopsis revealed by pseudo-heterozygosity," *Genome Biology*, vol. 24, no. 1, p. 44, Mar. 9, 2023, ISSN: 1474-760X. DOI: 10.1186/s13059-023-02875-3.

[137] W. Zhang, C. J. Thieme, G. Kollwig, *et al.*, "tRNA-Related Sequences Trigger Systemic mRNA Transport in Plants," *The Plant Cell*, vol. 28, no. 6, pp. 1237–1249, Jun. 1, 2016, ISSN: 1040-4651. DOI: 10.1105/tpc.15.01056.

[138] L. Yang, F. Machin, S. Wang, E. Saplaoura, and F. Kragler, "Heritable transgene-free genome editing in plants by grafting of wild-type shoots to transgenic donor rootstocks," *Nature biotechnology*, vol. 41, no. 7, pp. 958–967, Jan. 2023.

[139] S. Dasgupta, D. Y. Dayagi, G. Haimovich, *et al.*, "Global analysis of contact-dependent human-to-mouse intercellular mRNA and lncRNA transfer in cell culture," *eLife*, vol. 12, G. W. Yeo, V. Malhotra, and C. Brou, Eds., e83584, May 30, 2023, ISSN: 2050-084X. DOI: 10.7554/eLife.83584.

[140] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype," *Nature Biotechnology*, vol. 37, no. 8, pp. 907–915, 2019. DOI: 10.1038/s41587-019-0201-4.

[141] P. Danecek, J. K. Bonfield, J. Liddle, *et al.*, "Twelve years of samtools and bcftools," *GigaScience*, vol. 10, no. 2, Feb. 2021, ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008. eprint: https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf.

[142] M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, and S. L. Salzberg, "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and ballgown," *Nature Protocols*, vol. 11, no. 9, pp. 1650–1667, 2016. DOI: 10.1038/nprot.2016.095.

[143] C. Camacho, G. Coulouris, V. Avagyan, *et al.*, "BLAST+: Architecture and applications," *BMC Bioinformatics*, vol. 10, no. 1, p. 421, 2009. DOI: 10.1186/1471-2105-10-421.

[144] C. P. Robert. "The expected demise of the Bayes factor." arXiv: 1506.08292 [stat]. (Jul. 27, 2015), pre-published.

[145] A. Ly, J. Verhagen, and E.-J. Wagenmakers, "Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology," *Journal of Mathematical Psychology*, Bayes Factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments, vol. 72, pp. 19–32, Jun. 1, 2016, ISSN: 0022-2496. DOI: 10.1016/j.jmp.2015.06.004.

[146] K. Kamary, K. Mengersen, C. P. Robert, and J. Rousseau. "Testing hypotheses via a mixture estimation model." arXiv: 1412.2044 [stat]. (Dec. 31, 2018), pre-published.

[147] R. L. Wasserstein and N. A. Lazar, "The ASA Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, vol. 70, no. 2, pp. 129–133, Apr. 2, 2016, ISSN: 0003-1305. DOI: 10.1080/00031305.2016.1154108.

[148] R. Matthews, "The p-value statement, five years on," *Significance*, vol. 18, no. 2, pp. 16–19, 2021, ISSN: 1740-9713. DOI: 10.1111/1740-9713.01505.

[149] D. Lakens, F. G. Adolfi, C. J. Albers, *et al.*, "Justify your alpha," *Nature Human Behaviour*, vol. 2, no. 3, pp. 168–171, Mar. 2018, ISSN: 2397-3374. DOI: 10.1038/s41562-018-0311-x.

[150] V. Amrhein, S. Greenland, and B. McShane, "Scientists rise up against statistical significance," *Nature*, vol. 567, no. 7748, pp. 305–307, 7748 Mar. 2019. DOI: 10.1038/d41586-019-00857-9.

[151] S. Greenland, "Connecting simple and precise $P$ -values to complex and ambiguous realities (includes rejoinder to comments on "Divergence vs. decision $P$- values")," *Scandinavian Journal of Statistics*, vol. 50, no. 3, pp. 899–914, Sep. 2023, ISSN: 0303-6898, 1467-9469. DOI: 10.1111/sjos.12645.

[152] V. Wilde. "December Talk – Vera Wilde." (Dec. 31, 2023), [Online]. Available: https://verawil.de/2023/12/december-talk/ (visited on 09/02/2024).

[153] *FireShonks 2023 - Chat Control: Mass Screenings, Massive Dangers*, scriptwriter V. Wilde, Dec. 30, 2023. [Online]. Available: https://www.youtube.com/watch?v=z-Gi5mEFSq8 (visited on 09/02/2024).